Contents

Sponsors

Platinum Sponsor

d-Ear
得意音通技术
d-Ear Technologies

Gold Sponsor

CSLT GROUPING
Center for Speech and Language Technologies

Silver Sponsor

声原智能
Frontier Acoustics

ChipIntelli 启英泰伦

speechocean
海天瑞声

AISHELL

CVTE Dream·Future
视源股份

Academic Support

清華大學
TSINGHUA UNIVERSITY
1911

西北民族大學
1950
NORTHWEST MINZU UNIVERSITY

APSIPA
Signal and Information Processing

蘭州交通大學
1958
LANZHOU JIAOTONG UNIVERSITY

IEEE
Signal Processing Society ®

Schedule at a Glance

18th November 2019 (Monday) - Day 0								
Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
9:30-11:30	**Tutorial 1** Tensor Component Analysis. Yipeng Liu	**Tutorial 2** Speech Enhancement based on Deep Learning and Intelligibility Evaluation. Yu Tsao and Fei Chen	**AutoWare Workshop** Alexander Carballo and Qingguo Zhou					
11:30-13:00	**Lunch Break (not served)**							
13:30-15:30	**Tutorial 3** Fundamental and Progress of Deep Learning Based Statistical Parametric Speech Synthesis. Zhenhua Ling	**Tutorial 4** Few-Shot Learning, Adversarial Learning, and Their Applications. Jen-Tzung Chien and Zhanyu Ma	**AutoWare Workshop** Alexander Carballo and Qingguo Zhou					
15:30-16:00	**Coffee Break**							
16:00-18:00	**Tutorial 5** Fundamental and Progress of Immersive Visual Media Communications. Yo-Sung Ho	**Tutorial 6** Signal Enhancement for Consumer Products. Akihiko Sugiyama	**AutoWare Workshop** Alexander Carballo and Qingguo Zhou					
18:00-19:00	**Executive Committee Meeting**							
19:00-21:00	**Welcome Reception**							

19th November 2019 (Tuesday) - Day 1								
Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
8:30-9:00	**Open Ceremony** (Yonghong Li) (Grand Ballroom, Crowne Plaza Hotel, 3F)							
9:00-10:00	**Keynote 1:** Transfer Learning: from Bayesian Adaptation to Teacher-Student Modeling. Chin-Hui Lee (Haizhou Li, Professor of National University of Singapore) (Grand Ballroom, Crowne Plaza Hotel, 3F)							
10:00-10:20	**Coffee Break**							
10:20-12:00	**Overview 1** *Kazushi Ikeda, Jianquan Liu, Chung-Hsien Wu* (Homer Chen)	**TUE-AM1-SS1** Convergence of 5G with SDN/NFV/Cloud (Po-Chiang Lin, Wen-Ping Lai)	**TUE-AM1-SS2** Signal Processing for Big Data and Its Security (Yuhong Liu)	**TUE-AM1-SS3** Machine Learning for Small-sample Data Analysis (Zhanyu Ma,Jen-Tzung Chien, Ruiping Wang, Xiaoxu Li)	**TUE-AM1-O1** Voice conversion (Nobuaki Minematsu)	**TUE-AM1-O2** Speech Synthesis (Rohan Kumar Das)	**TUE-AM1-O3** Speech Recognition (Hemant Patil)	**TUE-AM1-SS4** Recent Advances in Biometric Security (Koichi Ito , Tetsushi Ohki)
12:00-13:20	**Lunch Break**							
	Friend Lab1 Machine Learning TC	**SPS Meeting**	**SIPTM Meeting**	**IVM Meeting**	**MSF Meeting**			
13:20-15:00	**POSTER 1** (Zhiyi Yu) （Gansu International Conference Centre(GICC), 3F）							**Conf. Board Meeting**

19th November 2019 (Tuesday) - Day 1								
Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
15:00-16:40	**Overview 2** Hsin-Min Wang, Chang-Su Kim, Sanghoon Lee (Nam Ik Cho)	**Overview 2** Hsin-Min Wang, Chang-Su Kim, Sanghoon Lee (H. Vicky Zhao , Yan Chen)	**TUE-PM2-O1** Speech Emotion Recognition (Chi-Chun Lee)	**TUE-PM2-O2** Speaker Recognition (Yanhua Long)	**TUE-PM2-O3** Image Processing (Isao Echizen)	**TUE-PM2-O4** Speech Synthesis (Jun Du)	**TUE-PM2-O5** Speech Recognition (Xueliang Zhang)	**TUE-PM2-O6** Speech Enhancement (Meng Sun)
Coffee Break								
17:00-18:40	**Industrial Forum** Intelligent Signal and Information Processing in Industries	**TUE-PM3-SS1** Special Learning under Limited Samples Scenarios (Ganggang Dong, Yinghua Wang, Bo Chen)	**TUE-PM3-SS2** Signal Processing in Behavior Analysis (Kazushi Ikeda, Li-Wei Kang)	**TUE-PM3-SS3** Latest Progress on Fractional Signal Processing Theory and Application (Bing-Zhao Li, Xiaolong Chen, Yong Guo)	**TUE-PM3-SS4** High Performance Image and Video Processing and Applications (Kosin Chamnongthai)	**TUE-PM3-O1** Speaker Recognition (Dong Wang)	**TUE-PM3-O2** Speech Recognition (Hsin-Min Wang)	**TUE-PM3-O3** Speech Enhancement (Changchun Bao)
19:00-22:00	**Buffet**							**BOG Meeting**

20th November 2019 (Wednesday) - Day 2								
Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
9:00-10:00	**Keynote 2:** Digital Retina – Improvement of Cloud Artificial Vision System from Enlighten of HVS Evolution. Wen Gao (Hitoshi Kiya, Professor, Tokyo Metropolitan University) (Multi-function Office, GICC 4F)							

20th November 2019 (Wednesday) - Day 2								
Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
10:00-10:20	**Coffee Break**							
10:20-12:00	**WED-AM1-O1** Paralinguistics in Speech and Language (Yu Tsao)	**WED-AM1-O2** Neural Signal Processing (Kazushi Ikeda)	**WED-AM1-O3** Computer Vision (Ce Zhu)	**WED-AM1-O4** Language Learning (Aijun Li)	**WED-AM1-O5** Dialog System (Qin Jin)	**WED-AM1-O6** Adaptive Signal Processing (Mau-Luen Tham)	**WED-AM1-O7** Signal Processing Methods (Shinsuke Ibi)	**WED-AM1-O8** Image Processing (Sanghoon Lee)
12:00-13:20	**Lunch Break**							
	Friend Lab2 Women in APSIPA	**SLA Meeting**	**BioSIPS Meeting**	**DL Meeting**	**WCN Meeting**			
13:20-15:00	**POSTER 2** (Xiangui Kang) （Gansu International Conference Centre(GICC), 3F）						**Tech. Board Meeting**	**Editorial Board Meeting**
15:00-16:40	**Overview 3** Hsueh-Ming Hang, Homer Chen, Jiwu Huang (Yoong-Choon Chang)	**WED-PM2-SS1** Advanced Topics on High-dimensional Data Analytics and Processing (Supavadee Aramvith , Shogo Muramatsu)	**WED-PM2-SS2** Lightweight Signal Processing and Machine Learning for Embedded Applications (Hakaru Tamukoh, Hiroshi Tsutsui)	**WED-PM2-SS3** High Performance Video Processing and Image Identification (Junyong Deng)	**WED-PM2-SS4** Physical and Wireless Environment Recognition Based on Signal Processing (Osamu Takyu)	**WED-PM2-SS5** Robust Rich Audio Analysis (Xiao-Lei Zhang)	**WED-PM2-O1** Signal Processing Methods (Mingyi He)	**WED-PM2-O2** Image Processing (Zhonghua Sun)
16:40-17:00	Coffee Break							

20th November 2019 (Wednesday) - Day 2

Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**
17:00-18:40	**WED-PM3-SS1** Recent Advances in Fingerprinting and Data Hiding (Minoru Kuribayashi, David Megías Jiménez)	**WED-PM3-SS2** Recent Advances in Speaker Recognition, Speaker Diarization and Language Recognition (Qingyang Hong, Lin Li)	**WED-PM3-SS3** Deep Generative Models for Media Clones and Its Detection (Fuming Fang, Zhenzhong Kuang, Xin Wang)	**WED-PM3-SS4** Recent Trends in Signal Processing & Machine Learning - Acoustic & Biomedical Applications (Kiyoshi Nishikawa, Felix Albu, Akhtar, Muhammad Tahir)	**WED-PM3-SS5** Information Security for Digital Content (Xiangui Kang, KokSheik Wong, Linna Zhou)	**WED-PM3-SS6** Second Language Speech Perception and Production[1] (Ying Chen, Jian Gong)	**WED-PM3-SS7** Recent Topics on Signal and Information Processing for Active Control of Sound (Yoshinobu Kajikawa, Chuang Shi)	**WED-PM3-SS8** Advanced Signal Processing for 5G Communication (Na Chen, Minoru Okada)
18:40-19:10	**APSIPA General Assembly**							
19:10-22:00	**Banquet**							

21th November 2019 (Thursday) - Day 3

Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
9:00-10:00	**Keynote 3**: Applying Deep Learning in Non-native Spoken English Assessment. Kate Knill (Hongwu Yang, Professor, Northwest Normal University) (Multi-function Office, GICC 4F)								

21th November 2019 (Thursday) - Day 3									
Time and Venue	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**
10:00-10:20	**Coffee Break**								
10:20-12:00	**Overview 4** *Seishi Takamura, Ying Loong Lee, Mau-Luen Tham* (Dong Wang)	**THU-AM1-SS1** Advanced Signal Processing and Machine Learning for Audio and Speech Applications (Shoji Makino, Hiroshi Saruwatari)	**THU-AM1-SS2** Multilingual Speech and Language Processing (Zhiyuan Tang, Dong Wang, GuanyuLi, Mijiti Ablimit)	**THU-AM1-SS3** Deep Learning Systems for Cloud, Fog, and Edge Computing, and Applications (Jia-Ching Wang)	**THU-AM1-SS4** Technologies for A Maximized Experience of Multi-dimensional Content: from 2D, 3D Modeling to An Objective Assessment (Sanghoon Lee, Chia-Hung Yeh)	**THU-AM1-SS5** Recent Trends in Computational Intelligence (Chern Hong Lim, Mei Kuan Lim)	**THU-AM1-SS6** Multi-source Data Processing and Analysis: Models, Methods and Applications (Ping Han, Qiuping Jiang, Runmin Cong, Chongyi Li)	**THU-AM1-SS7** Second Language Speech Perception and Production[2] (Ying Chen, Jian Gong)	**Friend Lab3**
12:00-12:30	**Closing Ceremony**								
12:30-13:30	**Lunch Break**								
14:00-18:00	**Tour Activities**								
19:00-21:00	**Buffet**								

Message from General Chairs

In 2011, when the third APSIPA Annual Summit and Conference (ASC) entered China for the first time, it was in Xi'an, the starting city in the ancient Silk Road. Eight years later, the eleventh APSIPA ASC is held in China for the second time. This time is Lanzhou, the second stop in the ancient Silk Road, and a very special city full of brilliant historical culture, amazing natural scenery and nice meals of multiple nations. More than 500 attendees will discuss diverse topics related to signal and information processing, under the theme of Intelligent Signal and Information Processing.

Founded in 2009, APSIPA aims to promote broad spectrum of research and education activities in signal processing, information technology, and communications in Asia-Pacific Region. The annual conferences have been held previously in Sapporo (2009), Singapore (2010), Xi'an (2011), Los Angeles (2012), Kaohsiung (2013), Siem Reap (2014), Hong Kong (2015), Jeju (2016), and Kuala Lumpur (2017) and Hawaii (2018).

To make APSIPA ASC 2019 a unique conference that all participants will have very unique experiences, members of the whole organizing team have been making great efforts together in many dimensions, including technical arrangement and local arrangement.

With their efforts, it is our great pleasure to see rapid and significant growth of APSIPA at the 11th APSIPA ASC in Lanzhou, when several records are broken, including the largest number of submissions, the lowest acceptance rate (such that papers with even better quality have been selected for the conference) and the largest number of attendees.

We would like to thank the APSIPA Officers, technical program committee members, organizing committee members, keynote speakers, authors, reviewers, participants, and local volunteers. We would also like to thank the sponsors who give us generous financial supports. Without you, the conference cannot be successful.

Once again, welcome to Lanzhou and enjoy the beautiful western China!

General Chairs:

Thomas Fang ZHENG

Tsinghua University, China

Hongzhi YU

Northwest Minzu University, China

Jianwu DANG

Lanzhou Jiaotong University, China

Wan-Chi SIU

The Hong Kong Polytechnic University, China

Hitoshi KIYA

Tokyo Metropolitan University, Japan

Message from APSIPA president

On behalf of the Asia-Pacific Signal and Information Processing Association (APSIPA), it is my great pleasure and honor to welcome each one of you to the APSIPA Annual Summit and Conference 2019 (APSIPA ASC 2019) held in Lanzhou, China. Lanzhou is located on the banks of the Yellow River, and it was historically a major link on the Northern Silk Road.

APSIPA ASC 2019 will be the 11th conference organized by APSIPA. APSIPA ASC 2018 was held as the 10th anniversary conference in its birthplace, Hawaii, where the idea of APSIPA was discussed for the first time during IEEE ICASSP 2007. On behalf of the Association, we take this opportunity to thank everyone who has contributed their dedicated effort to make APSIPA a great success. I am happy to start a new stage of our Association in this historical city with many attendees from a lot of countries. I am sure that APSIPA will continue to thrive and evolve, and will provide a greater leadership to the signal and information processing R&D community.

I am glad to know that the number of submissions is the largest one in our past conferences. I sincerely thank the conference organizers, leading by the General Co-Chairs: Prof. Thomas Fang Zheng, Prof. Hongzhi Yu, Prof. Jianwu Dang and Prof. Wan-Chi Siu. I would like also to thank the keynote speakers, the invited speakers, the authors and all the participants for their support.

I hope that everyone attending the conference will enjoy the conference and Lanzhou, and make new friends on this occasion. Many senior researchers and educators with very rich experience are a part of the APSIPA community. APSIPA will offer members the opportunity to network and make connections with such senior people. We look forward to working with many active people, and to furthering technological progress in the Asia-Pacific region.

Hitoshi Kiya

President (2019-2020), APSIPA

Message from Technical Program Chairs

It is our great pleasure to welcome you to 2019 APSIPA ASC in Lanzhou. Under the theme of intelligent signal and information processing, we have designed a strong program, and hope you will enjoy it.

This year, the Technical Program Committee received 272 regular paper submissions from 15 countries. This is the highest number in APSIPA. Among these submissions, 187 papers were accepted. The acceptance rate is about 68.8%. In addition, we have approved 28 special session proposals, and accepted 164 special session papers. The top-two submission countries are China and Japan, which contributed 78% of the submissions.

The paper reviews for both regular and special sessions represent great efforts of 563 reviewers. A total of 1469 reviews were received, which represents 4.1 independent reviews for each paper in average. The regular paper sessions and the special session papers together form 51 oral sessions and 2 poster sessions.

This year, we are very pleased to have three distinguished keynotes given by world-famous scholars: Prof. Chin-Hui Lee (Georgia Tech University), Prof. Wen Gao (Beijing University), and Prof. Kate Knill (Cambridge University). They will share their insights on transfer learning, cloud artificial vision system , and language assessment.

As part of the APSIPA ASC tradition, the Friend Labs gathering is held during lunch breaks. In this year, we have three friend labs: friend lab for ML TC, friend lab for Women in APSIPA, and friend lab in China. We also organized four overview sessions that feature overview talks given by leading experts on diverse topics. Thanks to Prof. Homer Chen, Nam Ik Cho, Yoong Choon Chang, Guan-Ming Su for the organization. We also have a one-day workshop on Autoware, organized by Prof. Alexander Carballo and Prof. Qingguo Zhou.

We have six tutorials offered by leading experts in the community. The optics cover tensor component analysis, speech enhancement, speech synthesis, few-shot learning and adversarial learning, immersive visual media communications, and signal enhancement for consumer products.

This year we have organized a very special event, the APSIPA Winter School. Our goal is to promote the research on signal and information technologies in the residence of APSIPA annual conference, and benefit students of the local universities. The theme of this inaugural event is “Speech and AI”. We invited six speakers, and most of them are supported by the APSIPA distinguished lecture program. Thanks to Prof. Woon Seng Gan for the initiative, and thanks to Prof. Axu Hu for the local organization.

Another innovation this year is the portrait poster show. We invited the authors of each paper to present an A3-size small poster to introduce their research and their institutes. These portrait posters are put in the corridor outside our conference rooms, which may give you some interesting information, e.g., opening positions of other groups. You can enjoy them in the coffee time.

As another innovation, we have designed a WeChat program to assist attendees to manage the conference. You can search papers, check the ongoing sessions, set alarms for interested sessions and papers, and receive notifications from the local organizer. You can download the WeChat App and scan the bar code everywhere in our conference venue to access these functions.

Finally, I'd like to take this opportunity to thank Prof. Chung-Nan Lee, the VP TC of APSIPA and all the TC chairs for organizing the review process. Most thanks to the special session organizers and our 563 reviewers. In order to give more time to our authors to prepare the manuscripts, we have delayed the submission deadline for a couple of weeks, which makes the reviewing process very stringent. Nevertheless, we finally met our goal and delivered high-quality review.

Once again, thanks for attending APSIPA and enjoy your stay in Lanzhou!

APSIPA ASC Technical Program Co-Chairs,

Regular sessions: Dong Wang. Jiun-In Guo, Zhiyi Yu, Kazushi Ikeda, Sung Chan Jun, Mingyi He, Yu Tsao, Supavadee Aramvith, Sanghoon Lee, Shinsuke Ibi, Naveed Ul Hassan, Woon-Seng Gan, Isao Echizen

Special sessions: Qin Jin, Yoshinobu Kajikawa, Chung-Nan Lee, Lap Pui Chau, Hsin-Min Wang, Kosin Chamnongthai, Askar Hamdulla

Organizing Committee

Honorary Co-Chairs

Sadaoki Furui (Toyota Technological Institute at Chicago, USA)
K. J. Ray Liu (University of Maryland, USA)
C.-C. Jay Kuo (University of Southern California, USA)
Haizhou Li (National University of Singapore, Singapore)

General Co-Chairs

Thomas Fang Zheng (Tsinghua University)
Hongzhi Yu (Northwest Minzu University)
Jianwu Dang (Lanzhou Jiaotong University)
Wan-Chi Siu (The Hong Kong Polytechnic University)
Hitoshi Kiya (Tokyo Metropolitan University, Japan)

TPC Co-Chairs

Dong Wang (Tsinghua University)
Jiun-In Guo (National Chiao-Tung University)
Zhiyi Yu (Fudan University)
Kazushi Ikeda (Nara Institute of Science and Technology, Japan)
Sung Chan Jun (Gwangju Institute of Science and Technology, Korea)
Mingyi He (Northwestern Polytechnical University)
Yu Tsao
Supavadee Aramvith (Chulalongkorn University, Thailand)
Sanghoon Lee (Yonsei University, Korea)
Shinsuke Ibi (Osaka University, Japan)
Naveed Ul Hassan (Lahore University of Management Sciences, Pakistan)
Woon-Seng Gan (NTU, Singapore)
Isao Echizen

Tutorial Co-Chairs

Yo-Sung Ho (Gwangju Institute of Science and Technology, Korea)
Hsueh-Ming Hang (National Chiao-Tung University)
Kong-Aik Lee (NEC, Japan)

Plenary Co-Chairs

Helen Meng (Chinese University of Hong Kong)
Kai Yu (Shanghai Jiaotong University)
Waleed Abdulla (University of Auckland, New Zealand)

Overview Session Co-Chairs

Homer Chen (National Taiwan University)
Nam Ik Cho (Seoul National University, Korea)
Yoong Choon Chang (University Tunku Abdul Rahman, Malaysia)
Guan-Ming Su

Special Session Co-Chairs

Qin Jin (Renmin University of China)
Yoshinobu Kajikawa (Kansai University, Japan)
Chung-Nan Lee (National Sun Yat-sen University)
Lap Pui Chau (Nanyang Technological University, Singapore)
Hsin-Min Wang (Academia Sinica)
Kosin Chamnongthai (King Mongkut's University of Technology Thonburi, Thailand)
Askar Hamdulla (Xinjiang University)

Publicity Co-Chairs

Changchun Bao (Beijing University of Technology)
Ce Zhu (University of Electronic Science and Technology of China)
Ambikairajah Eliathamby (UNSW, Australia)
Tan Lee (Chinese University of Hong Kong)
Woon-Seng Gan (NTU, Singapore)
Koichi Shinoda (Tokyo Institute of Technology)
Yo-sung Ho (Gwangju Institute of Science Technology, Korea)
Ming-Ting Sun (University of Washington, USA)

Publication Co-Chairs

Tatsuya Kawahara (Kyoto University, Japan)
Jiangyan Yi (Chinese Academy of Science)

Local Chairs

Yonghong Li (Northwest Minzu University)
Guanyu Li (Northwest Minzu University)
Yangping Wang (Lanzhou Jiaotong University)
Shipeng Xu (Gansu Institute of Political Science and Law)

Finance Chairs

Mingxing Xu (Tsinghua University)
Zhang Li (Tsinghua University)

Secretary / Web Co-Chairs

Lantian Li (Tsinghua University)
Zhiyuan Tang (Tsinghua University)
Yating Peng (Tsinghua University)

Industrial Forum Co-Chairs

Mingyi He (Northwestern Polytechnical University)
Lei Jia (Baidu)
Ming Zhou (Microsoft Research Asia)

Registration Co-Chairs

Mingxing Xu (Tsinghua University)
Askar Rozi (Tsinghua University)
Xunying Liu

Sponsorship Co-Chairs

Lei Xie (Northwest Polytechnic University)
Hui Bu (AI Shell)

Keynote Speeches

Transfer Learning: from Bayesian Adaptation to Teacher-Student Modeling

Dr. Chin-Hui Lee

School of ECE, Georgia Tech, USA

Time: Tuesday, Nov 19, 9:00-10:00

Place: Grand Ballroom, Crowne Plaza Hotel, 3F

Chair: Haizhou Li, Professor, National University of Singapore

Abstract

Transfer learning is referred to as a process of distilling knowledge learned in one task and utilizing it in another related task. In machine learning, transfer learning and domain adaptation are often synonymous, and they are designed to combat catastrophic forgetting of not remembering much of what had already been learned in the transfer process. When using generative models, such as probability distributions to characterize observed data with a set of parameters to be transferred, a Bayesian formulation is often adopted to combine knowledge summarized in prior distributions of the parameters and likelihood of newly observed adaptation data to establish a posterior distribution of the parameters to be optimized. Recently we had extended Bayesian adaptation to discriminative models, such as deep neural networks, and obtained a similar effectiveness. Another emerging approach, known as teacher-student (T-S) modeling, is to summarize what had been learned in a teacher model and what to be transferred to in a student model with similar or different architectures. An objective function characterizing the discrepancies between behaviors of the teacher and student models is then optimized for the student model on a set of adaptation data. Generative adversarial networks have also been used to preform adaptation data augmentation. Such a T-S learning framework facilitates a versatile variety of scenarios and applications. In this talk, we will present technical dimensions in transfer learning and highlight its potential opportunities.

Speaker's Biography

Chin-Hui Lee is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had accumulated 20 years of industrial experience ending in Bell Laboratories, Murray Hill, as a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the IEEE and a Fellow of ISCA. He has published over 500 papers and 30 patents, with more than 45,000 citations and an h-index of 80 on Google Scholar. He received numerous awards, including the Bell Labs President's Gold Award in 1998. He won the SPS's 2006 Technical Achievement Award for “Exceptional Contributions to the Field of Automatic Speech Recognition”. In 2012 he gave an ICASSP plenary talk on the future of automatic speech recognition. In the same year he was awarded the ISCA Medal in scientific achievement for “pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition”.

Digital Retina – Improvement of Cloud Artificial Vision System from Enlighten of HVS Evolution.

Prof. Wen Gao

Department of Computer Science and Technology, Peking University, China

Time: Wednesdsay, Nov 20, 9:00-10:00

Place: Multi-function Office, GICC 4F

Chair: Hitoshi Kiya, Professor, Tokyo Metropolitan University

Abstract

Edge computing is hop topics recently, and the smart city wave seems to be making more and more video devices in cloud vision system upgraded from traditional video camera into edge video device. However, there are some arguments on how much intelligence the device should be with, and how much the cloud should keep. Human visual system (HVS) took millions of years to reach its present highly evolved state, it might not be perfect yet, but much better than any of exist computer vision system. Most artificial visual system are consisted of camera and computer, like eye and brain for human, but with very low level pathway between two parts, comparing to human being. The pathway model of human being between eye and brain is quite complex, but energy efficient and comprehensive accurate, evolved by natural selection. In this talk, I will discuss a new idea about how we can improve the cloud vision system by HVS-like pathway model, which is called digital retina, to make the cloud vision system being more efficient and smart. The digital retina is with three key features, and the detail will be given in the talk.

Speaker's Biography

Wen Gao now is a Boya Chair Professor at Peking university. He also serves as the president of China Computer Federation (CCF) from 2016. He received his Ph.D. degree in electronics engineering from the University of Tokyo in 1991. He joined with Harbin Institute of Technology from 1991 to 1995, and Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS) from 1996 to 2005. He joined the Peking University since 2006. Prof. Gao works in the areas of multimedia and computer vision, topics including video coding, video analysis, multimedia retrieval, face recognition, multimodal interfaces, and virtual reality. His most cited contributions are model-based video coding and feature-based object representation. He published seven books, over 280 papers in refereed journals, and over 700 papers in selected international conferences. He is a fellow of IEEE, a fellow of ACM, and a member of Chinese Academy of Engineering.

Applying Deep Learning in Non-native Spoken English Assessment

Dr. Kate Knill

Automatic Language Teaching and Assessment Institute (ALTA), Cambridge University, UK

Time: Thursday, Nov 21, 9:00-10:00

Place: Multi-function Office, GICC 4F

Chair: Hongwu Yang, Professor, Northwest Normal University

Abstract

Over 1.5 billion people worldwide are using and learning English as an additional language. This has created a high and growing demand for certification of learners' proficiency, for example for entry to university or for jobs. Automatic assessment systems can help meet this need by reducing human assessment effort. They can also enable learners to monitor their progress with informal assessment when and wherever they choose. Traditionally automatic speech assessment systems were based on read speech so what the candidate said was (mostly) known. To properly assess a candidate's spoken communication ability, however, the candidate needs to be assessed on free, spontaneous, speech. The text is, of course, unknown in such speech, and we don't speak in fluent sentences. we hesitate and stop and restart. Added to this any automatic system has to handle a wide variety of accents and pronunciations for learners across first languages and highly variable audio recording quality. Together this makes non-native spoken English assessment a challenging problem. To help meet the challenge deep learning has been applied to a number of sub-tasks. This talk will look at some examples of how deep learning is helping to create automatic systems capable of free speaking spoken English assessment. These will include: 1) efficient ASR systems, and ensemble combination, for non-native English; 2) prompt-response relevance for off-topic response detection; 3) task-specific phone “distance” features for assessment and L1 detection; 4) grammatical error detection and correction for learner English. Deep learning techniques used in the above, include: recurrent sequence models; sequence ensemble distillation (teacher-student training); attentions mechanisms; and Siamese networks.

Speaker’s Biography

Dr. Kate Knill is a Principal Research Associate at the Department of Engineering and the Automatic Language Teaching and Assessment Institute (ALTA), Cambridge University. Kate was sponsored by Marconi Underwater Systems Ltd for her 1st class B.Eng. (Jt. Hons) degree in Electronic Engineering and Maths at Nottingham University and a PhD in Digital Signal Processing at Imperial College. She has worked for 25 years on spoken language processing, developing automatic speech recognition and text-to-speech synthesis systems in industry and academia. As an individual researcher and a leader of multi-disciplinary teams as Languages Manager, Nuance Communications, and Assistant Managing Director, Toshiba Research Europe Ltd, Cambridge Research Lab, she has developed speech systems for over 50 languages and dialects. Her current research focus is on applications for non-native spoken English language assessment and learning and detection of speech and language disorders. She is Secretary of the International Speech Communication Association (ISCA) and a member of the Institution of Engineering and Technology (IET) and Institute of Electrical and Electronic Engineers (IEEE).

Special Sessions

APSIPA ASC 2019 features a special sessions track for emerging research topics that ran throughout the conference in parallel to other tracks. Special sessions provide the opportunity for a more in-depth look at the subject matter presented. APSIPA ASC 2019 strongly encourages organization of Special Sessions to complement regular programs and to help bring together leading researchers and engineers from around the world to present the state-of-the-art research works. Special Session proposals may cover topics that relate to regular technical program tracks of APSIPA ASC 2019, but are not limited to those. Proposals were reviewed by the special session committee based on the relevance, innovation, content and quality. All papers submitted to the special sessions were went through the same review process as regular papers. The special session organizers and all invited speakers were required to register for the conference.

Special Session List

No.	Target track	Name	Organizer	Contact
1	BioSiPS	Signal Processing in Behavior Analysis	Kazushi Ikeda, Li-Wei Kang	kazushi@is.naist.jp
2	IVM	Advanced Topics on High-dimensional Data Analytics and Processing	Supavadee Aramvith, Shogo Muramatsu	supavadee.a@chula.ac.th, shogo@eng.niigata-u.ac.jp
3	IVM	Technologies for A Maximized Experience of Multi-dimensional Content: from 2D, 3D Modeling to An Objective Assessment	Sanghoon Lee, Chia-Hung Yeh	slee@yonsei.ac.kr, chyeh@ntnu.edu.tw
4	IVM	Recent Trends in Computational Intelligence	Chern Hong Lim, Mei Kuan Lim	lim.chernhong@monash.edu
5	IVM	High Performance Image and Video Processing and Applications	Kosin Chamnongthai	kosin.cha@kmutt.ac.th, kosin.chamnongthai.2017@gmail.com
6	IVM	Multi-source Data Processing and Analysis: Models, Methods and Applications	Ping Han, Qiuping Jiang, Runmin Cong, Chongyi Li	runmincong@gmail.com
7	Machine Learning	Machine Learning for Small-sample Data Analysis	Zhanyu Ma, Jen-Tzung Chien, Ruiping Wang, Xiaoxu Li	mazhanyu@gmail.com
8	MSF	Recent Advances in Fingerprinting and Data Hiding	Minoru Kuribayashi,	kminoru@okayama-u.ac.jp

			David Megías Jiménez	
9	MSF	Signal Processing for Big data and Its Security	Yuhong Liu	yhliu@scu.edu
10	MSF	Information Security for Digital Content	Xiangui Kang, KokSheik Wong,Linna Zhou	isskxg@mail.sysu.edu.cn, wong.koksheik@monash.edu, zhoulinna@tsinghua.edu.cn
11	MSF	Deep Generative Models for Media Clones and Its Detection	Fuming Fang, Zhenzhong Kuang , Xin Wang	fang@nii.ac.jp, zzkuang@nanase.comm.eng.osaka-u.ac.jp, wangxin@nii.ac.jp
12	MSF	Recent Advances in Biometric Security	Koichi Ito, Tetsushi Ohki	ito@aoki.ecei.tohoku.ac.jp, ohki@inf.shizuoka.ac.jp
13	SIPTM	Recent Trends in Signal Processing & Machine Learning - Acoustic & Biomedical Applications	Kiyoshi Nishikawa, Felix Albu, Akhtar, Muhammad Tahir	knishikawa@m.ieice.org, felix.albu@valahia.ro, akhtar@ieee.org
14	SIPTM	Latest Progress on Fractional Signal Processing Theory and Applications	Bing-Zhao Li, Xiaolong Chen,Yong Guo	li_bingzhao@bit.edu.cn
15	SIPTM	Special Learning under Limited Samples Scenarios	Ganggang Dong, Yinghua Wang	dongganggang@xidian.edu.cn
16	SIPTM	Signal Processing for Crowd Science	H. Vicky Zhao, Yan Chen	vzhao@tsinghua.edu.cn
17	SLA	Recent Advances in Speaker Recognition, Speaker Diarization and Language Recognition	Qingyang Hong, Lin Li	qyhong@xmu.edu.cn, lilin@xmu.edu.cn
18	SLA	Robust Rich Audio Analysis	Xiao-Lei Zhang	xiaolei.zhang@nwpu.edu.cn
19	SLA	Advanced Signal Processing and Machine Learning for Audio and Speech Applications	Shoji Makino, Hiroshi Saruwatari	maki@tara.tsukuba.ac.jp, hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp
20	SLA	Multilingual Speech and Language Processing	Zhiyuan Tang, Dong Wang, GuanyuLi, Mijiti Ablimit	tangzy@cslt.org, wangdong99@mails.tsinghua.edu.cn, guanyu-li@163.com, mijit@xju.edu.cn
21	SLA	Second Language Speech Perception and Production	Ying Chen, Jian Gong	ychen@njust.edu.cn
22	SLA	Recent Topics on Signal and	Yoshinobu	kaji@kansai-u.ac.jp,

		Information Processing for Active Control of Sound	Kajikawai, Chuang Shi	shichuang@uestc.edu.cn
23	SPS	Advanced Signal Processing for 5G Communication	Na Chen, Minoru Okada	chenna@is.naist.jp, 381725806@qq.com
24	SPS	Lightweight Signal Processing and Machine Learning for Embedded Applications	Hakaru Tamukoh, Hiroshi Tsutsui	tamukoh@brain.kyutech.ac.jp, hiroshi.tsutsui@ist.hokudai.ac.jp
25	SPS	High Performance Video Processing and Image Identification	Junyong Deng	djy@xupt.edu.cn
26	SPS	Deep Learning Systems for Cloud, Fog, and Edge Computing, and Applications	Jia-Ching Wang	jcw@csie.ncu.edu.tw
27	WCN	Physical and Wireless Environment Recognition Based on Signal Processing	Osamu Takyu	takyu@shinshu-u.ac.jp
28	WCN	Convergence of 5G with SDN/NFV/Cloud	Po-Chiang Lin, Wen-Ping Lai	pclin@saturn.yzu.edu.tw, wpl@saturn.yzu.edu.tw

Session Index

18th November 2019 (Monday) - Day 0

19th November 2019 (Tuesday) - Day 1

20th November 2019 (Wednesday) - Day 2

21th November 2019 (Thursday) - Day 3

Tutorials

Tensor Component Analysis

Prof. Yipeng Liu

Time: Monday, Nov 18, 9:30-11:30

Place: A1

Speech Enhancement based on Deep Learning and Intelligibility Evaluation

Dr. Yu Tsao and Dr. Fei Chen

Time: Monday, Nov 18, 9:30-11:30

Place: A2

Fundamental and Progress of Deep Learning Based Statistical Parametric Speech Synthesis

Prof. Zhenhua Ling

Time: Monday, Nov 18, 13:30-15:30

Place: A1

Few-Shot Learning, Adversarial Learning, and Their Applications

Prof. Jen-Tzung Chien and Prof. Zhanyu Ma

Time: Monday, Nov 18, 13:30-15:30

Place: A2

Fundamental and Progress of Immersive Visual Media Communications

Prof. Yo-Sung Ho

Time: Monday, Nov 18, 16:00-18:00

Place: A1

Signal enhancement for consumer products

Dr. Akihiko Sugiyama (a.k.a. Ken Sugiyama)

Time: Monday, Nov 18, 16:00-18:00

Place: A2

Abstract

TUE-AM1-SS1
Convergence of 5G with SDN/NFV/Cloud

Time: Tuesday, Nov 19, 10:20-12:00

Place: A2

Chairs: Po-Chiang Lin, Wen-Ping Lai

TUE-AMI-SS1.1: Dynamic Threshold for DDoS Mitigation in SDN Environment

Guo-Chih Hong, Chung-Nan Lee and Ming-Fneg Lee

National Sun Yat-sen University

Software-Defined Networking (SDN) is one of the key technologies of 5th generation mobile networks (5G). However, like the traditional network architecture, SDN is also vulnerable to the Distributed Denial of Service (DDoS) attack. This paper explores the dynamic threshold for DDoS attack in the SDN environment. Through the characteristics of SDN, we propose a feasible DDoS detection and defense mechanism. The proposed mechanism calculates the entropy of the network environment by the collected traffic status, and derives a dynamic threshold according to the network conditions to determine whether the environment is subject to DDoS attacks. In the event of a DDoS attack, the proposed mechanism discards the traffic from the malicious nodes to the victim nodes with a flow entry. In addition, if no DDoS attacks occur in the environment, the proposed system can disperse the traffic of the SDN switch, thereby balance the traffic load in the environment.

TUE-AMI-SS1.2: Large-Scale and High-Dimensional Cell Outage Detection in 5G Self-Organizing Networks

Po-Chiang Lin

Yuan Ze University

In this paper, we investigate the cell outage detection in Self-Organizing Networks. The purpose of cell outage detection is to automatically detect whether there exist some failures or degradation in the base stations, such that users could not obtain mobile services, or the obtained mobile services do not fulfill their requirements. The cell outage detection in 5G is with great challenge. The deployment of future 5G mobile communication networks would be heterogeneous and ultra-dense. The mobile communication environments are very complicated. They include the multipath transmission, fading, shadowing, interference, and so on. Users' mobility and usage pattern also vary. In such environments, the mobile data would be large-scale and high-dimensional. Traditional small-scale and low-dimensional anomaly detection methods would be unsuitable. Moreover, operational mobile communication networks should be normal almost all the time. Cell outage would be seldom. Therefore, the normal data and anomaly data would be imbalanced. In this paper, we formulate the cell outage detection problem as an anomaly detection problem. We propose an cell outage detection method using the autoencoder, which is a neural network that is trained by unsupervised learning. The network could be trained in advance even when the cell outage data is still not available. Moreover, the autoencoder is also useful for denoising. This proposed method could thus automatically detect the cell outage in complicated and time-varying mobile wireless communication environments. Comprehensive system-level simulations validate the performance of the proposed method.

TUE-AMI-SS1.3: Implementation of multiple routing configurations on software-defined networks with P4

Kouji Hirata and Takuji Tachibana

Kansai University, University of Fukui

In order to maintain high availability of communication networks, we should recover failures immediately after the failures occur. As one of techniques achieving such fast failure recovery, multiple routing configurations (MRC) have been proposed. In MRC, multiple backup routing configurations are prepared in advance to fast repair a single link/node failure. When a failure occurs during data transmission using a normal routing configuration, MRC changes the routing configuration to another routing configurations that does not use the failed point. Thus, MRC can realize fast recovery within few tens of milliseconds and continue the data transmission. In this paper, we implement MRC on software-defined networks with P4 (Programming Protocol-independent Packet Processeros). P4 is a programming language that enables us to define the behavior of the data plane of network equipment. We examine the MRC implementation, using Mininet.

TUE-AMI-SS1.4: Evaluation of countermeasure against future malware evolution with deterministic modeling

Koki Shimizu, Yuya Kumai, Kimiko Motonaka, Tomotaka Kimura and Kouji Hirata

Kansai University, Doshisha University

Recently, machine learning technologies have dramatically evolved. Accordingly, the concept of self-evolving botnets has been introduced, which discover vulnerabilities of hosts by distributed machine learning using the computational resources of infected hosts, and infect other hosts by attacks using the discovered vulnerabilities. The infectability of the self-evolving botnets is too strong compared with conventional botnets, so that such new botnets will become the serious threat to future network society including 5G and IoT environments. In this paper, we consider a volunteer model that discovers unknown vulnerabilities earlier than self-evolving botnets by distributed computing using volunteer hosts' resources and repairs the vulnerabilities. We propose deterministic modeling for the volunteer model. Through numerical calculations, we evaluate the performance of the volunteer model against self-evolving botnets.

TUE-AMI-SS1.5: NUMAP: NUMA-aware Multi-core Pinning and Pairing for Network Slicing at the 5G Mobile Edge

Wen-Ping Lai and Kuan-Chun Chiu

Yuan Ze University

Based on the concept of network functions virtualization (NFV) by adopting the virtualized micro-service-based event-driven model, this paper studies the system resource allocation

problems of deploying network/service slices on the mobile edge server, performing as a pivotal control/data/information hub in between radio-access and core networks or even the Internet clouds, in the coming era of 5G digital transformation. A non-uniform memory access (NUMA)-aware multi-core pinning-and-pairing method, called NUMAP, for network/service slicing with respect to different traffic levels is proposed to improve the system performance of a light-weight EPC slice (vEPCLW) on top of an x86 Dell PowerEdge Server (R740) for the MEC cloudlet platform. This server is equipped with two CPU sockets, each containing 12 physical cores sharing the same local memory bank, denoted as a NUMA node; namely remote memory access time to another NUMA node is longer than the local one. The novelty of the proposed NUMAP method lies in the fact that it is aware of three important pairing schemes based on their pairing distances, namely inter-node-pair, intra-node-pair and hyper-threaded-pair, and NUMAP can thus serve as a New Map for multi-core assignment. Preliminary experimental results show that the NUMAP algorithm outperforms the default one which is based on symmetric multi-processing (SMP).

TUE-AM1-SS2
Signal Processing for Big Data and Its Security

Time: Tuesday, Nov 19, 10:20-12:00

Place: A3

Chair: Yuhong Liu

TUE-AMI-SS2.1: Approach using Transforming Structural Data into Image for Detection of Malicious MS-DOC Files based on Deep Learning Models

Shaojie Yang, Wenbo Chen, Shanxi Li and Qingxiang Xu

Lanzhou University

Malicious MS-DOC file has a long history in cybersecurity and has rapid growth with tremendous appearance of advanced persistent threat (APT) attacks. Due to its obfuscation and complexities, regular detection methods are not ideal, and the specific detection methods are limited, either. This paper presents a new approach for malware detection of MS-DOC files. Inspired by analysis of MS-DOC files and tremendous success made by convolutional neural network (CNN) in the field of feature identification, especially image identification, a new approach including data extraction and conversion is designed to identify MS-DOC malicious files and benign files. Based on three CNN models, experiment results show that the accuracy rate of detection for test dataset reaches 94.09%, and in simulated zero-day malware detection experiment, the average accuracy rate reaches 94.70%. The approach proves the feasibility of MS-DOC malicious file detection based on convolutional neural network and proposes a new idea to detect zero-day MS-DOC malware.

TUE-AMI-SS2.2: Edge Mining on IoT Devices Using Anomaly Detection

Kavin Kamaraj, Behnam Dezfouli and Yuhong Liu

Santa Clara University

With continuous monitoring and sensing, millions of Internet of Things sensors all over the world generate tremendous amounts of data every minute. As a result, recent studies start to raise the question as whether to send all the sensing data directly to the cloud (i.e., direct transmission), or to preprocess such data at the network edge and only send necessary data to the cloud (i.e., preprocessing at the edge). Anomaly detection is particularly useful as an edge mining technique to reduce the transmission overhead in such a context when the frequently monitored activities contain only a sparse set of anomalies. This paper analyzes the potential overhead-savings of machine learning based anomaly detection models on the edge in three different IoT scenarios. Our experimental results prove that by choosing the appropriate anomaly detection models, we are able to effectively reduce the total amount of transmission energy as well as minimize required cloud storage. We prove that Random Forest, Multilayer Perceptron, and Discriminant Analysis models can viably save time and energy on the edge device during data transmission. K-Nearest Neighbors, although reliable in terms of prediction accuracy, demands exorbitant overhead and results in net time and energy loss on the edge device. In addition to presenting our model results for the different IoT scenarios, we provide guidelines for potential model selections through analysis of involved tradeoffs such as training overhead, prediction overhead, and classification accuracy.

TUE-AMI-SS2.3 A Hybrid Feature Selection Algorithm Applied to High-dimensional Imbalanced Small-sample Data Classification

Fang Feng, Qingquan Lv, Mingsong Wang, Xuhui Yang, Qingguo Zhou and Rui Zhou

State Grid Gansu Electric Power Research Institute, School of Information Science and Engineering, Lanzhou University

With the rapid development of microarray technology and interdisciplinary science, it is possible for microarray technology to be used to predict diseases. Microarray technology has the advantages of high speed, high efficiency and reliability in disease prediction. However,microarray data not only has the characteristics of high dimension, small sample size, but also often with imbalance problems, which brings a lot of difficulties to researchers. In view of the above problems, it is proposed in this paper a Filter-Wrapper hybrid feature selection algorithm Union Information Gini Cost-sensitive Feature selection General Vector Machine (UIG-CFGVM) to tackle the high-dimensional imbalanced small-sample problem.The improved hybrid algorithm is as follows: Firstly, the most common features are removed by the proposed hybrid filter algorithm UIG, which is obtained by Information Gain (Info)and Gini Index (Gini). Secondly, Cost-sensitive Feature selection General Vector Machine (CFGVM) is used as Wrapper method to further improve the performance of the algorithm. The experimental results show that the

proposed algorithm UIG-CFGVM has better classification performance in seven biomedical high-dimensional imbalanced small-sample datasets compared with other similar algorithms.

TUE-AMI-SS2.4: Blockchain-based Complete Self-tallying E-voting Protocol

Yikang Lin and Peng Zhang

ATR Key Laboratory of National Defense Technology

Electronic voting (E-voting) protocol is that voters can vote according to their wishes, and then the voting authority is responsible for collecting the votes and counting the final voting result.With the development of Blockchain, we tend to combine it with E-voting and propose Blockchain-based complete self-tallying E-voting protocol, which is more secure than E-voting protocol based on centralized servers.In our protocol, Blockchain acts as bulletin board, and ``Efficient One-out-of-T" zero knowledge proof (ZKP) is proposed to support multi-candidate voting. Moreover, the issues of abortive and adaptive are solved in our protocol.The security analysis shows that our protocol meets the security requirements, and it can be applied to small-scale and anonymous private scenario such as Corporate Board Voting. The performance analysis demonstrates that the proposed ZKP has low time consumption.

TUE-AMI-SS2.5: Image Compression with Deeper Learned Transformer

Licheng Xiao, Hairong Wang and Nam Ling

Santa Clara University

Deep learning is known for its flexibility and infinite potential to approximate any function. Is it possible to approximate image compression using deep learning? The answer is yes. This article compares three major deep learning techniques used in image compression now and proposed an approach with deeper learned transformer and improved optimization goal, which achieved improved peak signal-to-noise ratio (PSNR) and multi-scale structural similarity (MS-SSIM) under very low bits per pixel (bpp). Experimental results show that the proposed approach outperformed BPG (RGB 4:4:4) in natural scene images compression, and is capable to handle arbitrary image shapes, which makes it applicable to practical image compression workloads.

TUE-AMI-SS2.6: Modeling the Views of WeChat Articles by Branching Processes

Lin Zhang and Yuhong Liu

Beijing University of Posts and Telecommunications, Santa Clara University

In cyber security, the temporal patterns of information propagation in online social platforms is of crucial importance, especially for false information detection and defense. The mechanism of information propagation is the fundamental of online security, such as predicting, promoting or suppressing information dissemination. The dynamics of information popularity in online social networks opens the possibilities for understanding the mechanism of information spreading. In this paper, we study the temporal patterns associated with online contents generated by WeChat subscription accounts. In order to reveal the popularity dynamics of WeChat articles, we formulate a mathematical model using branching process to reveal the mechanism of views of WeChat articles. Specifically, a non-Markovian age-dependent branching process is introduced for the purpose of considering patterns of human behavior and the tree-like dissemination structures. Different from the Markovian case with exponentially distributed lifetime of particles in the branching system, the heavy-tailed power law distributed inter event time is one of the key ingredients for our model. Moreover, the limitation of attention is also considered in our model. We demonstrate our approach on the real data of WeChat articles' popularity time series. The branching model is successful in presenting the temporal patterns of the real evolution. Our findings offer insights into the temporal patterns of information popularity in online social networks, which provides references for further prediction and control of information propagation concerning cyber security.

TUE-AM1-SS3
Machine Learning for Small-sample Data Analysis

Time: Tuesday, Nov 19, 10:20-12:00

Place: A4

Chairs: Zhanyu Ma, Jen-Tzung Chien, Ruiping Wang, Xiaoxu Li

TUE-AMI-SS3.1: Lightweight models for weather identification

Congcong Wang, Pengyu Liu, Kebin Jia and Siwei Chen

Beijing University of Technology

At present, the recognition of weather phenomena mainly depends on the weather sensors and the weather radar. However, large-scale deployment of meteorological observation equipment for intensive weather monitoring is difficult because it is expensive and difficult to maintain. Moreover, convolutional neural networks (CNNs) can also be used to identify weather phenomena, but existing methods require high computing power of equipment, making it difficult to deploy in practice. Therefore, designing a lightweight model that can be deployed in a small device with weak computing power is crucial for intensive weather monitoring. In this paper, we study the shortcomings of some existing lightweight models. By comparing the disadvantages of these models, a new lightweight model is proposed. In addition, considering the number of existing weather datasets are too small to meet real monitoring needs, so we produced a dataset with a more complex variety of weather phenomena. Through the experiments, the proposed method can save more than 25 times memory usage with only 1.55% accuracy lost compared with the best CNNs method which achieves state-of-the-art performance among the other lightweight models.

TUE-AMI-SS3.2: Stochastic Fusion for Multi-stream Neural Network in Video Classification

Yu-Min Huang, Huan-Hsin Tseng and Jen-Tzung Chien

University of Michigan, National Chiao Tung University

Spatial image and optical flow provide complementary information for video representation and classification. Traditional methods separately encode two stream signals and then fuse them at the end of streams. This paper presents a new multi-stream recurrent neural network where streams are tightly coupled at each time step. Importantly, we propose a stochastic fusion mechanism for multiple streams of video data based on the Gumbel samples to increase the prediction power. A stochastic backpropagation algorithm is implemented to carry out a multi-stream neural network with stochastic fusion based on a joint optimization of convolutional encoder and recurrent decoder. Experiments on UCF101 dataset illustrate the merits of the proposed stochastic fusion in recurrent neural network in terms of interpretation and classification performance.

TUE-AMI-SS3.3: Dynamic Attention Loss for Small-sample Image Classification

Jie Cao, Yinping Qiu, Dongliang Chang, Xiaoxu Li and Zhanyu Ma

Lanzhou University of Technology, Beijing University of Posts and Telecommunications

Convolutional Neural Networks (CNNs) have been successfully used in various image classification tasks and gradually become one of the most powerful machine learning approaches nowadays. To improve the capability of model generalization and performance on small-sample image classification, a new trend is learning discriminative features via CNNs. The idea of this paper is to decrease the confusion between categories to extract discriminative features and enlarge inter-class variance, especially for classes which have indistinguishable features. The idea of this paper is to decrease the confusion between categories to extract discriminative features and enlarge inter-class variance, especially for classes which have indistinguishable features. In this paper, we propose a loss function named "Dynamic Attention Loss" (DAL), introducing confusion rate-weighted soft label (target) as the controller of similarity measurement of two classes. The similarity of DAL is dynamically evaluated during each iteration to adapt the model. Experimental results demonstrate that compared with Cross-Entropy Loss and Focal Loss, the proposed DAL achieved a better performance on LabelMe dataset and Caltech101 dataset.

TUE-AMI-SS3.4: Mixed Attention Mechanism for Small-Sample Fine-grained Image Classification

Xiaoxu Li, Jijie Wu, Dongliang Chang, Zhanyu Ma, Jie Cao and Weifeng Huang

Lanzhou University of Technology, Beijing University of Posts and Telecommunications, South–to–North Water Diversion Middle Route Information Technology Co.

Fine-grained image Classification is an important task in computer vision. The main challenge of the task are that intra-class similarity is large and that training data points in each class are insufficient for training a deep neural network. Intuitively, if we can learn more discriminative features and more detailed features from fined-grained images, the classification performance can be improved. Considering that channel attention can learn more discriminative features, spatial attention can learn more detailed features, this paper proposes a new spatial attention mechanism by modifying Squeeze-and-Excitation block, and a new mixed attention by combining the channel attention and the proposed spatial attention. Experimental results on two small-sample fine-grained image classification datasets demonstrate that on both VGG16 network and ResNet-50 network, the proposed two attention mechanisms achieve good performance, and outperform other referred fine-grained image classification methods.

TUE-AMI-SS3.5: A Loss With Mixed Penalty for Speech Enhancement Generative Adversarial Network

Jie Cao, Yaofeng Zhou, Hong Yu, Xiaoxu Li, Zhanyu Ma and Dan Wang

Lanzhou University of Technology, Ludong University, Beijing University of Posts and Telecommunications, South–to–North Water Diversion Middle Route Information Technology Co.

Speech enhancement based on generative adversarial network can overcome the problems of many classical speech enhancement methods, such as relying on the first-order statistics of the signal and ignoring the phase mismatch between the noisy and the clean signal. However, GANs are hard to train and have the vanishing gradients problem which may lead to poor generated samples.In this paper, we propose to use relativistic average generative adversarial network with a mixed penalty for speech enhancement. The mixed penalty can minimize the distance between generations and the clean samples more effectively. Experimental results on Valentini 2016 and Valentini 2017 dataset show that the proposed loss can make the training of neural network be more stable, and achieves good performance in both objective and subjective evaluation of speech quality.

TUE-AMI-SS3.6: Small-smaple Image Classification by Combining Prototype and Margin Learning

Xiaoxu Li, Liyun Yu, Dongliang Chang, Zhanyu Ma, Jie Cao and Nian Liu

Lanzhou University of Technology, Beijing University of Posts and Telecommunications, South–to–North Water Diversion Middle Route Information Technology Co.

Image classification is a fundamental and important task in the field of computer vision and artificial intelligence. In recent years, the image classification based on deep learning on large-scale dataset has made breakthrough progress. However, image classification with small-sample training data still exits big challenge. The main difficulty is that deep neural network overfit easily small-sample data and has big variance. Ensemble learning is a good way to overcome overfitting and reduce the variance of model, however, the existing ensemble methods based on deep neural network still could overfit on small-sample image data due to big randomness of deep neural network. In this paper, we propose a new ensemble method for small-sample image classification. The proposed method conclude two branches, in which one branch is a classifier base on prototype learning, and the other is a classifier base on margin learning. The experimental results on two small-sample image datasets, the LabelMe dataset and the Caltech101 dataset, show that the proposed method has better performance and performs more stably than other referred methods.

TUE-AMI-SS3.7: Recurrent Neural Network for Web Services Performance Forecasting, Ranking and Regression Testing

Muhammad Hasnain, Chern Hong Lim, Muhammad Fermi Pasha and Imran Ghani

Monash University, Indiana University of Pennsylvania

Accurate estimation of web services performance, which is critical to ensure the consumers satisfaction on web services is still a challenging task due to the dynamic, and personalized requirements of different individuals. Efficient estimation of web services performance can lead to a better ranking of webservices.

Regression testing is then performed on the ranked web services to ensure that existing functionality of the web services is not impacted through evolution in the web services. Soft computing techniques are highly resource consuming, and more complex for practitioners. Moreover, they show complex approximation with a low propagation, which can be improved by using the advanced deep neural networks. Previously proposed web services performance estimation and analysis have been never considered from the deep neural network. To address the problem of efficient estimation of web services performance, gated recurrent unit (GRU) has been proposed with the use of time slice quality of service (QoS) data of web services. The GRU model can analyze QoS values obtained from different sets of users in different timestamps. The proposed approach has been evaluated on the web services dataset and comparison indicates that the proposed approach shows the better prediction and estimation than the state of the art approaches.

TUE-AM1-O1
Voice conversion

Time: Tuesday, Nov 19, 10:20-12:00

Place: A5

Chairs: Nobuaki Minematsu

TUE-AMI-O1.1: Non-parallel Voice Conversion with Controllable Speaker Individuality using Variational Autoencoder

Tuan Vu Ho and Masato Akagi

Japan Advanced Institute of Science and Technology

We propose a flexible non-parallel voice conversion (VC) system that is capable of both performing speaker adaptation and controlling speaker individuality. The proposed VC framework aims to tackle the inability to arbitrarily modify voice characteristics in the converted waveform of conventional VC model. To achieve this goal, we use the speaker embedding realized by a Variational Autoencoder (VAE) instead of one-hot encoded vectors to represent and modify the target voice's characteristics. Neither parallel training data, linguistic label nor time alignment procedure is required to train our system. After training on a multi-speaker speech database, the proposed VC system can adapt an arbitrary source speaker to any target speaker using only one sample from a target speaker. The speaker individuality of converted speech can be controlled by modifying the speaker embedding vectors; resulting in a fictitious speaker individuality. The experimental results showed that our proposed system is similar to conventional non-parallel VAE-based VC and better than the parallel Gaussian Mixture Model (GMM) in both perceived speech naturalness and speaker similarity; even when our system only uses one sample from target speaker. Moreover, our proposed system can convert a source voice to a fictitious target voice with well perceived speech naturalness of 3.1 MOS.

TUE-AMI-O1.2: SINGAN: Singing Voice Conversion with Generative Adversarial Networks

Berrak Sisman, Karthika Vijayan, Minghui Dong and Haizhou Li

National University of Singapore, I2R

Singing voice conversion (SVC) is a task to convert the source singer's voice to sound like that of the target singer, without changing the lyrical content. So far, most of the voice conversion studies mainly focus only on the speech voice conversion that is different from singing voice conversion. We note that singing conveys both lexical and emotional information through words and tones. It is one of the most expressive components in music and a means of entertainment as well as self expression. In this paper, we propose a novel singing voice conversion framework, that is based on Generative Adversarial Networks (GANs). The proposed GAN-based conversion framework, that we call SINGAN, consists of two neural networks: a discriminator to distinguish natural and converted singing voice, and a generator to deceive the discriminator. With GAN, we minimize the differences of the distributions between the original target parameters and the generated singing parameters. To our best knowledge, this is the first framework that uses generative adversarial networks for singing voice conversion. In experiments, we show that the proposed method effectively converts singing voices and outperforms the baseline approach.

TUE-AMI-O1.3: Experimental investigation on the efficacy of Affine-DTW in the quality of voice conversion

Gaku Kotani, Hitoshi Suda, Daisuke Saito and Nobuaki Minematsu

The University of Tokyo

In this paper, the performance of Affine-DTW, which performs appropriate time alignment between source and target features in voice conversion (VC), is experimentally and thoroughly investigated. In traditional VC, parallel data are often required to train a mapping model between source and target features. While VC with non-parallel data is also studied to avoid collecting parallel data, the quality of its converted speech is still inferior to the traditional one with parallel data. One approach to further progress in VC is exploiting both parallel and non-parallel data, the former of which is pre-stored and the latter of which is assumed to be easily collected. In this case, it is still worthwhile to study time-alignment techniques to obtain appropriate alignment of parallel data. Affine-DTW is a technique in which dynamic time warping (DTW) and coarse conversion based on affine transformation are iteratively performed. In Affine-DTW, time alignment and parameters of affine transformation can be analytically calculated so that it can be easily adopted as pre-processing in VC. However, the influence on the performance of trained models based on the obtained alignments has not been well investigated experimentally. Hence, this paper investigates the performance of Affine-DTW in terms of quality improvement of converted speech in traditional VC methods based on Gaussian mixture models,

non-negative matrix factorization and neural networks. Experimental results show that Affine-DTW obtains appropriate alignments and the naturalness improvement of converted speech in subjective assessments is observed in trained models based on the alignments.

TUE-AMI-O1.4: Non-parallel Many-to-many Singing Voice Conversion by Adversarial Learning

Jinsen Hu, Chunyan Yu and Faqian Guan

Fuzhou University

With the rapid development of deep learning, although speech conversion had made great progress, there are still rare researches in deep learning to model on singing voice conversion, which is mainly based on statistical methods at present and can only achieve one-to-one conversion with parallel training datasets. So far, its application is limited. This paper proposes a generative adversarial learning model, MSVC-GAN, for many-to-many singing voice conversion using non-parallel datasets. First, the generator of our model is concatenated by the singer label, which denotes domain constraint. Furthermore, the model integrates self-attention mechanism to capture long-term dependence on the spectral features. Finally, switchable normalization is employed to stabilize network training. Both the objective and subjective evaluation results show that our model achieves the highest similarity and naturalness not only on the parallel speech dataset but also on the non-parallel singing dataset.

TUE-AMI-O1.5: Evaluation of the Lombard effect model on synthesizing Lombard speech in varying noise level environments with limited data

Thuan Van Ngo, Rieko Kubo and Masato Akagi

Japan Advanced Institute of Science and Technology

Lombard speech is intelligible speech produced by humans in noises. In this study, we focus on mimicking Lombard speech from natural neutral speech under backgrounds with varying noise levels to increase its intelligibility in these noises. Other approaches map corresponding speech features from the neutral speech to Lombard speech, which can only apply for an individual noise level, and cannot reveal feature tendencies. Instead, we implement a Lombard effect model to continuously estimate feature values with varying noise levels. The techniques, which are based on coarticulation, a source-filter model with MRTD and spectral-GMM, are used to easily modify features of the neutral speech to obtain their tendencies. Finally, these features are synthesized by STRAIGHT vocoder to obtain Lombard speech. The mimicking quality is evaluated in subjective listening experiments on similarity, naturalness, and intelligibility. The evaluation results show that the proposed method could convert neutral speech into Lombard speech in varying noise levels, which obtains comparable results with the state-of-the-art method.

TUE-AMI-O1.6: DNN-based Voice Conversion with Auxiliary Phonemic Information to Improve Intelligibility of Glossectomy Patients' Speech

Hiroki Murakami, Sunao Hara and Masanobu Abe

Okayama University

In this paper, we propose using phonemic information in addition to acoustic features to improve the intelligibility of speech uttered by patients with articulation disorders caused by a wide glossectomy. Our previous studies showed that voice conversion algorithm improves the quality of glossectomy patients' speech. However, losses in acoustic features of glossec to my patients' speech are so large that the quality of the reconstructed speech is low. To solve this problem, we explored potentials of several additional information to improve speech intelligibility. One of the candidates is phonemic information, more specifically Phoneme Labels as Auxiliary input (PLA). To combine both acoustic features and PLA, we employed a DNN-based algorithm. PLA is represented by a kind of one-of-k vector, i.e., ”one-of-k” is not always ”1.0”, but has a weight value (<1.0) that gradually changes in time axis. The results showed that the proposed algorithm reduced the mel-frequency cepstral distortion for all phonemes, and almost always improved intelligibility. Notably, the intelligibility was largely improved in phonemes /s/ and /z/, mainly because the tongue is used to sustain constriction to produces these phonemes. This indicates that PLA works well to compensate the lack of a tongue.

TUE-AM1-O2
Speech Synthesis

Time: Tuesday, Nov 19, 10:20-12:00

Place: A6

Chair: Rohan Kumar Das

TUE-AMI-O2.1: Speech-like Emotional Sound Generator by WaveNet

Kento Matsumoto, Sunao Hara and Masanobu Abe

Okayama University

In this paper, we propose a new algorithm to generate Speech-like Emotional Sound (SES). Emotional information plays an important role in human communication, and speech is one of the most useful media to express emotions. Although, in general, speech conveys emotional information as well as linguistic information, we have undertaken the challenge to generate sounds that convey emotional information without linguistic information, which results in making conversations in human-machine interactions more natural in some situations by providing non-verbal emotional vocalizations. We call the generated sounds ``speech-like", because the sounds do not contain any linguistic information. For the purpose, we propose to employ WaveNet as a sound generator conditioned by only emotional IDs. The idea is quite different from WaveNet Vocoder that synthesizes speech using spectrum information as auxiliary features. The biggest advantage of the idea is to reduce the amount of emotional speech data for the training. The proposed algorithm consists of two steps. In the first step, WaveNet is trained to obtain phonetic features using a large speech database, and in the second step, WaveNet is re-trained using a small amount of emotional speech. Subjective listening evaluations showed that the SES could convey emotional information and was judged to sound like a human voice.

TUE-AMI-O2.2: DNN-based Statistical Parametric Speech Synthesis Incorporating Non-negative Matrix Factorization

Shunsuke Goto, Daisuke Saito and Nobuaki Minematsu

The University of Tokyo

This paper proposes a novel approach of DNN-based statistical parametric speech synthesis where non-negative matrix factorization (NMF) is effectively utilized. In statistical parametric speech synthesis, Mel-frequency cepstrum is often employed for acoustic features. However, it represents a spectral envelope as a linear combination of fixed envelope curves (sines and cosines), and the envelope predicted by a DNN-based acoustic model loses its fine structure. On the other hand, in NMF, multiple spectral envelopes (spectrogram) are decomposed into two factors; spectral bases and their activity patterns (activation). Since the obtained bases keep the fine structure of envelopes, the remaining factor, i.e. activation can be employed for acoustic features. Due to its sparseness, the spectral envelope obtained by the predicted activation also keeps fine structure. In this study, activation derived from NMF is utilized for spectral representation, and DNN-based text-to-speech synthesis incorporating NMF is proposed. In addition, this framework can potentially incorporate some applications of NMF, such as bandwidth expansion, voice conversion, or noise reduction. In this study, bandwidth expansion is achieved, and experimental results demonstrate that the proposed method can generate more natural spectral parameters especially in 48kHz sampling rate, and that 16kHz-to-48kHz bandwidth expansion, where natural synthetic speech is produced, is achieved in the proposed framework.

TUE-AMI-O2.3: Efficient quantization of vocoded speech parameters without degradation

Masanori Morise and Genta Miyashita

Meiji University, University of Yamanashi

This paper introduces quantization algorithms for 3 speech parameters: fundamental frequency (F0), spectral envelope, and aperiodicity. In full-band speech (speech with a sampling frequency above 40 kHz), the dimensions of the spectral envelope and the aperiodicity can be reduced to 50 and 5 dimensions based on previous studies. This paper compares the quantization coding without degradation with speech synthesized by the speech parameters without coding. Efficient quantization would be effective for a study that uses graphics processing unit (GPU) computing because recent GPUs support 16-bit floating-point computing. We did two subjective evaluations. The first evaluation determined the appropriate quantization bits in each speech parameter. We obtained the 9 bit values in F0, 13 bit values in the spectral envelope, and 3 bit values in the aperiodicity. The second evaluation verified the effectiveness of our proposed coding. The results showed that our proposed algorithm achieved almost all the same sound quality as the speech parameters without coding.

TUE-AMI-O2.4: Speaker-independent Spectral Mapping for Speech-to-Singing Conversion

Xiaoxue Gao, Xiaohai Tian, Rohan Kumar Das, Yi Zhou and Haizhou Li

National University of Singapore

Speech-to-Singing (STS) conversion aims at converting one's reading speech into his/her singing vocal. The prior work has mainly focused on transforming the prosody of speech to singing, however, there exist prominent differences between the spectra of speech and singing, which need to be transformed as well. In this paper, we propose to make use of parallel multi-speaker speak-sing data and i-vectors extracted from corresponding speech to train a speaker-independent spectral mapping model for STS conversion. The converted singing spectra are then used together with prosody features to synthesize the target singing. We investigate the effectiveness of i-vector based average model adaptation to model the differences between speech and singing spectra for a specific speaker. The proposed model does not require parallel speak-sing data from target speakers during training. The experimental results conducted on NUS-48E and NUS-HLT-SLS database indicate that the proposed approach significantly outperforms baselines in terms of both quality and similarity.

TUE-AMI-O2.5: A Prosodic Mandarin Text-to-Speech System Based on Tacotron

Chuxiong Zhang, Sheng Zhang and Haibing Zhong

Data Science Research Center, Jiangsu Jinling Science and Technology Group Limited

The Tacotron performs well in English speech synthesis and successfully aligns two arbitrary sequences from different domain in an automatic way. However, to introduce TacotronintoMandarinChineseText-to-Speech(TTS),aprosody system is needed for generating more natural speech. This paper proposes a practical method to involve the prosodic annotation into Tacotron training for Mandarin Chinese synthesis system. A prosody model predicting the prosodic boundaries from the given text serves as the front-end system in our approach, followed by a Tacotronsynthesissystemtrainedwithwell-labeledTTSdatabase containing the prosodic annotations. Under subjective evaluation in terms of the prosody, results show that the synthesis system performs better by adding the prosodic system as the front-end system for Tacotron.

TUE-AM1-O3
Speech Recognition

Time: Tuesday, Nov 19, 10:20-12:00

Place: A7

Chair: Hemant Patil

TUE-AMI-O3.1: Distilling Knowledge for Distant Speech Recognition via Parallel Data

Jiangyan Yi and Jianhua Tao

National Laboratory of Pattern Recognition， CASIA

In order to improve the performance of distant speech recognition tasks, this paper proposes to distill knowledge from the close-talking model to the distant model using parallel data. The close-talking model is called the teacher model. The distant model is called the student model. The student model is trained to imitate the output distributions of the teacher model. This constraint can be realized by minimizing the Kullback-Leibler (KL) divergence between the output distribution of the student model and the teacher model. Experimental results on AMI datasets show that the best student model achieves up to 8.5% relative word error rate (WER) reduction when compared with the conventionally-trained baseline models.

TUE-AMI-O3.2: Batch Normalization based Unsupervised Speaker Adaptation for Acoustic Models

Jiangyan Yi and Jianhua Tao

National Laboratory of Pattern Recognition， CASIA

This paper proposes a simple yet effective unsupervised speaker adaptation approach for batch normalization based deep neural network acoustic models. The basic idea of this approach is to recompute means and variances in all batch normalization layers over the test data for every speaker. Thus the distribution of the test data can be close to the training data. This approach doesn't need to adjust any trainable parameters of the acoustic model. Experiments are conducted on CHiME-3 datasets. The results show that the proposed adaptation obtains improvement on the real test set by 2.17 % relative average word error rate (WER) reduction when compared with the scaling and shifting factors (SSF) adaptation. Combining our proposed MV adaptation with the SSF adaptation obtains further improvement.

TUE-AMI-O3.3: Robust Speech Recognition based on Multi-Objective Learning with GRU Network

Ming Liu, Yujun Wang, Zhaoyu Yan, Jing Wang and Xiang Xie

Beijing Institute of Technology, Xiaomi Inc.

This paper proposes a new scheme to execute the task of speech enhancement (SE) for recognition based on multi-objective learning method which uses three objectives in the gated recurrent unit (GRU) network training procedure. The first objective is the main target for the expected SE task by directly mapping the noisy log-power spectrum (LPS) features to clean Mel-frequency cepstral coefficients (MFCC) features. The second one is an auxiliary target to help improving the main one by learning additional information from the back-end acoustic model (AM). The third one is also an auxiliary target achieved by learning some information from mapping noisy LPS to clean LPS. The two auxiliary structures could help the original structure to optimize the network parameters by correcting the errors. This approach imposes more constraints on direct feature mapping and information passing from the acoustic model to the network, enabling the enhanced network to better serve the AM. The experimental results show that the new multi-objective scheme with joint feature mapping and the posterior probability learning method improves the performance of SE. And this scheme significantly lowers the Character Error Rate (CER) of the AM compared to the baseline deep neural network (DNN) network.

TUE-AMI-O3.4: Revisiting Dynamic Adjustment of Language Model Scaling Factor for Automatic Speech Recognition

Hiroshi Sato, Takafumi Moriya, Yusuke Shinohara, Ryo Masumura, Takaaki Fukutomi, Kiyoaki Matsui, Takanori Ashihara, Yoshikazu Yamaguchi and Yushi Aono

NTT, NTT TechnoCross

Automatic speech recognition (ASR) systems use the language model scaling factor to weight the probability output by the language model and balance it against those from other models including acoustic models. Although the conventional approach is to set the language model scaling factor to a constant value to suit a given training dataset to maximize overall performance, it is known that the optimal scaling factors varies depending on individual utterances. In this work, we propose a way to dynamically adjust the language model scaling factor to a single utterance. In the proposed method, a recurrent neural network (RNN) based model is introduced to predict optimum scaling factors given ASR results from a training dataset. Some studies have already tackled this utterance dependency in the 2000s, yet few have improved the quality of ASR due to the difficulty in directly modeling the relationship between a series of acoustic features and the optimal scaling factor; a recent breakthrough in RNN technology has now made this feasible. Experiments on a real-world dataset show that the dynamic optimization of the language model scaling factor can improve ASR quality and that the proposed method is effective.

TUE-AMI-O3.5: A Language Model-Based Design of Reduced Phoneme Set for Acoustic Model

Shuji Komeiji and Toshihisa Tanaka

Tokyo University of Agriculture and Technology

A language model-based design of reduced phoneme set for acoustic model is proposed.
In the case where the amount of training data is too small to train each phoneme model, the reduction of the phoneme set can lead to a reduced discriminative model of phonemes, which can increase homophones that yield degradation of speech recognition. The proposed approach enables us to reduce phonemes preventing the degradation, regarding pronunciation/word sequence confusion rate calculated from n-grams in a language model. In an experiment, the phoneme set designed with proposed approach was applied to Japanese large vocabulary speech recognition system. The word error rate with full 39 phonemes set was 9.5%, while the error rate with the 10 phonemes set designed with the proposed approach was 11.1%. The degradation was able to be prevented within 2%.

TUE-AMI-O3.6: Audio Codec Simulation based Data Augmentation for Telephony Speech Recognition

Thi Ly Vu, Zhiping Zeng, Haihua Xu and Eng-Siong Chng

Nanyang Techonological University

Real telephony speech recognition task is challenging due to 1) diversified channel distortions and 2) limited access to the real data because of the data privacy consideration. In this paper, assuming no real telephony data available, we employ diversified audio codecs simulation based data augmentation method to train telephony speech recognition system. Specifically, we assume only wide-band 16 kHz data available, and we first down-sample the 16 kHz data to the 8 kHz data; we then pass the down-sampled data through various categories of audio codecs to simulate the real channel distortion. As a result, we train our speech recognition with such distorted data. To analyze the effectiveness of different audio codec simulation methods, we classify them into three main categories according to their distortion severity, in terms of their spectrogram analysis. We conduct experiments on various real telephony test sets to show the effectiveness of the proposed data augmentation method. The result shows that the real data is more close with highly distorted simulation data, since the model with highly distorted data reduce the Word-Error-Rate 7.28% - 12.78% compared to the baseline.

TUE-AM1-SS4
Recent Advances in Biometric Security

Time: Tuesday, Nov 19, 10:20-12:00

Place: A8

Chairs: Koichi Ito, Tetsushi Ohki

TUE-AMI-SS4.1: Security and Efficiency of Biometric Template Protection for Identification

Wataru Nakamura, Yosuke Kaga, Masakazu Fujio and Kenta Takahashi

Hitachi

The Biometric Template Protection (BTP) Technology includes Cancelable Biometrics(CB), Biometric Cryptosystem (BC), and Biometric Signature (BS). CB schemes cannot satisfy security requirements on irreversibility and spoofing difficulty if enrolled data leak from multiple entities. On the other hand, it is considered that well-implemented BC schemes such as fuzzy extractor or BS schemes such as fuzzy signature satisfy the security requirements. However, when these schemes are naively applied to identification, computation and communication cost increases. In this paper, we define efficiency requirements based on computation and communication cost for Biometric Template Protection for Identification (BTPI). Then, we show that BTPI schemes based on conventional BTP schemes cannot satisfy requirements on security and efficiency simultaneously. Next, we propose a novel BTPI scheme obtained by combining fast CB and secure BC or BS. The proposed scheme achieves both requirements under certain assumptions on the publicity of biometric features used for the proposed scheme.

TUE-AMI-SS4.2: Security enhancement for touch panel based user authentication on smartphones

Daiki Izumoto and Yasushi Yamazaki

The University of Kitakyusyu

With the rapid spread of smartphones, user authentication for privacy protection is becoming increasingly important. Pattern lock is one of the most typical user authentication methods using a touch panel on smartphones. However, despite its high usability, it is vulnerable to shoulder surfing and smudge attacks. Therefore, to improve security of touch panel based user authentication on smartphones while maintaining usability, we propose a method that combines the function of pattern lock and handwritten biometrics and demonstrate its effectiveness through simulation experiments.

TUE-AMI-SS4.3: Efficient Spoofing Attack Detection against Unknown Sample using End-to-End Anomaly Detection

Tetsushi Ohki, Vishu Gupta and Masakatsu Nishigaki

Shizuoka University

With the evolution of a high precision sensor, printing machine, and manufacturing machine, spoofing attacks become a significant threat to the biometric systems. In order to mitigate the threats of diverse and unexpected attacks, conventional spoofing attack detection methods which aim to detect a specific attack are not sufficient. In this study, we propose a end-to-end machine learning technique which can model biometric information with the complicated structure of high dimension by a probability distribution. The proposed system can recognize whether inputted sample is spoofing one or not even if it is an unknown attack.

TUE-AMI-SS4.4: Eye-blink based Personal Authentication Using Time-series Directional Features and Waveform Features

Keisuke Takano and Hironobu Takano

Toyama Prefectural University

In this study, we propose the personal authentication method using characteristics of eye blink and investigate the feature that is effective for authentication. The time-series variations of gradient directional features and the waveform features extracted from the time-series gradient intensity are adopted as the features for recognition. The matching of registration and recognition features is performed with Euclidean distance and the dynamic time warping (DTW). From the experimental results, we found that individual differences are most likely to appear at the iris peripheral region and also they are most likely to appear at opening eye.

TUE-AMI-SS4.5: Personal Authentication with Eye Movement Features During PIN Input

Shion Tagawa and Hironobu Takano

Toyama Prefectural University

In recent years, personal authentication is used in various situations. Among them, the personal authentication using biometric information is becoming widespread. Conventional biometric methods have a spoofing problem. Personal authentication using eye movement is considered to be more difficult to spoof. Biometrics research using eye movement has been conducted in the case where the eye movement range is relatively. In this study, we aim at development of the personal authentication method using eye movement in the case of narrow range of eye movement such as smartphone and ATM operation. Therefore, in this paper, we investigated effective features of eye movement in the case where the range of eye movement is narrow. The authentication accuracy was evaluated using equal error rate (EER). In addition, we examined whether the authentication accuracy could be improved by score level fusion. In the experimental results, it was found that dynamic time warping (DTW) has the best authentication accuracy when the range of eye movement is narrow, and the authentication accuracy was improved by score level fusion.

TUE-AMI-SS4.6: Attribute Estimation Using Multi-CNNs from Hand Images

Yi-Chun Lin, Yusei Suzuki, Hiroya Kawai, Koichi Ito, Hwann-Tzong Chen and Takafumi Aoki

National Tsing-Hua University, Tohoku University

The human hand is one of the primary biometric traits in person authentication. A hand image also includes a lot of attribute information such as gender, age, skin color, accessory, and etc. Most conventional methods for hand-based biometric recognition rely on one distinctive attribute like palmprint and fingerprint. The other attributes such as gender, age, skin color and accessory, which is known as soft biometrics, are expected to help identify individuals but are rarely used for identification. This paper proposes an attribute estimation method using multi-convolutional neural network (CNN) from hand images. We specially design new multi-CNN architectures dedicated to estimating multiple attributes from hand images. We train and test our models using 11K Hands, which consists of more than 10,000 images with 7 attributes and ID. The experimental results demonstrate that the proposed method exhibits efficient performance on attribute estimation.

POSTER 1

Time: Tuesday, Nov 19, 13:20-15:00

Place: Gansu International Conference Centre(GICC), 3F

Chairs: Zhiyi Yu

POSTER 1.1: Dictionary based Compression Type Classification using a CNN Architecture

Hyewon Song, Beom Kwon, Seongmin Lee and Sanghoon Lee

Yonsei University

As digital devices which are capable of viewing contents easily such as mobile phones and tablet PCs have become widespread, the number of digital crimes using these digital contents also increases. Usually, the data which can be the evidence of crimes is compressed and the header of data is damaged to conceal the contents. Therefore, it is necessary to identify the characteristics of the entire bits of the compressed data to discriminate the compression type not using the header of the data. In this paper, we propose a method for distinguishing 16 dictionary-based compression types. We utilize 5-layered CNN for classification of compression type using the Spatial Pyramid Pooling (SPP) layer. We evaluate our proposed method on the Wikileaks Dataset, which is a text file database. The average accuracy of 16 dictionary-based compression algorithms is 99%. We expect that our proposed method will be useful for providing evidence for Digital Forensics.

POSTER 1.2: Investigation of Monaural Front-End Processing for Robust Speech Recognition Without Retraining or Joint-Training

Zhihao Du, Xueliang Zhang and Jiqing Han

School of Computer Science and Technology, Inner Mongolia University

There are two effective approaches to improve the performance of an automatic speech recognizer with the front-end processing under noisy condition, one is retraining the acoustic model with the enhanced features, the other is joint-training the acoustic model with the front-end processing model. However, in real life, the automatic speech recognition (ASR) systems are always located in cloud servers but the front-end processing models run locally, which results in the impracticality of the retraining and joint-training strategy for ASR. In this paper, we investigate whether the independent front-end processing can directly improve the performance of a speech recognizer without retraining and joint-training. Three common-used enhancement methods are evaluated in different time-frequency (T-F) domains.Our experiments on CHiME-3 reveal that, with appropriate T-F domains and enhancement methods, the front-end processing can make 35.30% and 11.78% relative word-error-rate (WER) reduction for the Gaussian Mixed Model based (GMM-based) and Deep Neural Network based (DNN-based) recognizer, respectively. For the DNN-based ASR system, we propose using masking-based methods in log-fbank domain to do front-end processing. We find that masking based methods, in general, are better than spectral mapping based methods with respect to WER reduction. In addition, the phases of noisy speech are useless and even harmful to reduce the WER. For generalization capability, the front-end processing can improve the multi-conditional trained ASR system under both matched and unmatched noise condition.

POSTER 1.3: A sequential prediction method of quasi-periodicity based on Gaussian process state space model

Akira Tamamori and Tomoko Matsui

Aichi Institute of Technology, Institute of Statistical Mathematics

In this paper, we develop a sequential prediction method of quasi-periodicity based on Gaussian process state space model. We introduce a latent variable to represent the phase hidden in the quasi-periodic phenomenon. The proposed prediction method adopts the predictive distribution for the starting point of the next period. The hyperparameters of the Gaussian process can be inferred by using particle Markov Chain Monte Carlo method. By using basal body temperature data of 17 female subjects, we evaluated the performance of the proposed method in the prediction accuracy of menstrual cycle length. The results showed that the prediction accuracy was improved compared with the conventional method, which adopts a predicted day after a fixed period from the last menstruation.

POSTER 1.4: Automatic Ontology Population Using Deep Learning for Triple Extraction

Ming—Hsiang Su, Chung—Hsien Wu and Po—Chen Shih

NCKU

Ontology is a kind of representation used to represent knowledge in a form that computers can derive the content meaning. The purpose of this work is to automatically populate an ontology using deep neural networks for updating an ontology with new facts from an input knowledge resource. In this study for automatic ontology population, a bi-LSTM-based term extraction model based on character embedding is proposed to extract the terms from a sentence. The extracted terms are regarded as the concepts of the ontology. Then, a multi-layer perception network is employed to decide the predicates between the pairs of the extracted concepts. The two concepts (one serves as subject and the other as object) along with the predicate form a triple. The number of occurrences of the dependency relations between the concepts and the predicates are estimated. The predicates with low occurrence frequency are filtered out to obtain precise triples for ontology population. For evaluation of the proposed method, we collected 46,646 sentences from Ontonotes 5.0 for training and testing the bi-LSTM-based term extraction model. We also collected 404,951 triples from ConceptNet 5 for training and testing the multilayer perceptron-based triple extraction model. From the experimental results, the proposed method could extract the triples from the documents, achieving 74.59% accuracy for ontology population.

POSTER 1.5: Transfer Learning for Punctuation Prediction

Karan Makhija, Ho Thi Nga and Chng Eng Siong

Birla Institute of Technology and Science, Nanyang Technological University

The output from most of the Automatic Speech Recognition system is a continuous sequence of words without proper punctuation. This decreases human readability and the performance of downstream natural language processing tasks on ASR text. We treat the punctuation restoration task as a sequence tagging task and propose an architecture that uses pre-trained BERT embeddings. Our model improves the state of art on the IWSLT dataset. We achieve an overall F1 of 81.4 % on the joint prediction of period, comma and question mark.

POSTER 1.6: A LSTM-Based Joint Progressive Learning Framework for Simultaneous Speech Dereverberation and Denoising

Xin Tang, Jun Du, Li Chai, Yannan Wang, Qing Wang and Chin-Hui Lee

University of Science and Technology of China, Tencent Technology (Shenzhen) Co., Georgia Institute of Technology

We propose a joint progressive learning (JPL) framework of gradually mapping highly noisy and reverberant speech features to less noisy and less reverberant speech features in a layer-by-layer stacking scenario for simultaneous speech denoising and dereverberation. As such layers are easier to learn than mapping highly distorted speech features directly to clean and anechoic speech features, we adopt a divide-and-conquer learning strategy based on a long short-term memory (LSTM) architecture, and explicitly design multiple intermediate target layers. Each hidden layer of the LSTM network is guided by a step-by-step signal-to-noise-ratio (SNR) increase and reverberant time decrease. Moreover, post-processing is applied to further improve the enhancement performance by averaging the estimated intermediate targets. Experiments demonstrate that the proposed JPL approach not only improves objective measures for speech quality and intelligibility, but also achieves a more compact model design when compared to the direct mapping and two-stage, namely denoising followed dereverberation approaches.

POSTER 1.7: Location-Independent Multi-Channel Acoustic Scene Classification Using Blind Dereverberation, Blind Source Separation, and Model Ensemble

Ryo Tanabe, Takashi Endo, Yuki Nikaido, Kenji Ichige, Phong Nguyen, Yohei Kawaguchi and Koichi Hamada

Hitachi Ltd.

This paper presents a location-independent multichannel acoustic scene classification (ASC) system that avoids spatial overfitting. Generally, ASC suffers from noise and reverberation in real environments. In addition, the ASC performance is decreased by overfitting a dataset, which is the result of learning from acoustic transfer functions enclosed in the dataset. To resolve these problems, we present a location-independent multi-channel ASC system using blind dereverberation, blind sound source separation, pre-trained model-based classifiers, and model ensemble. Experimental results on the DCASE 2018 Task 5 dataset indicate that the proposed system, with an F1 score of 88.4%, outperforms the baseline system. Also, the results indicate that although no one specific function improves the performance dramatically, all functions complement each other through the model ensemble.

POSTER 1.8: Phonetic-Attention Scoring for Deep Speaker Features in Speaker Verification

Lantian Li, Zhiyuan Tang, Ying Shi and Dong Wang

Tsinghua University

Recent studies have shown that frame-level deep speaker features can be derived from a deep neural network with the training target set to discriminate speakers by a short speech segment. By pooling the frame-level features, utterance-level representations, called d-vectors, can be derived and used in the automatic speaker verification (ASV) task. This simple average pooling, however, is inherently sensitive to the phonetic content of the utterance. An interesting idea borrowed from machine translation is the attention-based mechanism, where the contribution of an input word to the translation at a particular time is weighted by an attention score. This score reflects the relevance of the input word and the present translation. We can use the same idea to align utterances with different phonetic contents. This paper proposes a phonetic-attention scoring approach for d-vector systems. By this approach, an attention score is computed for each frame pair. This score reflects the similarity of the two frames in phonetic content, and is used to weigh the contribution of this frame pair in the utterance-based scoring. This new scoring approach emphasizes the frame pairs with similar phonetic contents, which essentially provides a soft alignment for utterances with any phonetic contents. Experimental results show that compared with the naive average pooling, this phonetic-attention scoring approach can deliver consistent performance improvement in ASV tasks of both text-dependent and text-independent.

POSTER 1.9: Dementia Detection by Analyzing Spontaneous Mandarin Speech

Zhaoci Liu, Zhiqiang Guo, Zhenhua Ling, Shijin Wang, Lingjing Jin and Yunxia Li

University of Science and Technology of China, iFLYTEK Research, Shanghai Tongji Hospital

The Chinese population has been aging rapidly resulting in the largest population of people with dementia. Unfortunately, current screening and diagnosis of dementia rely on the evidences from cognitive tests, which are usually expensive and time consuming. Therefore, this paper studies the methods of detecting dementia by analyzing the spontaneous speech produced by Mandarin speakers in a picture description task. First, a Mandarin speech dataset contains speech from both healthy controls and patients with mild cognitive impairment (MCI) or dementia is built. Then, three categories of features, including duration features, acoustic features and linguistic features, are extracted from speech recordings and are compared by building logistic regression classifiers for dementia detection. The best performance of identifying dementia from healthy controls is obtained by fusing all features and the accuracy is 81.9% in a 10-fold cross-validation. The importance of different features is further analyzed by experiments, which indicate that the difference of perplexities derived from language models is the most effective one.

POSTER 1.10: Dynamic-attention based Encoder-decoder model for Speaker Extraction with Anchor speech

Hao Li, Xueliang Zhang and Guanglai Gao

Inner Mongolia University

Speech plays an important role in human-computer interaction. For many real applications, an annoying problem is that speech is often degraded by interfering noise. Extracting target speech from background interference is a meaningful and challenging task, especially when interference is also human voice. This work addresses the problem of extracting target speaker from interfering speaker with a short piece of anchor speech which is used to obtain the target speaker identify. We propose a encoder-decoder neural network architecture. Specifically, the encoder transforms the anchor speech to a embedding which is used to represent the identity of target speaker. The decoder utilizes the speaker identity to extract the target speech from mixture. To make a acoustic-related speaker identity, The dynamic-attention mechanism is utilized to build a time-varying embedding for each frame of the mixture. Systematic evaluation indicates that our approach improves the quality of speaker extraction.

POSTER 1.11: Classification of causes of speech recognition errors using attention-based bidirectional long short-term memory and modulation spectrum

Jennifer Santoso, Takeshi Yamada and Shoji Makino

University of Tsukuba

In this paper, we address the problem of classifying four common utterance characteristics related to the utterance speed, which cause speech recognition errors. We previously proposed bidirectional long short-term memory (BLSTM) as a classifier and the modulation spectrum as an acoustic feature. However, the performance of is still insufficient, since BLSTM classified the utterance characteristics from the overall utterance, while most of the recognition errors resulted from utterance characteristics occurring in only a small part of utterance. In this paper, we propose an approach to enhance classifier by using attention mechanism (attention-based BLSTM). Attention-based BLSTM enables the classifier to weight

each frame according to its importance instead of directly measuring overall information from the speech. Furthermore, we investigate the correspondence of utterance characteristics to different modulation spectrum block lengths. To evaluate the performance of the proposed method, we conducted a classification experiment on Japanese conversational speeches with four different utterance characteristics: 'fast', 'slow', 'filler', and 'stutter'. As a result, the proposed method improved the F-score by 0.033--0.129 compared with the previously proposed method using BLSTM. This result confirms the effectiveness of attention-based BLSTM in classifying cause of errors based on utterance characteristics.

POSTER 1.12: Semi-Coprime Microphone Arrays for Estimating Direction of Arrival of Speech Sources

Jiahong Zhao and Christian Ritz

University of Wollongong

This paper evaluates the performance of semi-coprime microphone arrays (SCPMAs) for speech source direction of arrival (DOA) estimation based on the steered response power – phase transform (SRP-PHAT) algorithm. The SCPMA is an extension of the coprime microphone array (CPMA), which combines the outputs of three sub-arrays to reduce the impact of spatial aliasing and achieves performance comparable to that obtained from arrays using much larger numbers of microphones. The proposed approach considers two different processors to calculate the outputs from the sub-arrays and adapts the SRP-PHAT approach to these arrays. Simulations are conducted under anechoic and reverberant scenarios in a noisy room. Beam pattern and array gain results indicate that the SCPMA works better than the conventional CPMA at reducing the peak side lobe (PSL) level and total side lobe area while increasing the capability of amplifying the desired target signal and restraining noise from all other directions for typical frequencies of speeches. DOA Estimation results also show that the SCPMA achieves accurate DOA estimates in anechoic and low reverberant conditions, which is comparable to the equivalent full ULA, while the large side lobes in the beam pattern of the SCPMA lead to less accurate results in the highly reverberant environment.

POSTER 1.13: Speaker-discriminative Embedding Learning via Affinity Matrix for Short Utterance Speaker Verification

Junyi Peng, Rongzhi Gu, Yuexian Zou and Wenwu Wang

Peking University, University of Surrey

Text-independent short utterance speaker verification (TI-SUSV) task remains more challenging compared to the full-length utterance SV task due to inaccurately estimated feature statistics or insufficient distinguishable speaker embeddings. It is noted that recently developed end-to-end SV systems (E2E-SV) achieve the state-of-the-art on several datasets, which directly learn a mapping from speech features to the compact fixed length speaker embeddings. In this study, following the E2E-SV pipeline, we strive to further improve the accuracy of TI-SUSV task. Our research is based on two intuitive ideas: better speech feature representation for SUs and better training loss function to obtain more discriminative embeddings. Specifically, a bi-directional gated recurrent unit network with residual connection (Res-BGRU) is firstly designed to improve feature representation capability. Secondly, a novel affinity loss is proposed where the mini-batch data has been manipulated to obtain more supervision information. In details, a speaker identity affinity matrix formed by one-hot speaker identity vectors is taken as the supervisor of the speaker embedding affinity matrix to obtain better inter-speaker separability and intra-speaker compactness. Experimental results on the Voxceleb1 dataset show that our system outperforms a conventional i-vector and x-vector system on TI-SUSV.

POSTER 1.14: Prosodic Structure Prediction using Deep Self-attention Neural Network

Yao Du, Zhiyong Wu, Shiyin Kang, Dan Su, Dong Yu and Helen Meng

Graduate School at Shenzhen, Tencent AI Lab, The Chinese University of Hong Kong

Prosodic structure prediction is a key part of the text analysis front-end of the text-to-speech (TTS) system. It predicts prosodic boundary tags given the input text context, which is essential to the naturalness of synthesized speech. Conventional methods such as conditional random fields (CRF) and recurrent neural network (RNN) have been successfully applied to this task. However, the lack of modeling temporal dependencies at different scopes (the short-term dependency as well as the long-span dependency across the entire sentence) limits their performance. In this paper we examine a new application of self-attention to the prosodic structure prediction task. The self-attention mechanism can capture the dependencies between two arbitrary words at any distance in the sentence. Experimental results show that the proposed approach outperforms the strong baseline CRF model with an absolute improvement of 3.4% in total accuracy.

POSTER 1.15: Resolve Cross-chunk Permutation through Chunklevel Speaker Embedding for Blind Speech Separation

Rongzhi Gu, Junyi Peng, Yuexian Zou and Dong Yu

Peking University, Tencent AI Lab

Speaker-independent speech separation (SI-SS) refers to recovering speech of unknown speakers from multi-speaker mixtures. The well-known deep clustering (DC) based SI-SS methods cast the speech separation problem into a clustering problem in an embedding space, where time-frequency (T-F) features are encoded as high-dimensional vectors (T-F embeddings). In training stage, the T-F embeddings from the same speaker are trained to be close to each other, otherwise far away. In prediction stage, the T-F embeddings are partitioned into clusters by K-Means, where each cluster corresponds to an unknown speaker from the mixture. To reduce the latency, the T-F embeddings are usually extracted on short speech chunks rather than utterances, which unfortunately leads to a cross-chunk permutation (CCP) problem. In this study, we focus on solving this CCP problem by using the speaker labels as the auxiliary supervision information to train a deep model to map the T-F embeddings of one cluster to one chunk-level speaker embedding (CL-SE). Therefore, in prediction stage, the generated CL-SEs are used to calculate the similarity between each cluster over consecutive chunks. As a result, the speech chunks with the more similar CL-SEs are concatenated to yield the complete utterances. The evaluation is conducted on the well-known WSJ0-2mix and the signal-to-distortion ratio (SDR) is adopted for performance evaluation. Noted that we obtain 41% SDR gain over DC baseline and up to 32% over other speaker-aware methods in open conditions.

POSTER 1.16: Acceleration of rank-constrained spatial covariance matrix estimation for blind speech extraction

Yuki Kubo, Norihiro Takamune, Daichi Kitamura and Hiroshi Saruwatari

The University of Tokyo, National Institute of Technology, Kagawa College

In this paper, we propose new accelerated update rules for rank-constrained spatial covariance model estimation, which efficiently extracts a directional target source in diffuse background noise. The naive update rule requires heavy computation such as matrix inversion or matrix multiplication. We resolve this problem by expanding matrix inversion to reduce computational complexity; in the parameter update step, we need neither matrix inversion nor multiplication. In an experiment, we show that the proposed accelerated update rule achieves 83 times faster calculation than the naive one.

POSTER 1.17: Improving Automatic Jazz Melody Generation by Transfer Learning Techniques

Hsiao-Tzu Hung, Chung-Yang Wang, Yi-Hsuan Yang and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taiwan AI Labs, Research Center for IT Innovation

In this paper, we tackle the problem of transfer learning for Jazz automatic generation. Jazz is one of representative types of music, but the lack of Jazz data in the MIDI format hinders the construction of a generative model for Jazz. Transfer learning is an approach aiming to solve the problem of data insufficiency, so as to transfer the common feature from one domain to another. In view of its success in other machine learning problems, we investigate whether, and how much, it can help improve automatic music generation for under-resource musical genres. Specifically, we use a recurrent variational autoencoder as the generative model, and use a genre-unspecified dataset as the source dataset and a Jazz-only dataset as the target dataset. Two transfer learning methods are evaluated using six levels of source-to-target data ratios. The first method is to train the model on the source dataset, and then fine-tune the resulting model parameters on the target dataset. The second method is to train the model on both the source and target datasets at the same time, but add genre labels to the latent vectors and use a genre classifier to improve Jazz generation. The evaluation results show that the second method seems to perform better overall, but it cannot take full advantage of the genre-unspecified dataset.

POSTER 1.18: Speech Loss Compensation by Generative Adversarial Networks

Yupeng Shi, Nengheng Zheng, Yuyong Kang and Weicong Rong

Shenzhen University

Speech loss, including frequency loss and packet loss, can lead to significant speech distortion in many Internet-based speech communication services. In this study, a generative adversarial networks (GANs) structure, which takes deep convolutional neural networks (CNN) as the generator and discriminator components, is adopted as a general framework for speech loss compensation. Network settings are modified for real-time communications. A set of experiments are conducted to evaluate the performance of the GANs-based framework for both bandwidth expansion (BWE) and packet loss concealment (PLC)

at several simulated loss conditions. Experimental results demonstrate that the proposed system achieves better performance, with respective to 4 objective metrics, in both BWE and PLC compared to the baseline systems.

POSTER 1.19: Snoring sound classification using multiclass classifier under actual environments

Keisuke Nishijima and Ken'ichi Furuya

Oita University

The problem with conventional snoring sound identification methods is that their performance declines when the snoring sound is identified in the actual environment. Therefore, it is necessary to cope with the stationary and nonstationary environmental sounds that cause the decrease. In this research, we tried to cope with stationary environmental sounds by spectrum subtraction method for noise suppression. Non-stationary environmental sounds were regarded as one class for each type of environmental sound. We tried to identify the snoring sounds by multikernel learning, which is a multiclass extension of a support vector machine and by multilayer perceptron, which is a kind of neural network.

POSTER 1.20: Nonlinear Echo Cancellation Based on Polyphase Filter Bank

Meng Liang, Zhong-Hua Fu, Xiang Zhao, Jinglei Zhou and Haikun Wang

Northwestern Polytechnical University, Xi'an IFLYTEK Hyper Brain Information Technology Co., School of Electronics and Information, Xi'an Polytechnic University

In hands-free telephone systems and mobile communication devices, it is often desirable for the devices to operate at large sound volume,which can result in obvious nonlinear acoustic echoes due to overload of small loudspeaker.These nonlinear echoes can't be fully eliminated by linear AEC (Acoustic Echo Cancellation) algorithm, so the conversation quality is affected seriously. Since the nonlinear echoes contain additional harmonics in high frequency, which breaks the linear relation required for fullband linear AEC, these harmonics, however, becomes additive noise in subband system. Therefore, in this paper, a subband AEC method based on poly-phase filter-bank is proposed. It is found that ERLE (Echo Return Loss Enhancement) of the proposed method outperforms the fullband counterpart constantly in nonlinear situation, especially with tone-like signals, where ERLE improves more than 15 dB. The results are validated through both simulation and real signals.

POSTER 1.21: Question Mark Prediction By Bert

Yunqi Cai and Dong Wang

Tsinghua University

Punctuation resotration is important for Automatic Speech Recognition and the down-stream applications, e.g., speech translation. Despite the continuous progress on punctuation restoration, discriminating question marks and periods remains very hard. This difficulty can be largely attributed to the fact that interrogatives and narrative sentences are mostly characterized and distinguished by long-distance syntactic and semantic dependencies, which are cannot well modeled by existing models (e.g., RNN or n-gram). In this paper we propose to solve this problem by the self-attention mechanism of the Bert model.Our experiments demonstrated that compared the best baseline, the new approach improved the F1 score of question mark prediction from 30% to 90%.

POSTER 1.22: Part-Based Bilinear CNN For Person Re-Identification

Li Li, Jianwu Dang, Yangping Wang and Song Wang

Lanzhou Jiaotong University

Abstract— Aiming at the problems of image misalignment and the weak discriminative feature of Person Re-Identification, based on the fine-grained network bilinear CNN, a multi-part Re-ID network is proposed. The branch network is used to learn the part features to reduce the influence of the misalignment problem of the datasets image on the Re-ID effect, and the compact bilinear pooling used for the fusion of each part of the branch network to generate discriminative feature. Weighted values of block feature and global feature loss are used to optimize the network. The validity of the proposed network structure is verified on the dataset CUHK03 and Market-1501. The results show that the proposed model has higher average recognition accuracy than traditional algorithms and other similar network models.

POSTER 1.23: Polyphonic Voicing Optimization for Automatic Music Completion

Christoph M. Wilk and Shigeki Sagayama

Meiji University, The University of Tokyo

In this paper, we present a new algorithm for automatic music completion. We have proposed automatic music completion as the class of music composition assistance problems of generating a complete piece of music given fragments of musical ideas input by a user. These fragments include partial melodies in multiple voices or parts of the underlying harmony progression. Therefore, it is a generalization of common problems such as melody harmonization or harmony constrained melody generation, but also includes problems with constraints in multiple domains, i.e. multiple voices and harmony. We present a new polyphonic voicing model for automatically completing four-part chorales. It is based on a hidden Markov model and what we call correction factors. These factors are trainable functions that efficiently capture the context of a voicing in order to account for a multitude of music theoretical rules without having to resort to a rule-based system. We observed improvement over our old model with regards to metrics that are derived from music theory for polyphonic voicing, and also invite the reader to try our algorithm themselves at http://160.16.202.131/music_completion_apsipa.

POSTER 1.24: Online Layered Multiple Object Tracking Using Residual-Residual Networks

Bo-Cheng Jiang, Chung-Nan Lee

National Sun Yat-sen University, CSE

When dealing with multiple object tracking in the real world, it faces several challenges: (a) The number of targets to be tracked will change over time, (b) The data association of the target at different times will be affected by occlusion, (c) The problem of estimating the continuous state of all targets and deciding whether the targets leave the screen and then stop tracking. In this paper a novel multiple object tracking method that consists of a residual-residual network and a four-layer data association scheme. The residual-residual network combines a deep residual classification network and a deep residual feature network. The deep residual classification network is used to remove unwanted background noise from the frame and corrects the target position of the missing ones. It can accurately track the position of multiple targets, link their positions in each time period, combine layered target data association method, and stepwise pair the trajectories and target candidates according the features generated from the deep residual feature network. Experiments using MOT16 that is a multi-object tracking database, show that the proposed method leads most existing researches in several evaluation criteria including the accuracy, speed and false positive.

POSTER 1.25: Rail Surface Defect Recognition Method Based on AdaBoost Multi-classifier Combination

Biao Yue, Yangping Wang, Yongzhi Min, Wenrun Wang and Jiu Yong

Experimental Teaching Cener on Computer Science, Lanzhou Jiaotong University,

School of Automation and Electrical Engineering, Lanzhou Jiaotong University

Rail surface defects have the characteristics of various types and complex morphological characteristics. It is difficult to obtain accurate classification results only by using a single classification method. Therefore, this paper presents a rail surface defect recognition method based on AdaBoost multi-classifier combination. Firstly, defect attributes are described by extracting geometric shape and gray level features of defect area, and Relief algorithm is used to select defect features and filter out features unrelated to classification. Then by using AdaBoost multi-classifier combination method and taking CART decision tree as a weak classification algorithm to design a combined classifier, rail surface defects classification is realized. The results show that this method can effectively identify three common types of defects: rail surface peeling block, tread crack and fish scale peeling crack.

POSTER 1.26: Focal Loss for End-to-end Short Utterances Chinese Dialect Identification

Qiuxian Zhang, Jiangyan Yi, Jianhua Tao, Mingliang Gu and Yong Ma

Jiangsu Normal University, Institute of Automation

Short utterances dialect identification is a challenging task because of the substantial similarity between dialects. The previous cross-entropy loss function does not consider the category and probability of prediction error, which result in insensitivity to easily misclassified and unbalanced samples. To solve this problem, we propose to use an improved cross-entropy loss function, namely focal loss, introducing category weights and tunable focusing parameter to improve the classification accuracy. Experiments are carried out on AI Dialect Contest database. The results demonstrate that our proposed end-to-end model trained with focal loss achieves better performance than the model trained with cross-entropy loss function.

POSTER 1.27: Speech representation based on tensor factor analysis and its application to speaker recognition and language identification

Daisuke Saito, So Suzuki and Nobuaki Minematsu

The University of Tokyo

This paper proposes a novel approach to speech representation for both speaker recognition and language identification by characterizing the entire feature space by a tensor. In conventional studies of both tasks, i-vector is commonly used as the state-of-the-art representation. Here, i-vector extraction can be regarded as projection of utterance-based GMM supervector onto a low-dimensional space. In this paper, for the aim of explicit modeling of the correlation among mean vectors of a GMM, an utterance is not modeled as its GMM-based supervector but as its matrix and the entire set of utterances is modeled as its tensor. By applying tensor factor analysis, we obtain a new representation for an input utterance. Experimental evaluations for speaker recognition and language identification show that our proposed approach has effectiveness especially for the speaker recognition task.

POSTER 1.28: Optimizing Learned Object Detection on Point Clouds from 3D Lidars Through Range and Sparsity Information

Jacob Lambert, Eijiro Takeuchi and Kazuya Takeda

Nagoya University

There has been an emergence of research developing 3D lidar-based object detection as the sensor is seeing increasing use in autonomous systems. Supervised learning methods analogous to those used on images were shown to be very effective for object detection in both 2D and 3D representations of lidar data. However, image-based methods have been applied perhaps naively to this new sensing modality, with lidar-specific improvements only recently introduced. This paper shows that the relationship between object appearance and distance from lidar sensor, as well as the impact this has on training label quality, is an important consideration when training 3D lidar-based object detection models. Dataset filtering approaches are developed and evaluated on a state-of-the-art framework, and shown to improve detection accuracy on the KITTI dataset. Including range information directly in the network input was also shown to improve detection capabilities.

POSTER 1.29: Through the Eyes of Viewers: A Comment-Enhanced Media Content Representation for TED Talks Impression Recognition

Huan-Yu Chen, Yun-Shao Lin and Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University

Developing computational frameworks for personalized content query and recommendation has sparked numerous research into automatic indexing and retrieval of multimedia data. Assessing viewer impression as an appropriate index of media content is especially important as it links directly to the audience preferences toward media content. Most of the existing machine learning frameworks rely on modeling the media contents solely without considering the potential usefulness of user feedback in order to assess the viewer impressions. In this work, we develop a cross-modal network that projects the multimodal media content through the viewer's comment space in order to learn a joint (content and viewer) embedding space to perform viewer impression recognition. Specifically, we gather a large corpus of TED talks including viewer's online comments for each of the presentation video. Our proposed cross-modal projection network achieves 80.8%, 79.5%, and 80.8% of unweighted average recall (UAR) in binary classification tasks for three different viewer impression ratings (i.e., inspiring, persuasive, and funny, respectively). Our experiments demonstrate intuitively that online user comments reflect the viewerimpression the most, but an interesting finding shows that it is important to project the content's information into the user comment space, i.e., through the eyes of the comment, in order to obtain an improved recognition accuracy as compared to simply concatenating content and comment features directly.

POSTER 1.30: End-to-end autonomous driving based on the convolution neural network model

Yuanfang Zhao and Yunli Chen

Beijing University of Technology

Abstract—In the control algorithm of autopilot system, the Deep Learning method plays a vital role. Since the convolutional neural network (CNN) model used in automatic driving has a huge amount of parameters and the training results are prone to over-fitting, an excellent model is necessary. In this paper, an end-to-end control method was proposed to apply a convolutional neural network with a new network structure to control the steering angle and speed of the vehicle and reach the goal of automatic

vehicle driving. The experimental results show that it not only greatly reduces the number of parameters, but also keeps the error rate of the experimental results at the low level.

POSTER 1.31: DSNet: An Efficient CNN for Road Scene Segmentation

Hsueh–Ming Hang, Ping–Rong Chen, Sheng–Wei Chan and Jing–Jhih Lin

National Chiao Tung University, Industrial Technology Research Institute

Road scene understanding is a critical component in an autonomous driving system. Although the deep learning-based road scene segmentation can achieve very high accuracy, its complexity is also very high for developing real-time applications. It is challenging to design a neural net with high accuracy and low computational complexity. To address this issue, we investigate the advantages and disadvantages of several popular CNN architectures in terms of speed, storage and segmentation accuracy. We start from the Fully Convolutional Network (FCN) with VGG, and then we study ResNet and DenseNet. Through detailed experiments, we pick up the favorable components from the existing architectures and at the end, we construct a light-weight network architecture based on the DenseNet. Our proposed network, called DSNet, demonstrates a real-time testing (inferencing) ability (on the popular GPU platform) and it maintains an accuracy comparable with most previous systems. We test our system on several datasets including the challenging Cityscapes dataset (resolution of 1024 ×512) with an mIoU of about 69.1% and runtime of 0.0147 second per image on a single GTX 1080Ti. We also design a more accurate model but at the price of a slower speed, which has an mIoU of about 72.6 % on the CamVid dataset.

POSTER 1.32: Phone-Aware Multi-task Learning and Length Expanding for Short-Duration Language Recognition

Miao Zhao, Rongjin Li, Shijiang Yan, Zheng Li, Hao Lu, Shipeng Xia, Qingyang Hong and Lin Li

Xiamen University

In the language recognition, the phonetic information has shown great potential for neural network to learn the high-level representations. In this paper, we explore two significant aspects to improve the system performance under the short-duration condition on the oriental language recognition (OLR) challenge. Firstly, we propose to learn the language information and phonetic information jointly with multi-task learning. The classified networks can learn the extra phonetic representation from a frame-level phone-task and extract the language embedding at the segment level. Furthermore, we propose to introduce length expanding strategy to provide supplemental information of short-duration utterances, by dithering the short duration evaluation utterances at different speeds. The evaluation results of the 3rd OLR Challenge showed that our proposed methods obtained the best results on the short-duration condition.

POSTER 1.33: SDBF-Net: Semantic and Disparity Bidirectional Fusion Network for 3D Semantic Detection on Incidental Satellite Images

Zhibo Rao, Mingyi He, Zhidong Zhu, Yuchao Dai and Renjie He

Northwestern Polytechnical University

In this paper, we propose a conceptually simple, flexible, and general framework for the semantic stereo task on incidental satellite images. Our method efficiently detects the objects in an incidental satellite image for generating a high-quality segmentation map, and more accurately match the left-right incidental satellite images for obtaining a more accurate disparity map at the same time. The method, called semantic and disparity bidirectional fusion network (SDBF-Net), consists of three main modules: the Semantic Segmentation Module (SSM), the Stereo Matching Module (SMM), and the Fusion Module (FM). The semantic segmentation module takes advantage of the capacity of global context information by extending the receptive field to produce the initial segmentation map. The stereo matching module applies the 3D convolutional operation to regularize the feature map of left-right images to generate the initial disparity map. The fusion module fuses the initial segmentation and disparity map to obtain the refined segmentation and disparity map. Extensive quantitative and qualitative evaluations on the US3D dataset demonstrate the superiority of our proposed SDBF-Net approach, which outperforms state-of-the-art semantic stereo approaches significantly.

POSTER 1.34: A Study on Angular Based Embedding Learning for Text-independent Speaker Verification

Zhiyong Chen, Zongze Ren and Shugong Xu

Shanghai Institute for Advanced Communication and Data Science, Shanghai University

Learning a good speaker embedding is important for many automatic speaker recognition tasks, including verification, identification and diarization. The embeddings learned by softmax are not discriminative enough for open-set verification tasks. Angular based embedding learning target can achieve such discriminativeness by optimizing angular distance and adding margin penalty. We implement several different popular angular margin embedding learning methods in this work and explicitly compare their performance on Voxceleb speaker recognition dataset. Observing the fact that the use of the angular based embedding learning is not helpful for inter-class separability, we also implement an inter-class regularization as a complement for angular based loss. We explore the effectiveness of these methods for learning a discriminative embedding space in ASV task with several experiments. Among all the systems we implement, the angular based embedding learning manage to achieve at most 16.5% improvement on equal error rate (EER) and 18.2% improvement on minimum detection cost function comparing with baseline softmax systems.

POSTER 1.35: Speaker Embedding Extraction with Multi-feature Integration Structure

Zheng Li, Hao Lu, Jianfeng Zhou, Lin Li and Qingyang Hong

Xiamen University

Recently x-vector has achieved a promising performance of speaker verification task and becomes one of the mainstream systems. In this paper, we analyzed the feature engineering based on the x-vector structure, and proposed a multi-feature integration method to further improve the feature representation of speaker characteristic. The proposed multifeature integration method could be implemented in two ways, with the symmetric branches and the asymmetric branches, respectively, to incorporate different types of acoustic features in one neural network. While each branch processed one type of acoustic features on the frame level, and the outputs of the two branches for each frame were spliced together as a super vector before being input into the statistics pooling layer. The experiments were executed on the VoxCeleb1 data set, and the results showed that the proposed multi-feature integration method obtained a 22.8% relative improvement over the baseline in EER value.

POSTER 1.36: AN EFFECTIVE ROAD EXTRACTION METHOD from REMOTE SENSING IMAGES BASED on SELF-ADAPTIVE THRESHOLD FUNCTION

Zhuozheng Wang, Meng Zhang and Wei Liu

Beijing University of Technology

In the field of remote sensing imagery, road extraction is one of the key technologies supporting for Landuse Landcover classification. In this paper, a new semantic segmentation neural network named SAT U-Net is proposed for road extraction from remote sensing imagery. The new improved network replaces the sigmoid layer in the U-Net with a self-adaptive threshold method proposed to self-adaptively adjust the road thresholds for segmentation results of U-Net. The proposed method is combined with the strength of U-Net architecture to retain the complete road spatial features, thus overcomes the problem of unconnected and blurry roads in the segmentation results. To prove the effectiveness and utility of the proposed network, it was experimented on the test set of a public road dataset and compared with U-Net in five different environments. Experimental results demonstrate that the proposed method is superior to U-Net and presents clearer and more complete road structures.

POSTER 1.37: Structure Growth for Small-Footprint Speech Recognition

Jiayao Wu, Zhiyuan Tang and Dong Wang

Huazhong University of Science and Technology, Tsinghua University

Modern speech recognition (ASR) is based on large-scale deep neural nets (DNNs) with various architectures. For small-footprint applications running on low-power chips, however, the size of the DNNs must be extremely constrained. In this case, training a generalizable acoustic model is not feasible, especially when the acoustic conditions are diverse. Most of existing approaches to small-footprint networks start from a large net and reduce its scale by pruning. In this paper, we investigate a reverse idea: starting from a small net and increasing it gradually. This structure-growth approach follows a `general to specific' principle and grows the net gradually. We start from the AdaBoost algorithm that builds specific nets for error-prone data, and then propose a new ConBoost that builds specific nets for specific conditions. Our experiments on a small-footprint ASR task demonstrated that both AdaBoost and ConBoost outperform the baseline and other comparative methods including bagging and double-net retraining. Furthermore, ConBoost performs better than AdaBoost.

POSTER 1.38: On Energy Compaction of 2D Saab Image Transforms

Na Li, Yongfei Zhang, Yun Zhang and C.-C. Jay Kuo

Shenzhen Institutes of Advanced Technology, School of Computer Science and Engineering, University of Southern California

The block Discrete Cosine Transform (DCT) is commonly used in image and video compression due to its good energy compaction property. The Saab transform was recently proposed as an effective signal transform for image understanding. In this work, we study the energy compaction property of the Saab transform in the context of intra-coding of the High Efficiency Video Coding (HEVC) standard. We compare the energy compaction property of the Saab transform, the DCT, and the Karhunen-Loeve transform (KLT) by applying them to different sizes of intra-predicted residual blocks in HEVC. The basis functions of the Saab transform are visualized. Extensive experimental results are given to demonstrate the energy compaction capability of the Saab transform.

TUE-PM2-SS1
Signal Processing for Crowd Science

Time: Tuesday, Nov 19, 15:00-16:40

Place: A2

Chairs: H. Vicky Zhao, Yan Chen

TUE-PM2-SS1.1: Measuring the Hazard of Malicious Nodes in Information Diffusion over Social Networks

Hangjing Zhang, Yuejiang Li, Yang Hu, Yan Chen and H. Vicky Zhao

University of Science and Technology of China, Tsinghua University

Social networks have become prevalent in our daily life: people learn, discuss, and spread different kinds of information through social networks every day. While bringing a lot of convenience, the prevalence of social networks creates the security challenge. The information released and/or spread by malicious nodes can be wrong, misleading or even virus, which may lead to bad influences and severe consequences. Therefore, understanding the information diffusion process and the hazard impact of malicious nodes' behaviors to the information diffusion is critical. In this paper, we utilize the evolutionary game theory to measure the hazard influence of malicious nodes to the information diffusion over social networks, by investigating the information diffusion dynamics and evolutionary stable strategies. Finally, simulations are conducted to validate the theoretic analysis and illustrate the impact of the malicious nodes.

TUE-PM2-SS1.2: Controlling Information Diffusion with Irrational Users

Benliu Qiu, Yuejiang Li, Yan Chen and H. Vicky Zhao

University of Science and Technology of China, Tsinghua University

Understanding how information propagates over networks is critical to the development of online social networking. Different from existing works on modeling the rational behaviors in information diffusion, in this paper, we focus on the study of how to control the information propagation over networks through the use of irrational users. With the help of graphical evolutionary game theory, we analyze how the irrational users influence the rational neighbors and thus the whole rational networks. Simulation results verify our theoretic analysis, and show that with a few irrational users, the number of rational users adopting the forwarding strategy can be significantly increased, i.e., we are able to control the information diffusion with irrational users.

TUE-PM2-SS1.3: Modeling Multi-source Information Diffusion: A Graphical Evolutionary Game Approach

Hong Hu, Yuejiang Li, Hong Zhao and Yan Chen

Dept. of Automation and Inst. for Aritificial Intelligence, Tsinghua University, BNRist Center, School of Info.&Comm. Engr., Univ. of Electronic Science and Technology of China

Modeling of information diffusion over social networks is of crucial importance to better understand how the avalanche of information overflow affects our social life and economy, thus preventing the detrimental consequences caused by rumors and motivating some beneficial information spreading. However, most model-based works on information diffusion either consider the spreading of one single message or assume different diffusion processes are independent of each other. In real-world scenarios, multi-source correlated information often spreads together, which jointly influences users' decisions. In this paper, we model the multi-source information diffusion process from a graphical evolutionary game perspective. Specifically, we model users' local interactions and strategic decision making, and analyze the evolutionary dynamics of the diffusion processes of correlated information, aiming to investigate the underlying principles dominating the complex multi-source information diffusion. Simulation results on synthetic and Facebook networks are consistent with our theoretical analysis. We also test our proposed model on Weibo user forwarding data and observe a good prediction performance on real-world information spreading process, which demonstrates the effectiveness of the proposed approach.

TUE-PM2-SS1.4: A Universal Intelligence Measurement Method Based on Meta-analysis

Zheming Yang and Wen Ji

Institute of Computing Technology, Chinese Academy of Sciences

The multiple factors of intelligence measurement are critical in the intelligent science. The intelligence measurement is typically built at a model based on the multiple factors. The different digital self is generally difficult to measure due to the uncertainty between multiple factors. Effective methods for the universal intelligence measurement are therefore important for the different digital self. In this paper, we propose a universal intelligence measurement method based on meta-analysis. Firstly, we get study data through keywords in database and delete the low-quality data. Secondly, after encoding the data, we compute the effect value by Odds ratio, Relatve risk and Risk difference. Then we test the homogeneity by Q-test and analysis the bias by funnel plots. Thirdly, we select the Fixed Effect and Random Effect as statistical model. Finally, simulation results confirm that our method can effectively solve the multiple factors of different digital self. Especially for the intelligence of human, machine, company, government and institution.

TUE-PM2-SS1.5: Modeling Content Interaction in Information Diffusion with Pre-trained Sentence Embedding

Qinyuan Ye, Yuejiang Li, Yan Chen and H. Vicky Zhao

Tsinghua University, University of Science and Technology of China

Social networks have become indispensable parts of our daily life, and therefore understanding the process of information diffusion over social networks is a meaningful research topic. Usually, multiple pieces of information do not spread in isolation; rather, they interact with each other throughout the diffusion process. This paper aims to quantify these interactions by modeling users' forwarding behavior after reading a series of information. Inspired by several successful components prevalent in recent research of deep learning, i.e., long short term memory (LSTM) network and bi-directional encoder representation from transformers (BERT), we designed IMM Enhanced model and InfoLSTM model. In our experiments on real-world Weibo dataset, both models significantly outperform baselines such as the prior IMM model and IP model, with IMM Enhanced model improving 23.52% and InfoLSTM model improving 32.56% in F1 score (absolute value) compared to that of baseline IMM model. In addition, we visualize the dataset and the parameters learned in IMM Enhanced model, which further enables us to discuss the relationship between text similarity and information diffusion interaction with case studies.

TUE-PM2-O1
Speech Emotion Recognition

Time: Tuesday, Nov 19, 15:00-16:40

Place: A3

Chair: Chi-Chun Lee

TUE-PM2-O1.1: Pain versus Affect? An Investigation in the Relationship between Observed Emotional States and Self-Reported Pain

Fu-Sheng Tsai, Yi-Ming Weng, Chip-Jin Ng and Chi-Chun Lee

Department of Electrical Engineering, National Tsing Hua University, Department of Emergency Medicine, Tao-Yuan General Hospital, Department of Emergency Medicine, Chang Gung Memorial Hospital

Pain is an internal sensation intricately intertwined with individual affect states resulting in a varied expressive behaviors multimodally. Past research have indicated that emotion is an important factor in shaping one's painful experiences and behavioral expressions. In this work, we present a study into understanding the relationship between individual emotional states and self-reported pain-levels. The analyses show that there is a significant correlation between observed valence state of an individual and his/her own self-reported pain-levels. Furthermore, we propose an emotion-enriched multitask network (EEMN) to improve self-reported pain-level recognition by leveraging the rated emotional states using multimodal expressions computed from face and speech. Our framework achieves accuracy of 70.1% and 52.1% in binary and ternary classes classification. The method improves a relative of 6.6% and 13% over previous work on the same dataset. Further, our analyses not only show that an individual's valence state is negatively correlated to the pain-level reported, but also reveal that asking observers to rate valence attribute could be related more to the self-reported pain than to rate directly on the pain intensity itself.

TUE-PM2-O1.2: Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters

Yuxuan Xi, Pengcheng Li, Yan Song, Yiheng Jiang and Lirong Dai

University of Science and Technology of China

Despite considerable recent progress in deep learning methods for speech emotion recognition (SER), performance is severely restricted by the lack of large-scale labeled speech emotion corpora. For instance, it is difficult to employ complex neural network architectures such as ResNet which, accompanied by large-sale corpora like VoxCeleb and NIST SRE, have proven to perform well for the related speaker verification (SV) task. In this paper, a novel domain adaptation method is proposed for the speech emotion recognition (SER) task, which aims to transfer related information from a speaker corpus to an emotion corpus. Specifically, a residual adapter architecture is designed for the SER task where ResNet acts as a universal model for general information extraction. An adapter module then trains limited additional parameters to focus on modeling deviation for the specific SER task. To evaluate the effectiveness of the proposed method, we conduct extensive evaluations on benchmark IEMOCAP and CHEAVD 2.0 corpora. Results show significant improvement, with overall results in each task outperforming or matching state-of-the-art methods.

TUE-PM2-O1.3: Speech Emotion Recognition Using Speech Feature and Word Embedding

Bagus Tris Atmaja, Masato Akagi and Kiyoaki Shirai

JAIST

Emotion recognition can be performed automatically from many modalities. This paper presents a categorical speech emotion recognition using speech feature and word embedding. Text features can be combined with speech features to improve emotion recognition accuracy, and both features can be obtained from speech. Here, we use speech segments, by removing silences in an utterance, where the acoustic feature is extracted for speech-based emotion recognition. Word embedding is used as an input feature for text emotion recognition and a combination of both features is proposed for performance improvement purpose. Two unidirectional LSTM layers are used for text and fully connected layers are applied for acoustic emotion recognition. Both networks then are merged by fully connected networks in early fusion way to produce one of four predicted emotion categories. The result shows the combination of speech and text achieve higher accuracy i.e. 75.49% compared to speech only with 58.29% or text only emotion recognition with 68.01%. This result also outperforms the previously proposed methods by others using the same dataset on the same modalities.

TUE-PM2-O1.4: Dimensional Emotion Recognition from Speech Using Modulation Spectral Features and Recurrent Neural Network

Zhichao Peng, Zhi Zhu, Masashi Unoki, Jianwu Dang and Masato Akagi

Japan advanced institute of science and technology, fairy devices Inc.

Dimensional emotion recognition from speech is used to track the dynamics of emotions for natural interaction with human. It is mainly studied from two aspects, one is how to select the appropriate acoustic features and duration to extract salient emotional features; the other is how to capture the dynamic change of emotions from feature sequences. Mel Frequency Cepstrum Coefficients (MFCCs) are commonly used features in speech emotion. However, MFCCs cannot reflect the dynamic characteristics of speech signals very well. Previous studies have indicated that temporal modulation cues are good at capturing the temporal dynamic cues for speech perception and understanding. In this paper, we propose a dimensional emotion recognition system using modulation spectral features (MSFs) and Recurrent Neural Networks (RNNs). The MSFs are obtained from temporal modulation cues, which are produced from auditory front-ends by auditory filtering of speech signals and modulation filtering of the temporal envelope in a cascade manner. Eventually, the MSFs are feed into RNNs to extract continuous temporal-dynamics information. Our experiments of predicting valence and arousal involving the RECOLA database demonstrated that the proposed system significantly outperforms the baseline systems, improved 17% in arousal predictions and 29.5% in valence predictions.

TUE-PM2-O1.5: Adversarial Data Augmentation Network for Speech Emotion Recognition

Lu Yi and Man—Wai Mak

The Hong Kong Polytechnic University

Insufficient data is a common issue in training deep learning models. With the introduction of generative adversarial networks (GANs), data augmentation has become a promising solution to this problem. This paper investigates whether data augmentation can help improve speech emotion recognition. Unlike conventional GANs, we train a GAN with an autoencoder, where the input to the discriminator comes from the bottleneck layer of the autoencoder and the output of the generator. The synthetic samples can be obtained from the decoder, using the output of the generator as the decoder's input. The combined net- work, namely adversarial data augmentation network (ADAN), can generate samples that share common latent representation with the real data. Evaluations on EmoDB and IEMOCAP show that using OpenSmile features as input, the ADAN can produce augmented data that make an ordinary SVM classifier outperforms an RNN classifier with local attention and make a DNN competitive to some state-of-the art systems.

TUE-PM2-O2
Speaker Recognition

Time: Tuesday, Nov 19, 15:00-16:40

Place: A4

Chair: Yanhua Long

TUE-PM2-O2.1: VAE-based Domain Adaptation for Speaker Verification

Xueyi Wang, Lantian Li and Dong Wang

China University of Mining & Technology, Tsinghua University

Deep speaker embedding has achieved satisfactory performance in speaker verification (SV). By enforcing the neural model to discriminate the speakers in the training set, deep speaker embedding (called `x-vectors') can be derived from the hidden layers. Despite its good performance, the present embedding model is highly domain sensitive, which means that it is often worked well for domains which are similar with the training set (in-domain), but performance degradation is often observed in mismatched domains (out-of-domain). In this paper, we present a domain adaptation approach based on Variational Auto-Encoder (VAE). This model transforms x-vectors to a regularized latent space; within this latent space, a small amount of data from the target domain is sufficient to adapt the entire system to the desired domain. Our experiments demonstrated that by this VAE-adaptation approach, speaker embeddings can be easily transformed to the target domain, leading to noticeable performance improvement.

TUE-PM2-O2.2: Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019

Xingliang Cheng, Mingxing Xu and Thomas Fang Zheng

Tsinghua University

Automatic Speaker Verification (ASV) technology is vulnerable to various kinds of spoofing attacks, including speech synthesis, voice conversion, and replay. Among them, the replay attack is easy to implement, posing a more severe threat to ASV. The constant-Q cepstrum coefficient (CQCC) feature is effective for detecting the replay attacks, but it only utilizes the magnitude of constant-Q transform (CQT) and discards the phase information. Meanwhile, the commonly used Gaussian mixture model (GMM) cannot model the reverberation present in far-field recordings. In this paper, we incorporate the CQT and modified group delay function (MGD) in order to utilize the phase of CQT. Also, we present a simple 2d-convolution multi-branch network architecture for replay detection, which can model the distortion both in the time and frequency domains. The experiment shows that the proposed CQT-based MGD feature outperforms traditional MGD feature, and performance can be further improved by combining both magnitude-based and phase-based feature. Our best fusion system achieves 0.0096 min-tDCF and 0.39% EER on ASVspoof 2019 Physical Access evaluation set. Comparing with the CQCC-GMM baseline system provided by the organizer, the min-tDCF is relatively reduced by 96.09% and EER is relatively reduced by 96.46%. Our system is submitted to the ASVspoof 2019 Physical Access sub-challenge and won 1st place.

TUE-PM2-O2.3: Cross-Domain Speaker Recognition using Cycle-Consistent Adversarial Networks

Yi Liu, Bairong Zhuang, Zhiyu Li and Takahiro Shinozaki

Tokyo Institute of Technology

Speaker recognition systems often suffer from severe performance degradation due to the difference between training and evaluation data, which is called domain mismatch problem. In this paper, we apply adversarial strategies in deep learning techniques and propose a method using cycle-consistent adversarial networks for i-vector domain adaptation. This method performs an i-vector domain transformation from the source domain to the target domain to reduce the domain mismatch. It uses a cycle structure that reduces the negative influence of losing speaker information in i-vector during the transformation and makes it possible to use unpaired dataset for training. The experimental results show that the proposed adaptation method improves recognition performance of a conventional i-vector and PLDA based speaker recognition system by reducing the domain mismatch between the training and the evaluation sets.

TUE-PM2-O2.4: SHNU Anti-spoofing Systems for ASVspoof 2019 Challenge

Zhimin Feng, Qiqi Tong, Yanhua Long, Shuang Wei, Chunxia Yang and Qiaozheng Zhang

Shanghai Normal University

This paper presents an experimental analysis of SHNU anti-spoofing systems for the ASVspoof 2019 challenge. This challenge focuses on countermeasures for three major attack types, namely those stemming from the advanced technology of TTS, VC and replay spoofing attacks. According to the type of attacks, the challenge is divided into two independent sub-challenges, the logical access (LA) and physical access (PA). Results of different anti-spoofing technologies on both sub-challenges are reported. Furthermore, the same countermeasures are also evaluated on two previous challenges, the ASVspoof 2015 and 2017. Experiments on cross-databases show that, it appears hard to generalize the classifiers trained from ASVspoof 2019 LA and PA databases to the previous challenges. The generalization ability of anti-spoofing technologies to different, new and unknown conditions is still very challenging. In addition, the effectiveness of different acoustic features are also examined and reported. Finally, we investigated the linear and an interfusing score-level fusion methods to individual systems to achieve better performance.

TUE-PM2-O2.5: Clustering-Based Score Normalization for Speaker Verification

Bin Gu, Wu Guo, Jian Sun and Yao Liu

National Engineering Laboratory for Speech and Language Information Processing,

China General Technology Research Institute

Score normalization can improve speaker verification (SV) performance by adjusting the distribution of test scores to follow a normal distribution. In this paper, all of the imposter scores for the target speakers is first obtained from the normalization cohorts; then, these scores are clustered by an unsupervised clustering algorithm, and Gaussian mixture models (GMMs) are used to fit the score distribution. The mean and the standard deviation of the Gaussian component with the maximum mean value is used in the SV score normalization method. Experiments are carried out on the NIST SRE 2016 test set. Compared with conventional score normalization methods, the proposed method can effectively improve SV performance.

TUE-PM2-O2.6: Triplet Based Embedding Distance and Similarity Learning for Text-independent Speaker Verification

Zongze Ren, Zhiyong Chen and Shugong Xu

Shanghai Institute for Advanced Communication and Data Science

Speaker embeddings become growing popular in the text-independent speaker verification task. In this paper, we propose two improvements during the training stage. The improvements are both based on triplet cause the training stage and the evaluation stage of the baseline x-vector system focus on different aims. Firstly, we introduce triplet loss for optimizing the Euclidean distances between embeddings while minimizing the multi-class cross entropy loss. Secondly, we design an embedding similarity measurement network for controlling the similarity between the two selected embeddings. We further jointly train the two new methods with the original network and achieve state-of-the-art. The multi-task training synergies are shown with a 9% reduction equal error rate (EER) and detected cost function (DCF) on the 2016 NIST Speaker Recognition Evaluation (SRE) Test Set.

TUE-PM2-O3
Image Processing

Time: Tuesday, Nov 19, 15:00-16:40

Place: A5

Chair: Isao Echizen

TUE-PM2-O3.1: Automatic Handwriting Verification and Suspect Identification for Chinese Characters Using Space and Frequency Domain Features

Wei-Cheng Liao and Jian-Jiun Ding

National Taiwan University

Automatic handwriting verification is to identify whether the script was written by a person himself or forged. Compared to related works about handwriting verification, the proposed algorithm adopts the features in both the time domain and the frequency domain. Moreover, in addition to distinguishing the forged manuscript from the genuine one, the proposed algorithm can also identify the suspect. The proposed algorithm is robust to writing instruments. In addition to the information of the luminance of the script, we also adopt the energy distribution on the 2-D frequency domain, the Pearson product-moment correlation coefficient (PPMCC) with genuine scripts, and vital information on characterized script points. Simulations show that the proposed method outperforms many advanced methods, including the deep-learning based method and manual identification by human beings. The proposed algorithm can well identify the script even if it is forged after several times of practice.

TUE-PM2-O3.2: Robust Change Detection in High Resolution Satellite Images with Geometric Distortions

Dongkwon Jin, Kyungsun Lim and Chang-Su Kim

Korea University

A robust change detection algorithm for high resolution satellite images, which are not perfectly registered, is proposed in this work. To achieve this goal, a change detection technique for registered images and an image registration technique are employed in a cooperative way. Specifically, we use not only hand-crafted features but also change detection results to match keypoints extracted from two images. We then align the images using the matching pairs of keypoints. Finally, we obtain a change map from the aligned images. These steps of image registration and change detection are alternately iterated until the convergence. Experimental results demonstrate that proposed algorithm outperforms the conventional change detection technique significantly, when there are geometric distortions between temporal satellite images.

TUE-PM2-O3.3: MSDC-Net: Multi-Scale Dense and Contextual Networks for Stereo Matching

Zhibo Rao, Mingyi He, Yuchao Dai, Zhidong Zhu, Bo Li and Renjie He

Northwestern Polytechnical University

Disparity prediction from stereo images is essential to computer vision applications such as autonomous driving, 3D model reconstruction, and object detection. To more accurately predict disparity map, a novel deep learning architecture (called MSDC-Net) for detecting the disparity map from a rectified pair of stereo images is proposed. Our MSDC-Net contains two modules: the multi-scale fusion 2D convolution module and the multi-scale residual 3D convolution module. The multi-scale fusion 2D convolution module exploits the potential multi-scale features, which extracts and fuses the different scale features by Dense-Net. The multi-scale residual 3D convolution module learns the different scale geometry context from the cost volume which aggregated by the multi-scale fusion 2D convolution module. Experimental results on Scene Flow and KITTI datasets demonstrate that our MSDC-Net significantly outperforms other approaches in the non-occluded region.

TUE-PM2-O3.4: Classification of Polarimetric SAR Image based on Improved Fuzzy Clustering

Zheng Cheng, Ping Han, Binbin Han and Jiahui Sun

Civil Aviation University of China

This paper presents an improved fuzzy clustering approach for Polarimetric SAR image by incorporating neighborhood information. Firstly, polarimetric scattering characteristics of the terrain in PolSAR image

are used to generate appropriate initial centers to avoid the issue that FCM is sensitive to random class centers. Then to further enhance the robustness to speckle noise, the conventional robust fuzzy C-mean clustering approach is improved. The work mainly exists in two aspects: (1) The revised Wishart distance is adopted as the data distance measure instead of Euclidean distance to assign a label to each pixel. (2) A weighted fuzzy membership is established by considering local spatial distance and class membership between the central pixel and its neighborhood simultaneously. Finally, the real polarimetric SAR data is utilized for the validation of the proposed unsupervised classification method. Experimental results demonstrate the superiority of the proposed method over the comparisons.

TUE-PM2-O3.5: Intensity-aware GAN for Single Image Reflection Removal

Nien-Hsin Chou Chou, Li-Chung Chuang and Ming-Sui Lee

National Taiwan University

Single image reflection removal is a challenging task in computer vision. Most existing approaches rely on carefully handcrafted priors to solve the problem. Contrast to the optimization-based methods, an intensity-aware GAN with dual generators is proposed to directly estimate the function which transforms the mixture image into the reflection image itself. From the observation that the reflection layer has more discriminating power in the region with low intensity than that in the region with high intensity, the proposed architecture better describes the characteristic of the model. Moreover, a reflection image synthesis method based on the screen blending model is also presented. Experimental results demonstrate that the results of reflection removal are satisfactory in real cases while comparing with state-of-the-art methods.

TUE-PM2-O3.6: CNN with ICA-PCA-DCT Joint Preprocessing for Hyperspectral Image Classification

Aamir Naveed Abbasi and Mingyi He

School of Electronics and Information, Northwestern Polytechnical University

In this paper a simpler convolutional neural network with a joint preprocessing is proposed for hyperspectral image classification. Primarily the spectral dimension of raw hyperspectral data cube is reduced in a unique fashion by using PCA and DCT such that the data is reduced effectively but having much information intact for classification task. The raw data cube is divided into two small spectrally reduced cubes, the first cube (PCA cube) is a simple PCA based dimension reduction considering few top principal components and the second cube (PDCT cube) performing DCT as preliminary step which confined maximum energy into low frequencies and then subsequently applying PCA by selecting same number of principal components as in the first PCA cube. After that both PCA and PDCT cubes are fused together, furthermore ICA is carried out on fused data cube to make the output classes much independent for next steps. In the final pre-processing step, the ICA performed data cube is divided into small square patches having labeled center pixel and a fixed size neighboring pixels by considering that in hyperspectral image neighboring pixels are highly correlated and having same class label. These square patches are fed into our simpler convolutional neural network which effectively and automatically extract the suitable features for our classification prediction job. The results validated that our acclaimed model which mainly exploit a novel pre-processing tactic and simpler but effective CNN performs enormously well in comparison to the other compared models and can be used as an effectual classification model for hyperspectral images in particular.

TUE-PM2-O4
Speech Synthesis

Time: Tuesday, Nov 19, 15:00-16:40

Place: A6

Chair: Jun Du

TUE-PM2-O4.1: Dongxiang speech systhesis based on statistical parameter method

Man Wang, Fangkun Qi, Hongwu Yang and Jingwen Sun

College of Physics and Electronic Engineering

Dongxiang language is a kind of text-free Chinese dialect. Therefore, it is difficult to employ traditional text-to-speech(TTS) technology to realize a Dongxiang dialect TTS system. The paper realized a Dongxiang dialect speech synthesis using Hidden Markov Model-based statistical parametric (HMM), Deep Neural Network (DNN)-based method and speaker adaptive training method by analyzing the linguistic features and acoustic characteristics of Dongxiang language. The experimental results show that, in the case of a certain corpus, the DNN-based speaker adaptive speech synthesis method can achieve better performance than the other two methods and can synthesize more natural speech.

TUE-PM2-O4.2: Human-in-the-loop speech-design system and its evaluation

Daichi Kondo and Masanori Morise

University of Yamanashi, Meiji University

We propose human-in-the-loop (HITL) speech-design system with an interface. General text-to-speech (TTS) systems generate the speech waveform from the input text without the need for manual modification.In particular, end-to-end TTS systems can synthesize speech as naturally as human speech. However, it is difficult for users to modify the speech parameters without degrading sound quality. The purpose of this study was to enable collaboration between the user and a deep neural network (DNN) to develop a system with which a user can control the speech parameters without sound-quality degradation. The main problem to be solved is to improve the quality of the speech-parameters generated from the speech parameters designed by the user. We developed several acoustic models with DNNs to meet the purpose of this study. We carried out a subjective evaluation to determine the effectiveness of the proposed system. The subjective score regarding Muffledness improved by using the proposed system compared with speech processed using a TTS system that involves signal-processing without a DNN.

TUE-PM2-O4.3: High-quality waveform generator from fundamental frequency, spectral envelope, and band aperiodicity

Masanori Morise and Takuro Shono

Meiji University, University of Yamanashi

This paper introduces a waveform generation algorithm from three speech parameters (fundamental frequency (F0), spectral envelope, and band aperiodicity). The conventional speech analysis/synthesis system based on a vocoder mainly has a waveform generator based on pitch synchronous overlap and add (PSOLA). Since it uses the fast Fourier transform (FFT) to generate the vocal cord vibration, the processing speed is proportional to the F0. The algorithm also uses the spectral representation of the aperiodicity, whereas the band aperiodicity is mainly used in speech synthesis applications such as statistical parametric speech synthesis. We propose a waveform generation algorithm that reduces the computational cost and memory usage without degrading the synthesized speech. The algorithm utilizes excitation signal generation by directly using the band aperiodicity. The computational cost in a certain period is fixed because the excitation signal is filtered and processed by the overlap-add (OLA) algorithm. We used the re-synthesized speech to perform two evaluations for the processing speed and sound quality. The results showed that the sound quality of speech synthesized was almost the same by our proposed algorithm as by the conventional algorithm. The proposed algorithm can also reduce computational cost and memory usage.

TUE-PM2-O4.4: A Study on Acoustic Parameter Selection Strategies to Improve Deep Learning-Based Speech Synthesis

Hyeonjoo Kang, Young-Sun Joo, Inseon Jang, Chunghyun Ahn and Hong-Goo Kang

Yonsei University, ETRI

In this paper, we investigate the variation in the performance of a deep learning-based speech synthesis (DLSS) system based on the configuration of output acoustic parameters. Our method is mainly

applicable for vocoding-based statistical parametric speech synthesis (SPSS), which has advantages in low-resource scenarios. Given the independence assumption of the source-filter model for the spectral and fundamental frequency F0 parameters, we propose a reliable network architecture for training acoustic parameters. Particularly, the F0 parameter suffers from high fluctuation and an extremely low number of dimensions. To relieve these problems, we introduce a context-window approach. Furthermore, we apply data augmentation to the proposed structure to overcome a lack of training data, which is a frequent issue with multi-speaker TTS systems. Experimental results confirm the superiority of the proposed algorithm over conventional ones in both single-speaker and multi-speaker TTS setups.

TUE-PM2-O4.5: End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training

Pengfei Wu, Zhenhua Ling, Lijuan Liu, Yuan Jiang, Hongchuan Wu and Lirong Dai

University of Science and Technology of China, iFLYTEK Research

This paper proposes an end-to-end emotional speech synthesis (ESS) method which adopts global style tokens (GSTs) for semi-supervised training. This model is built based on the GST-Tacotron framework. The style tokens are defined to present emotion categories. A cross entropy loss function between token weights and emotion labels is designed to obtain the interpretability of style tokens utilizing the small portion of training data with emotion labels. Emotion recognition experiments confirm that this method can achieve one-to-one correspondence between style tokens and emotion categories effectively. Objective and subjective evaluation results show that our model outperforms the conventional Tacotron model for ESS when only 5% of training data has emotion labels. Its subjective performance is close to the Tacotron model trained using all emotion labels.

TUE-PM2-O5
Speech Recognition

Time: Tuesday, Nov 19, 15:00-16:40

Place: A7

Chair: Xueliang Zhang

TUE-PM2-O5.1: End-to-end Tibetan Ando dialect speech recognition based on hybrid CTC/attention architecture

Jingwen Sun, Gang Zhou, Hongwu Yang and Man Wang

College of Physics and Electronic Engineering

End-to-end automatic speech recognition reduces the difficulty of building a speech recognition system through single network architecture. The tokenization, pronunciation dictionary and phonetic context-dependency trees required in the traditional deep learning-based speech recognition are omitted in this system to simplify the complex modeling process. This paper proposes a method to realize Tibetan Ando dialect speech recognition with end-to-end speech recognition model based on hybrid connectionist temporal classification (CTC)/attention. A bidirectional long short-term memory network (BLSTM) is used for the encoder network through 80 mel-scale filter-bank coefficients alone with pitch features form total 83-dimensionals acoustic features to train the network. We compared proposed method with the methods only based on CTC architecture and the structure only based on attention architecture by adjusting CTC weight of the system. The result shows that the hybrid model can obtain optimal weight to achieves the highest recognition rate of 64.5% when the CTC weight is 0.2.

TUE-PM2-O5.2: Multiple fixed beamformers with a spacial Wiener-form postfilter for far-field speech recognition

Sining Sun, Shuran Zhou, Mei—Yuh Hwang, Lei Xie, Qin Li and Xin Lei

Northwestern Polytechnical University, Mobvoi AI Lab

Far-field speech recognition is becoming a hot topic in research and industrial applications. In this paper, in order to improve far-field speech recognition performance, we propose to use multiple fixed beamformers with a spacial Wiener-form postfilter (MFB-SWP) to suppress noise and interference. Our proposed method consists of two parts, beamforming and postfilter estimation. First, multiple fixed beamformers are designed and each of them aims at one specific direction. Next the target speech is estimated using the fixed beamformer aiming to the target direction, and the noise and interference signals are estimated using the remaining beamformers. After that, we calculate a spacial Wiener-form time-frequency and frame-level gains, as postfilter to further reduce the residual noise and interference. Compared with a single fixed beamformer, the proposed MFB-SWP method can suppress noise and interference significantly. It is also computationally more efficient comparing with other adaptive beamforming methods. Our experiments showed that proposed method achieved 16-50% relative character error rate (CER) reduction compared with using the single fixed beamformer only.

TUE-PM2-O5.3: Teacher-Student BLSTM Mask Model for Robust Acoustic Beamforming

Zhaoyi Liu and Yuexian Zou

Peking University

Microphone array beamforming has been approved to be an effective approach for suppressing adverse interferences. Recently, acoustic beamformers employing neural networks (NN) for time-frequency (T-F) mask prediction, termed as Mask-BF, have received tremendous interest and shown great benefits as a front-end for distant automatic speech recognition (ASR). However, our preliminary experiments using Mask-BF for ASR task show that the mask model trained with only simulated training data underperforms when the real-recording data appears in the testing stage, where a data mismatch problem occurs. In this study, we aim at reducing the impact of the data mismatch on the mask model. Our research is quite intuitive that the real-recording data can be used together with the simulated data to make the mask model more robust against data mismatch problem. Specifically, two bi-directional long short-term memory (BLSTM) models, are designed as a teacher mask model and a student mask model, respectively. The teacher mask model is trained with simulated data, and it is then employed to generate the soft mask labels for both simulated and real-recording data separately. Then, the simulated data and the real-recording data with generated soft mask labels form the new training data to train the student mask model. As a result, a novel T-S mask BF is developed accordingly. Our T-S mask BF is evaluated as a front-end for ASR on the CHiME-3 dataset. Experimental results show that the generalization ability of our T-S mask BF is enhanced where we obtain relative 4% word error rate (WER) reduction compared to the baseline Mask-BF in the real-recording test set.

TUE-PM2-O5.4: Using Convolution and Sequence-discriminative Training to Improving Children Speech Recognition

Meng Fanchang, Peng Shouye and Zhang Guohui

Beijing Century TAL Education Technology Co.

The conclusion that ASR for children's speech is especially difficult compared to adult was given by the robotics community from recent works. Challenges on Children's speech recognition mainly due to the increased variability in acoustic and linguistic correlates depending on a young age. This work focused on the recognition of oral English spoken by Chinese children aging six to twelve. Experiments were conducted on: (1) Speaker Normalization algorithms, including Cepstral Mean and Variance Nor-malization (CMVN) and Vocal Tract Length Normalization (VTLN) techniques; (2) Acoustic models adapting techniques, such as Maximum Likelihood Linear Transform (MLLT) and Speaker Adaptive Training (SAT) based on Constrained MLLR; (3) Different acoustic models, GMM-HMM, DNN-HMM, CNN-DNN; (4) Training criterion, with frame-level training such as Cross entropy (CE), and sequence-discriminative training (SDT) such as MMI, MPE and sMBR were conducted in this paper. The re-sults included: (1) with the increase of age, the variability of children's pronunciation decreased significantly; (2) the convolu-tion on the frequency axis and sequence-discriminative training (CNN-DNN-sMBR) has a great performance contribution (34.72%) to the variability of children over the baseline system.

TUE-PM2-O5.5: Speech Recognition Based on Deep Tensor Neural Network and Multifactor Feature

Yahui Shan, Min Liu, Qingran Zhan, Shixuan Du, Jing Wang and Xiang Xie

Beijing Institute of Technology

This paper presents a speech recognition system based on deep tensor neural network which uses multifactor feature as input feature of acoustic model. First, a deep neural network is trained to estimate articulatory feature from input speech, where the training data is MOCHA database[1]. Mel frequency cepstrum coefficients in conjunction with articulatory feature are used as multifactor feature. Deep tensor neural network which involves tensor interactions among neurons is used as the acoustic model in this system. Speech recognition results indicate that the multifactor feature helps in improving speech recognition performance not only under clean conditions but also under noisy background conditions; deep tensor neural network is more capable of modeling multifactor features because of its tensor interactions than deep neural network.

TUE-PM2-O5.6: Can We Simulate Generative Process of Acoustic Modeling Data? Towards Data Restoration for Acoustic Modeling

Ryo Masumura, Yusuke Ijima, Satoshi Kobashikawa, Takanobu Oba and Yushi Aono

NTT

In this paper, we present an initial study on data restoration for acoustic modeling in automatic speech recognition (ASR). In the ASR field, the speech log data collected during practical services include customers' personal information, so the log data must often be preserved in segregated storage areas. Our motivation is to permanently and flexibly utilize the log data for acoustic modeling even though the log data cannot be moved from the segregated storage areas. Our key idea is to construct portable models that can simulate the generative process of acoustic modeling data so as to artificially restore the acoustic modeling data. Therefore, this paper proposes novel generative models called acoustic modeling data restorers (AMDRs), that can randomly sample triplets of a phonetic state sequence, an acoustic feature sequence, and utterance attribute information, even if original data is not directly accessible. In order to precisely model the generative process of the acoustic modeling data, we introduce neural language modeling to generate the phonetic state sequences and neural speech synthesis to generate the acoustic feature sequences. Experiments using Japanese speech data sets reveal how close the restored acoustic data is to the original data in terms of ASR performance.

TUE-PM2-O6
Speech Enhancement

Time: Tuesday, Nov 19, 15:00-16:40

Place: A8

Chair: Meng Sun

TUE-PM2-O6.1: Noise Prior Knowledge Learning for Speech Enhancement via Gated Convolutional Generative Adversarial Network

Cunhang Fan, Bin Liu, Jianhua Tao, Jiangyan Yi, Zhengqi Wen and Ye Bai

Institute of Automation, Chinese Academy of Sciences

Speech enhancement generative adversarial network (SEGAN) is an end-to-end deep learning architecture, which only uses the clean speech as the training targets. However, when the signal-to-noise ratio (SNR) is very low, predicting clean speech signals could be very difficult as the speech is dominated by the noise. In order to address this problem, in this paper, we propose a gated convolutional neural network (CNN) SEGAN (GSEGAN) with noise prior knowledge learning to address this problem. The proposed model not only estimates the clean speech, but also learns the noise prior knowledge to assist the speech enhancement. In addition, gated CNN has an excellent potential for capturing long-term temporal dependencies than regular CNN. Motivated by this, we use a gated CNN architecture to acquire more detailed information at waveform level instead of regular CNN. We evaluate the proposed method GSEGAN on Voice Bank corpus. Experimental results show that the proposed method GSEGAN outperforms the SEGAN baseline, with a relative improvement of 0.7%, 28.2% and 43.9% for perceptual evaluation of speech quality (PESQ), overall Signal-to-Noise Ratio (SNRovl) and Segmental Signal-to-Noise Ratio (SNRseg), respectively.

TUE-PM2-O6.2: Domain Adversarial Training for Speech Enhancement

Nana Hou, Chenglin Xu, Eng Siong Chng and Haizhou Li

School of Computer Science and Engineering, Nanyang Technological University, Department of Electrical and Computer Engineering, National University of Singapore

The performance of deep learning approaches to speech enhancement degrades significantly in face of mismatch between training and testing. In this paper, we propose a domain adversarial training technique for unsupervised domain transfer, that 1) overcomes domain mismatch, and 2) provides a solution to the scenario where we only have noisy speech data, and we don't have clean-noisy parallel data in the new domain. Specifically, our method includes two parts that are jointly trained, 1) an enhancement net to map noisy speech to clean speech by indirectly estimating a mask with a spectrum approximation loss, and 2) a domain predictor to distinguish between domains. As the proposed approach is able to adapt to a new domain only with noisy speech data in target domain, we call it an unsupervised learning technique. Experiments suggest that our approach delivers voice quality comparable with other supervised learning techniques that require clean-noisy parallel data.

TUE-PM2-O6.3: Subjective Feedback-based Neural Network Pruning for Speech Enhancement

Fuqiang Ye, Yu Tsao and Fei Chen

Southern University of Science and Technology, Research Center for Information Technology Innovation

Speech enhancement based on neural networks provides performance superior to that of conventional algorithms. However, the network may suffer owing to redundant parameters, which demands large unnecessary computation and power consumption. This work aimed to prune the large network by removing extra neurons and connections while maintaining speech enhancement performance. Iterative network pruning combined with network retraining was employed to compress the network based on the weight magnitude of neurons and connections. This pruning method was evaluated using a deep denoising autoencoder neural network, which was trained to enhance speech perception under nonstationary noise interference. Word correct rate was utilized as the subjective intelligibility feedback to evaluate the understanding of noisy speech enhanced by the sparse network. Results showed that the iterative pruning method combined with retraining could reduce 50% of the parameters without significantly affecting the speech enhancement performance, which was superior to the two baseline conditions of direct network pruning with network retraining and iterative network pruning without

network retraining. Finally, an optimized network pruning method was proposed to implement the iterative network pruning and retraining in a greedy repetition manner, yielding a maximum pruning ratio of 80%.

TUE-PM2-O6.4: Compressed Multimodel Hierarchical Extreme Learning Machine for Speech Enhancement

Tassadaq Hussain, Yu Tsao, Hsin-Min Wang, Jia-Ching Wang, Sabato Marco Siniscalchi and Wen-Hung Liao

Institute of Information Science, Academia Sinica, Research Center for Information Technology Innovation (CITI) at Academia Sinica,Academia Sinica, National Central University, Kore University of Enna, National Chengchi University

Recently, model compression that aims to facilitate the use of deep models in real-world applications has attracted considerable attention. Several model compression techniques have been proposed to reduce computational costs without significantly degrading the achievable performance. In this paper, we propose a multimodal framework for speech enhancement (SE) by utilizing a hierarchical extreme learning machine (HELM) to enhance the performance of conventional HELM-based SE frameworks that consider audio information only. Furthermore, we investigate the performance of the HELM-based multimodal SE framework trained using binary weights and quantized input data to reduce the computational requirement. The experimental results show that the proposed multimodal SE framework outperforms the conventional HELM-based SE framework in terms of three standard objective evaluation metrics. The results also show that the performance of the proposed multimodal SE framework is only slightly degraded when the model is compressed through model binarization and quantized input data.

TUE-PM2-O6.5: Speech Enhancement Based on Deep Mixture of Distinguishing Experts

Xupeng Jia and Dongmei Li

Tsinghua University

In this work, we propose a new strategy for deep mixture of experts (DMoE) based speech enhancement. DMoE system is difficult to train due to the specific network structure and the necessity of carefully designed pre-training methods to guarantee good performance. We propose using distinguishing deep neural networks (DNNs) as experts, dealing with magnitude spectrogram and log-magnitude spectrogram respectively. The proposed method is compared with the state-of-art DMoE system utilizing hard expectation maximization (HEM) pre-training method. Speech enhancement experiments in 30 (5*6) noise and SNR conditions show the superiority of the proposed method over the baseline method. The average improvements obtained for matched conditions are 0.076 in perceptual evaluation of speech quality (PESQ), 1.824dB in segmental signal to noise ratio (segSNR) and 0.043 in short time objective intelligibility (STOI).

TUE-PM3-SS1
Special Learning under Limited Samples Scenarios

Time: Tuesday, Nov 19, 17:00-18:40

Place: A2

Chairs: Ganggang Dong, Yinghua Wang, Bo Chen

TUE-PM3-SS1.1: Semi-supervised Multimodal Emotion Recognition With Improved Wasserstein GANs

Jingjun Liang, Shizhe Chen and Qin Jin

Renmin University of China

Automatic emotion recognition has faced the challenge of lacking the large-scale human labeled dataset for model learning due to the expensive data annotation cost and inevitable label ambiguity. To tackle such challenge, previous works have explored to transfer emotion label from one modality to the other modality assuming that the supervised annotation does exist in one modality or explored semi-supervised learning strategies to take advantage of large amount of unlabeled data with the focus on a single modality. In this work, we address the multi-modal emotion recognition problem with the acoustic and visual modalities and propose a multi-modal network structure of the semi-supervised learning approach with an improved generative adversarial network CT-GAN. Extensive experiments conducted on a multi-modal emotion recognition corpus demonstrate the effectiveness of the proposed approach and prove that utilizing unlabeled data via GANs and combining multi-modalities both benefit the classification performance. We also carry out some detailed analysis experiments such as influence of unlabeled data quantity on recognition performance and impact of different normalization strategies for semi-supervised learning etc.

TUE-PM3-SS1.2: Robust Attack on Deep Learning based Radar HRRP Target Recognition

Yijun Yuan, Jinwei Wan and Bo Chen

National Laboratory of Radar Signal Processing in Xidian University

Deep learning methods have attracted increasing attention in the past years due to their powerful ability to learn useful features from the dataset automatically, and showed high recognition performance. But recent studies show that data-driven based methods are vulnerable to adversarial examples in the field of computer vision, resulting from small-magnitude perturbations added to the input. In this paper, we verified adversarial examples also exist in the field of HRRP-based radar automatic target recognition. And then we propose a noval adversarial attack algorithm, Robust HRRP Attack(RHA), which could generate robust adversarial perturbations in real-world. We show RHA decrease HRRP recognition performance significantly. As far as we know, this is the first paper which focus on the adversarial examples in the filed of HRRP-based radar automatic target recognition.

TUE-PM3-SS2
Signal Processing in Behavior Analysis

Time: Tuesday, Nov 19, 17:00-18:40

Place: A3

Chairs: Kazushi Ikeda, Li-Wei Kang

TUE-PM3-SS2.1: Development of a Chinese Depressed Speech Corpus Based on The Disturbed Effect of Self-processing

Xiaoyong Lu, Yanqin Li, Haizhen An and Tao Pan

Northwest Normal University, Lanzhou Resources and Environment Voc–Tech College

Depression has long been recognized as one of the leading causes of disability and burden worldwide. In psychology, it is well known that the self is not only the cognitive subject, but also the core of personality. And the high incidence of suicide and pervasive hopelessness in depressed individuals suggested that the self might be abnormal among them. In light of the psychological characteristics, we employ classical scientific psychology paradigms on abnormalities of self-related processing in depressed individuals to develop a Chinese depressed speech corpus. Eleven depressed individuals and ten healthy subjects, who are gender-balanced and age-balanced, were recruited to participate in this work. Currently we have preliminarily collected 6 and 2.5 hours of speech data respectively, with the results of preliminary analysis indicating that there exist abnormalities in the depressed speech. The study results will provide a new perspective and strategy for further study on the building and application of speech corpus in depression.

TUE-PM3-SS2.2: Statistical analysis on characteristic whisker movements observed in reward processing

Junichiro Yoshimoto, Jumpei Ozaki, Kohta Mizutani, Takashi Nakano, Kazushi Ikeda and Takayuki Yamashita

Nara Institute of Science and Technology, Osaka University, Nagoya University

Internal states of the brain can be often reflected as facial expressions. However, how animals show their facial expression is largely unexplored. Here, we focus on mice and investigate whether their whisker movements could be a facial expression of their internal states related to reward processing. We trained three mice for an auditory association task and filmed their whiskers during the task performance after enough learning. We found that approximately 5-8 Hz periodic whisking was commonly observed during reward-associated Go cue presentation. Such whisking rarely occurred in No-Go cue trials or in Go cue trials where the mice were not motivated to get a reward. Furthermore, after acquiring a reward, the mice whisked with a more protracted set-point. Using machine learning, we obtained computer algorithms that could accurately indicate reward-anticipating and reward-acquiring trials only from whisker time plots. Our analyses suggest that mice exhibit stereotypic whisker movements as a part of orofacial movements related to reward anticipation and acquisition.

TUE-PM3-SS2.3: Interaction Analysis in Hunting Behavior of Finless Porpoises

Mikiko Konda, Takatomi Kubo, Naruki Morimura and Kazushi Ikeda

Nara Institute of Science and Technology, Kyoto University

Finless porpoises (Neophocaena asiaeorintalis) usually live in a relatively small group. Aggregated porpoises prey on fish simultaneously and they do not show explicit allotted roles like other dolphins. However, whether each individual hunt fish independently from others is not clear. In this study, we tried to find dependency in porpoises' hunting using movies of their feeding taken by a drone.

TUE-PM3-SS2.4: Extraction of Biomolecular Signals Controlling Complex Behavior of Biological Cells

Yuichi Sakumura and Katsuyuki Kunida

Nara Institute of Science and Technology

Cell deformation and migration are one of the important biological phenomena related to various biological functions. In this study, when giving the enormous time-series data of cell morphological changes (protrusion and retraction of cell leading edge) and activities of regulatory molecules, we propose a novel data analysis method for extracting molecular activity patterns corresponding to

specific morphological changes based on the reverse correlation analysis. As an example of analysis, using time-series data of three representative regulatory molecules, we extracted the dynamical molecular activities for four types of the morphological change; sustained protrusion and retraction, and transient protrusion and retraction. As results, it was confirmed that extraction results consistent with previous molecular biological findings were obtained.

TUE-PM3-SS2.5: Physiological signals responses to normal and abnormal brake events in simulated autonomous car

Yaming Hu, Shun Nakamura, Tsuyoshi Yamanaka and Toshihisa Tanaka

Tokyo University of Agriculture and Technology; CorLab Inc., Innovative Technology Development Department, JATCO Ltd.

An autonomous driving technology brings a new challenge of communication between human and autonomous car. The development of this new technology has been devoted to safety, but comfortability for drivers has not been interested. The safety is necessary, but the comfortability should be considered to achieve a reliable technology. Thus we investigated how the mental state appears in physiological signals during autonomous driving. To this end, we explored the responses of physiological signals to normal and abnormal brake situations in a simulated autonomous car. It was assumed that a brake timing of drivers is different from each other. Thus, we created normal and abnormal brake scenes in autonomous driving simulation. Then, we recorded EEG and ECG signals during the normal and abnormal brake situation of autonomous driving. In the abnormal brake situation, after brake, an event-related desynchronization (ERD) of the frequency band of 8 to 13 Hz observed in some subjects. Those subjects showed higher tension by analysis RR interval of ECG: the ratio of LF power to HF power during abnormal brake situation was higher than normal brake situation.

TUE-PM3-SS2.6: A One-Dimensional Convolutional Neural Network Model for Automated Localization of Epileptic Foci

Boning Li, Xuyang Zhao, Qibin Zhao, Toshihisa Tanaka and Jianting Cao

Saitama Institute of Technology, RIKEN AIP, Tokyo University of Agriculture and Technology, Brain Science Institute,RIKEN

iEEG (intracranial electrocorticogram) is often used by clinical experts to determine the location of the epileptic focal in the treatment of epilepsy. However, assess the location of epileptic foci by using iEEG is time-consuming and strenuous for clinical experts. Technology for automated localization of the channel of epileptic focal is indispensable. Hence, we developed a one-dimensional convolutional neural network (1D-CNN) model, which can directly extract features and train model by the raw signals without preprocessing, and performed the classification of focal and nonfocal epileptic iEEG signals. Compared with other machine learning methods, the amount of parameter reduced significantly. Our developed model has yielded the classification accuracy of 85.14% in classifying the focal and nonfocal epileptic iEEG signals.

TUE-PM3-SS3
Latest Progress on Fractional Signal Processing Theory and Application

Time: Tuesday, Nov 19, 15:00-16:40

Place: A4

Chairs: Bing-Zhao Li, Xiaolong Chen, Yong Guo

TUE-PM3-SS3.1: Cohen's class time-frequency representation in linear canonical domains: definition and properties

Zhichao Zhang, Maokang Luo, Ke Deng and Tao Yu

Sichuan University

The traditional Cohen's class time-frequency representation is extended to the linear canonical domain by using a well-established closed-form instantaneous cross-correlation function (CICF) type of linear canonical transform (LCT) free parameters embedded approach, resulting in the CICF type of Cohen's class (CICFCC) that unifies some well-known Cohen's classes in linear canonical domains including the affine characteristic (AC), basis function (BF), convolution expression (CE) and instantaneous cross-correlation function (ICF) types of Cohen's classes, and can be considered as the Cohen's class's closed-form representation in linear canonical domains. A fundamental theory about the CICFCC's essential properties, such as marginal distribution, energy conservation, unique reconstruction, Moyal formula, complex conjugate symmetry, time reversal symmetry, scaling property, time shift property, frequency shift property, and LCT invariance, is also established for the requirement of future practical applications.

TUE-PM3-SS3.2: LFM Signal Detection and Estimation Based on Deep Convolutional Neural Network

Xiaolong Chen, Qiaowen Jiang, Ningyuan Su, Baoxin Chen and Jian Guan

Naval Aviation University

Linear frequency modulation (LFM) signal detection and estimation are important for radar, communication, or spectrum analysis etc.. As the generalized form of Fourier transform (FT), Fractional FT (FRFT) has good energy aggregation ability for LFM signal and can reflect the Doppler variation, which is suitable for LFM signal detection and estimation. However, it needs two-dimensional parameters searching and for multiple signals it requires searching one by one and easily affected by strong signals with poor resolution. In this paper, the convolutional neural network (CNN) is applied for replacing the FT and FRFT and used for signal frequency signal and LFM signal detection and estimation. The pre-trained CNN model can establish the relations among various single frequency signal or LFM signal and the two dimensional parameters domain. By simulation, it is found that the CNN based method can also achieve the function of FRFT and has the advantages of high precision and resolution. And it is proved that the CNN based method can achieve good recognition performance even at lower signal-to-noise ratio (SNR) combined with the denoising method. The proposed method would provide a novel solution for radar moving target detection, as well as speech intelligent signal processing, sonar signal processing, etc..

TUE-PM3-SS3.3: Nonuniform fast linear canonical transform

Yan—Nan Sun and Bing—Zhao Li

Jiangsu University, Beijing institude of techonlogy

The linear canonical transform (LCT) is a generalized form of the Fourier transformation.
It has been shown to be one of the most powerful tools in applied mathematics, signal processing and optics fields. The aim of this paper is to present a nonuniform fast linear canonical transform (NFLCT), which emerges in many areas of physics and engineering. The proposed algorithm generalizes the fast linear canonical transform to the case of non-integer frequencies on the interval $[-b\pi,b\pi]$. The algorithm requires $O(N \log N + N\log(1/\epsilon))$ arithmetic operations where $\epsilon$ is the precision of computations and $N$ is the number of nodes. The efficiency of the approach is illustrated by simulations.

TUE-PM3-SS3.4: Digital implementation of Hilbert Transform in the LCT domain associated with FIR filter

Deng Bing, Huang Qingshun and Zhang Lin

Naval Aviation University

In this paper, digital implementation of Hilbert Transform in the LCT domain is proposed based on FIR filter. First, they are described about the definition of linear canonical transform, Hilbert transform, and analytical signal. Then, the Implementation principle is analyzed about Hilbert transform in the LCT domain. Finally, the digital implementation method is proposed based on FIR filter.

TUE-PM3-SS3.5: Adaptive Matching Pursuit Method Based on Auxiliary Residual for Sparse Signal Recovery

Juan Zhao and Xia Bai

Beijing Institute of Technology

Greedy pursuit methods are widely used for compressive sensing (CS) and sparse signal recovery due to their low computational complexity. In this paper an adaptive matching pursuit is proposed, which is based on the backtracking-based adaptive orthogonal matching pursuit (BAOMP) and uses auxiliary residual to make correlation test to add more correct atoms per iteration. The proposed method can be regarded as an improved BAOMP. The simulation results show that it has better performance to those of some other greedy pursuit methods. Finally the experiment of CS-based ISAR imaging verifies the effectiveness of the proposed method.

TUE-PM3-SS3.6: Decomposition of Covariance Matrix Using Cascade of Trees

Navid Tafaghodi Khajavi and Anthony Kuh

University of Hawaii

We are looking at the statistical model approximation for jointly Gaussian random vectors. To do so, we are using a cascade of linear transformations that go beyond tree approximations. Here, we propose an algorithm which incorporates the Cholesky factorization method to compute the decomposition matrix and thus can approximate a simple graphical model using a cascade of the Cholesky factorization of the tree approximation transformations. The Cholesky decomposition keeps the sparsity pattern of the inverse decomposition and thus reduces computations for the tree structure linear transformation at each cascade stage of the algorithm. This is a different perspective on the approximation model, and algorithms such as Gaussian belief propagation can be used on this overall graph. We conclude with some simulation results.

TUE-PM3-SS4
High Performance Image and Video Processing and Applications

Time: Tuesday, Nov 19, 17:00-18:40

Place: A5

Chair: Kosin Chamnongthai

TUE-PM3-SS4.1: Handwritten Text Segmentation in Scribbled Document via Unsupervised Domain Adaptation

Junho Jo, Jae Woong Soh and Nam Ik Cho

Department of ECE

Supervised learning methods have shown promising results for the handwritten text segmentation in scribbled documents. However, many previous methods have handled the problem as a connected component analysis due to the extreme difficulty of pixel-level annotations. Although there is an approach to solve this problem by using synthetically generated data, the resultant model does not generalize well to real scribbled documents due to the domain gap between the real and synthetic dataset. To alleviate the problems, we propose an unsupervised domain adaptation strategy for the pixel-level handwritten text segmentation. This is accomplished by employing an adversarial discriminative model to align the source and target distribution in the feature space, incorporating entropy minimization loss to make the model discriminative even for the unlabeled target data. Experimental results show that the proposed method outperforms the baseline network both quantitatively and qualitatively. Specifically, the proposed adaptation strategy mitigates the domain shift problem very well, showing the improvement of baseline performance (IoU) from 64.617% to 85.642%.

TUE-PM3-SS4.2: Research on Cloud Recognition Technology Based on Transfer Learning

Chunyao Fang, Kebin Jia, Pengyu Liu and Liang Zhang

Beijing University of Technology

The cloud is an important part of the earth's thermodynamic balance and water and air cycle. At present, abundant achievements have been made in the research of satellite cloud image, while the recognition of ground-based cloud image has always been a difficulty in the field of pattern recognition, and the achievements are relatively limited. In this paper, based on the ground-based cloud map data set provided by standard weather stations, after data enhancement, 5 network models were trained by means of fine-tuning network parameters and freezing weights of different network layers, and 5 network migration configurations were used on the enhanced data set. Experimental results show that the fine-tuned Densenet network model is more suitable for this project, and the recognition accuracy can reach 94.42%.

TUE-PM3-SS4.3: Saliency Detection via Robust Seed Selection of Foreground and Background Priors

Muwei Jian, Ruihong Wang and Kin-Man Lam

Shandong University of Finance and Economics, Ocean University of China, The Hong Kong Polytechnic University

Recently, saliency detection has become a research hotspot in both the computer-vision and image-processing fields. Among the diverse saliency-detection approaches, those based on the foreground and background-based model can achieve promising performance. Reliable seed selection for the foreground and background priors is a critical step for successful saliency detection. In this paper, we firstly exploit the spatial distribution of the extracted directional patches to predict the centroid of each salient object in an image. Then, we adopt the located centroids as the visual-attention center of the whole image to compute the superpixel-based center prior, which can facilitate the subsequent seed selection for the foreground and background-prior generation. Finally, the two individual foreground-based and background-based saliency maps are combined together into a plausible and authentic saliency map. Extensive experimental assessments on publicly available datasets demonstrate the effectiveness of our proposed model.

TUE-PM3-SS4.4: A Hue Correction Scheme Based on Constant-Hue Plane for Color Image Enhancement

Yuma Kinoshita, Kouki Seo and Hitoshi Kiya

Tokyo Metropolitan University

In this paper, we propose a novel hue correction scheme based on constant-hue plane in the RGB color space for color image enhancement. A number of hue-preserving image enhancement methods have already been proposed. Although these methods can preserve hue, these methods cannot be applied to the state-of-the-art enhancement methods such as deep-learning based ones. We therefore generalize a hue-preserving method based on the constant-hue plane in this paper. This generalization derives our novel hue correction scheme. In the proposed scheme, any existing image enhancement method including deep-learning based ones can be used to enhance images. The hue distortion due to the enhancement is then removed by replacing the maximally saturated colors of an enhanced image with those of the corresponding input one. Experimental results show that the proposed scheme is effective to suppress the hue distortion due to two color enhancement methods including a deep-learning based one. Furthermore, objective quality evaluations demonstrate that the proposed scheme can maintain the performance of image enhancement methods.

TUE-PM3-SS4.5: A spatial domain secret image embedding technique with image authentication feature

S.K. Felix Yu, Zi Xin Xu, Yuk Hee Chan and Pak Kong Lun

The Hong Kong Polytechnic University, Chengdu University of Information Technology

In practical applications, it is required to provide a means to authenticate an information-embedded image such that its integrity can be guaranteed. However, conventional studies generally consider image hiding and image authentication as two different tasks. When both are required, the secret image and a fragile watermark are separately embedded into a cover image. In this paper, to address this issue, we propose a spatial domain image embedding scheme that can embed rich pictorial information and fragile watermark simultaneously into a cover image with the same technique to reduce the complexity and improve the efficiency.

TUE-PM3-SS4.6: Reconstruction of Multitone BTC Images using Conditional Generative Adversarial Nets

Jing-Ming Guo and Sankarasrinivasan Seshathiri

National Taiwan University of Science and Technology

Multitone Block Truncation Coding (MT-BTC) image is the superior version of halftone based BTC compression methods. The MT-BTC images are developed based on ordered dithering method which further utilize multitone dithering mask for image construction. During the transformation, the original image is processed in a block-wise manner and is approximated in terms of the maximum, minimum and their intermediate values of the respective block. In comparison with standard compressions techniques, MTBTC possess very unique representation and suffers from inherent halftone noises. In this paper, a simplified version of image to image translation architecture is developed based on cGAN's. To begin with, a MT-BTC database is developed using the latest multitone approach, and it comprise of around 10,000 images. Further, the proposed cGAN's model is optimized to perform with minimal layers and reduced parameters. The PatchGAN discriminator is adjusted to judge over the patch size of 64x64 which has good impact over quality improvements. From the comprehensive performance evaluation, it has been validated that the proposed approach can achieve consistent and improved reconstruction quality.

TUE-PM3-O1
Speaker Recognition

Time: Tuesday, Nov 19, 17:00-18:40

Place: A6

Chair: Dong Wang

TUE-PM3-O1.1: Data augmentation and post selection for improved replay attack detection

Yuanjun Zhao, Roberto Togneri and Victor Sreeram

School of Electrical

Vulnerabilities of the Automatic Speaker Verification (ASV) technology have been recognized and have generated much interest to design anti-spoofing detectors. Replay attacks pose a severe threat due to the relative difficulty for detection and the ease in mounting spoofing attacks. In this paper, a high performing spoofing detection countermeasure is presented. Deep Learning (DL) based speech embedding extractors and a novel data augmentation approach are combined to improve the detection performance. To select augmented samples with high quality and diversity and avoid the bias caused by human subjective perception, we propose the use of a Support Vector Machine (SVM) based post-filter. With the generated extra informative training data, problems of over-fitting and lack of generalization can be significantly alleviated. Experimental results measured by equal error rates (EERs) indicate a relative improvement of 30% on the development and evaluation subsets. This provides the motivation for the proposed audio data augmentation and also promotes the future research on generated samples selection in the application of speaker spoofing detection.

TUE-PM3-O1.2: Deep Segment Attentive Embedding for Duration Robust Speaker Verification

Bin Liu, Shuai Nie, Wenju Liu, Hui Zhang, Xiangang Li and Changliang Li

National Laboratory of Patten Recognition, Institute of Automation, Chinese Academy of Sciences, DiDi AI Labs, kingsoft AI lab

Deep learning based speaker verification usually uses a fixed-length local segment randomly truncated from an utterance to learn the utterance-level speaker embedding, while using the average embedding of all segments of a test utterance to verify the speaker, which results in a critical mismatch between testing and training. This mismatch degrades the performance of speaker verification, especially when the durations of training and testing utterances are very different. To alleviate this issue, we propose the deep segment attentive embedding method to learn the unified speaker embeddings for utterances of variable duration. Each utterance is segmented by a sliding window and LSTM is used to extract the embedding of each segment. Instead of only using one local segment, we use the whole utterance to learn the utterance-level embedding by applying an attentive pooling to the embeddings of all segments. Moreover, the similarity loss of segment-level embeddings is introduced to guide the segment attention to focus on the segments with more speaker discriminations, and jointly optimized with the utterance-level embeddings loss. Systematic experiments on DiDi Speaker Dataset, Tongdun and VoxCeleb show that the proposed method significantly improves system robustness and achieves the relative EER reduction of 18.3%, 50% and 11.54% , respectively.

TUE-PM3-O1.3: Sequential Speaker Embedding and Transfer Learning for Text-Independent Speaker Identification

Qian-Bei Hong, Chung-Hsien Wu, Ming-Hsiang Su and Hsin-Min Wang

National Cheng Kung University and Academia Sinica, National Cheng Kung University

In this study, an approach to speaker identification is proposed based on a convolutional neural network (CNN)-based model considering sequential speaker embedding and transfer learning. First, a CNN-based universal background model (UBM) is constructed and a transfer learning mechanism is applied to obtain speaker embedding using a small amount of enrollment data. Second, considering the temporal variation of acoustic features in an utterance of a speaker, this study generates sequential speaker embedding to capture temporal characteristics of speech features of a speaker. Experiments were conducted on the King-ASR series database for UBM training, and the LibriSpeech corpus was adopted for evaluation. The experimental results showed that the proposed method using sequential speaker embedding and transfer learning achieved an equal error rate (EER) of 6.89% outperforming the method based on x-vector and PLDA method (8.25%). Furthermore, we considered the effect of speaker number for speaker identification. When the number of enrolled speakers was from 50 to 1172, the identification accuracy of the proposed method was degraded from 82.99% to 73.26%, which outperformed the identification accuracy of the method using x-vector and PLDA which was dramatically degraded from 83.17% to 60.95%.

TUE-PM3-O1.4: Improving replay attack detection by combination of spatial and spectral features

Ryoya Yaguchi, Sayaka Shiota, Nobutaka Ono and Hitoshi Kiya

Tokyo Metropolitan University

In this paper, we propose a replay attack detection based on score fusion of spatial and spectral features-based systems. Recently, a replay attack detection (RAD) system using generalized cross-correlation (GCC) of a stereo signal has been proposed. The GCC is calculated from non-speech sections of input signals. It reported that the GCC-based method achieved high performance under several situations. However, since the performance of the GCC-based method depends on the situations, it is required to improve the performance without situation dependence. The GCC-based method uses spatial features, which utilize the different feature from spectral features. In this paper, we perform score fusion of the GCC-based and the spectral feature-based methods to improve the robustness of RAD systems. In the experiments, the proposed method achieved a relative error reduction of 69.5%, compared with a GCC-based single method under one of the hard tasks. And, the performance of score fusion systems improved without situation dependence.

TUE-PM3-O1.5: Multi-band Spectral Entropy Information for Detection of Replay Attacks

Yitong Liu, Rohan Kumar Das and Haizhou Li

National University of Singapore

Replay attacks have been proven to be a potential threat to practical automatic speaker verification systems. In this work, we explore a novel feature based on spectral entropy for detection of replay attacks. The spectral entropy is a measure to capture spectral distortions and flatness. It is found that the replay speech carries artifacts in the process of recording and playback. We hypothesize that spectral entropy can be a useful information to capture such artifacts. In this regard, we explore multi-band spectral entropy feature for replay attack detection. The studies are conducted on ASVspoof 2017 Version 2.0 database that deals with replay speech attacks. A baseline system with popular constant-Q cepstral coefficient (CQCC) feature is also developed. Finally, a combined system is proposed with multi-band spectral entropy and CQCC features that outperforms the baseline and validates the idea of multi-band spectral entropy feature.

TUE-PM3-O2
Speech Recognition

Time: Tuesday, Nov 19, 17:00-18:40

Place: A7

Chair: Hsin-Min Wang

TUE-PM3-O2.1: Knowledge Distillation from Multilingual and Monolingual Teachers for End-to-End Multilingual Speech Recognition

Jingyi Xu, Junfeng Hou, Yan Song, Wu Guo and Lirong Dai

University of Science and Technology of China

Attention-based encoder-decoder models significantly reduce the burden of developing multilingual speech recognition systems. By means of end-to-end modeling and parameters sharing, a single model can be efficiently trained and deployed for all languages. Although the single model benefits from jointly training across different languages, it should handle the variation and diversity of the languages at the same time. In this paper, we exploit knowledge distillation from multiple teachers to improve the recognition accuracy of the end-to-end multilingual model. Considering that teacher models learning from monolingual and multilingual data contain distinct knowledge of specific languages, we introduce multiple teachers including monolingual teachers of each language, and multilingual teacher to teach a same sized multilingual student model so that the multilingual student will learn various knowledge embedded in the data and intend to outperform multilingual teacher. Different from conventional knowledge distillation which usually relies on a linear interpolation for hard loss from true label and soft losses from teachers, a new random augmented training strategy is proposed to switch the optimization of the student model between hard or soft losses in random order. Our experiments on Wall Street Journal (English) and AISHELL-1 (Chinese) composed multilingual speech dataset show the proposed multiple teachers and distillation strategy boost the performance of the student significantly relative to the multilingual teacher.

TUE-PM3-O2.2: Learning Adaptive Downsampling Encoding for Online End-to-End Speech Recognition

Rui Na, Junfeng Hou, Wu Guo, Yan Song and Lirong Dai

University of Science and Technology of China

Attention based encoder-decoder models have shown promising performance for various sequence-to-sequence problems. However, for speech recognition, the very long input speech consumes a lot of computation and memory resource when performing encoding and soft attention over the input sequence. While fixed-rate downsampling is usually employed to reduce the computation steps, it fails to consider the variable durations of phonemes. Motivated by this, we propose a differentiable adaptive downsampling approach which encodes the input sequence with a recurrent layer by keeping crucial frames and discarding redundant frames adaptively in real-time. Therefore, the proposed downsampling approach can dynamically generate input hidden representations and is suitable for online end-to-end speech recognition. Experiments show that our proposed method can reduce phone error rate (PER) by 7.0% relative without loss of speed compared with fixed downsampling technique. In addition, the adaptive encoding makes the model robust to variable speed speech.

TUE-PM3-O2.3: Multi-task Learning for Acoustic Modeling Using Articulatory Attributes

Yueh–Ting Lee, Xuan–Bo Chen, Hung–Shin Lee, Jyh–Shing Roger Jang and Hsin–Min Wang

National Taiwan University, Institute of Information Science, Academia Sinica

In addition to the phone sequences, articulatory attributes in spoken utterances have demonstrated salient cues for supervised training of acoustic models in automatic speech recognition (ASR). In this paper, a multi-task learning (MTL) scheme for neural network-based acoustic modeling is proposed. It aims to simultaneously minimize the cross-entropy losses of the triphone-states and articulatory attributes, given their corresponding true alignments. Supposing the articulatory information associated with the physical process is not as abstract and composite as the phonetic descriptions, the layer-wise neuron sharing occurs only in the first few layers. Moreover, instead of the fully-connected feed-forward networks (FFNs), the well-known structure of time-delay neural networks (TDNNs) is adopted to efficiently model the long-term contexts of each acoustic input frame. The results of experiments on the MATBN Mandarin Chinese broadcast news corpus show that our proposed framework achieves relative character error rate

reductions of 3.3% and 5.7% over the non-MTL TDNN-based system and the MTL-FFN-based system, respectively.

TUE-PM3-O2.4: Hypothesis Correction Based on Semi-character Recurrent Neural Network for End-to-end Speech Recognition

Yuuki Tachioka

Denso IT Laboratory

End-to-end automatic speech recognition (ASR) has become popular because of its simple modeling, but it encounters out-of-vocabulary words more frequently than the conventional hybrid approaches that jointly use acoustic and language models. In particular, word-based end-to-end systems cannot output any words unseen in training data. To address this problem, character-based end-to-end systems have been proposed; however, they are susceptible to noise, and their output words are not necessarily correct in terms of language. This is because language constraints, such as lexicons and language models, are lacking in the decoding process. Thus, errors like misspellings occur frequently. In the field of natural language processing, to correct spelling errors, a semi-character recurrent neural network (scRNN) was proposed whose inputs are the counts of characters in a word and outputs are word ids. To apply scRNN to ASR, extensions are needed because scRNN focuses only on substitution errors. Here, to consider insertion and deletion errors, we introduce blank word symbols, similar to blank symbols in connectionist temporal classification, and word concatenation. Two different ASR tasks, a noisy ASR task and an ASR task with a large vocabulary, showed that scRNN with the proposed extension improved the word error rate.

TUE-PM3-O2.5: Hypersphere Embedding and Additive Margin for Query-by-example Keyword Spotting

Haoxin Ma, Ye Bai, Jiangyan Yi and Jianhua Tao

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Query-by-example (QbE) keyword spotting is convenient for users to define their own keywords, so it is useful in device control. However, conventional regular softmax, which is commonly used for training QbE models, has two limitations. First, the learned features are not discriminative enough. Second, norm variations of the unnormalized features affect computing cosine similarities. To address these issues, this paper introduces normalization and additive margin into residual networks for QbE keyword spotting. Features and weights are normalized on a hypersphere of fixed radius. Additive margin further helps to reduce the intra-class variations and increase inter-class differences. Based on public datasets AISHELL-1 and HelloNPU, we design three different test sets, namely in-vocabulary, out-of-vocabulary, and cross-corpus, to evaluate our proposed method. Experiments show that our proposed method can learn more discriminative embedding features. For totally unseen situation, our proposed method achieves a relative false rejection rate reduction of 46.60% when the false alarm rate is 2% in cross-corpus evaluation, compared with regular softmax.

TUE-PM3-O2.6: A Speech Enhancement Neural Network Architecture with SNR-Progressive Multi-Target Learning for Robust Speech Recognition

Nan Zhou, Jun Du, Yan-Hui Tu, Tian Gao and Chin-Hui Lee

University of Science and Technology of China, iFlytek Research, Georgia Institute of Technology

We present a pre-processing speech enhancement network architecture for noise-robust speech recognition by learning progressive multiple targets (PMTs). PMTs are represented by a series of progressive ratio masks (PRMs) and progressively enhanced log-power spectra (PELPS) targets at various layers based on different signal-to-noise-ratios (SNRs), attempting to make a tradeoff between reduced background noises and increased speech distortions. As a PMT implementation, long short-term memory (LSTM) is adopted at each network layer to progressively learn intermediate dual targets of both PRM and PELPS. Experiments on the CHiME-4 automatic speech recognition (ASR) task, when compared to unprocessed speech using multi-condition trained LSTM-based acoustic models without retraining, show that PRM-only as the learning target can achieve a relative word error rate (WER) reduction of 6.32% (from 27.68% to 25.93%) averaging over the RealData evaluation set, while conventional ideal ration masks severely degrade the ASR performance. Moreover, the proposed LSTM-based PMT network, with the best configuration, outperforms the PRM-only model, with a relative WER reduction of 13.31% (further down to 22.48%) averaging over the same test set.

TUE-PM3-O3
Speech Enhancement

Time: Tuesday, Nov 19, 17:00-18:40

Place: A8

Chair: Changchun Bao

TUE-PM3-O3.1: CycleGAN-based speech enhancement for the unpaired training data

Jing Yuan and Changchun Bao

Beijing University of Technology

Speech enhancement is an important task of improving speech quality in noise scenario. Many speech enhancement methods have achieved remarkable success based on the paired data. However, for many tasks, the paired training data is not available. In this paper, we present a speech enhancement method for the unpaired data based on cycle-consistent generative adversarial network (CycleGAN) that can minimize the reconstruction loss as much as possible. The proposed model employs two discriminators and two generators to preserve speech components and reduce noise so that the network could map features better for the unseen noise. In this method, the generators are used to generate the enhanced speech, and two discriminators are employed to discriminate real inputs and the outputs of the generators. The experimental results showed that the proposed method effectively improved the performance compared to traditional deep neural network (DNN) and the recent GAN-based speech enhancement methods.

TUE-PM3-O3.2: Phase Unwrapping Based Speech Enhancement

Rui Cheng and Changchun Bao

Beijing University of Technology

Speech enhancement is a vital technology for reducing the noise in speech communication. Most speech enhancement methods only estimate magnitude spectrum of clean speech from noisy speech and combine noisy phase spectrum to recover the enhanced speech. In this paper, considering the importance of recovering the phase of clean speech in speech enhancement, a phase recovery method of speech is proposed by combining phase unwrapping and deep neural network (DNN). By integrating the recovered phase of clean speech into conventional magnitude enhancement methods, the performance is improved effectively. The verification is conducted by several types of noises at different signal-to-noise ratio (SNR) levels. The experimental results also confirmed that the recovered phase of clean speech resulted in an obvious improvement on the speech quality and intelligibility compared to the noisy phase.

TUE-PM3-O3.3: End-to-End Speech Enhancement Using Fully Convolutional Networks with Skip Connections

Dujuan Wang and Changchun Bao

Beijing University of Technology

The purpose of speech enhancement is to extract useful speech signal from noisy speech. The performance of speech enhancement has been improved greatly in recent years with fast development of the deep learning. However, these studies mainly focus on the frequency domain, which needs to complete time-frequency transformation and the phase information of speech is ignored. Therefore, the end-to-end (i.e. waveform-in and waveform-out) speech enhancement was investigated, which not only avoids fixed time-frequency transformation but also allows modelling phase information. In this paper, a fully convolutional network with skip connections (SC-FCN) for end-to-end speech enhancement is proposed. Without the fully connected layers, this network can effectively characterize local information of speech signal, and better restore high frequency components of waveform using lesser number of the parameters. Meanwhile, because of existence of skip connections in different layers, it is easier to train deep networks and the problem of gradient vanishing can also be tackled. In addition, these skip connections can obtain more details of speech signal in different convolutional layers, which is beneficial for recovering the original speech signal. According to our experimental results, the proposed method can recover the waveform better.

TUE-PM3-O3.4: Single Channel Speech Enhancement Using Temporal Convolutional Recurrent Neural Networks

Jingdong Li, Hui Zhang, Xueliang Zhang and Changliang Li

Inner Mongolia University, Kingsoft AI Laboratory

Speech enhancement aims to separate clean speech from noisy signals. In the past decades, deep learning-based models have shown promising performance in this task. Most methods are based on estimating time-frequency (T-F) representation of target speech directly or indirectly. The most popularly used T-F representation is short-time Fourier transform (STFT), which can be decomposed to the magnitudeand phase spectrum. Most of the previous methods focus on estimating a more accurate magnitude and use the phase of noisy signals to reconstruct the waveform. Until recently, this strategy seems to reach a ceiling. Some methods attempt to estimate the phase of target speech, but fitting raw phase values is challenging. In this paper, we proposed a temporal convolutional recurrent network(TCRN) to conduct speech enhancement in time domain, whichis directly trained with the raw waveform as input and output. We employ several techniques to stabilize the training process. Experimental results show that the proposed model consistently outperforms existing speech enhancement approaches, in terms of speech intelligibility and quality.

TUE-PM3-O3.5: GSC Based Speech Enhancement with Generative Adversarial Network

Yao Zhou, Changchun Bao and Rui Cheng

Beijing University of Technology

At present, the technology of using microphone arrays for speech enhancement has been widely concerned, and the enhancement effect is excellent. The widely used Generalized Sidelobe Canceller (GSC) method can achieve good noise reduction for noisy speech in the additive noise acoustic environment, and achieve better intelligibility improvement. But there are also areas for improvement. In the lower branch of GSC, signal leakage caused by the estimation of the incident angle or the slight change of the position of the microphone array may cause the self-cancellation of target speech signal, thereby the severe speech distortion is caused. In this paper, the Generative Adversarial Network (GAN), which has broad application prospects in deep learning technology, replaces the lower branch of the traditional GSC structure, thus the self-cancellation of speech signals is avoided and improving the anti-error ability of the enhancement system is improved effectively.

WED-AM1-O1
Paralinguistics in Speech and Language

Time: Wednesday, Nov 20, 10:20-12:00

Place: A1

Chair: Yu Tsao

WED-AM1-O1.1: Real-time and interactive tools for voccal training based on an analytic signal with a cosine series envelope

Hideki Kawahara, Ken-Ichi Sakakibara, Eri Haneishi and Kaori Hagiwara

Wakayama University, Health Science University of Hokkaido, Showa University of Music

We introduce real-time and interactive tools for assisting vocal training. In this presentation, we demonstrate mainly a tool based on real-time visualizer of fundamental frequency candidates to provide information-rich feedback to learners. The visualizer uses an efficient algorithm using analytic signals for deriving phase-based attributes. We start using these tools in vocal training for assisting learners to acquire the awareness of appropriate vocalization. The first author made the MATLAB implementation of the tools open-source. The code and associated video materials are accessible in the first author's GitHub repository.

WED-AM1-O1.2: Likability Estimation of Call-center Agents by Suppressing Annotator Variability

Hosana Kamiyama, Atsushi Ando, Ryo Masumura, Satoshi Kobashikawa and Yushi Aono

NTT

This paper proposes a effective likability estimation technique for call-center agents. Most likability estimation models need numerous annotated speech samples to obtain high-quality training labels since the likability annotations often vary due to annotator disagreement. The performance of conventional likability estimation models is often poor since they do not adequately account for annotator variability which is the difference between each annotator's assessment reliability. Our approach suppresses the effect of annotator variability by taking into account the individual annotator's reliability, which is the probability of correctly assessing likability. To estimate target annotator-independent likability, we introduce a graphical model with annotator reliability and optimize the model by using the EM-algorithm. We also propose a new neural network architecture to improve the model's performance. The architecture has a layer that takes as input the probability of target annotator-independent likability and the probability of annotator reliability. To propagate the loss of likability estimation independent from annotator reliability, our proposal processes annotations via the proposed layer. Given just two annotations per call, our proposal yields better accuracy than either the baseline or conventional methods.

WED-AM1-O1.3: Urgent Voicemail Detection Focused on Long-term Temporal Variation

Hosana Kamiyama, Atsushi Ando, Ryo Masumura, Satoshi Kobashikawa and Yushi Aono

NTT

This paper proposes a effective urgent speech detection for voicemails focused on speech rhythm.
Previous techniques use short-term features with millisecond scale (such as fundamental frequency, loudness and spectral features), and conventional techniques for urgent speech detection use also features obtained from entire speech (such as average speech rate). However, the features obtained from entire speech are too over-smoothed to explain the difference between urgent and non-urgent speech. We found that there was a difference between urgent and non-urgent speech in temporal variability related to speech rhythm. To handle the temporal variability of speech rhythm, the proposal extracts long-term temporal features. The long-term temporal features are envelope modulation spectrum and temporal statistics of Mel-frequency cepstrum coefficient with 1 sec scale. To use both features with different time scales, the proposed method integrates the long-term temporal features and the short-term features on neural networks. Our proposal yields better accuracy than the conventional methods (which uses e features obtained from entire speech); it achieves a 50.0% reduction in the error rate.

WED-AM1-O1.4: Learning Contextual Representation with Convolution Bank and Multi-head Self-attention for Speech Emphasis Detection

Liangqi Liu, Zhiyong Wu, Runnan Li, Jia Jia and Helen Meng

Tsinghua University

In speech interaction scenarios, speech emphasis plays an important role in conveying the underlying intention of the speaker. For better understanding of user intention and further enhancing user experience, techniques are employed to automatically detect emphasis from the user's input speech in human-computer interaction systems. However, even for state-of-the-art approaches, challenges still exist: 1) the various vocal characteristics and expressions of spoken language; 2) the long-range temporal dependencies in the speech utterance. Inspired by human perception mechanism, in this paper, we propose a novel attention-based emphasis detection architecture to address the above challenges. In the proposed approach, convolution bank is utilized to extract informative patterns of different dependency scope and learn various expressions of emphasis, and multi-head self-attention mechanism is utilized to detect local prominence in speech with the consideration of global contextual dependencies. Experimental results have shown the superior performance of the proposed approach, with 2.62% to 3.54% relative improvement on F1-measure compared with state-of-the-art approaches.

WED-AM1-O1.5: Effect of Relative Frequency of Lexical Meanings on Accessing Lexical Ambiguities: Evidence from the Coordinator 'and'

Xiaoqun Dong and Xueqin Zhao

Nanjing University of Science and Technology

Lexical ambiguity is a common phenomenon in English. Research on the resolution of lexical ambiguity began since 1970s, and has developed several theories on how comprehenders settle on a single meaning [5, 9, 30, 32]. Many studies have investigated the effects of relative meaning frequency and other factors on lexical ambiguity resolution [6, 18, 27], while the research subjects are mainly content words. Whether there are effects of relative meaning frequency on accessing coordinators keeps unclear. The present study takes the coordinator 'and' as the research subject to explore the effect of relative meaning frequency on lexical access via a lexical decision task and further investigate whether related meanings of 'and' lead to confusions in lexical access. In the experiment, 21 participants who are advanced Chinese EFL learners were requested to choose one of the two meanings for 'and' which connects two clauses in a complex sentence, and the accuracy and reaction time were collected. It was found that relative meaning frequency did influence accessing meanings of coordinator 'and'—the higher the relative meaning frequency, the shorter the response time, and the relatedness between meanings led to confusions in lexical access. These results confirm the effect of relative meaning frequency on accessing meanings of coordinators and reveal the importance of distinguishing the related meanings.

WED-AM1-O2
Neural Signal Processing

Time: Wednesday, Nov 20, 10:20-12:00

Place: A2

Chair: Kazushi Ikeda

WED-AM1-O2.1: Training data expansion for classification between normal and abnormal lung sounds

Naoki Umeno, Masaru Yamashita, Hiroyuki Takada and Shoichi Matsunaga

Nagasaki University

In this paper, we investigate the effectiveness of training data expansion methods to distinguish between normal and abnormal lung sounds. Acoustic characteristics of lung sounds vary according to auscultation points. In conventional classification methods, acoustic models were usually trained using only lung sounds recorded at the same auscultation points to that of evaluation data. This results in a small amount of training data and, thus, hinders the achievement of a high classification rate. To overcome this problem, we performed training data expansion by selecting the lung sounds, which are expected to be useful for generating acoustic models with higher classification performance, among sound samples recorded at other auscultation points. We investigated the two types of selection approach: selection based on the similarity of acoustic features in sound samples and selection based on the confidence measure represented by the difference between the acoustic likelihood for a normal or abnormal respiratory candidate. Our experiments showed that both selection types have the potential to increase the classification performance between normal and abnormal lung sounds, as well as the classification performance between healthy and unhealthy subjects.

WED-AM1-O2.2: Hybrid Convolutional Recurrent Neural Networks Outperform CNN and RNN in Task-state EEG Detection for Parkinson's Disease

Xinjie Shi, Tianqi Wang, Lan Wang, Hanjun Liu and Nan Yan

School of Software Engineering, University of Science and Technology of China, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Sun Yat-sen University

In hospitals, clinicians or neurologists diagnose brain-related disorders such as Parkinson's disease (PD) by analyzing electroencephalograms (EEG). However, conventional EEG-based diagnosis for PD relies on handcrafted feature extraction, which is laborious and time-consuming. With the emergence of deep learning, automated analysis of EEG signals can be realized by exploring the inherent information in data, and outputting the results of classification from the hidden layer. In the present study, four deep learning algorithm architectures, including two convention deep learning models (convolutional neural network, CNN; and recurrent neural network, RNN) and two hybrid convolutional recurrent neural networks (2D-CNN-RNN and 3D-CNN-RNN), were designed to detect PD based on task-state EEG signals. Our results showed that the hybrid models outperformed conventional ones (fivefold average accuracy: 3D-CNN-RNN 82.89%, 2D-CNN-RNN 81.13%, CNN 80.89%, and RNN 76.00%) as they combine the strong modeling power of CNN in temporal feature extraction, and the advantage of RNN in processing sequential information. This study represents the first attempt to use hybrid convolutional recurrent neural networks in classifying PD and normal take-state EEG signals, which carries important implications to the clinical practice.

WED-AM1-O2.3: Investigation of speech-planning mechanism based on eye movement and EEG

Jinfeng Huang, Bin Zhao, Jianwu Dang and Minbo Chen

Japan Advanced Institute of Science and Technology

A major concern in the speech production research is how speakers make a plan for speech articulation, in which the latent time is an important index for evaluating the planning process. Previous researches on this topic using either isolated words or phrases found that the word length and familiarity could affect the latent time of speech planning. However, in continuous sentence processing, semantic prediction was found to be more influential from our previous eye movement investigation. To probe further into the underlying neural causes, this study combined eye movement and EEG techniques to analyze the behavior-locked brain activities during the speech planning process in a sentence reading task. The results showed that the latent time decreases gradually with ongoing reading process as the context

information got richer. And the subjects tend to look ahead prior to the articulation of the current word. Functional network analyses for the visual and semantic processing were consistent with the behavior results and suggested that the look-ahead phenomenon is a companying effect of speech coarticulation, and as speech prediction becomes easier, the latent time for speech planning tend to be shortened.

WED-AM1-O2.4: Towards Generation of Visual Attention Map for Source Code

Takeshi D. Itoh, Takatomi Kubo, Kiyoka Ikeda, Yuki Maruno, Yoshiharu Ikutani, Hideaki Hata, Kenichi Matsumoto and Kazushi Ikeda

Nara Institute of Science and Technology, Kyoto Women’s University

Program comprehension is a dominant process in software development and maintenance. Experts are considered to comprehend the source code efficiently by directing their gaze, or attention, to important components in it. However, reflecting importance of components is still a remaining issue in gaze behavior analysis for source code comprehension. Here we show a conceptual framework to compare the quantified importance of source code components with gaze behavior of programmers. We use “attention” in attention models (e.g., code2vec) as the importance indices for source code components and evaluate programmers' gaze locations based on the quantified importance. In this report, we introduce the idea of our gaze behavior analysis using the attention map, and the results of a preliminary experiment.

WED-AM1-O2.5: Voice Conversion by Dual-Domain Bidirectional Long Short-Term Memory Networks with Temporal Attention

Xiaokong Miao, Meng Sun and Xiongwei Zhang

Army Engineering University

Voice conversion(VC) is a method for seeking to convert one speaker’s voice into another person’s voice while maintaining the content unchanged. One of the key steps is to construct a mapping of features from the source speaker to the target speaker. Given the strong ability to model contextual information, bidirectional long short-term memory network (BLSTM) is usually taken as the mapping tool. In this paper, an improved version of BLSTM, dual-domain BLSTM, is considered as a baseline where dynamic time warping (DTW) is conventionally taken as a tool to align speech frames. In order to alleviate the negative impacts of temporal misalignment in DTW, a casual attention mechanism is introduced to improve the dual-domain BLSTM. Experiments demonstrated the effectiveness of the proposed approach by yielding lower mel-ceptral distances and higer MOS scores than the baselines.

WED-AM1-O3
Computer Vision

Time: Wednesday, Nov 20, 10:20-12:00

Place: A3

Chair: Ce Zhu

WED-AM1-O3.1: Vision-based Localization with Monocular Camera for Light-rail System

Kebin Jia, Tingxian Wang and Meng Yao

Beijing University of Technology

As an emerging localization method, vision-based localization methods have been widely used in localization systems for the vehicle. By considering the practical requirements like the high accuracy, real-time performance and easy installation, this paper designs a set of real-time localization system for urban light rail based on the monocular camera. This system is divided into two parts: offline and online. To solve the problem of the high similarity of light rail scene under a high frame rate, we proposed a new scene recognition method based on local key regions and key frames. This method not only guarantees the accuracy of matching but also satisfies the real-time requirement of the system. The offline module uses the unsupervised method to extracts the region with high saliency from each reference frame as the key region and selects the key frame based on it. The online module can quickly match the current frame and reference frame within the retrieval range provided by the key frame, by calculating the binary feature with a low correlation in the key regions. While meeting the high-precision needs of the light rail system, it significantly improves real-time performance. This paper uses both the public test dataset in Nordland and the challenging Hong Kong light rail dataset. The experiment results show that the precision of the system can reach more than 90% in extreme situations such as large-area scene occlusion.

WED-AM1-O3.2: Taxi Drivers' Smoking Behavior Detection in Traffic Monitoring Video

Siwei Chen, Kebin Jia, Pengyu Liu and Xunping Huang

Beijing University of Technology

Taxi is an indispensable part of the urban public transportation industry. However, the existence of illegal operating behaviors brings security risks to passengers and affects the city's civilization. The paper studies the detection methods of taxi drivers' smoking behavior during operation in the surveillance video environment. Firstly, the Vibe algorithm is used to detect the motion foreground of the video. Then the Haar-Adaboost algorithm is used to identify the taxi and extract the target detection area. The HSV color model is used to extract the color characteristics of the smoking smoke, and separate the target area from the moving objects which is similar to smoking smoke. Finally, combined changes in shape and movement characteristics of smoking smoke to further eliminate the interference objects, and the final detection results are obtained. The experimental results show that this algorithm has high real-time and accuracy, which is beneficial to off-site law enforcement in intelligent traffic management.

WED-AM1-O3.3: Integrating Action-aware Features for Saliency Prediction via Weakly Supervised Learning

Jiaqi Feng, Shuai Li, Yunfeng Sui, Lingtong Meng and Ce Zhu

University of Electronic Science & Technology of China, Research Center of Second Research Insititute of CAAC

Deep learning has been widely studied for saliency prediction. Despite the great performance improvement introduced by deep saliency models, some high-level concepts that contribute to the saliency prediction, such as text, objects of gaze and action, locations of motion, and expected locations of people, have not been explicitly considered. This paper investigates the objects of action and motion, and proposes to use action-aware features to compensate deep saliency models. The action-aware features are generated via weakly supervised learning using an extra action classification network trained with existing image based action datasets. Then a feature fusion module is developed to integrate the action-aware features for saliency prediction. Experiments show that the proposed saliency model with the action-aware features achieves better performance on three public benchmark datasets. More experiments are further conducted to analyze the effectiveness of the action-aware features in saliency prediction. To the best of our knowledge, this study is the first attempt on explicitly integrating objects of action and motion concept into deep saliency models.

WED-AM1-O3.4: Efficient and Robust Pseudo-Labeling for Unsupervised Domain Adaptation

Hochang Rhee and Nam Ik Cho

Seoul National University

Unsupervised domain adaptation is to transfer knowledge from an annotated source domain to a fully-unlabeled target domain. The conventional methods consider the data which exceed a certain threshold of confidence as pseudo-labeled data for the target domain, and thus choosing the appropriate threshold affects the target performance. In this paper, we propose a new confidence-based weighting scheme for obtaining pseudo-labels and an adaptive threshold adjustment strategy to provide sufficient and accurate pseudo-labels throughout the training. To be precise, our confidence-based weighting scheme generates pseudo-labels having a different contribution based on the confidence, which maintains sufficiency and accuracy of pseudo-labels. Also, the proposed adaptive threshold adjustment method chooses the threshold according to the degree of adaptation of a network to the target domain, and thus obviates the need for an exhaustive search for the appropriate threshold. Experimental results on a digit classification task show that the proposed methods efficiently utilizes the pseudo-labels to preserve sufficiency and accuracy.

WED-AM1-O3.5: A Multi-Objective Optimization Perspective for Joint Consideration of Video Coding Quality

Wei Gao

Peking University

Traditional efforts for video coding usually focus on either rate-distortion (R-D) performance or quality smoothness (QS), which may not effectively achieve the desirable visual quality of experience for reconstructed videos. Single-objective optimization cannot guarantee the performance of the other one. In this paper, we would like to introduce a new perspective for the better joint consideration of video coding quality, where both R-D and QS are simultaneously evaluated, and then the traditional video coding optimization problem is converted from single objective to multiple objectives. We provide the details to demonstrate the establishment method of the joint video coding quality (JVCQ), which is given as the comprehensive analytical model to jointly considers spatial and temporal artifacts. The FixedQP results and consumed bits are referred to construct the comparable JVCQ results. Finally, by collecting JVCQ results for four different target bit rates and five different weighting strategies, we can obtain the rate-averaged JVCQ (RJVCQ), weighting-averaged JVCQ (WJVCQ) and rate-weighting-averaged JVCQ (RWJVCQ), respectively. By comparing the different rate control (RC) algorithms, experiments validate the consistency and applicability of the proposed JVCQ with the original two separate evaluation metrics, and the proposed RC algorithms can have better performances. Besides the video coding quality evaluation, the group of JVCQs can be used to guide the optimization process of video coding to achieve gains on both evaluation metrics simultaneously.

WED-AM1-O3.6: Spherical Position Dependent Rate-distortion Optimization for 360-degree Video Coding

Yuyang Liu, Hongwei Guo, Ce Zhu and Yipeng Liu

University of Science and Technology of China, University of Electronic Science and Technology of China & Honghe University

360-degree video in spherical format cannot be well handled by the conventional video coding tools. Currently, most of the 360-degree video coding methods first project the spherical video content onto a 2-dimensional plane and then compress the projected video using a conventional video codec. However, the projection conversion process will cause an irreversible conversion error, which indicates that the reconstruction quality of the projected video cannot fully represent that of the spherical video. In view of this, this paper proposes a spherical position dependent rate-distortion optimization (RDO) approach for 360-degree video coding. During the RDO process, spherical reconstruction quality is taken into consideration and calculated according to the spherical position of the pixels in each coding unit (CU). Furthermore, the Lagrangian multiplier and quantization parameter are adjusted accordingly. The proposed method is implemented on HEVC reference software HM-16.7. Experimental results show that the proposed method can achieve better coding performance, compared with HM-16.7.

WED-AM1-O4
Language Learning

Time: Wednesday, Nov 20, 10:20-12:00

Place: A4

Chair: Aijun Li

WED-AM1-O4.1: Is average RMSE appropriate for evaluating acoustic-to-articulatory inversion?

Qiang Fang

Institute of Linguistics, Chinese Academy of Social Sciences

Acoustic-to-articulatory inversion has potential application in number of fields. For decades, average root mean square error and Pearson correlation coefficient are the most prevalent quantities adopted to evaluate the performance of acoustic-to-articulatory inversion. Various inversion methods have been developed to less the average root mean square error, but very few studies explored whether the average root mean square error is appropriate for evaluating and comparing the performance of different inversion methods. In this study, we attempt to tackle this issue by comparing not only the average root mean square error but also channel root mean square error of each articulatory channel, and the root mean square error of the critical and non-critical portions of each articulatory channel for methods within and between different groups. It is found that: i) the root mean square error of each articulatory channel, and the root mean square error of the critical and non-critical portions of each articulatory channel decrease while the average root mean square error decrease if the AAI methods belong to the same group; ii) exceptions are found if the inversion methods belong to different categories; iii) the average root mean square error is dominated by that of non-critical portions of articulatory channels. This suggests that new methods, which pay more attention to the performance of acoustic-to-articulatory inversion on critical articulators and facilitate the comparison of performance of methods belonging to different categories, should be developed in the future.

WED-AM1-O4.2: Normalization of GOP for Chinese Mispronunciation Detection

Wenwei Dong and Yanlu Xie

Beijing Language and Culture University

Goodness of Pronunciation (GOP) is a kind of Computer-Assisted Pronunciation Training (CAPT) technique that can provide language learners with scoring feedback, and its accuracy easily suffers from the performance of model alignment and phone classification. In order to reduce the influence of those aspects, this paper proposes two ways to normalize GOP scores. The first is to separate the GOP calculation of Chinese Initials and those of Chinese Finals. The second is to use the corresponding native pronunciation score as a template to scale the non-native one. In 2-hours test set of Japanese speaking Chinese corpus, the experiment results show the average relative improvement of Diagnose Accuracy (DA) in the approach one is 16.9%, and 28.7% in scaling approach comparing to the traditional scoring method. The combination of those two methods achieves the best performance. The result is 35.9% of average relative improvement. Experimental results demonstrate the effectiveness of the two methods.

WED-AM1-O4.3: Disfluency Detection Based on Speech-Aware Token-by-Token Sequence Labeling with BLSTM-CRFs and Attention Mechanisms

Tomohiro Tanaka, Ryo Masumura, Takafumi Moriya, Takanobu Oba and Yushi Aono

NTT

This paper presents a new method for token-bytoken sequence labeling that can leverage not only lexical information but also speech information without any alignments. Our motivation is to detect disfluencies such as fillers and word fragments robustly from spontaneous speech. Disfluency detection is often modeled as a token-by-token sequence labeling using a transcribed text via automatic speech recognition. However, utilizing the lexical information alone is not sufficient because the disfluencies cause changes to speech information. One problem is that the speech and the transcribed text need to be aligned when we handle speech and lexical information simultaneously. This prevents introducing speech information to the disfluency detection. To solve this problem, we propose a method for token-by-token sequence labeling, one that can simultaneously use lexical and speech information without requiring any alignments. To this end, we introduce attention mechanisms into a method for neural sequence labeling based on bi-directional long shortterm memory recurrent neural network conditional random fields. The attention mechanisms enable us to find the term of disfluencies from speech automatically. Our experimental results show that the proposed method using acoustic and prosodic features improves the labeling accuracy compared with that using lexical features alone.

WED-AM1-O4.4: Unsupervised Pronunciation Fluency Scoring by infoGan

Wenwei Dong and Yanlu Xie

Beijing Language and Culture University

Pronunciation fluency scoring (PFS) is a primary task in computer-aided second language (L2) learning. Most of existing PFS algorithms are based on supervised learning, where human-labeled scores are used to train the scoring model. However, the human labeling is rather costly and tends to be biased. In order to tackle this problem, we propose an unsupervised learning approach, where an infoGan model is constructed to infer latent speech codes, and then these codes are used to build a classifier that distinguishes native and foreign speech. We found that this native-foreign classifier can generate good utterance-based fluency scores

WED-AM1-O4.5: Study on the Tones Biases of Mandarin Speaker in Amdo Tibetan Areas Based on Statistics

Gan Zhenye, Jiao Yi, Yang Hongwu, Zhao Guangying and Song Zhimeng

Northwest Normal University

Tone learning is a major difficulty when students in Amdo Tibetan Area learn mandarin.This paper uses experimental phonetics, comparative analysis, and theory and methods of biases analysis to investigate and analyze the pronunciation of mandarin in the Amdo Tibetan area.The perception experiment and similarity experiment were used to analyze the tone of mandarin in the Amdo Tibetan area.The experimental results show that in the process of learning mandarin, the students in Amdo Tibetan area learn from the highest to the lowest in order of tone4, tone1, tone3 and tone2.The most prone to sound in the process of speaking mandarin is tone4, and the most common pronunciation in their biased pronunciation is tone4. The similarity detection has a higher diagnostic accuracy rate, and tone1 has the best effect. The tone2 and tone3 are easy to judge the correct pronunciation as the biased pronunciation, and tone4 is easy to judge the biases pronunciation as the correct pronunciation.

WED-AM1-O5
Dialog System

Time: Wednesday, Nov 20, 10:20-12:00

Place: A5

Chair: Qin Jin

WED-AM1-O5.1: Automatically Annotate TV Series Subtitles for Dialogue Corpus Construction

Leilan Zhang and Qiang Zhou

Tsinghua University

In recent years, the scarcity of dialogue corpus is becoming the bottleneck of Chinese dialogue generation systems. Although subtitles provide favorable material to construct dialogue corpus because of their abundance and diversity, lacking speaker information makes it hard to extract dialogues from subtitles directly. To utilize these resources, we proposed an improved method to automatically annotate bilingual TV subtitles with speaker and scene tags using their corresponding scripts. First, tags of speakers and scene boundaries in the scripts are mapped to the subtitles through an information retrieval method. Then, the mapping errors are detected with a convolutional network and corrected by heuristic strategies to improve the annotation quality. We applied this method on 779 bilingual subtitle files of 4 TV series and obtained a Chinese dialogue corpus Tv4Dialog containing 260674 utterances. The experiment result shows that our method can achieve an accuracy of 94.62% on speaker tag annotation, improving nearly 12% on the previous state-of-the-art result.

WED-AM1-O5.2: Topic Segmentation for Dialogue Stream

Leilan Zhang and Qiang Zhou

Tsinghua University, RIIT

Topic segmentation, which aims to divide a document into topic blocks, is a fundamental task in natural language processing. Most of the previous researches focus on written text rather than dialogue text. However, dialogue text has its unique characteristic and is more challenging in topic segmentation. The existing neural models for topic segmentation are usually built on RNN or CNN, which are competent in written text but has a poor performance in dialogue text. We argue that a better segmentation result for dialogue text requires a better semantic representation of sentences. In this paper, we formulate topic segmentation as a sequence labeling task and propose a model based on BERT and TCN (Temporal Convolutional Network) to accomplish the task. We also present three datasets, including two dialogue datasets and a news dataset, to evaluate the model's performance. Compared to the previous best model, our model shows an absolute performance improvement of 8%-17% in F1 scores. Moreover, we explore the impact of importing speakers on dialogue text segmentation, the experiment result shows that the additional speaker information could effectively improve the segmentation performance.

WED-AM1-O5.3: Joint Learning of Conversational Temporal Dynamics and Acoustic Features for Speech Deception Detection in Dialog Games

Huang-Cheng Chou, Yi-Wen Liu and Chi-Chun Lee

Department of Electrical Engineering

Deception is an intended action of a deceiver to make an interrogator believe something is true (or false) that the deceiver believes to be false (or true) as a purposeful mechanism to share a mix of truthful and deceptive experiences when being asked to respond to questions. Conventionally, automatic deception detection from speech is regard as a recognition task modeled only using the deceiver's acoustic cues and does not include temporal conversation dynamics between the interlocutors, i.e., ignoring the potential deception-related cues when the two interlocutors coordinate such a back-an-forth interaction. In this paper, we propose a joint learning framework to detect deception by simultaneously considering variations and patterns of the conversation using both interlocutor's acoustic features and their conversational temporal dynamics. Our proposed model achieves an unweighted average recall (UAR) of 74.71% on a recently collected Chinese deceptive corpus of dialog games. Further analyses reveal that the interrogator behaviors are correlated to the deceiver's deception behaviors, and including the conversational features provides enhanced deception detection power.

WED-AM1-O5.4: Type of Response Selection utilizing User Utterance Word Sequence, LSTM and Multi-task Learning for Chat-like Spoken Dialog Systems

Kengo Ohta, Ryota Nishimura and Norihide Kitaoka

National Institute of Technology, Tokushima University, Toyohashi University of Technology

This paper describes a method of automatically selecting types of responses, such as back-channel responses, changing the topic or expanding the topic, in conversational spoken dialog systems by using an LSTM-RNN-based encoder-decoder framework and multi-task learning. In our dialog system architecture, response utterances are generated after the response type is explicitly determined in order to generate more appropriate and cooperative response than the conventional end-to-end approach which generate response utterances directly. As a response type selector, an encoder and two decoders share states of hidden layers and are trained with the interpolated loss function of the two decoders. One of the decoders is for selecting types of responses and the other is for estimating the word sequence of the response utterances. In an evaluation experiment using a corpus of dialogs between elderly people and an interviewer, our proposed method achieved better performance than the standard method using single-task learning, especially when the amount of training data was limited.

WED-AM1-O5.5: Prosodic Cues in the Interpretation of Echo Questions in Chinese Spoken Dialogues

Aijun Li, Gan Huang and Zhiqiang Li

Institute of Linguistics of CASS and Graduate School of CASSU, School of Chinese Culture and Communication, Department of Modern and Classical Languages, University of San Francisco

This study examines the effect of prosodic cues in the disambiguation of five discourse-pragmatic functions of echo questions and the corresponding statements in Chinese spoken dialogues. Data were collected in a “role-play” format to mimic different communicative functions of echo questions in real-life situations. Statistical analyses were performed on both global and local F0 variations associated with intonation patterns in echo questions and corresponding statements. Results showed that boundary tone features alone are not good predictors in distinguishing echo questions and statements; variations in intonation patterns are related to the different discourse-pragmatic functions that echo questions serve; echo questions and statements, as well as different discourse-pragmatic functions of echo questions, can be distinguished on the basis of global variations of prosodic features such as overall F0 slope and average F0, combined with local changes due to boundary tone features; and when information about morpho-syntactic structures and boundary tone features were included in the analysis, the accuracy of discriminant analysis was at 76.5%~94.1% for statements and echo questions, and at 57.6%~83.5% for different discourse-pragmatic functions. The accuracy dropped to 70.9% (2 groups) and 40.9% (6 groups) when morpho-syntactic structural information was not included, indicating that structural and contextual information contributed 30% and 60% respectively.

WED-AM1-O6
Adaptive Signal Processing

Time: Wednesday, Nov 20, 10:20-12:00

Place: A6

Chair: Mau-Luen Tham

WED-AM1-O6.1: A DOA Estimation Method of coherent and uncorrelated sources based on Nested Arrays

Fanghao Cheng and Julan Xie

UESTC

This paper presents a novel way to estimate the DOA of coherent and uncorrelated sources for nested array. At first, by using the vectorization of the covariance matrix of the received data, the LASSO algorithm with a modified dictionary matrix based on compressed sensing is applied to estimate a vector containing all signal powers. Then, the estimated vector is reconstructed into the covariance matrix of coherent and uncorrelated sources. A peak searching to estimate DOAs can be performed by using the diagonal elements of this reconstructed matrix. In this presented method, the advantage of fully utilizing the degree of freedom of nested arrays is preserved. Theoretical analysis and simulation results show the effectiveness of the proposed algorithm.

WED-AM1-O6.2: A DOA Estimation Method in the presence of unknown mutual coupling based on Nested Arrays

Fanghao Cheng and Julan Xie

UESTC

A novel DOA method is proposed to deal with the DOA estimation in the presence of the unknown mutual coupling for nested arrays. By using a new expression of the steering matrix in the presence of mutual coupling, a novel expression of the receiving data vector in the virtual array field is available. Then, based on a modified direction matrix constructed with block matrix, which relates to space discretized sampling grid, the sparse Bayesian compressive sensing method applies to estimate a vector, which contains the signal powers information and the mutual coupling information. The problem of off-grid DOAs is also considered for sparse Bayesian compressive sensing. Based on the estimated vector, a peak searching is performed to estimate the initial DOA. Finally, the estimation of DOA is modified to initial estimate plus off-grid error value. The advantage of fully utilizing the degree of freedom of nested arrays is preserved in this proposed algorithm. Moreover, no complicated calculation is needed to obtain the mutual coupling coefficients or rearrange the position of array element. Theoretical analysis and simulation results show the effectiveness of the proposed algorithm.

WED-AM1-O6.3: An Alternative Solution to the Dynamically Regularized RLS Algorithm

Feiran Yang, Felix Albu and Jun Yang

Institute of Acoustics, Chinese Academy of Sciences, Valahia University of Targoviste

The recursive least-squares (RLS) algorithm should be explicitly regularized to achieve a satisfactory performance when the signal-to-noise ratio is low. However, a direct implementation of the involved matrix inversion results in a high complexity. In this paper, we present a recursive approach to the matrix inversion of the dynamically regularized RLS algorithm by exploiting the special structure of the correlation matrix. The proposed method has a similar complexity to the standard RLS algorithm. Moreover, the new method provides an exact solution for a fixed regularization parameter, and it has a good accuracy even for a slowly time-varying regularization parameter. Simulation results confirm the effectiveness of the new method.

WED-AM1-O6.4: A Norm Constraint Lorentzian Algorithm Under Alpha-stable Measurement Noise

Xinqi Huang, Yingsong Li and Felix Albu

Harbin Engineering University, Valahia University of Targoviste

An l0-norm constraint Lorentzian (L0-CL) algorithm is proposed for adaptive sparse system identification to combat impulsive noise. The L0-CL algorithm is derived via exerting an l0-norm penalty on the coefficients in the cost function, which is equivalent to add a zero-attractor in the iterations. The zero-attractor attracts the coefficients to zero during the iterations. By the way, the L0-CL algorithm can achieve lower mean square error (MSE) for estimating the sparse systems. The simulation results presented in this paper demonstrate that the proposed algorithm has superior performance in both convergence rate and steady-state behavior by identifying the sparse systems in the impulsive noise environment.

WED-AM1-O6.5: Random Signal Estimation by Ergodicity associated with Linear Canonical Transform

Liyun Xu

Shanxi University

The linear canonical transform (LCT) provides a general mathematical tool for solving problems in optical and quantum mechanics. For random signals, which are bandlimited in the LCT domain, the linear canonical correlation function and the linear canonical power spectral density form a LCT pair. The linear canonical translation operator, which is used to define the convolution and correlation functions, also plays a significant role in the analysis of the random signal estimation. Firstly, the eigenfunctions which are invariant under the linear canonical translation and the unitarity property of it are discussed. All of these connect the LCT sampling theorem and the von Neumann ergodic theorem in the sense of distribution, which will develop an estimation method for the power spectral density of a chirp stationary random signal from one sampling signal in the LCT domain.

WED-AM1-O7
Signal Processing Methods

Time: Wednesday, Nov 20, 10:20-12:00

Place: A7

Chair: Shinsuke Ibi

WED-AM1-O7.1: Study on Pre-warning Model of Railway Signal System with Fuzzy Analytical Hierarchy Process

Shanpeng Zhao, Shaoxiang Zhao, Youpeng Zhang and Zhengjie Xu

Lanzhou Jiaotong University, Zhuzhou crrc times electric co.

The safety issue of high speed railway signal system is very important. According to the railway signal system enterprises' safe characteristics, this paper analyzed the risk factors of railway signal system and studied the safety problem of “human-machine-environment” system. The risk factors of railway signal system are divided into natural geological factors, personnel factors, equipment factors and management factors. Based on the fuzzy mathematics and these factors, the safety pre-warning model of railway signal system is constructed by using the fuzzy AHP method. Through the use of pre-warning model, we can find the risk sources and hidden danger of railway signal system in time. This pre-warning model is applied to the railway signal system. The results show that this pre-warning model is effective to prevent the accident.

WED-AM1-O7.2: Energy Management in Energy Harvesting Wireless Networks: A Reinforcement Learning Framework

Chengrun Qiu, Yang Hu and Yan Chen

University of Science and Technology of China

In this paper, we propose a novel energy management algorithm based on the reinforcement learning to optimize the net bit rate in energy harvesting (EH) networks. By utilizing deep deterministic policy gradient (DDPG), the proposed algorithm is applicable for the continuous states and realizes the continuous energy management. With only one day's real solar data and the simulative channel data for training, the proposed algorithm shows excellent performance in the validation with about 800 days length of real solar data. Compared with the state-of-the-art algorithms, the proposed algorithm achieves better performance in terms of long-term average net bit rate.

WED-AM1-O7.3: Calibration of Position and Orientation between Cameras without Common Field of View Using Cooperative Target

Yongzhi Min and Jie Hu

School of Automation and Electrical Engineering, Lanzhou Jiaotong University

In online surface settlement monitoring system of image-based ballastless track, a transfer station consisting of a multi-camera without a common field of view is often used. Because measurement error will accumulate as the number of transfer increases, position and orientation relation betweeen cameras should be calibrated. Traditional calibration methods based on total stations are pretty complicated. We found that the hand-eye calibration of the robot is equivalent to this problem. Firstly , the multi-camera is moved twice in small step and the multi-camera captures the cooperative target image at three different positions and image physical coordinates of four mark points in each image are extracted. Secondly , in order to solve the extrinsic parameters of the camera, we use a special P4P algorithm for feature points distributed in a square. Lastly, the matrix rearrangement method is used to solve the hand-eye transformation matrix according to the extrinsic parameters of the cameras. The experimental results show that the measurement accuracy of this method is same as the traditional method. Moreover , the method features simple operation , less calculation task and high precision position and orientation.

WED-AM1-O7.4: Frequency Decomposition Model of Popularity Evolution in Online Social Media

Shufeng Duan, Ligu Zhu, Yujing Shi, Lei Zhang and Bo Hui

Communication University of China, Shijiazhuang TieDao University, HeBei Provincial Public Security Department, BaZhou Public Security Bureau

The effective analysis of information diffusion popularity in online social media plays an important role in accurate mastering the development trend of public opinion and maintaining the normal order of society.

The existing quantitative methods are mainly based on time-domain, which fails to comprehensively analyze the network dynamics, the contributing factor with the generalization. To this end, an assumption of waveform synthesis in the frequency-domain is presented, that is, the waveform of popularity evolution in the time-domain without noise is a linear combination of many sinusoidal waveforms with different frequencies and a single period via stretching and translation transforming. The transformation parameters can be utilized to analyze contributing factor and their variety corresponds to the network dynamic because such parameters are relative to the motive power and burst-time. Therefore, a waveform decomposition algorithm is designed for actual popularity evolution in this paper. The data of the diffusion waveform is standardized and normalized, then the greedy algorithm is utilized to decompose the waveform, and the results include two parts: the parameters of sub-waveforms and the value of SNR(Signal-to-Noise Ratio). Finally, the standardization and normalization benchmarks are gained via the Baseline Dataset which is the popularity evolution dataset in Weibo, and three different types of granularity are chosen to decompose the diffusion waveform in both the Baseline Dataset and the SpecialNews Dataset which includes entertainment news only. The results are compared to prove the generalization potential and feasibility of the model.

WED-AM1-O7.5: Acoustic-Domain Self-Interference Cancellation for Full-Duplex Underwater Acoustic Communication Systems

Yanyan Wang, Yingsong Li, Lu Shen and Yuriy Zakharov

Harbin Engineering University, University of York

In full-duplex (FD) underwater acoustic communication (FD-UWAC) systems, the self-interference (SI) will affect the communication performance. Till now, there is no solution for active cancellation of the wide-band SI in the acoustic domain. In this paper, we propose such a solution with two transducers, a primary transducer and a secondary transducer. The acoustic signal emitted by the secondary transducer is generated to cancel the SI signal received at the hydrophone from the primary transducer. The performance of the proposed scheme is investigated by simulation. We use the Waymark UWA simulator that allows the virtual signal transmission in various acoustic environments. The simulation results demonstrate that the proposed scheme can provide an effective acoustic SI cancellation for FD-UWAC systems, in terms of the mean square error and bit error ratio.

WED-AM1-O8
Image Processing

Time: Wednesday, Nov 20, 10:20-12:00

Place: A8

Chair: Sanghoon Lee

WED-AM1-O8.1: Background Modeling Algorithm for Multi-feature Fusion

Zhicheng Guo, Jianwu Dang, Yangping Wang and Jing Jin

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou Jiaotong University

In order to improve the accuracy of foreground target detection and establish a stable background model, this paper proposes a multi-feature fusion background modeling algorithm, which initializes the background model with the spatial correlation between the first frame pixel and the domain pixel, and quickly establishes the background. model. A multi-feature sample set consisting of pixel values, update frequency, update time, and adaptive dynamic coefficients is updated with temporal correlation of subsequent intra-pixels. According to the multi-feature sample set, the background complexity is adjusted to adjust the update speed of the model in different regions, which effectively improves the ghost phenomenon of the foreground target and reduces the false holes in the target and the false foreground in the background. The test results of multiple sets of data sets show that the proposed algorithm improves the adaptability and robustness of foreground target detection in scenarios with high dynamic changes.

WED-AM1-O8.2: Image Haze Removal By Adaptive CycleGAN

Yi–Fan Chen, Amey Patel and Chia–Ping Chen

National Sun Yat–sen University, Indian Institute of Technology Indore

We introduce our machine-learning method to remove the fog and haze in image. Our model is based on CycleGAN, an ingenious image-to-image translation model, which can be applied to de-hazing task. The datasets that we used for training and testing are creatd according to the atmospheric scattering model. With the change of the adversarial loss from cross-entropy loss to hinge loss, and the change of the reconstruction loss from MAE loss to perceptual loss, we improve the performance measure of SSIM value from 0.828 to 0.841 on the NYU dataset. With the Middlebury stereo datasets, we achieve an SSIM value of 0.811, which is significantly better than the baseline CycleGAN model.

WED-AM1-O8.3: Remote Sensing Image Scene Classification Based on SURF Feature and Deep Learning

Jinxiang Liang, Jianwu Dang, Yangping Wang, Jingyu Yang and Zhenhai Zhang

Lanzhou Jiaotong University, Lanzhou Bocai Technology Co.

Remote sensing image scene classification is one of the key points in remote sensing image interpretation. The traditional remote sensing image scene classification feature performance is not strong, and the deep learning extraction semantic feature process is complex. This paper proposes a fusion feature remote sensing image scene classification method which is based on artificial feature and deep learning semantic feature. Firstly, the SURF feature of remote sensing image is extracted and encoded by the VLAD algorithm. The semantic feature of remote sensing image is extracted by transfer learning. Then the feature reduction is performed by PCA algorithm and feature fusion is performed. Finally, the scene classifier is trained by using the random forest algorithm. The experimental results show that the classification accuracy and Kappa coefficient of this method are higher and the method is effective.

WED-AM1-O8.4: Research on the Improved Retinex Algorithm for Low Illumination Image Enhancement

Shaoquan Wang, Deyong Gao, Yangping Wang and Song Wang

Lanzhou Jiaotong University

Low-illumination images are generally low-quality images. The retinex algorithm can cause halo artifacts and loss of details in processing. Therefore, an improved Retinex algorithm is proposed. Firstly, the HSI color space which is more in line with the human visual characteristics is selected instead of the RGB image, that is, the luminance component I is processed. Then, the illuminance image is estimated by using a guided filter that fuses the edge detection operator, and the edge detection operator can be better positioned. At the edge, an illuminance image with rich edge information can be obtained; after

obtaining the illuminance image, the reflected image can be obtained by the Retinex principle, the obtained reflected image is subjected to low-rank decomposition, and the low-rank property of the image is used to suppress the enlarged halo and the enhancement process. Noise; finally, the visual effect is further improved by local contrast enhancement. Experiments show that the algorithm can effectively improve the brightness and contrast of the image, preserve the details of the image, and also suppress the noise interference in the enhancement process. The subjective visual effect and objective evaluation results of the image have also been greatly improved.

WED-AM1-O8.5: Encrypted JPEG image retrieval using histograms of transformed coefficients

Peiya Li and Zhenhui Situ

Jinan University, The Hong Kong Polytechnic University

This work proposes an encrypted JPEG image retrieval mechanism based on the histograms of transformed coefficients. With this scheme, JPEG image is encrypted during its compression process by using other orthogonal transforms for blocks' transformation, rather than 8×8 DCT. Then the encrypted images are transferred to and stored in the cloud server. When receiving an encrypted query image from the authorized user, the server calculates the histograms of transformed coefficients located at different frequency positions. By computing the distance between the histograms of encrypted query image and database cipherimages, encrypted images with plaintext content similar to the query image are retured to the authorized user for decryption. Experiments are conducted to show that our scheme can provide effective cipherimage retrieval service, while ensure format compliance and compression friendly.

WED-AM1-O8.6: CNN-based bit-depth enhancement by the suppression of false contour and color distortion

Changmeng Peng, Luting Cai, Zhizhong Fu and Xiaofeng Li

UESTC:University of Electronic Science and Technology of China

Although 10-bit monitors are getting popular, most of the available media sources are 8-bit. The inconsistence between the low-bit-depth media sources and high-bit-depth monitors should be properly solved to make full use of the high-bit-depth equipments. Simply converting low-bit-depth images/videos to high-bit-depth ones via zero-padding would result in false contour artifacts in smooth region, which greatly degrades the visual quality. In this paper, a novel auto-encoder like CNN model is proposed to convert low-bit-depth images to high-bit-depth ones. Our method can significantly suppress false contour by the use of vgg loss(mean square error computed on pre-trained VGG-19 feature maps). However, significant color distortion would be found in some results if only vgg loss is used. In order to suppress color distortion, range loss is proposed which restrains the difference between the resultant pixel values and the zero-padded ones within the range of [0, S), where S is the requantization step. Benefit from the novel network model and the designed loss function consisting of range loss and vgg loss, the proposed method has comparable objective metric with state-ofthe-art. In particular, our method achieves better visual quality by significantly suppressing false contour artifacts and color distortion. Those conclusions are proved by experiments, and our code can be found at https://github.com/pengcm/BE-AUTO-ext.

POSTER 2

Time: Wednesday, Nov 20, 13:20-15:00

Place: Gansu International Conference Centre(GICC), 3F

Chair: Xiangui Kang

POSTER 2.1: Effective training End-to-End ASR systems for low-resource Lhasa dialect of Tibetan language

Lixin Pan, Sheng Li, Longbiao Wang and Jianwu Dang, Japan

Tianjin University, National Institute of Information & Communications Technology (NICT), Japan Advanced Institute of Science and Technology

The Lhasa dialect is the most important Tibetan dialect and has the largest number of speakers in Tibet and massive written scripts in the long history. Studying how to apply speech recognition techniques to Lhasa dialect has special meaning for preserving Tibet's unique linguistic diversity. Previous research on Tibetan speech recognition focussed on selecting phone-level acoustic modeling units and incorporating tonal information but paid less attention to the problem of limited data. In this paper, we focus on training End-to-End ASR systems for Lhasa dialect using transformer-based models. To solve the low-resource data problem, we investigate effective initialization strategies and introduce highly compressed and reliable sub-character units for acoustic modeling which have never been used before. We jointly training the transformer-based End-to-End acoustic model with two different acoustic unit sets and introduce an error-correction dictionary to further improve the system performance. Experiments show our proposed method can effectively modeling low-resource Lhasa dialect compared to DNN-HMM baseline systems.

POSTER 2.2: Joint Training ResCNN-based Voice Activity Detection with Speech Enhancement

Tianjiao Xu, Hui Zhang and Xueliang Zhang

Inner Mongolia University

Voice activity detection (VAD) is considered as a solved problem in noise-free condition, but it is still a challenge task in low signal-to-noise ratio (SNR) noisy condition. Intuitively, reducing noise will improve the VAD. Therefore, in this study, we introduce a speech enhancement module to reduce noise. Specifically, a convolutional recurrent neural network (CRN) based encoder-decoder speech separation module is trained to reduce noise. Then the low-dimensional features code from its encoder together with the raw spectrum of noisy speech are feed into a deep residual convolutional neural network (ResCNN) based VAD module. The speech separation and VAD modules are connected and trained jointly. To balance the training speed of the two modules, an empirical dynamic gradient balance strategy is proposed. Experimental results show that the proposed joint-training method has obvious advantages in generalization ability.

POSTER 2.3: A Rescoring Method Using Web Search and Word Vectors for Spoken Term Detection

Haruka Tanji, Kazunori Kojima, Hiroaki Nanjo, Shi-Wook Lee and Yoshiaki Itoh

Iwate Prefectural University, Kyoto University, Tokyo Institute of Technology

We propose a rescoring method using words related to a query obtained by Web search and word vectors for spoken term detection (STD). In this paper, we assume that words associated with the topic in speech data and co-occurring with the query are called “words related to the query”, and that the related words appear multiple times in the speech data. To identify the words related to the query, we introduce distributed expression of words obtained by Word2vec, and first convert each word in the word-recognition results of speech data into a word vector. Each word vector is then compared with a word vector of the query. Words related to the query are determined by calculating the degree of similarity between the two word vectors. However, a word vector of an out-of-vocabulary (OOV) query cannot be obtained in this manner, since OOV queries do not appear in word-recognition results. For such OOV queries, we perform a Web search using the query, whereupon texts including the query are extracted. Recognition results of the speech data and the extracted texts are then combined and used for training of Word2vec. In this manner, a word vector of the OOV query can be obtained. Distances to all candidates in the document, including words related to the query, are used advantageously. Experiments are conducted to evaluate the performance of the proposed method using open test collections of the NTCIR-10 and NTCIR-12 workshops. For retrieval accuracy, an improvement of 3.2 points in mean average precision was achieved using the proposed method.

POSTER 2.4: Derivative of instantaneous frequency for voice activity detection using phase-based approach

Binh Thien Nguyen, Yukoh Wakabayashi, Takahiro Fukumori and Takanobu Nishiura

Ritsumeikan University

In this paper, we consider the use of the phase spectrum in speech signal analysis. In particular, we propose a phase-based voice activity detection by using the derivative of instantaneous frequency. Preliminary experiments reveal that the distribution of this feature can indicate the presence or absence of speech. We evaluated the performance of the proposed method in comparison with the conventional amplitude-based method. In addition, we considered a combination of the amplitude-based and phase-based methods in a simple manner to demonstrate the complementarity of both spectra. The experimental results confirmed that the phase information can be used to detect voice activity with at least 62% accuracy. The proposed method showed better performance compared to the conventional amplitude-based method in the case when a speech signal was corrupted by white noise at low signal-to-noise ratio (SNR). A combination of two methods achieved even higher performance than each of them separately, in limited conditions.

POSTER 2.5: Voice Activity Detection Based on Time-Delay Neural Networks

Ye Bai, Jiangyan Yi, Jianhua Tao, Zhengqi Wen and Bin Liu

Institute of Automation, Chinese Academy of Sciences

Voice activity detection (VAD) is an important preprocessing part of many speech applications. Context information is important for VAD. Time-delay neural networks (TDNNs) capture long context information with a few parameters. This paper investigates a TDNN based VAD framework. A simple chunk based decision method is proposed to smooth raw posteriors and decide border points of utterances. To evaluate decision performance, a metric intersection-over-union (IoU) is introduced from image object detection. The experiment results are evaluated on Wall Street Journal (WSJ0) corpus. Frame classification performance is measured by area under the curve (AUC) and equal error rate (EER). Compared with long short-term memory baseline, the TDNN based system achieves a 41.26% EER relative reduction on average in matched noise condition, and relative improvement of average AUC is 3.82%. Proposed decision method achieves an 18.74% IoU relative improvement on average compared with moving average method on average.

POSTER 2.6: Investigation of Neural Network Approaches for Unified Spectral and Prosodic Feature Enhancement

Wei–Cheng Lin, Yu Tsao, Hsin–Min Wang and Fei Chen

Academia Sinica, Southern University of Science and Technology

Most speech enhancement (SE) systems focus on the spectral feature or raw-waveform enhancement. However, many speech-related applications rely on other features other than the spectral features, such as the intensity and fundamental frequency (f0). Therefore, a unified feature enhancement for different types of features is worth investigating. In this work, we train our neural network (NN)-based SE system in a manner that simultaneously minimizes the spectral loss and preserves the correctness of the intensity and f0 contours extracted from the enhanced speech. The idea is to introduce an NN-based feature extractor to the SE framework that imitates the feature extraction of Praat. Then, we can train the SE system by minimizing the combined loss of the spectral feature, intensity, and f0. We investigate three bidirectional long short-term memory (BLSTM)-based unified feature enhancement systems: fixed-concat, joint-concat, and multi-task. The results of the experiments on the Taiwan Mandarin hearing in a noise test dataset (TMHINT) demonstrate that all three systems show improved intensity and f0 extraction accuracy without sacrificing the perceptual evaluation of the speech quality and short-time objective intelligibility scores compared with the baseline SE system. Further analysis of the experimental results shows that the improvement mostly comes from better f0 contours under difficult conditions such as low signal-to-noise ratio and nonstationary noises. Our work demonstrates the advantage of the unified feature enhancement and provides new insights for SE.

POSTER 2.7: Improve Data Utilization with Two-stage Learning in CNN-LSTM-based Voice Activity Detection

Tianjiao Xu, Hao Li, Hui Zhang and Xueliang Zhang

Inner Mongolia University

Voice activity detection (VAD) is essential for the speech signal processing system. Convolutional long short-term memory deep neural network (CLDNN), which consists of a CNN and an LSTM, has shown excellent improvement in VAD. However, the training data of the CLDNN must be sequence data

because of the LSTM. To improve data utilization, we proposed a two-stage training strategy. Specifically, the first stage trains the CNN on shuffled frame-level data to get high-level feature expression, individually. The second stage trains the LSTM to model the speech continuity. We show that our method has obvious advantages in discriminative ability and generalization ability than compared approaches in different scale of training data, especially in small datasets. The proposed method achieves over 2.89% relative improvement than the original CLDNN on noise matched condition and over 1.07% on unmatched condition.

POSTER 2.8: Allpass Modeling of Phase Spectrum of Speech Signals for Formant Tracking

Karthika Vijayan, K Sri Rama Murty and Haizhou Li

National University of Singapore, Indian Institute of Technology Hyderabad

Formant tracking is a very important task in speech applications. Most of the current formant tracking methods bank on peak picking from linear prediction (LP) spectrum of speech, which suffers from merged/spurious peaks in LP spectra, resulting in unreliable formant candidates. In this paper, we present the significance of phase spectrum of speech in refining the formant candidates from LP analysis. The short-time phase spectrum of speech is modeled as phase response of an allpass (AP) system, where the coefficients of AP system are initialized with LP coefficients and estimated with an iterative procedure. This technique refines the initial formants from LP analysis using phase spectrum of speech through an AP analysis, thereby accom- plishing fusion of information from magnitude and phase spectra. The group delay of the resultant AP system exhibits unambiguous peaks at formants and, delivers reliable formant candidates. The formant trajectories obtained by selection of formants from these candidates are reported to be more accurate than those obtained from LP analysis. The fused information from magnitude and phase spectra has rendered relative improvements of 25%, 15% and 18% in tracking accuracy of first, second and third formants, respectively, over those from magnitude spectrum alone.

POSTER 2.9: A Study on Mispronunciation Detection Based on Fine-grained Speech Attribute

Minghao Guo, Rui Cai, Wei Wang, Jinsong Zhang, Yanlu Xie and Lin Binghuai

Beijing Language and Culture University, MIG

Over the last decade, several studies have investigated speech attribute detection (SAD) for improving computer assisted pronunciation training (CAPT) systems. The pre-defined speech attribute catagories either is IPA or language-dependent catogories, which cannot handle multiple languages mispronunciation detection. In this paper, we propose a fine-grained speech attribute (FSA) modeling method, which defines types of Chinese speech attribute by combining Chinese phonetics with the international phonetic alphabet (IPA). To verify FSA, a large scale Chinese corpus was used to train Time-delay neural networks (TDNN) based on speech attribute models, and tested on Russian learner data set. Experimental results showed that all FSA's accuracy on Chinese test set is about 95% on average, and the diagnosis accuracy of the FSA-based mispronunciation detection achieved a 2.2% improvement compared to that of segment-based baseline system. Besides, as the FSA is theoretically capable of modeling language-universal speech attributes, we also tested the trained FSA-based method on native English corpus, which achieved about 50% accuracy rate.

POSTER 2.10: Robust Camera Model Identification Based on Richer Convolutional Feature Network

Zeyu Zou, Yunxia Liu, Wenna Zhang, Yuehui Chen, Yunli Zang, Yang Yang and Ngai-Fong Law

University of Jinan, Integrated Electronic Systems Lab Co., Shandong University, The Hong Kong Polytechnic University

Based on convolutional neural network (CNN), the problem of robust patch level camera model identification is studied in this paper. Firstly, effective feature representation is achieved by concatenating a multiscale residual prediction module as well as the original RGB channels. Motivated by exploration of multi-scale characteristic, the multiscale residual prediction module automatically learn the residual images to avoid the subsequent CNN being affected by the scene content. Furthermore, color channel information is integrated for enhanced diversity of CNN inputs. After that, a modified richer convolutional feature network is proposed for robust camera model identification by fully exploiting the learnt features. Finally, the effectiveness of the proposed method is verified with abundant experimental results.

POSTER 2.11: A Robust Method for Blindly Estimating Speech Transmission Index using Convolutional Neural Network with Temporal Amplitude Envelope

Suradej Duangpummet, Jessada Karnjana, Waree Kongprawechnon and Masashi Unoki

Japan Advanced Institute of Science and Technology, National Science and Technology Development Agency, Sirindhorn International Institute of Technology, Thammasat University, Japan Advanced Institute of Science and Technology

We have developed a robust scheme for blindly estimating the speech transmission index (STI) based on a convolutional neural network (CNN) with temporal envelope features. When assessing the quality of acoustics in a room where there are people present, STI needs to be estimated without measuring the room impulse response (RIR) or using a modulation transfer function (MTF). This can be problematic because a blind method based on the MTF has low accuracy when the stochastic models of RIR and the background noise are mismatched to real sound environments. We improve the accuracy of STI estimation in noisy reverberant spaces by using a CNN that takes the entire temporal amplitude envelope of an observed speech signal as its input. Simulations were performed to evaluate the proposed scheme and results showed that it can maintain the appropriate accuracy under various realistic room acoustic conditions with an average RMSE of 0.12 and correlation of 0.87. These results demonstrate that the proposed scheme can robustly and blindly estimate STIs in noisy reverberant environments.

POSTER 2.12: Compressing Speech Recognition Networks with MLP via Tensor-Train Decomposition

Dan He and Yubin Zhong

Guangzhou University

Deep neural networks (DNNs) have produced state-of-the-art performance in automatic speech recognition (ASR).This success is often associated with a large DNN structure with millions or even billions of parameters. Such large-scale networks take large disk space and require huge computational resources at run-time, therefore not suitable for applications in mobile or wearable devices. In this paper, we investigate a compression approach for DNNs based on Tensor-Train (TT) decomposition and apply it to the ASR task. Our results on the TIMIT database reveals that the compressed networks can maintain the performance of the original full-connected network, while greatly reducing the number of parameters. In particular, we found that the rate of model size decreasing is much larger than the rate of WER (word error rate) increasing, which means that the performance loss caused by the TT-based compression can be well compensated by the model size reduction.Moreover, how many layers and which layer can be substituted by TT is application dependent and should be carefully designed according to the application scenario.

POSTER 2.13: Generalized Combined Nonlinear Adaptive Filters for Nonlinear Acoustic Echo Cancellation

Wenxia Lu, Lijun Zhang, Jie Chen and Jingdong Chen

Northwestern Polytechnical University

This paper proposes a new and generalized algorithm of combination of nonlinear adaptive filters (CNAF) for nonlinear acoustic echo cancellation (NAEC). In contrast to combining filters in a conventional parallel manner, the candidate filters are organized to form in a network structure with two subnetworks. The nodes in each subnetwork serve as linear and nonlinear
adaptive filters respectively. A generalized CNAF (GCNAF) is then obtained by linking the nodes in the network and using the diffusion adaptation strategy. The proposed GCNAF algorithm allows information exchange and sharing among the nodes, so as to maximally optimize the performance of the combined filters. Simulations with noise and speech signals demonstrate the effectiveness of the proposed GCNAF for the NAEC problem.

POSTER 2.14: Automatic Prosodic Structure Labeling using DNN-BGRU-CRF Hybrid Neural Network

Yao Du, Zhiyong Wu, Shiyin Kang, Dan Su, Dong Yu and Helen Meng

Graduate School at Shenzhen, Tsinghua University, Tencent AI Lab, The Chinese University of Hong Kong

The speech corpus with labeled prosodic structure information is crucial for text-to-speech (TTS) synthesis to train a reliable model that can generate high quality natural syn- thetic speech. Traditional manual prosodic structure labeling is laborious and time-consuming and may encounter inconsistency problem caused by different annotators. Automatic prosodic labeling is thus desirable, which can not only speech up the labeling process, but also keep the labeling results from the inconsistency problem. This paper presents a DNN-BGRU-CRF hybrid neural network, which aggregates the advantages of deep neural network, bidirectional gated recurrent units and conditional random fields, to label three-level prosodic structure boundaries. It exploits both text and acoustic cues in a neural network framework. Experimental results demonstrate the effec-tiveness of the proposed model.

POSTER 2.15: Monaural Singing Voice Separation Using Fusion-Net with Time-Frequency Masking

Feng Li, Kaizhi Qian, Mark Hasegawa-Johnson and Masato Akagi

Japan Advanced Institute of Science and Technology, University of Illinois at Urbana-Champaign

Monaural singing voice separation has received much attention in recent years. In this paper, we propose a novel neural network architecture for monaural singing voice separation, Fusion-Net, which is combining U-Net with the residual convolutional neural network to develop a much deeper neural network architecture with summation-based skip connections. In addition, we apply time-frequency masking to improve the separation results. Finally, we integrate the phase spectra with magnitude spectra as the post-processing to optimize the separated singing voice from the mixture music. Experimental results demonstrate that the proposed method can achieve better separation performance than the previous U-Net architecture on the ccMixter database.

POSTER 2.16: Identification of Alzheimer's disease Patients based on Oral Speech Features

Qing Zhou, Yong Ma, Benyan Luo, Mingliang Gu and Zude Zhu

School of Linguistic Sciences and Arts, School of Physics and Electronic Engineering, Jiangsu Normal University, Department of Neurology, the First Affiliated Hospital of Medical School of Zhejiang University

Screening Alzheimer's disease (AD) patients quickly and non-invasively is of great challenge in the field of clinical medical. In this study, a method based on oral speech features for AD patients identification was proposed. AD (27 people), MCI (Mild Cognitive Impairment, 42 people) and HCs (Healthy Controls, 25 people) were recruited to make a detailed description of the Cookie Theft picture. Linguistic features and acoustic features were extracted manually and automatically respectively from the speech. Based on these features, Support Vector Machine (SVM) classifier was adopted to model and identify AD patients. The results based on linguistic features and acoustic features reached an accuracy of 94.2% and 93.62% respectively. The results suggested that a validated oral task could be further used with automatic algorithm in AD identification. This study is the first study to classify Chinese AD patients with linguistic features and acoustic features, sending important message for rapid AD early screening based on a quick ecological oral task.

POSTER 2.17: Mixture of CNN Experts from Multiple Acoustic Feature Domain for Music Genre Classification

Yang Yi, Kuan-Yu Chen and Hung-Yan Gu

National Taiwan University of Science and Technology

Nowadays, deep learning achieves successful results in many research fields, one of the well-known contributions is convolution neural network (CNN) in computer vision. In the field of music information retrieval (MIR), audio spectrogram can carry a great deal of information about the music content so as to be a robust visual representation for music signal. Recently, many research literatures show that CNN has ability to capture indicative acoustic patterns from spectrogram input, and make remarkable performance on MIR-related tasks such as music genre classification (MGC). In this paper, we continue the line of research to explore different type of spectrograms, such as the harmonic and percussive components generated from the original spectrogram, to emphasize different characteristics of music genre for MGC task. Besides, modulation from time domain of original spectrogram is also used, which is containing temporal dynamics of music signal. To jointly leverage all of these features, in this paper, a mixture of experts (MoE) system is proposed. More formally, a set of MGC models can be derived by using the various spectrogram-based statistics. Since these models capture different characteristics in music, we treat each model as an individual expert. Accordingly, a neural mixture model is introduced to collect and compile the predictions from the expert models, and then to output a final decision for a given music to be predicted. In a nutshell, our major contributions in this paper are at least twofold. On one

hand, we comprehensively examine several spectrogram-based features for the MGC task. On the other hand, a neural-based MoE system, which can dynamically decide the weighting factor for each expert system, is proposed to enhance the performance of MGC task. Experimental results demonstrate that the proposed framework not only can achieve success results than individual expert model, but has ability to provide comparable classification accuracy to the SOTA systems.

POSTER 2.18: Utterance-level Permutation Invariant Training with Latency-controlled BLSTM for Single-channel Multi-talker Speech Separation

Lu Huang, Gaofeng Cheng, Pengyuan Zhang, Yi Yang, Shumin Xu and Jiasong Sun

Tsinghua University, Institute of Acoustics, Chinese Academy of Sciences, North China Power Engineering CO.

Utterance-level permutation invariant training (uPIT) has achieved promising progress on single-channel multi- talker speech separation task. Long short-term memory (LSTM) and bidirectional LSTM (BLSTM) are widely used as the separation networks of uPIT, i.e. uPIT-LSTM and uPIT-BLSTM. uPIT-LSTM has lower latency but worse performance, while uPIT-BLSTM has better performance but higher latency. In this paper, we propose using latency-controlled BLSTM (LC-BLSTM) during inference to fulfill low-latency and good-performance speech separation. To find a better training strategy for BLSTM- based separation network, chunk-level PIT (cPIT) and uPIT are compared. The experimental results show that uPIT outperforms cPIT when LC-BLSTM is used during inference. It is also found that the inter-chunk speaker tracing (ST) can further improve the separation performance of uPIT-LC-BLSTM. Evaluated on the WSJ0 two-talker mixed-speech separation task, the absolute gap of signal-to-distortion ratio (SDR) between uPIT-BLSTM and uPIT-LC-BLSTM is reduced to within 0.7 dB.

POSTER 2.19: Bidirectional Temporal Convolution with Self-Attention Network for CTC-Based Acoustic Modeling

Jian Sun, Wu Guo, Bin Gu and Yao Liu

University of Science and Technology of China, China General Technology Research Institute

Connectionist temporal classification (CTC) based on recurrent (RNNs) or convolutional neural networks (CNNs) is a method for end-to-end acoustic modeling. Inspired by the recent success of the self-attention network (SAN) in machine translation and other domains such as images, we apply the SAN to CTC acoustic modeling in this paper. SAN has powerful capabilities for capturing global dependencies, but it cannot model the sequential information and local interactions of utterances. The bidirectional temporal convolution with self-attention network (BTCSAN) is proposed in order to capture both the global and local dependencies of utterances. Furthermore, the down- and upsampling strategies are adopted in the proposed BTCSAN in order to achieve computational efficiency and high recognition accuracy. Experiments are carried out using the King-ASR-117 Japanese corpus. The proposed BTCSAN can obtain a 15.87% relative improvement in the CER over the BLSTM-based CTC baseline.

POSTER 2.20: Query-by-Example Spoken Term Detection using Attentive Pooling Networks

Kun Zhang, Zhiyong Wu, Jia Jia, Helen Meng and Binheng Song

Graduate School of Shenzhen, Tsinghua University, The Chinese University of Hong Kong

Query-by-example spoken term detection (QbE-STD) is attractive because it's a key technology for retrieving and browsing spoken content without transcribing them into text. Several end-to-end models based on encoder architecture have been proposed for QbE-STD, in which the input pair, spoken query and audio segment, are first projected into fixed-length vector representations by feature extraction module and then similarity measure module is used to output detection score based on the representations. Attention mechanism has been applied into the feature extractor; however, traditional approach calculates attention vector for audio segment only, which makes it a one-way attention mechanism. In this paper, we present a novel feature extraction module based on two-way attention mechanism, called attentive pooling networks, for end-to-end QbE-STD. The main idea is to learn a similarity measure over the projected input pair and extract information in a way that two input items can directly influence the computation of each other's representation. Evaluations on the LibriSpeech corpus and cross-linguistic audio archive confirm the effectiveness of our proposed approach compared to the traditional ones.

POSTER 2.21: Novel Adaptive Generative Adversarial Network for Voice Conversion

Maitreya Patel, Mihir Parmar, Savan Doshi, Nirmesh Shah and Hemant Patil

Graduate School of Shenzhen, Dhirubhai Ambani Institute of Information and Communication Technology, Arizona State University

Voice Conversion (VC) converts the speaking style of a source speaker to the speaking style of a target speaker by preserving the linguistic content of a given speech utterance. Recently, Cycle Consistent Adversarial Network (CycleGAN), and its variants have become popular for non-parallel VC tasks. However, CycleGAN uses two different generators and discriminators. In this paper, we introduce a novel Adaptive Generative Adversarial Network (AdaGAN) for non-parallel VC task, which effectively requires single generator, and two discriminators for transferring the style from one speaker to another while preserving the linguistic content in the converted voices. To the best of authors' knowledge, this is the first study of its kind to introduce a new Generative Adversarial Network (GAN)-based architecture (i.e., AdaGAN) in machine learning literature, and the first attempt to apply this architecture for non-parallel VC task. In this paper, we compared the results of the AdaGAN w.r.t. state-of-the-art CycleGAN architecture. Detailed subjective and objective tests are carried out on the publicly available VC Challenge 2018 corpus. In addition, we perform three statistical analysis which show effectiveness of AdaGAN over CycleGAN for parallel-data free one-to-one VC. For inter-gender and intra-gender VC, We observe that the AdaGAN yield objective results that are comparable to the CycleGAN, and are superior in terms of subjective evaluation. A subjective evaluation shows that AdaGAN outperforms CycleGAN-VC in terms of naturalness, sound quality, and speaker similarity. AdaGAN was preferred 58.33% and 41% time more over CycleGAN in terms of speaker similarity and sound quality, respectively.

POSTER 2.22: Many-to-many Cross-lingual Voice Conversion with a Jointly Trained Speaker Embedding Network

Yi Zhou, Xiaohai Tian, Rohan Kumar Das and Haizhou Li

National University of Singapore

Among various voice conversion (VC) techniques, average modeling approach has achieved good performance as it benefits from training date of multiple speakers, therefore, reducing the reliance on training date from the target speaker. Many existing average modeling approaches rely on the use of i-vector to represent the speaker identity for model adaptation. As such i-vector is extracted in a separate process, it is not optimized to achieve the best voice conversion quality for the average model. To address this problem, we propose a low dimensional trainable speaker embedding network that augments the primary VC network for joint training. We validate the effectiveness of the proposed idea by performing a many-to-many cross-lingual VC, which is one of the most challenging tasks in VC. We compare the i-vector scheme with the speaker embedding network in the experiments. It is found that the proposed system effectively improve the speech quality and speaker similarity.

POSTER 2.23: A MULTI-SCALE FULLY CONVOLUTIONAL NETWORK FOR SINGING MELODY EXTRACTION

Ping Gao, Cheng—You You and Tai—Shih Chi

National Chiao Tung University

The melody extraction can be considered as a sequence-to-sequence task or a classification task. Many recent models based on semantic segmentation have been proven very effective in melody extraction. In this paper, we built up a fully convolutional network (FCN) for melody extraction from polyphonic music. Inspired by the state-of-the-art architecture of the semantic segmentation, we constructed the encoder in a dense way and designed the decoder accordingly for audio processing. The combined frequency and periodicity (CFP) representation, which contains spectral and cepstral information, was adopted as the input feature of the proposed model. We conducted performance comparison between the proposed model and several methods on various datasets. Experimental results show the proposed model achieves state-of-the-art performance with less computation and fewer parameters.

POSTER 2.24: End-to-end Tibetan Speech Synthesis Based on Phones and Semi-syllables

Guanyu Li, Lisai Luo, Chunwei Gong and Shiliang Lv

Northwest Minzu University

Tibetan speech synthesis based on end-to-end is studied and implemented. Due to the 2D architecture of Tibetan characters, it is not convenient to treat the letters sequences as the input of the model. The experiment is conducted based on phones and semi-syllables respectively. During training and testing, the text is segmented into a sequence of syllables first, then syllables are transformed into phones and semi-syllables as the input sequence of the model. The results demonstrate that the encoding and decoding alignment effect of Tibetan speech synthesis based on phones is better than that based on

semi-syllables. In addition, the results demonstrate that the Highway network neural network in the structure plays a key role in the convergence of the model.

POSTER 2.25: Reversible Data Hiding in PDF Document Exploiting Prefix Zeros in Glyph Coordinates

Neelesh Nursiah, Koksheik Wong and Minoru Kuribayashi

Monash University, Okayama University

In the contemporary world of information technology, PDF (Portable Document Format) has become the de facto document standard which allows users to exchange and view electronic documents across various platforms. PDF is the most widely exchanged document format since Internet gained popularity. Although PDF has a good authenticity system by making use of digital signatures, the file format is still susceptible to copyright infringement as there are many libraries available on the Internet to bypass the digital signature of a PDF. Therefore, claiming ownership for PDF has become a paramount issue that needs to be addressed. This paper proposes the idea of hiding data in the glyph positioning coordinate value. To suppress bit stream size increment, the reverse zero-run length coding technique is adopted. Experiments are conducted to verify the basic performance of the proposed data hiding method. In the best case scenario, 0.62 bits of data can be embedded into each Byte of the PDF file. The injected leading zeros can be removed to restore the original PDF file.

POSTER 2.26: Multiple-Operation Image Anti-Forensics with WGAN-GP Framework

Jianyuan Wu, Zheng Wang, Hui Zeng and Xiangui Kang

Sun Yat-sen University, School of Computer Science & Tech., School of Data and Computer Science, Sun Yat-sen University

A challenging task in the field of multimedia security involves concealing or eliminating the traces left by a chain of multiple manipulating operations, i.e., multiple-operation anti-forensics in short. However, the existing anti-forensic works concentrate on one specific manipulation, referred as single-operation anti-forensics. In this work, we propose using the improved Wasserstein generative adversarial networks with gradient penalty (WGAN-GP) to model image anti-forensics as an image-to-image translation problem and obtain the optimized anti-forensic models of multiple-operation. The experimental results demonstrate that our multiple-operation anti-forensic scheme successfully deceives the state-of-the-art forensic algorithms without significantly degrading the quality of the image, and even enhancing quality in most cases. To our best knowledge, this is the first attempt to explore the problem of multiple-operation anti-forensics.

POSTER 2.27: Improving code-switching speech recognition with data augmentation and system combination

Duo Ma, Haihua Xu, Guanyu Li and Eng Siong Chng

Northwest Minzu University, Nanyang Technological University

This paper is focusing on a study of comprehensive approaches to an improved code-switching speech recognition, using data augmentation and system combination methods. For data augmentation, we not only use speech speed perturbation based method, we also attempt to add diversified room impulse response based reverberant noise, as well as music, babel , and white noise based additive noise. It is found we still can achieve significant performance improvement with such noise-corrupted data augmentation methods, though our SEAME code-switching data belongs to a clean corpus. In addition to data augmentation methods, we also adopt minimum Bayesian risk based lattice combination method to further improve our recognition results. We achieve significant word error rate (WER) reduction on lattice combination with/without recurrent neural network language model based lattice rescoring. Compared with our previous efforts [7], we achieve up to 2.29% and 5.61% absolute WER reduction on the two dev sets respectively, while 4.83% and 8.04% absolute WER reduction after system combination.

POSTER 2.28: A Multi-feature Fusion Based Method For Urban Sound Tagging

Jisheng Bai, Chen Chen and Jianfeng Chen

Northwestern Polytechnical University

Noise pollution is one of the serious issues for citizens. Mapping urban noise is essential to improve the quality of life for residents and construction for smart cities. Yet, most cities lack effective classification or tagging methods to monitor urban noise. To tackle this challenge, we propose a multi-feature fusion based method for urban sound tagging (UST). This method combines various features and Convolutional Neural Networks (CNNs) to predict whether noise of pollution is present in a 10-second recording. Log-Mel, harmonic, short-time Fourier transform (STFT) and Mel Frequency Cepstral Coefficents (MFCC) spectrograms are fed into different CNN architectures. And a fusion method is applied to make the final

outputs. The proposed method is evaluated on the DCASE2019 task5 dataset and achieves a macro-AUPRC score of 0.68, outperforming the baseline system of 0.54.

POSTER 2.29: Activation Driven Synchronized Joint Diagonalization for Underdetermined Sound Source Separation

Taiki Izumi, Yuuki Tachioka, Shingo Uenohara and Ken'ichi Furuya

Oita University, Denso IT Laboratory

Blind sound source separation (BSS) is effective to improve the performance of various applications such as speech recognition. The condition of BSS can be divided into underdetermined conditions (number of microphones < number of sound sources) and overdetermined conditions (number of microphones ⩾ number of sound sources). Here, we focus on Synchronized Joint Diagonalization (SJD), which is a newly proposed BSS method and utilizes non-stationarity of a sound source signal. The advantage of SJD is faster separation and smaller number of parameters to be estimated. However, the application of SJD is limited to overdetermined conditions, and the performance of SJD is degraded in underdetermined conditions. In this paper, to solve these performance degradations, we propose an activation driven SJD, which uses a pre-estimated activation matrix. It is practical because activation estimation is easier than source separation. The effectiveness of the proposed method was validated by conducting BSS experiments. We confirmed that the performance of SJD can be improved in underdetermined conditions.

POSTER 2.30: Deep Neural Networks with Batch Speaker Normalization for Intoxicated Speech Detection

Weiqing Wang, Haiwei Wu and Ming Li

Duke Kunshan University, Sun Yat–sen University

Alcohol intoxication can affect people both physically and psychologically, and one's speech will also become different at certain levels. However, detecting the intoxicated state from speech is a challenging task. In this paper, we first implement the baseline model with ComParE feature and then explore the influence of the speaker information on the intoxication detection task. In addition, we apply a ResNet18 based end-to-end model to this task. The model contains three parts: a representation learning sub-network with Deep Residual Neural Network (ResNet) of 18-layers, a global average pooling (GAP) layer and a classifier of 2 fully connected layers. Since we cannot perform speaker z-normalization on the variant-length feature input, we employ the batch z-normalization to train the proposed model. It also achieves similar improvement like applying the speaker normalization to the baseline method. Experimental results show that speaker normalization on the baseline model and batch z-normalization on the ResNet18 based model provides 4.9% and 3.8% unweighted accuracy improvement respectively. The results show that speaker normalization can improve the performance of both baseline model and proposed model.

POSTER 2.31: Subtraction-Positive Similarity Learning

Liang He, Xianhong Chen, Can Xu and Jia Liu

Tsinghua University

Many methods evaluate the similarity between two vectors xand y by norm or metric learning. They need to get a subtraction vector x-y and then evaluate its length. However, only considering the length of subtraction vector and ignoring its position may lost a lot of information. In this paper, we propose to utilize the position information of subtraction vector to evaluate the similarity. As the subtraction vector between x and y can be expressed either by x-y or by y-x, its distribution is centrosymmetric and redundancy. Thus, only half of the subtraction vectors are chosen and named as subtraction positive vectors. The subtraction positive vectors from different classes or from the same class are then modeled by Gaussian mixture models or deep neural network. Experiments were carried out on speaker verification databases including NIST SRE08, SRE10 and NIST i-vector challenge 2014. Results demonstrate the effectiveness of the proposed method.

POSTER 2.32: A Novel Effective Dimensionality Reduction Algorithm for Water Chiller Fault Data

Zhuozheng Wang, Yingjie Dong and Wei Liu

Beijing University of Technology

The reliability of chiller is very important for the safe operation of refrigeration system. In order to solve the problem that the traditional linear discriminant analysis (LDA) based on L_2 norm is sensitive to outliers, this paper introduced a novel dimensionality reduction algorithm for chiller fault data set –

RSLDA. Firstly, L_(2,1) norm is used to extract the most discriminant features adaptively and eliminate the redundant features instead of L_2 norm. Secondly, an orthogonal matrix and a sparse matrix are introduced to ensure the extracted features contain the main energy of the raw features. In addition, the recognition rate of the nearest classifier is defined as the performance criteria to evaluate the effectiveness of dimensionality reduction. Finally, the reliability of algorithm was verified by experiences compared with other algorithms. Experimental results revealed that RSLDA not only improves robustness but also has a good performance in SSS problem of fault classification.

POSTER 2.33: Speech Prosody and Eye Movements in Processing Discourse Information: A Preliminary Study in Mandarin Chinese

Ying Chen, Wentao Xiao, Jie Cui and Hanyu Xu

School of Foreign Studies, Nanjing University of Science and Technology

This study investigates variations in speech prosody and eye movements and their potential correlations in processing discourse information of map direction in Mandarin Chinese. A production experiment was conducted to collect mean duration, F0, intensity of target words in speech prosody and fixation counts and fixation duration of target areas of interest in eye movements for statistical analyses. The results show fixation counts, fixation duration, and syllable duration of the target words decreased, syllable intensity increased, but syllable pitch remained intact as the information became old to the speaker in the discourse. Prosodic reduction of duration, F0, and intensity was found in speech repetition and in the processing of old information.

POSTER 2.34: An RNN and CRNN Based Approach to Robust Voice Activity Detection

Guan—Bo Wang and Wei—Qiang Zhang

Tsinghua University

In this paper, we propose a voice activity detection (VAD) system, which combines a convolutional recurrent neural network (CRNN) and a recurrent neural network (RNN). In order to improve the performance of our system in low signal-noise ratio conditions, we also add a speech-enhancement module, a one-dimensional dilation-erosion module, and a model ensemble module, all of which contribute significantly. We evaluate our proposed system on development dataset of Public Safety Communications (PSC) and Video Annotation for Speech Technologies (VAST) from NIST Open Speech Analytic Technologies 2019 (OpenSAT19). Compared to the baseline system, our proposed system achieves better performance, using OpenSAT19 official evaluation metrics.

POSTER 2.35: Study of Chinese Text Steganography using Typos

Linna Zhou and Derui Liao

Beijing University of Posts and Telecommunications, University of International Relations

Nowadays, with the Information Explosion and the rapid development of information technology, huge amounts of data are constantly being generated every day on the Internet. But most of the texts provided online is of a kind that usually contain many typos, which is very common among individual users, self-media, etc. However, disambiguation is human's talent, so these typos often do not frustrate human understanding the text, and sometimes it is even difficult to recognize some typos. This phenomenon appears both in English and Chinese, so it seems to be cross-lingual. Therefore, in such texts, it is not surprising that one can perform information-hiding by judiciously injecting typos. We studied Chinese typos in the text contents on Weibo or WeChat, and propose a text steganography method based on Chinese typos with the help of NLP, which can embed secret information by carefully injected typos and guarantee the security of the secret and the readability of the texts. Unlike format-based steganography algorithms, our algorithm can resist format adjustments, OCR re-inputs, etc. Furthermore, Weibo and WeChat platform contain many kinds of media, so by combining other algorithms, Cross-Media or even Cross-Social Network information hiding is practical.

POSTER 2.36: A Simple Gaussian Kernel Classifier with Automated Hyperparameter Tuning

Kosuke Fukumori and Toshihisa Tanaka

Tokyo University of Agriculture and Technology

This paper establishes a fitting method for a kernel logistic regression model that uses generalized Gaussian kernel and its parameter optimization method. Kernel logistic regression is a classification model that uses kernel methods effectively. This is one of the methods to construct an effective nonlinear system with a reproducing kernel Hilbert space (RKHS) induced from positive semi-definite kernels. Most classifiers that are combined with Gaussian kernel functions generally assume uncorrelatedness within the feature vectors. Thus, the Gaussian kernel consists of only two parameters (namely, mean and precision). In this paper, we propose a model using a generalized Gaussian kernel represented flexibly in each dimension of feature vector. In addition, the parameters of kernel are fully datadriven. For the fitting of proposed model, an $\ell 1$-regularization is introduced to supress the number of support vectors. A numerical experiment showed that the classification performance of the proposed model is almost the same as RBF-SVM even though the proposed model has a small number of support vectors.

POSTER 2.37: Exploring RNN-Transducer for Chinese Speech Recognition

Senmao Wang, Pan Zhou, Wei Chen, Jia Jia and Lei Xie

NWPU, Tsinghua University, Sogou Inc.

End-to-end approaches have drawn much attention recently for significantly simplifying the construction of an automatic speech recognition (ASR) system. RNN transducer (RNN-T) is one of the popular end-to-end methods. Previous studies have shown that RNN-T is difficult to train and a very complex training process is needed for a reasonable performance. In this paper, we explore RNN-T for a Chinese large vocabulary continuous speech recognition (LVCSR) task and aim to simplify the training process while maintaining performance. First, a new strategy of learning rate decay is proposed to accelerate the model convergence. Second, we find that adding convolutional layers at the beginning of the network and using ordered data can discard the pre-training process of the encoder without loss of performance. Besides, we design experiments to find a balance among the usage of GPU memory, training circle and model performance. Finally, we achieve 16.9% character error rate (CER) on our test set, which is 2% absolute improvement from a strong BLSTM CE system with language model trained on the same text corpus.

POSTER 2.38: Linguistic Steganography by Sampling-based Language Generation

Rui Yang and Zhen—Hua Ling

University of Science and Technology of China

Linguistic steganography aims to hide secret messages within text carriers. In this paper, we propose a linguistic steganography method by means of sampling-based language generation. Comparing with deterministic text generation using beam-search, the sampling-based approach increases the redundancy of generated texts and benefits the hiding of information. The arithmetic coding (AC) algorithm is adopted to embed messages in our proposed method. Its performance is compared with fixed-length coding (FLC) and variable-length coding (VLC) which were designed for embedding messages during deterministic text generation. Besides, the KL divergence and temperature based strategies are designed to control the embedding rates of FLC, VLC and AC respectively. Experiments using a story generation model show that AC performed better than FLC and VLC when embedding messages during sampling-based text generation. With an embedding rate of 1.45 bits/word, our AC-based steganography method achieved ideal imperceptibility, and the subjective quality of its generated text is as good as the non-steganography one.

WED-PM2-SS1
Advanced Topics on High-dimensional Data Analytics and Processing

Time: Wednesday, Nov 20, 15:00-16:40

Place: A2

Chairs: Supavadee Aramvith, Shogo Muramatsu

WED-PM2-SS1.1: Long-term 3D Registration Method Based on LCT Tracking and Improved ORB Detection

Jiu Yong, Yangping Wang, Xiaomei Lei and Fang Yong
Lanzhou Jiaotong Univeristy, Gansu Meteorological Service, Sichuan University

Aiming at the complex environment such as fast moving of registered area, occlusion, illumination change and the high requirement of real-time and precision of feature detection in the 3D registration of augmented reality system, a long-term 3D registration method based on LCT tracking and improved ORB detection is proposed in this paper. Firstly, the reliability of LCT algorithm in long-term tracking is used to track the area to be registered in augmented reality; secondly, ORB algorithm with excellent real-time performance is improved by setting adaptive thresholds, number of feature points and distance thresholds to optimize the dense area of image feature points. Parallel algorithm is used to retain feature points with larger eigenvalues, and discrete difference feature is used to enhance illumination unevenness. The stability of ORB operator under uniform change can solve the problem of low precision of feature detection and poor anti-jamming ability. Finally, the 3D registration matrix is calculated by using the detected feature points to enhance the real world. The simulation results show that the LCT algorithm has high reliability in long-term tracking and registration. Compared with ORB algorithm, the improved ORB algorithm improves the precision of feature detection by about 22%. It effectively improves the real-time and precision of feature matching in augmented reality system. The performance of the long-term 3D registration method based on LCT tracking and improved ORB detection is excellent, which improves the robustness, stability and practicability of the augmented reality system.

WED-PM2-SS1.2: Restoration of Minute Light Emissions Observed by Streak Camera Based on N-CUP Method

Masahiro Tsumori, Shinichiro Nagai, Ryosuke Harakawa, Toru Sasaki and Masahiro Iwahashi
Nagaoka University of Technology

To observe high-speed phenomena such as discharge plasma, it is necessary to restore minute light emissions from an image observed by a streak camera, which includes multiple light emissions at each time. There has been proposed CUP method for restoring minute light emissions via a compressed sensing scheme; however, there is a case in which artefacts occur in the restoration results depending on initial values of the optimization for restoration. To overcome this limitation, N-CUP method that enables successful restoration of minute light emissions is proposed in this paper. N-CUP method estimates initial values suitable for the optimization by iteratively performing CUP method. Through simulation using image datasets emulating phenomena of fundamental light emissions, it was confirmed that N-CUP method obtained successful restoration results.

WED-PM2-SS1.3: A Prefatory Study on Data Channelling Mechanism towards Industry 4.0

Yiqi Tew, Kheng Hui Ng and Mum Wai Yip
Tunku Abdul Rahman University College

Data are increasing in volume, variety and velocity in this Internet of things and big data era. It applies from industry (or manufacturing) process monitoring control to video surveillance analysis to track human and machines activities. Therefore, fast and accurate approaches in data channelling are needed to effectively deal with these big data. This paper presents practical methods to manage and transfer the data from industry manufacturing site to a centralized data processing hub. In this hub, data are transformed into understandable information, which can assist human in understanding and monitoring manufacturing situation autonomously. These data are be collected and channelled to desired location to be analyzed through Open Platform Communication Unified Architecture (OPC UA). Industrial protocols and standards are used to interpret the data channelling methods and tested on several industrial machines. Result shows that size of data and number of OPC UA Client that connects to OPC Server affects the data channelling speed.

WED-PM2-SS1.4: Successive Stripe Artifact Removal Based on Robust PCA for Millimeter Wave Automotive Radar Image

Weiwei Shan, Shogo Muramatsu, Akira Oshima and Hiroyoshi Yamada

Niigata University

This study proposes a stripe artifact removal method based on robust principal component analysis (RPCA) for millimeter wave (MW) automotive radar images. With the development of MW radar detection technology, there is a demand for installing obstacle detectors on vehicular for safety. From this background, the authors developed squint-mode synthetic aperture radar (SAR) with MW (MW-SAR) as a high-resolution imaging technique. For synthesizing radar images, a back-projection algorithm (BPA) is adopted because of its real-time processing nature with high accuracy. However, SAR images obtained with Single Input and Single Output (SISO) systems are prone to be contaminated by a stripe-shaped artifact and can affect to the obstacle detection performance. Thus, to reduce the structured noise, this paper proposes successive RPCA on the assumption that the stripe artifacts and obstacle reflection are low-rank and sparse, respectively. As a solver, the alternating direction method of multipliers (ADMM) is adopted. The main contribution of this work is to initialize the ADMM state by taking account of the similarity of low-rank components between adjacent segments. Through simulations with experimental data, the significance of the proposed method is verified.

WED-PM2-SS1.5: Anomaly Event Detection Using Generative Adversarial Network for Surveillance Videos

Thittaporn Ganokratanaa, Supavadee Aramvith and Nicu Sebe

Chulalongkorn University, University of Trento

Anomalous event detection is advantageous for real-time video surveillance systems in terms of safety and security. Current works mostly run offline and struggle with abnormal event detection in crowded scenes. We propose unsupervised anomaly event detection using Generative Adversarial Network (GAN) with Optical Flow to obtain spatiotemporal features in appearance and motion representations. In training, GAN is used to train only the normal event images to generate their corresponding optical flow image. Hence, in testing, since the model knows only the normal patterns, any unknown events are considered as the anomaly event which can be detected by subtracting the pixels between the generated and the real optical flow images. We implement on the publicly available benchmark datasets and compare with state-of-the-art methods. Experiment results show that our model is effective for anomaly event detection in real-time surveillance videos.

WED-PM2-SS2
Lightweight Signal Processing and Machine Learning for Embedded Applications

Time: Wednesday, Nov 20, 15:00-16:40

Place: A3

Chairs: Hakaru Tamukoh, Hiroshi Tsutsui

WED-PM2-SS2.1: Semi-supervised Training of Acoustic Models Leveraging Knowledge Transferred from Out-of-Domain Data

Tien-Hong Lo and Berlin Chen

National Taiwan Normal University

More recently, a novel objective function of discriminative acoustic model training, namely lattice-free MMI (LF-MMI), has been proposed and achieved the new state-of-the-art in automatic speech recognition (ASR). Although LF-MMI shows excellent performance in a wide array of ASR tasks with supervised training settings, there is a dearth of work on investigating its effectiveness in the scenario of unsupervised or semi-supervised training. On the other hand, semi-supervised (or self-training) of acoustic model suffers from the problem that it is hard to estimate a good model when only a limited amount of correctly transcribed data is made available. It is also generally acknowledged that the performance of discriminative training is vulnerable to correctness of speech transcripts employed for training. In view of the above, this paper explores two novel extensions to LF-MMI. The first one is to distill knowledge (acoustic training statistics) from a large amount of out-of-domain data to better estimate the seed models for use in semi-supervised training. The second one is to make effective selection of the un-transcribed target domain data for semi-supervised training. A series of experiments conducted on the AMI benchmark corpus demonstrate the gains from these two extensions are pronounced and additive, which also reveals their effectiveness and viability.

WED-PM2-SS2.2: Training Data Reduction using Support Vectors for Neural Networks

Toranosuke Tanio, Kouya Takeda, Jeahoon Yu and Masanori Hashimoto

Osaka University

In the field of machine learning, deep learning is widely used to improve versatility and accuracy by deepening the network. Deep learning can achieve higher expression ability compared to conventional models but requires large amounts of data and time for training. To tackle this issue, we propose a training data reduction method using support vectors that are closest data to the classification boundary obtained by support vector machine. In this research, we use the training data consisting of support vectors to training neural networks and evaluate the effect. In the evaluation experiment, we confirmed that it is possible to reduce the number of training data by about 12% and reduce the learning time of neural network by about 9.5% by using ResNet, a model of deep learning, and the CIFAR-10 data set.

WED-PM2-SS2.3: Distilling Knowledge for Non-Neural Networks

Shota Fukui, Jaehoon Yu and Masanori Hashimoto

Osaka University

Deep neural networks (NNs) have shown high inference performance in the field of machine learning, but at the same time, researchers require their speeding-up and miniaturization methods due to the computational complexity. Distillation is drawing attention as one of the ways to overcome this problem. NNs usually have better expression power than its learning ability. Distillation bridges the gap between expressive power and learnability by training a small NN with additional information obtained from a larger already trained NN. This gap does not exist only in neural networks but also in other machine learning methods such as support vector machine, random forest, and gradient boosting decision tree. In this research, we propose a distillation method using information extracted from NNs for non-NN models. Experimental results show that distillation can improve the accuracies of other machine learning methods, and especially, the accuracy of SVM increases by 2.80%, 90.15% to 92.95%.

WED-PM2-SS2.4: Proposal of Minimization Problem Based Lightness Modification Method Considering Visual Characteristics of Protanopia and Deuteranopia

Meng Meng and Go Tanaka

Nagoya City University

Color is important for humans to receive visual information such as from traffic lights and maps. However sometimes important color-based information cannot be correctly received by some visually impaired people such as protanopes and deuteranopes. It is necessary to understand their difficulties in perceiving color and provide solutions to improve their quality of life. In our paper, a lightness modification method that can convey the color-based information to protanopes and deuteranopes is proposed. By considering color differences in an input image, the proposed method modifies the output lightness of an image for protanopes and deuteranopes to preserve its visual detail. The proposed method only changes the lightness without changing the hue, resulting in the output image having natural colors. In experiments, we use six color blindness test images with different color distributions and compare the proposed method with three existing methods. The experiment results and their quantitative evaluation show that our proposed method is reliable and effective.

WED-PM2-SS2.5: An Evaluation of Stack Light Indicator Color Detection System Using Web Cameras for Automatic Production Lines

Hiroshi Tsutsui, Kentaro Yamada, Akihiro Sudo and Yoshikazu Miyanaga

Hokkaido University, DENSO Hokkaido Corporation

In production lines, manufacturing devices are operated in a pipelined manner to produce desired products. The pipeline stops when an error occurs at any manufacturing device. In order to prevent a decrease in production efficiency, it is necessary to detect the abnormality earlier. So, in this paper, it aims at early detection and record of the abnormal occurrence by judging the color of the stack light indicator informing the abnormality installed in each device. We utilize web cameras for this purpose, considering its versatility. In this paper, we show an evaluation result of a system for detecting stack light indicator color using web cameras.

WED-PM2-SS3
High Performance Video Processing and Image Identification

Time: Wednesday, Nov 20, 15:00-16:40

Place: A4

Chair: Junyong Deng

WED-PM2-SS3.1: Multi-Task and Multi-Level Detection Neural Network Based Real-Time 3D Pose Estimation

Dingli Luo, Songlin Du and Takeshi Ikenaga
Waseda University & University of Electronic Science and Technology of China, Waseda University

3D pose estimation is a core step for human-computer interaction and human action recognition. However, time-sensitive applications like virtual reality also need this task to achieve real-time speed. This paper proposes a multi-task and multi-level neural network architecture with a high-speed friendly 3D human pose representation. Based on this, we build a real-time multi-person 3D pose estimation system with a single RGB image as input. The network estimates 3D poses from the input image directly by the multi-task design and keeps both accuracy and speed by the multi-level detection design. By evaluation, we show our system achieves the 21 fps on RTX 2080 with only 33 mm accuracy lose compared with related works. We also provide network visualization to prove our network work as we design. This work shows the possibility for a single RGB image based 3D pose estimation system to achieve real-time speed, which is a basement for building a low-cost 3D motion capture system.

WED-PM2-SS3.2: A Fast Inter-view Mode Selection Algorithm Based on Video Array Processor

Yu Wang, Xueting Li, Yun Zhu and Feilong He
Xi'an University of Posts and Telecommunications

In the encoding depth map of 3D-HEVC, a fast block search algorithm is proposed to solve the problem of fixed block pattern of inter-view prediction algorithm. Variable block sizes mode is used to select prediction mode by comparing depth values, and then parallel mapping is carried out based on video array processor. Compared with HTM16.2 software implementation, the average encoding time of the proposed algorithm is reduced by more than 100 times. Resource usage fell by 27.21 percent when hardware processing speeds was generally consistent.

WED-PM2-SS3.3: NPFONoC: A Low-loss, Non-blocking, Scalable Passive Optical Interconnect Network-on-Chip Architecture

Junyong Deng, Haoyue Wu, Rui Shan, Yiwen Fu, Xinchuang Liu and Ping Wang
Xi'an University of Posts & Telecommunications

With the increase of inter-core communication requirements for large-scale processors, optical interconnection on-chip is an important means of multi-core processor communication. At present, high-blocking, large delay, and high insertion loss is the bottleneck of large-scale processor inter-core communication. This paper proposes a non-blocking, low-loss, scalable passive optical interconnection network-on-chip structure (NPFONoC). In this structure, the 2*2 optical switch unit network-on-chip designed by wavelength division multiplexing technology and passive optical interconnect micro-ring resonator self-resonance characteristics, it is easily expanded into 16*16, 32*32, 64*64 optical networks structure and achieve non-blocking communication simultaneously. The number of waveguides and micro-ring resonators in the optical interconnection on-chip structure are important parameters affecting the insertion loss of the network structure. In the 16*16 optical interconnection network structure, NPFONoC has great advantage in the number of micro-rings compared with the λ-route, GWOR, Crossbar and new topology structure, with reduction rate of 90.9%, 90.9%, 75%, and 20% respectively. By detecting the performance parameters of the 8*8 optical interconnection network structure on the OMNET++ platform, the results show that the average insertion loss of NPFONoC is smaller than λ-route, GWOR, Crossbar and Mesh structures by 11.6%, 3%, 16.7%, 4.8%.

WED-PM2-SS3.4: Parallelization Design of Motion Compensation Algorithm Based on Reconfigurable Video Array Processor

Xiaoyan Xie, Xiang Lei, Jinna Zhou, Yun Zhu and Lin Jiang
Xi'an University of Posts & Telecommunications

The motion compensation algorithm in High Efficiency Video Coding (HEVC) has a large number of interpolation calculations at the same time, and it is difficult to achieve flexible switching of different coding blocks, which puts higher requirements on its computational efficiency and control logic. In order to solve such problems, the data is divided according to the characteristics of the algorithm, and the motion compensation algorithm is mapped onto the reconfigurable array structure, so that the previous serial algorithm can be processed in parallel. According to the data overlapping relationship between the next reference block and the current reference block of the encoding process, the data multiplex idea is used to reduce the number of reading reference pixels from the external storage, thereby shortening the reading time of the next reference block data. At the same time, according to the reconfigurable structural features, flexible switching of the algorithm variable block mode is designed to improve flexibility. Finally, parallel processing is performed according to the data rule of motion compensation algorithm and a large number of interpolation characteristics, which improves the computational efficiency of the algorithm. In this paper, a 16×16 processing unit (PE) is used to dynamically process a 4×4-64×64 block size. On the Virtex-6 FPGA attached to the BEE4cube, the reference block update speed is increased by 39.9%; in the case of an array size of 16 PEs. In parallel, the degree of parallelism can reach 16, which has better flexibility while achieving higher execution efficiency.

WED-PM2-SS3.5: SCRA: A Hybrid Deterministic Routing Algorithm for Aging-Resilient Network-on-Chip

Bowen Zhang, Huaxi Gu and Ruiqi Guo

Xidian University

Network-on-Chip (NoC) has been proposed as a promising interconnection candidate solution for its high network bandwidth, low communication energy consumption and good parallel transmission capability. However, future many-cores processor will face aging problems such as negative bias temperature instability (NBTI), hot-carrier injection (HCI) and electro-migration (EM). These aging problems will cause switching delay and critical path depravation under imbalanced loads, which leads to bad system reliability. In this paper, a deterministic aging-resilient hybrid routing algorithm called SCRA (source-based configuration router algorithm) is proposed to evenly distribute packet flow over entire network and relieve the aging problems in NoC. In SCRA, a flow distribution model is used to achieve the best uniformity of network communications by combing the complementary characteristics of XY and YX routing algorithm. With the simulation and analysis results, SCRA can realize better uniformity and incremental longevity on the premise of ensuring accessibility and achieves acceptable network communication performance when compared with the single dimensional order routing algorithm.

WED-PM2-SS4
Physical and Wireless Environment Recognition Based on Signal Processing

Time: Wednesday, Nov 20, 15:00-16:40

Place: A5

Chair: Osamu Takyu

WED-PM2-SS4.1: Access Decision based on Secure Capacity for prevention to CSI Impersonation of Untrusted Relay

Ryota Sugimoto and Osamu Takyu

Shinshu University

A physical layer network coding is a highly efficient data exchanging scheme between two nodes through a relay. However, if the untrusted relay is assumed, it impersonates the channel state information (CSI) for exploiting the data through relay. This paper pays attention to the protocol of CSI impersonation and proposes the wireless access decision based on secure capacity. The computer simulation shows the proposed access decision suppresses the exploitation of data through relay as well as increases the secure capacity against the untrusted relay

WED-PM2-SS4.2: Recognition and countermeasure to Hidden terminal problem by packet analysis in wireless LAN

Akinori Kamio, Osamu Takyu, Mai Ohta, Takeo Fujii, Fumihito Sasamori and Shiro Handa

Shinshu University, Fukuoka University, The University of Electro-Communications,

A carrier sense multiple access with collision avoidance (CSMA/CA) is an access protocol of wireless LAN (WLAN). Since the carrier sensing is missed, the WLAN systems simultaneously access the channel and then packet collision occurs. It is a hidden terminal problem. In this paper, the recognition of the hidden terminal problem is performed by packet analysis. For suppressing the packet collision under the hidden terminal problem, the proposed countermeasure uses the modulation and coding set with high order modulation and low coding rate. Since the length of the packet is shorten, the probability of packet collision is reduced and thus the throughput and delay performances are improved

WED-PM2-SS4.3: Machine Learning-Aided Indoor Positioning Based on Unified Fingerprints of Wi-Fi and BLE

Shunsuke Tsuchida, Takumi Takahashi, Shinsuke Ibi and Seiichi Sampei

Osaka University, Doshisha University

This paper deals with an indoor positioning with the aid of machine learning based on the received power strength indication (RSSI) fingerprints of beacon signals of both Wi-Fi and Bluetooth low energy (BLE). In fingerprint positioning, a site-survey is conducted in advance to build the radio map which can be used to match radio signatures with specific locations, thus, it can take the impacts of empirical indoor environments into consideration. However, even if the physical positional relationship in the indoor environment is static, the observed RSSI values are dynamically fluctuated according to the probabilistic wireless channels. Unfortunately, it is difficult to analytically capture the stochastic behavior of RSSI in real-environments, and the accuracy of position estimation is degraded due to the model errors. To tackle this challenging problem, machine learning-based logistic regression is applied to fingerprint positioning with the RSSI data set (available as big data). Additionally, by exploiting an unified fingerprint generated from both Wi-Fi and BLE beacon signals, further performance improvement in the estimation accuracy is possible, owing to the transmit diversity effects. The experimental results show the validity of proposed positioning scheme with the unified Wi-Fi and BLE fingerprint.

WED-PM2-SS4.4: An Overloaded SC-CP IoT Signal Detection Method via Sparse Complex Discrete-Valued Vector Reconstruction

Kazunori Hayashi, Ayano Nakai-Kasai and Ryo Hayakawa

Osaka City University, Kyoto University

The paper proposes low-latency signal detection methods for overloaded MU-MIMO (Multi-User Multi-Input Multi-Output) SC-CP (Single Carrier block transmission with Cyclic Prefix) using convex optimization approach for uplink IoT (Internet of Things) environments where a lot of IoT terminals are served by a base station having less number of antennas than that of IoT terminals. The proposed method detects overloaded IoT signals via convex optimization approach named sum of complex sparse regularizers (SCSR) taking advantage of both the discreteness and the sparsity of the SC-CP IoT signal. Simulation results demonstrate the validity of the proposed method.

WED-PM2-SS4.5: NOMA Based UAV Relay Communication Protocol in Cellular Network

Jumpei Kawakami, Hendrik Lumbantoruan and Koichi Adachi

The University of Electro-Communications

Introducing unmanned aerial vehicles (UAVs) into wireless communication systems has recently gained a lot of attention. UAV provides many advantages such as shorter communication distance and a higher probability of having line-ofsight (LoS) condition due to its dynamic positioning. In this paper, UAV is deployed as a relay station because it can give superior performance due to the high probability of having LoS channel compared to a fixed ground relay station. However, the achievable throughput is decreased as relay communication requires twice of time resources in direct communication. To tackle this problem, non-orthogonal multiple access (NOMA) based communication protocol is proposed in this paper. Furthermore, as a specific problem for UAV, interference from neighbouring base stations (BSs) is large due to LoS channel between UAV and neighbouring BSs. In order to eliminate the interference from neighbouring BSs, UAV relay is equipped with array antenna. The simulation results elucidate that the throughput improvement of the proposed protocol over the conventional protocol.

WED-PM2-SS5
Robust Rich Audio Analysis

Time: Wednesday, Nov 20, 15:00-16:40

Place: A6

Chair: Xiao-Lei Zhang

WED-PM2-SS5.1: Improving the Spectra Recovering of Bone-Conducted Speech via Structural SIMilarity Loss Function

Zheng Changyan, Jibin Yang, Xiongwei Zhang, Meng Sun and Kun Yao

Army Engineering University

Bone-conducted (BC) speech is immune to background noise, but suffers from low speech quality due to the severe loss of high-frequency components. The key of BC speech enhancement is to restore the missing parts in the spectra. However, even with advanced deep neural networks (DNN), some of the recovered components still lack expected spectro-temproal structures. Mean Square Error loss function (MSE) is the typical choice for supervised DNN training, but it can only measure the distance of the spectro-temporal points and and is not able to evaluate the similarity of structures. In this paper, Structural SIMilarity loss function (SSIM) originated from image quality assessment is proposed to train the spectral mapping model in BC speech enhancement, and to our best knowledge, it is the first time that SSIM is deployed in DNN-based speech signal processing tasks. Experimental results show that compared with MSE, SSIM can acquire better objective results and obtain spectra with spectro-temporal structures more similar with the target one. Some adjustments of hyper-parameters in SSIM are made due to the difference between natural image and magnitude spectrogram, and the optimal choice of them are suggested. In addition, the effects of three components in SSIM are analyzed individually, aiming to help further study on the applications of this loss function in other speech signal processing tasks.

WED-PM2-SS5.2: Dilated-Gated Convolutional Neural Network with A New Loss Function on Sound Event Detection

Ke-Xin He, Wei-Qiang Zhang, Jia Liu and Yao Liu

Tsinghua University, China General Technology Research Institute

In this paper, we propose a new method for rare sound event detection. Compared with conventional Convolu-tional Recurrent Neural Network (CRNN), we devise a Dilated-Gated Convolutional Neural Network (DGCNN) to improve the detection accuracy as well as computational efficiency.Furthermore, we propose a new loss function. Since frame-level predictions will be post processed to get final prediction, continuous false alarm frames will lead to more insertion errors than single false alarm frame. So we adopt a discriminative penalty term to the loss function to reduce insertion errors. Our method is tested on the dataset of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge task 2. Our model can achieve a F-score of 91.3% and error rate of 0.16 on the evaluation dataset while baseline achieves a F-score of 87.5% and error rate of 0.23.

WED-PM2-SS5.3: Augmented Strategy For Polyphonic Sound Event Detection

Bolun Wang, Zhong-Hua Fu and Hao Wu

School of Computer Science, Northwestern Polytechnical University, Xi'an IFLYTEK Hyper Brain Information Technology Co.

Sound event detection is an important issue for many applications like audio content retrieval, intelligent monitoring, and scene-based interaction. The traditional studies on this topic are mainly focusing on identification of single sound event class. However, in real applications, several sound events usually happen concurrently and with different durations. That leads to a new detection task on polyphonic sound event classification along with event time boundaries. In this paper, we propose an augmented strategy for this task, which faces challenges of a large amount of unbalanced and weakly labelled training data. Specifically, the strategy includes data augmentation to enrich training set to eliminate data unbalance, a new loss function that combines cross entropy and F-score, and model fusion to integrate the powers of different classifiers. The performance of the strategy is validated on DCASE2019 dataset, and both the event and segment detections are significantly improved over the baseline system.

WED-PM2-SS5.4: Domain Adaptation Neural Network for Acoustic Scene Classification in Mismatched Conditions

Rui Wang, Mou Wang, Xiao-Lei Zhang and Susanto Rahardja

Northwestern Polytechnical University

Acoustic scene classification is a task of predicting the acoustic environment of an audio recording. Because the training and test conditions in most real world acoustic scene classification problems do not match, it is strongly needed to develop domain adaptation methods to solve the cross-domain problem. In this paper, we propose a domain adaptation neural network (DANN) based acoustic scene classification system. Specifically, we first extract two kinds of acoustic features, i.e. mel-spectrogram and hybrid constant-Q transform, which have been proven to be effective in previous studies. Then, we train a DANN to project the training and test domains into one common space where the acoustic scenes are categorized jointly. To boost the overall performance of the proposed method, we further train an ensemble of convolutional neural network (CNN) models with different parameter settings on different acoustic features respectively. Finally, we fuse the DANN and CNN models by averaging the output of the models. We have evaluated the proposed system on the subtask B of the DCASE 2019 ASC challenge, which is a closed-set classification problem whose audio recordings were recorded by mismatched devices. Experimental results demonstrate the effectiveness of the proposed system on the acoustic scene classification problem in mismatched conditions.

WED-PM2-SS5.5: Boosting Spatial Information for Deep Learning Based Multichannel Speaker-Independent Speech Separation In Reverberant Environments

Ziye Yang and Xiao-Lei Zhang

Northwestern Polytechnical University

Recently, supervised speaker-independent speech separation methods, such as deep clustering and permutation invariant training, have demonstrated better performance than conventional unsupervised speech separation methods. However, their performance drops sharply in reverberant environments. To solve the problem, we propose a multi-channel speech separation algorithm that fully explores spatial information. It first extracts a spatial feature, named interaural phase difference (IPD), as one of the input features of the single-channel deep clustering algorithm. Then, it uses the deep clustering as the noise estimation component of the deep-learning-based beamforming. The novelty of the proposed algorithm lies in that it extends the spatial-feature-based deep clustering to a multichannel algorithm which boosts the performance by exploring spatial information at both the input and output of deep clustering. Its advantages have two aspects. First, the spatial feature IPD significantly improves the robustness of deep clustering in reverberant environments. Second, the deep-clusteing-based beamforming, which is a linear algorithm, suffers less nonlinear distortions than the single-channel deep clustering. We have compared the proposed algorithm with the single-channel deep clustering algorithm, spatial-feature-based multi-channel deep clustering with IPD, and deep-clustering-based beamforming without IPD in reverberant environments. Experimental results show that the proposed algorithm performs significantly better than the comparison methods.

WED-PM2-SS5.6: Hybrid Constant-Q Transform Based CNN Ensemble for Acoustic Scene Classification

Mou Wang, Rui Wang, Xiao-Lei Zhang and Susanto Rahardja

Northwestern Polytechnical University

Acoustic scene classification (ASC) has attracted much attention in recent years. In this paper, we present a hybrid constant-Q transform (HCQT) based convolutional neural network (CNN) system for ASC. Specifically, we first extract Mel-spectrogram, its harmonic-percussive source separation, and HCQT from each recording as the acoustic features. Then, we feed the three features into three CNNs respectively. Finally, we develop several methods to integrate the outputs of the CNNs, including averaging, weighted averaging, and random forests. The novelty of the proposed system lies in the following two respects. First, we introduce HCQT to the study of ASC for the first time, to our knowledge. HCQT combines two CQTs with different spectral resolutions together for remedying the loss of the high-frequency bins of traditional CQT. Second, we investigate different fusion strategies of the CNN models thoroughly.

We evaluated the proposed system in the DCASE 2019 challenge. Experimental results show that HCQT is more effective than the conventional CQT. Furthermore, the accuracies of our system on the validation and leaderboard datasets are 77.3% and 79% respectively, which outperforms the two comparison baseline systems significantly.

WED-PM2-SS5.7: Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection

Jiakang Li, Meng Sun and Xiongwei Zhang

Army Engineering University

With the development of spoofing technologies, automatic speaker verification (ASV) systems have encountered serious challenges on security. In order to address this problem, many anti-spoofing countermeasures have been explored. There are two intuitive recipes to protect an ASV system from spoofing. The first one is to use a cascaded structure where spoofing detection is performed firstly and ASV is subsequently conducted only on the attempts which have passed the spoofing detection. The other one is to perform spoofing detection and ASV jointly. The discriminate reliably of the joint system has been proven to be more advantageous than cascaded systems with traditional methods, not only in accuracy, but also in convenience and computational efficiency. In this paper, we proposed a multi-task learning approach based on deep neural network to make a joint system of ASV and anti-spoofing. The performance of different acoustic features and structures of deep neural networks has been investigated on the ASVspoof 2017 version 2.0 dataset. The experimental results showed that the joint equal error rate (EER) of our approach was reduced by 0.55% compared to a joint system with Gaussian back-end fusion baseline.

WED-PM2-O1
Signal Processing Methods

Time: Wednesday, Nov 20, 15:00-16:40

Place: A7

Chair: Mingyi He

WED-PM2-O1.1: Frequency domain variant of Velvet noise and its application to acoustic measurements

Hideki Kawahara, Ken–Ichi Sakakibara, Mitsunori Mizumachi, Hideki Banno, Masanori Morise and Toshio Irino

Health Science University of Hokkaido, Kyushu Institute of Technology, Meijo Universitty, Wakayama University

We propose a new family of test signals for acoustic measurements such as impulse response, nonlinearity, and the effects of background noise. The proposed family complements difficulties in existing families, the Swept-Sine (SS), pseudo-random noise such as the maximum length sequence (MLS). The proposed family uses the frequency domain variant of the Velvet noise (FVN) as its building block. An FVN is an impulse response of an all-pass filter and yields the unit impulse when convolved with the time-reversed version of itself. In this respect, FVN is a member of the time-stretched pulse (TSP) in the broadest sense. The high degree of freedom in designing an FVN opens a vast range of applications in acoustic measurement. We introduce the following applications and their specific procedures, among other possibilities. They are as follows. a) Spectrum shaping adaptive to background noise. b) Simultaneous measurement of impulse responses of multiple acoustic paths. d) Simultaneous measurement of linear and nonlinear components of an acoustic path. e) Automatic procedure for time axis alignment of the source and the receiver when they are using independent clocks in acoustic impulse response measurement. We implemented a reference measurement tool equipped with all these procedures. The MATLAB source code and related materials are open-sourced and placed in a GitHub repository.

WED-PM2-O1.2: Fast & Efficient Delay Estimation Using Local All-Pass & Kalman Filters

Beth Jelfs and Christopher Gilliam

RMIT University

Delay estimation is a common problem in signal processing which can be particularly challenging when the delay is time-varying and the recorded signals are non-stationary. In this paper we present a method for time-varying delay (TVD) estimation which is suitable for real-time non-stationary applications. The proposed method combines local all-pass filters (LAP) with a Kalman filter. By using measurement fusion to combine the outputs of several LAP filters in the Kalman filter we can accurately track time-varying delays whilst allowing for for fast and efficient parallel computation. Illustrative simulations demonstrate the effectiveness of the proposed approach.

WED-PM2-O1.3: Speech Demodulation-based Techniques for Replay and Presentation Attack Detection

Madhu Kamble, Aditya Krishna Sai Pulikonda, Maddala Venkata Siva Krishna, Ankur Patil, Rajul Acharya and Hemant Patil

DA–IICT, IIIT Vadodara, DA–IICT Gandhinagar

Automatic Speaker Verification (ASV) system is vulnerable to various kinds of spoofing attacks. The combination of different speech demodulation techniques, Hilbert Transform(HT), Energy Separation Algorithm (ESA), and its Variable length version Variable ESA (VESA) is investigated for replay Spoof Speech Detection (SSD) task. In particular, the feature sets are developed using Instantaneous Amplitude and Instantaneous Frequency (IA-IF) components of narrowband filtered speech signals obtained from linearly-spaced Gabor filterbank. We observed relative effectiveness of these demodulation techniques on two spoof speech databases, i.e., BTAS 2016 and ASVspoof 2017 version 2.0 challenge database that focus on the presentation and replay attacks, respectively. From different demodulation techniques the results obtained are comparable for both the databases. For VESA demodulation technique, we found that with dependency index (DI) = 2 gave relatively better performance compared to the other DIs on both the databases for SSD task. All the demodulation technique-based feature sets gave lower Equal Error Rate (EER) than their baseline system for both the databases.

WED-PM2-O1.4: Spectrum Sensing Algorithm Based on LSTM and Its Implementation of Multiple USRP

Huachao Lu and Zhijin Zhao

Hangzhou Dianzi University

Aiming at the problem that the fusion rules of cooperative spectrum sensing have great impact on performance, a cooperative sensing algorithm based on LSTM, which is implemented on multiple USRPs is proposed. The received signal has different sequence characteristics when the primary user signal is present or absent. LSTM is used to extract the temporal characteristics of each primary user' s signal sequence, and the fully connected layer is used to fuse the features in the fusion center, then softmax is used to classify fusion features. A number of USRPs and a host are built a spectrum sensing system, and the LSTM model obtained by offline training is used to perform online real-time detection. The system can effectively detect the primary user signal.

WED-PM2-O1.5: Quality of Experience using Deep Convolutional Neural Networks and future trends

Woojae Kim, Jaekyung Kim and Sanghoon Lee

Yonsei University

The development of immersive display technology enables to represent the details of contents more naturally by providing a more realistic viewing environment while increasing immersion. In parallel, quality of experience (QoE) has been dealt with and discussed from both academy and industry to grade consumer products from the quality perspective. However, for quantification of QoE, it is very challengeable to analyze the human perception more accurately, even if it has been studied in many decades. Currently, there is no solid methodology to verify human perception as a closed-form objectively due to the limitation of human perception analysis. Recently, the deep convolutional neural network (CNN) has emerged as a core technology while breaking most performance records in the area of artificial intelligence via intensive training in accordance with the massive dataset. The main motivation of this paper lies in finding new insight into human perception analysis for QoE evaluation through visualization of intermediate node values. This new QoE assessment approach enables us to figure out the human visual sensitivity without using any prior knowledge. Toward the end, we provide a novel clue of how to obtain visual sensitivity, which is expected to be essentially applied for future QoE applications. In addition, we discuss future applications in QoE assessment with respect to the display types.

WED-PM2-O1.6: Dynamic Adjustment of Railway Emergency Plan Based on Utility Risk Entropy

Qian Ren and Zhenhai Zhang
Lanzhou Jiaotong University

Raising the speed of high-speed railway train provides great convenience for people to travel, but once an emergency occurs, the consequences are incalculable. Because of the uncertainties in the development process of railway emergencies, emergency decision-making often needs to be adjusted according to the changes of the state of the incident, that is, dynamic adjustment. Aiming at the dynamic adjustment of railway emergency plan, the emergency decision-making process is divided into several stages according to the key nodes. At each stage, the perceived utility value under the combination of the corresponding scheme and the scenario is obtained, and the utility risk function is derived by combining the utility value with its occurrence probability. Considering the utility risk under different situations of the same scheme, the utility risk entropy of the emergency response scheme is obtained, and the best scheme at the current moment is selected. Finally, an example is given to verify the effectiveness of the proposed method.

WED-PM2-O2
Image Processing

Time: Wednesday, Nov 20, 15:00-16:40

Place: A8

Chair: Zhonghua Sun

WED-PM2-O2.1: Stereo Matching and Image Inpainting Based on Binocular Camera

Yibo Du, Kebin Jia and Chang Liu

Beijing University of Technology

Abstract—Stereo matching is one of the key technologies in the field of computer vision. The depth map obtained by stereo matching contains the three-dimensional information of the scene. The use of depth map is of great significance in the three-dimensional reconstruction of the map and the autonomous navigation of the robot. Aiming at the accuracy and speed of stereo matching, this paper applies a semi-global stereo matching method to match corrected left and right perspective images. Because there are noise points and holes in the matched disparity map, which affect the image quality, a sample block filling method which combines mean filtering and point-by-point scanning is proposed to repair the image. Then a gradient priority selection mechanism is proposed to maintain the edge structure of the object in the process of restoration. Experimental results show that the proposed method is good for the restoration of holes and noises in disparity maps, and the processing speed is improved by about 30% compared with the traditional Criminisi algorithm.

WED-PM2-O2.2: Optimization-Based Fundus Image Decomposition for Diagnosis Support of Diabetic Retinopathy

Daichi Kitahara, Swathi Ananda and Akira Hirabayashi

Ritsumeikan University, NMAM Institute of Technology

Diabetes mellitus often leads to a serious eye disease called diabetic retinopathy, which is one major cause of blindness among adults. Since this blindness can be prevented if the diabetic retinopathy is detected at an early stage and appropriate medical treatment is provided, routine screening tests with fundus images are very important. However, as the number of diabetic patients increases, the routine screening tests are becoming big burdens for ophthalmologists. To reduce these burdens, in this paper, we propose a diagnosis support method by using convex optimization. The proposed method decomposes a green channel fundus image into a basic image composed of non-disease parts, a positive image including exudates, and a negative image including hemorrhages. Numerical experiments show the effectiveness of our method.

WED-PM2-O2.3: An Integrated CNN-based Post Processing Filter For Intra Frame in Versatile Video Coding

Ming Ze Wang, Shuai Wan, Hao Gong, Yuanfang Yu and Yang Liu

Northwestern Polytechnical University, Guangdong OPPO Mobile Telecommunications Corp.

Versatile Video Coding (H.266/VVC) standard achieves up to 30% bit-rate reduction while keeping the same quality compared with H.265/HEVC. To eliminate various coding artifacts like blocking, blurring, ringing, and contouring effects, etc., three in-loop filters have been incorporated in H.266/VVC. Recently, convolutional neural network (CNN) has attracted tremendous attention and achieved great success in many image processing tasks. In this paper, we focus on CNN-based filtering in video coding, where a single model solution for post-loop filtering is designed to replace the current in-loop filters. An architecture is proposed to reduce the artifacts of video intra frames, which take advantage of useful information such as partitioning modes and quantization parameters (QP). Different from existing CNN-based approaches, which generally need to train different models for different QP and only suitable for luma component, the proposed filter can well adapt to different QP, i.e. various levels of degradation of frames, and all components (i.e., luma and chroma) are jointly processed. Experiment results show that the proposed CNN post-loop filter not only can replace the de-blocking filter (DBF), sample adaptive offset (SAO) and adaptive loop filter (ALF) in H.266/VVC, and also outperforms them, leading to 6.46%, 10.40%, 12.79% BD-rate savings for Y, Cb and Cr, respectively, under all intra configuration.

WED-PM2-O2.4: Parameter-free Image Segmentation Based on Extreme Learning Machine

Hongwei Zhang, Liuai Wu and Yanchun Yang

Lanzhou jiaotong University

For the problem of spending much time on adapting parameters, a parameter-free image segmentation method based on extreme learning machine (ELM) is proposed. Firstly, each image is segmented as superpixels by simple linear iterative clustering (SLIC) with different parameters. Secondly, each superpixel segmentation result is combined with some rules, and initial segmentation results are obtained. Each initial segmentation result is evaluated, and the parameter with the best performance is selected as its class. Thirdly, in order to construct the training sets of ELM, the cooccurrence of each image is constructed, and some of its attributes are calculated as its features, and a parameter-free framework is learned by ELM. The experimental results show that the proposed method in this paper gets better segmentation results, which is closer to human annotation than other methods.

WED-PM2-O2.5: Automatic Fundus Image Segmentation for Diabetic Retinopathy Diagnosis by Multiple Modified U-Nets and SegNets

Swathi Ananda, Daichi Kitahara, Akira Hirabayashi and K. R. Udaya Kumar Reddy

NMAM Institute of Technology, Ritsumeikan University

Diabetes mellitus leads to damage of the retina by a high blood sugar level. This disease is called diabetic retinopathy (DR), and it is one major cause of blindness among working-aged people. DR affects about 80% of patients who have had diabetes for twenty years or more. The longer a period of diabetes is, the higher the risk of developing DR is. In order to prevent the blindness caused by DR, accurate DR diagnosis from a retinal fundus image is important. Recently, deep learning techniques play a significant role in the field of computer vision. When we apply deep learning to segmentation of damaged parts in fundus images, two major problems arise. One is that the number of available data is insufficient to train a deep neural network. The other is that the sizes of the damaged parts are quite different depending on the type of the damage, which leads to low segmentation accuracy for small damages. These two problems make the fundus image segmentation challenging. In this paper, we propose a segmentation method using multiple deep neural networks. To train the deep neural networks from a small number of data, we use data augmentation as a preprocessing and adopt the Dice coefficient with binary cross entropy as a loss function. Moreover, to improve the segmentation accuracy for small damages, e.g., microaneurysms, we construct one individual network for each type of the damage. In experiments, the networks are trained from IDRiD dataset and tested for MESSIDOR dataset. We compare and discuss the accuracy of the proposed method with modified U-Nets and SegNets.

WED-PM2-O2.6: Kernel Prediction Network for Detail-Preserving High Dynamic Range Imaging

Haesoo Chung, Yoonsik Kim, Junho Jo, Sang-Hoon Lee and Nam Ik Cho

Seoul National University

Generating a high dynamic range (HDR) image from multiple exposure images is challenging in the presence of significant motions, which usually causes ghost artifacts. To alleviate this problem, previous methods explicitly align the input images before merging the controlled exposure images. Although recent works try to learn the HDR imaging process using a convolutional neural network (CNN), they still suffer from ghosting or blurring artifacts and missing details in extremely under/overexposed areas. In this paper, we propose an end-to-end framework for detail-preserving HDR imaging of dynamic scenes. Our method employs a kernel prediction network and produces per-pixel kernels to fully utilize every pixel and its neighborhood in input images for the successful alignment. After applying the kernels to the input images, we generate a final HDR image using a simple merging network. The proposed framework is an end-to-end trainable method without any preprocessing, which not only avoids ghosting or blurring artifacts but also hallucinates fine details effectively. We demonstrate that our method provides comparable results to the state-of-the-art methods regarding qualitative and quantitative evaluations.

WED-PM3-SS1
Recent Advances in Fingerprinting and Data Hiding

Time: Wednesday, Nov 20, 17:00-18:40

Place: A1

Chairs: Minoru Kuribayashi, David Megías Jiménez

WED-PM3-SS1.1: Efficient Decentralized Tracing Protocol for Fingerprinting System with Index Table

Minoru Kuribayashi and Nobuo Funabiki

Okayama University

Due to the burden at a trusted center, a decentralized fingerprinting system has been proposed by delegating authority to an authorized server so that the center does not participate in the tracing protocol. As a fingerprinting code is used to retain a collusion resistance, the calculation of correlation score for each user is required to identify illegal users from a piratec copy. Considering the secrecy of code parameters, the computation must be executed by a seller in an encrypted domain to realize the decentralized tracing protocol. It requires much computational costs as well as the communication costs between the center and a seller because encrypted database (DB) is necessary for the computation. In this paper, we propose a method to reduce such costs by using the ElGamal cryptosystem over elliptic curve instead of the Paillier cryptosystem used in the conventional scheme. Our experimental results indicate that the time consumption becomes almost 100 times shorter and the size of encrypted DB reduced by a factor of 7/32 under 112-bit security level. The encrypted DB is further compressed by introducing an index table

WED-PM3-SS1.2: Hand Gesture Recognition with Ensemble Time-Frequency Signatures Using Enhanced Deep Convolutional Neural Network

Xiang Feng, Qun Song, Qingfang Guo, Duo Liu, Zhanfeng Zhao and Yinan Zhao
Weifang Medical University, Harbin Institute of Technology

Hand gesture recognition using radar has been widely applied to control electronic appliances, military appliances and so on. In this paper, we investigate the feasibility of recognizing hand gestures using fused multiple time-frequency signatures, which ensembles micro-Doppler signatures, range-time signatures and angle-time signatures on spectrograms, with an Enhanced Deep Convolutional Neural Network (EDCNN). Several typical gestures included Tick, Double pushing, Rotating clockwise, and Rotating counterclockwise, were measured using Mm-wave radar and their spectrograms investigated. Therein EDCNN was employed to classify the spectrograms, with 80% of the data utilized for training and the remaining 20% for validation. Simulation said that the classification accuracy of the proposed method was found to be 96.2%.

WED-PM3-SS1.3: Blockchain-based P2P multimedia content distribution using collusion-resistant fingerprinting

Amna Qureshi and David Megías

Universitat Oberta de Catalunya , Internet Interdisciplinary Institute (IN3)

Due to the popularization of low-cost broadband Internet access, the amount of the digital data that is illegally redistributed is growing, making content creators and owners lose their income. Fingerprinting, a watermarking-based technology for embedding buyer identifications in legally distributed contents, has emerged as a promising approach to fight illegal redistribution. On the other hand, the blockchain technology is also shaping up to handle the challenge of digital copyright protection. With blockchain, media producers can authorize and manage their copyrights on a public ledger. In this paper, we present a blockchain-based distribution system which blends different technologies (collusion-resistant fingerprinting, perceptual hash functions, and a peer-to-peer file distribution network) to provide copyright protection, collusion resistance, atomic payment, piracy tracing, transparency, proof-of-delivery, revocable privacy (to a buyer), and dispute resolution. The paper also analyzes several security and privacy compromising attacks and countermeasures.

WED-PM3-SS1.4: Consideration of a Selecting Frame of Finger-Spelled Words from Backhand View

Kosin Chamnongthai, Ponlawat Chophuk and Kanjana Pattanaworapn

King Mongkut's University of Technology Thonburii, Bangkok University

To understand finger alphabet from backhand sign video, there are many redundant video frames between consecutive alphabets and among video frames of an alphabet. These redundant video frames cause loss in finger alphabet understanding, and should be considered to delete. This paper proposes a method to select significant video frames of sign for finger-spelled words of each letter to make more information from backhand view. In this method, finger-spelled words video is divided into frames, and each frame is converted to a binary image by an automatic threshold, and a binary image change to contour frames. Then, we apply the located centroid as the center of the contour image frame to calculate the distance to all boundaries of image frames. After that, all distances of each frame are presented as signature signals that identify each frame, and these values are used with the selected frame equation to select a significant frame. Finally, 1D Signature signal as their feature is extracted from selected frames. In the evaluation of our proposed method, 6 samples of finger-spelled words of the American Sign Language (ASL) are used to select a significant frame, and Hidden Markov Models (HMM) is used to classify the words. The accuracy of this method is approximately 96.67%.

WED-PM3-SS2
Recent Advances in Speaker Recognition, Speaker Diarization and Language Recognition

Time: Wednesday, Nov 20, 17:00-18:40

Place: A2

Chairs: Qingyang Hong, Lin Li

WED-PM3-SS2.1: Triplet-Center Loss Based Deep Embedding Learning Method for Speaker Verification

Yiheng Jiang, Yan Song, Jie Yan, Lirong Dai and Ian McLoughlin

University of Science and Technology of China, University of Science and Technology of China, School of Computing, University of Kent

In this work, we introduce an effective loss function, i.e., triplet-center loss, to improve the performance of deep embedding learning methods for speaker verification (SV). The triplet-center loss is combination of triplet loss and center loss so that it shares superiorities of these two loss functions. Comparing with the widely used softmax loss, the main advantage of triplet-center loss is that it learns a center for each class, and it requires distances between samples and centers from the same class are closer than those from different classes. To evaluate the performance of triplet-center loss, we conduct extensive experiments on noisy and unconstrained dataset, i.e., Voxceleb. The results show that triplet-center loss significantly improves the performance of SV. Specifically, it reduces equal error rate (EER) from softmax loss by 11.6%, 10.4% in cosine scoring and PLDA backend, respectively.

WED-PM3-SS2.2: Speaker Clustering with Penalty Distance for Speaker Verification with Multi-Speaker Speech

Rohan Kumar Das, Jichen Yang and Haizhou Li

National University of Singapore

Speaker verification in a multi-speaker environment is an emerging area of research. Speaker clustering, that separates multiple speakers, can be effective if a predetermined threshold or the number of speakers present in a multi-speaker utterance is given. However, the problem at hand does not provide the leverage for either of the factors. This work proposes to handle such a problem by introducing a penalty distance factor in the pipeline of traditional clustering techniques. The proposed framework first uses traditional clustering techniques to form speaker clusters for a given number of speakers. We then compute the penalty distance based on Bayesian information criterion that is used for merging alike clusters in a multi-speaker utterance. The studies are conducted on speakers in the wild (SITW) and recent NIST SRE 2018 databases that contain multi-speaker conversational speech in noisy environments. The results show the effectiveness of the proposed penalty distance based refinement in such a scenario.

WED-PM3-SS2.3: Geometric Discriminant Analysis for I-vector Based Speaker Verification

Can Xu, Xianhong Chen, Liang He and Jia Liu

Tsinghua University

Many i-vector based speaker verification use linear discriminant analysis (LDA) as a post-processing stage. LDA maximizes the arithmetic mean of the Kullback-Leibler (KL) divergences between different pairs of speakers. However, for speaker verification, speakers with small divergence are easily misjudged. LDA is not optimal because it does not emphasize on enlarging small divergences. In addition, LDA makes an assumption that the i-vectors of different speakers are well modeled by Gaussian distributions with identical class covariance. Actually, the distributions of different speakers can have different covariances. Motivated by these observations, we explore speaker verification with geometric discriminant analysis (GDA), which uses geometric mean instead of arithmetic mean when maximizing the KL divergences. It puts more emphasis on enlarging small divergences. Furthermore, we study the heteroscedastic extension of GDA (HGDA), taking different covariances into consideration. Experiments on i-vector machine learning challenge indicate that, when the number of training speakers becomes smaller, the relative performance improvement of GDA and HGDA compared with LDA becomes larger. GDA and HGDA are better choices especially when training data is limited.

WED-PM3-SS2.4: Extraction of Noise-Robust Speaker Embedding Based on Generative Adversarial Networks

Jianfeng Zhou, Tao Jiang, Qingyang Hong and Lin Li

Xiamen University

In the field of speaker verification, the speaker systems based on x-vector framework are widely used in many scenarios. However, it suffers from the performance degradation caused by noise disturbance. In this paper, we firstly analyzed the noisy robustness of x-vector by training the networks using a mixture dataset which includes clean data and corrupted data. Then, we proposed a novel adversarial strategy against noise interference and extracted the noise-robust speaker embedding with x-vector. The proposed adversarial method named as triplenet GAN employs three connected networks: a generator network (G), a discriminator network (D) and a classifier network (C). The spectral coefficients of clean and noisy speech utterances are fed to the G, of which the structure is nearly the same as x-vector. The outputs of G are transferred in a parallel way to D and C. And the labels of D are set binary for clean data and corrupted data, while the labels of C are set corresponding to speaker identities, which aims to learn the speaker embedding features invariant to noise. Finally, we executed the experiments with different variants of triple-net GAN to verify the denoising capability of the proposed adversarial method. Experimental results on Librispeech corpus demonstrate that our proposed method could achieve a better performance under the noisy environments.

WED-PM3-SS2.5: DKU-Tencent Submission to Oriental Language Recognition AP18-OLR Challenge

Haiwei Wu, Weicheng Cai, Ming Li, Ji Gao, Shanshan Zhang, Zhiqiang Lyu and Shen Huang

School of Electronics and Information Technology, Sun Yat-sen University, Data Science Research Center, Tencent Research, Duke Kunshan University

In this paper, we describe our submitted DKU-Tencent system for the oriental language recognition AP18-OLR Challenge. Our system pipeline consists of three main components, including data augmentation, frame-level feature extraction, and utterance-level modeling. First, we perform speed perturbation to increase the diversity and amount of training data. Second, we extract several kinds of frame-level features, including the hand-crafted acoustic features as well as the deep phonetic features. Third, we aggregate the frame-level features into fixed-dimensional utterance-level representation through i-vector and x-vector modelings. We also propose a deep residual network to obtain the utterance-level language posteriors in an end-to-end manner. Our submitted primary system achieves Cavg of 0.0499, 0.0146, and 0.0135 for the corresponding short-utterance, confusing language and open-set tasks on the evaluation set.

WED-PM3-SS2.6: Margin Matters: Towards More Discriminative Deep Neural Network Embeddings for Speaker Recognition

Xu Xiang, Shuai Wang, Houjun Huang, Yanmin Qian and Kai Yu

AISpeech Co. Shanghai Jiao Tong University

Recently, speaker embeddings extracted from a speaker discriminative deep neural network (DNN) yield better performance than the conventional methods such as i-vector. In most cases, the DNN speaker classifier is trained using cross entropy loss with softmax. However, this kind of loss function does not explicitly encourage inter-class separability and intra-class compactness. As a result, the embeddings are not optimal for speaker recognition tasks. In this paper, to address this issue, three different margin based losses which not only separate classes but also demand a fixed margin between classes are introduced to deep speaker embedding learning. It could be demonstrated that the margin is the key to obtain more discriminative speaker embeddings. Experiments are conducted on two public text independent tasks: VoxCeleb1 and Speaker in The Wild (SITW). The proposed approach can achieve the state-of-the-art performance, with 25%~30% equal error rate(EER) reduction on both tasks when compared to strong baselines using cross entropy loss with softmax, obtaining 2.238% EER on VoxCeleb1 test set and 2.761% EER on SITW core-core test set, respectively.

WED-PM3-SS3
Deep Generative Models for Media Clones and Its Detection

Time: Wednesday, Nov 20, 17:00-18:40

Place: A3

Chairs: Fuming Fang, Zhenzhong Kuang, Xin Wang

WED-PM3-SS3.1: Any-to-one Face Reenactment Based on Conditional Generative Adversarial Network

Tianxiang Ma, Bo Peng, Wei Wang and Jing Dong

National Laboratory of Pattern Recognition, CASIA

Face reenactment refers to the process of transferring the expressions and postures of a given face to the target face. We present a novel Any-to-one Face Reenactment Model based on Conditional Generative Adversarial Network, which has a simple dual converter structure: Any-to-one Face Landmarks Map Converter(AFLC) and Landmark-to-face Converter based on Conditional Generative Adversarial Network(LFC). The former transfers any source face into the landmarks map of the target face, and the map has the expression and posture attributes of the source face. The latter has a generator that transfers the landmarks map of the target face into the realistic and identity-preserving target facial image. The whole model is purely learning-based without any 3D model, and can generate high quality transferred face comparable to the state-of-the-art. What's more the model is highly robust to wild faces, including various faces of different complexions, ages, and genders. We performed an ablation study on our proposed AFLC to verify its importance for face reenactment of any object. AFLC helps the overall model to achieve an effective facial reenactment.

WED-PM3-SS3.2: Discrimination between Handwritten and Computer-Generated Texts using a Distribution of Patch-Wise Font Features

Naoki Hamasaki, Kazuaki Nakamura, Naoko Nitta and Noboru Babaguchi

Osaka University

Recently developed deep generative models allow us to generate images of handwritten-like texts that closely resemble a target writer’ s actual handwriting. Although these models are applicable to useful systems (e.g. communication tools for hand-impaired people), they also could be misused for document forgery by a malicious user. To cope with this problem, in this paper, we propose a text-independent method for discriminating between the computer-generated texts (CGTs) and actual handwritten texts (HWTs). Our proposed method only takes a single text image and recognizes whether the image is a CGT or a HWT. Characters in HWT images have various shapes even when the same writer writes the same sentence. This property is difficult to perfectly mimic even by state-of-theart CGT generation methods. To capture this difference between HWTs and CGTs, we use the distribution of patch-wise font features. The proposed procedure for discriminating HWTs and CGTs is as follows: First, we divide a given text image into several patches and classify each patch into one of pre-determined standard font classes. Then we compute the histogram of the standard fonts, which is finally fed into the recognizer that discriminates between HWTs and CGTs. In our experiments, the proposed method achieved more than 96% of discrimination accuracy, which demonstrates the effectiveness of the proposed method.

WED-PM3-SS3.3: Generating Spoofing Tweets considering Points of Interest of Target User

Lim Jeongwoo, Nitta Naoko, Nakamura Kazuaki and Babaguchi Noboru

Osaka University

Personal information of legitimate users shared on social networking services (SNS) can be used for identity spoofing. The simplest approach is to clone the profile information of the target user. Recent deep learning techniques have enabled us to even automatically generate spoofing messages by imitating the past messages of the target user; however, such message generators can only be trained for target users who have posted sufficient number of messages to train the generator. Further, since the legitimate users actually exist in the real world, their messages are often related to the situations in the real world. Such relations to the real world have not been considered in generating the spoofing messages, which can be the cues for detecting the identity spoofing. This paper further examines the possibility of identity spoofing even for target users who have posted only a limited number of messages based on the assumptions that messages about semantically related points of interest (PoIs) in the real world can be similar regardless of users. Our proposed method firstly collects messages about various PoIs posted by arbitrary users and estimates the semantic topic of each PoI based on the content of its messages. A topic-based message generator trained on the collected messages can be commonly used to generate spoofing messages about PoIs in the real world according to the interest of each target user.

WED-PM3-SS3.4: Anonymization of Gait Silhouette Video by Perturbing Its Phase and Shape Components

Yuki Hirose, Kazuaki Nakamura, Naoko Nitta and Noboru Babaguchi

Osaka University

Nowadays there are a lot of videos containing walking people on the web (e.g. YouTube). These videos can cause a privacy issue because the walking people can be identified by silhouette-based gait recognition systems which have been rapidly advanced in recent years. To solve the issue, in this paper, we propose a method for anonymizing human gait silhouettes. A gait silhouette consists of a static component including the body shape and a dynamic component including postures. We refer to the former and the latter as a shape component and a phase component, respectively. The proposed method anonymizes given gait silhouettes as follows: First, each of the given silhouettes is decomposed into its shape and phase components. Next, both components are separately perturbed. Finally, a new gait silhouette is generated from the perturbed components. Owing to the perturbation, the original silhouettes become less informative in the static aspect as well as the dynamic aspect, by which the gait recognition performance is seriously degraded. In our experimental results, the accuracy was actually degraded from 100% to 30% or less, without yielding any unnatural appearance in the output anonymized gait silhouettes.

WED-PM3-SS3.5: An RGB Gait Anonymization Model for Low Quality Silhouette

Ngoc-Dung T. Tieu, Huy H. Nguyen, Fuming Fang, Junichi Yamagishi and Isao Echizen

SOKENDAI, National Institute of Informatics

Gait anonymization for protecting a person's identity against gait recognition while maintaining the naturalness of the gait is a new research direction. There have been few research introduced, but all of them were reported on high quality silhouette dataset. In this paper, we propose an RGB gait anonymization model for low quality silhouette gait, in which our model is able to generate natural, seamless anonymized gaits whose original contours are unable to be extracted correctly. Our model includes two main networks. The first one, which is a deep convolutional generative adversarial network, is to anonymize the original gait by adding a random noise vector to this gait. By training on a high quality silhouette dataset, this network can generate a high quality anonymized silhouette sequence from a low quality silhouette one. Restricting the input of the first network to binary silhouette sequence instead of color gait force it focus on anonymizing the gait rather than changing the body's color. The second one, which is an autoencoder network and follows the first one, aims to colorize the anonymized silhouette sequence generated by the first one with the color of the original gait. We used two metrics to measure the naturalness and success rate to evaluate the performance of our proposed method.

WED-PM3-SS4
Recent Trends in Signal Processing & Machine Learning - Acoustic & Biomedical Applications

Time: Wednesday, Nov 20, 17:00-18:40

Place: A4

Chairs: Kiyoshi Nishikawa, Felix Albu, Akhtar, Muhammad Tahir

WED-PM3-SS4.1: A Generalization of Laplace Nonnegative Matrix Factorization and Its Multichannel Extension

Hiroki Tanji, Takahiro Murakami and Hiroyuki Kamata

Meiji University

The aim of this paper is to generalize the statistical models of nonnegative matrix factorization (NMF) and multichannel NMF (MNMF). For the NMF and its multichannel extensions, various statistical models have been proposed to improve the model flexibility in the literature on signal separation. However, few studies have been done on the generalization which includes the model based on the Laplace distribution. Thus, we propose the generalized models of the NMF and the MNMF, which include the models based on the Gaussian distribution and the two types of the Laplace distributions, using the Bessel function distribution. To estimate unknown model parameters, we derive the update rules based on the majorization-minimization algorithm. The performances of the proposed NMF and MNMF are evaluated in fitting synthetic data and music signal separation, respectively.

WED-PM3-SS4.2: A Late Reverberation Power Spectral Density Aware Approach to Speech Dereverberation Based on Deep Neural Networks

Yuanlei Qi, Feiran Yang and Jun Yang

Key Laboratory of Noise and Vibration Research , Institute of Acoustics, Chinese Academy of Sciences

In recent years, a variety of speech dereverberation algorithms based on deep neural network (DNN) have been proposed. These algorithms usually adopt anechoic speech as their target output. Consequently, speech distortion might occur which impairs the speech intelligibility. As a matter of fact, early reflections can increase the strength of the direct-path sound and therefore have a positive impact on the speech intelligibility. In traditional speech dereverberation methods, early reflections are generally remained together with the direct-path sound. Based on these observations, we propose to adopt both direct-path sound and early reflections as the target DNN output in this paper. Moreover, we propose a late reverberation power spectral density (PSD) aware training strategy to further suppress the late reverberation. Experimental results demonstrate that the proposed DNN framework achieves significant improvement in objective measures even under mismatched conditions.

WED-PM3-SS4.3: A Comparison Study of GRAPPA and Generalized Series Methods for parallel MRI at high acceleration factor

Khanh Nguyen Thi Kieu and Hien Nguyen Minh

Vietnamese German University

A comparison study of GRAPPA and Generalized Series for parallel MRI images is performed and analyzed at high acceleration factor. Through extensive simulation experiments on various data sets, the conventional GRAPPA method proved its efficiency at low acceleration factor when acquiring few autocalibrating lines. However, at high acceleration factor with very sparse k−space sampling, the GS method proved to be more superior by significantly reducing noise level and achieving much higher accuracy. The key to success of GS method is sufficient amount of low frequency info contained in reference image. This can be achieved by collecting a sufficient the number of ACS lines. In this study, 15%-20% of total k−space lines was sufficient.

WED-PM3-SS4.4: Consideration on application of the concept of Saak transform to convolutional neural networks

Tomonori Maeda and Kiyoshi Nishikawa

Tokyo Metropolitan University

In this paper, we consider applying the concept of Saak (Subspace approximation with augmented kernels) transform to the convolutional neural networks (CNNs). In neural networks, several activation functions are known to enable the nonlinear input-output relation. Recently, the activation function known as ReLU (Rectified linear unit) is widely used in various networks based on CNNs for signal processing applications, e.g., image classification, image super resolution, etc. However, when ReLU is used, there is a possibility that loss of information will occur due to the characteristics of the activation functions, i.e., they set all negative values to 0. In CNNs, this process can be interpreted that the negative correlation will be ignored, and it may cause loss of important information in image processing. The Saak transform is an attempt to recover those negative correlation by using the augmented kernels. In this paper, we consider the structures of CNNs to utilize the concept of Saak. We apply the proposed structures to the image classification and confirm the effectiveness.

WED-PM3-SS4.5: A Norm Penalized Noise-free Maximum Correntropy Criterion Algorithm

Wanlu Shi, Yingsong Li and Felix Albu

Harbin Engineering University, Valahia University of Targoviste

l1-norm penalty and noise-free approach are considered in this paper to contribute to a maximum correntropy criterion (MCC) based algorithm. The introduced l1-norm constrained
noise-free MCC (L1-NFMCC) algorithm inherits the good behavior of MCC in non-Gaussian environments. The cost function of the L1-NFMCC algorithm is created by introducing l1-norm penalty into the traditional cost function of the MCC. In this regard, the L1-NFMCC algorithm can fully use the sparse characteristics which is existed in many real systems. In addition, the noise-free method which is also known as shrinkage technique is used in the L1-NFMCC algorithm to provide a variable convergence step (VCS). The VCS is obtained by minimizing the noise-free (NF) a posteriori error signal (APOES) with respect to the convergence step. As a consequence, the proposed L1-NFMCC algorithm holds an excellent mean square deviation (MSD) behavior. Meanwhile, it shows particularly good property in sparse system. Numerical simulations are utilized to investigate the superiority of the L1-NFMCC algorithm in non-Gaussian noises which show the validity of the L1-NFMCC algorithm.

WED-PM3-SS4.6: Differentiable Programming based Step Size Optimization for LMS and NLMS Algorithms

Kazunori Hayashi, Kaede Shiohara and Tetsuya Sasaki

Osaka City University

We propose TLMS (Trainable Least Mean Squares) and TNLMS (Trainable Normalized LMS) algorithms, which use different step size parameter at each iteration determined by machine learning approach. It has been known that LMS algorithm can achieve fast convergence and small steady-state error simultaneously by dynamically controlling the step size compared as a fix step size, however, in conventional variable step size approaches, the step size parameter has been controlled in rather heuristic manners. In this study, based on the concept of differential programming, we unfold the iterative process of LMS or NLMS algorithms, and obtain a multilayer signal-flow graph similar to a neural network, where each layer has a step size of each iteration of LMS or NLMS algorithm as an independent learnable parameter. Then, we optimize the step size parameters of all iterations by using a machine learning approach, such as the stochastic gradient descent. Numerical experiments demonstrate the performance of the proposed TLMS and TNLMS algorithms under various conditions.

WED-PM3-SS5
Information Security for Digital Content

Time: Wednesday, Nov 20, 17:00-18:40

Place: A5

Chairs: Xiangui Kang, KokSheik Wong, Linna Zhou

WED-PM3-SS5.1: Non-structured Pruning for Deep-learning based Steganalytic Frameworks

Qiushi Li, Zilong Shao, Shunquan Tan, Jishen Zeng and Bin Li

College of Computer Science and Software Engineering, College of Information Engineering, Shenzhen University

Image steganalysis aims to discriminate innocent cover images and those suspected stego images embedded with secret message. Recently, increasing advanced deep neural networks have been proposed and used in image steganalysis. Though those deep learning models can gain superior performance, they also result in redundancy of computational resource and memory storage. In this paper, we apply a non-structured pruning method to slim XuNet2 and SRNet --- the two state-of-the-art deep-learning framework in the field of JPEG image steganalysis. We obtain the priorities of the connections among neurons according to a certain criterion, then keep those significant weights and prune those nonsignificant ones in the meantime. We have conducted extensive experiments on BOSSBase and BOWS image dataset. The experimental results demonstrate that our proposed non-structured pruning method can significantly reduce the cost of computation and storage required by the original deep-learning frameworks without affecting their detection accuracy.

WED-PM3-SS5.2: Effective Source Camera Identification based on MSEPLL Denoising Applied to Small Image Patches

Wenna Zhang, Yunxia Liu, Zeyu Zou, Yunli Zang, Yang Yang and Bonnie Ngai—Fong Law

University of Jinan, Integrated Electronic Systems Lab Co., Shandong University, The Hong Kong Polytechnic University

Sensor Pattern Noise (SPN) has proven to be an effective fingerprint for source camera identification, while its estimation accuracy heavily relies on denoising algorithm. In this paper, an effective source camera identification scheme based on Multi-Scale Expected Patch Log Likelihood (MSEPLL) denoising algorithm is proposed, firstly. With enhanced prior modeling across multiple scales, MSEPLL can accurately restore the original image. As a consequence, estimated SPN is less influenced by image content. Secondly, the source camera identification problem is formulated by hypothesis testing, where normalized correlation coefficient is adopted for SPN detection. Finally, the effectiveness of the proposed method is verified by abundant experiments in terms of identification accuracy as well as receiver operating characteristic. Performance improvement is more prominent for small image patches, which is more conducive to real forensics applications.

WED-PM3-SS5.3: Filtering Adversarial Noise with Double Quantization

April Pyone Maung Maung, Yuma Kinoshita and Hitoshi Kiya

Tokyo Metropolitan University

Despite deep learning being powerful to solve challenging problems, they are vulnerable towards adversarial examples. To defend the adversarial blind spots in the deep learning, researchers have proposed various approaches. However, conventional adversarial training can reduce the accuracy significantly. In this paper, we propose a method to incorporate quantized images in both training and testing to maintain identical accuracy for both normal and adversarial examples. Specifically, the proposed method utilizes dithering during training and dithering and linear quantization as a mean of adversarial filtering during testing. We evaluated the proposed method with a well-known strong first-order adversary and also conducted experiments with the proposed mechanism on different bit depths. The results suggest that the proposed method achieves 87.14 % and 85.28 % accuracy for 2-bit and 1-bit dithered models for both normal and adversarial tests on the noise level of 8. In addition, due to having identical accuracy for both adversarial and normal tests, the proposed method can detect adversarial examples if the original test dataset is known. The code for the experiments is released on https://github.com/fugokidi/one-bit-quantization.

WED-PM3-SS5.4: Image Identification of Grayscale-Based JPEG Images for Privacy-Preserving Photo Sharing Services

Kenta Iida and Hitoshi Kiya

Tokyo Metropolitan University

We propose an image identification scheme for double-compressed JPEG images encrypted by a grayscale-based encryption method, where the encrypted JPEG images are referred to as grayscale-based JPEG images, that has been proposed for Encryption-then-Compression (EtC) systems with JPEG compression. The proposed scheme aims to identify encrypted JPEG images that are generated from an original JPEG image. To store images without any visual sensitive information on photo sharing services, grayscale-based JPEG images are generated by using the grayscale-based encryption method. The use of the grayscale-based JPEG images and feature vectors extracted from the JPEG images allows us to identify images not only recompressed multiple times but re-encrypted with different keys. In an experiment, the proposed scheme is shown to have a high identification performance, even when images are recompressed multiple times and re-encrypted with different keys.

WED-PM3-SS5.5: Privacy-Preserving Deep Neural Networks Using Pixel-Based Image Encryption Without Common Security Keys

Warit Sirichotedumrong, Yuma Kinoshita and Hitoshi Kiya

Tokyo Metropolitan University

We present a novel privacy-preserving scheme for deep neural networks (DNNs) that enables us not to only apply images without visual information to DNNs but to also consider the use of independent encryption keys, for both training and testing images for the first time. In this paper, a novel pixel-based image encryption method, which considers maintaining the properties of original images, is first proposed for privacy-preserving DNNs. For training, a DNN model is trained with images encrypted by using the proposed method under the use of independent keys. For testing, the model enables us to applied both encrypted images and plain images for image classification. Therefore, there is no need to manage the keys. In an experiment, the proposed method is applied to a well-known network, deep residual networks, for image classification. The experimental results demonstrate that the proposed method with independent encryption keys has robustness against ciphertext-only attack (COA) and can provide almost the same classification performance as that of using plain images. Moreover, the results confirm that the proposed scheme is able to classify plain images as well as encrypted images.

WED-PM3-SS5.6: Delving into the Methods of Coverless Image Steganography

Koi Yee Ng, Sim Ying Ong and Kok Sheik Wong

University of Malaya, Monash University

Conventional cover-based image steganography methods embed secret information by modifying the original state of a cover image. This type of algorithm leaves a trace of changes on output stego image and eventually leads to successful detection by common steganalysis tools. As a solution, a coverless image steganographic method is proposed, where no cover image is required for embedding secret information. In this paper, the conventional coverless image steganography methods are first reviewed and categorized into constructive and non-constructive-based methods. Next, these methods are summarized and analyzed, followed by a discussion about their advantages and drawbacks. Finally, the performance of the proposed methods are discussed using the common steganography evaluation metrics, including resistance to attack, embedding capacity, and perceptual image quality.

WED-PM3-SS5.7: Image Reconstruction from Local Descriptors Using Conditional Adversarial Networks

Haiwei Wu, Jiantao Zhou and Yuanman Li

University of Macau

Many applications rely on the local descriptors extracted around a collection of interest points. Recently, the security of local descriptors has been attracting increasing attention. In this paper, we study the possibility of image reconstruction from these descriptors, and propose a coarse-to-fine framework for the image reconstruction. By resorting to our gradually reconstructing network architecture, the novel multi-scale feature map generation algorithm, and the strategically designed loss functions, our proposed algorithm can recover the images with very high perceptual quality, even partial descriptors are provided only. Extensive experimental results are reported to show its superiority over the existing algorithms. Our study implies that the local descriptors contain surprisingly rich information of the original image. Users should pay more attention to sensitive information leakage when using local descriptors.

WED-PM3-SS6
Second Language Speech Perception and Production[1]

Time: Wednesday, Nov 20, 17:00-18:40

Place: A6

Chairs: Ying Chen, Jian Gong

WED-PM3-SS6.1: Computational perception of information foci produced by Chinese English learners and American English speakers

Juqiang Chen and Xuliang He

Western Sydney University, Nantong University

This study used computational perception, via SVM and Random Forest models, to examine phonetic features used by American English speakers (AE) and Chinese second language learners of English (CE1 with low proficiency and CE2 with high proficiency) in realizing different information foci. For all participant groups, the machine learning models achieved above chance level accuracy. Coda duration and the duration of the rising contour were two phonetic features that ranked top across three participant groups in terms of their importance to the models. The SVM models trained with the AE data classified different foci by CE1 and CE2 with above chance level accuracy, but English proficiency had little effect on the classification results.

WED-PM3-SS6.2: Acoustic Attributes of Citation Tones in Standard Chinese Produced by Prelingually Deaf Adults

Jie Hou, Yu Chen, Yutong Xing and Jianwu Dang

Tianjin University of Technology, Tianjin University

Based on perception judgment and acoustic analysis with the data of ten prelingually deaf adults (PDAs) and ten normal hearing adults (NHAs), the present paper investigated the performance of the four citation tones in Standard Chinese produced by prelingually deaf adults. Overall, the error rate of perception judgment for PDAs was 12.95%; however, the error rates of Tone 2 and Tone 3 were 30.70% and 19.85% respectively, which were much higher than that of Tone 4 and Tone1. As the results of acoustic analysis, although the general performance of the deaf females was significantly different from NHAs, they approached or reached the level of the control group on some parameters, while deaf males faced greater challenge than deaf females. Therefore, this study confirmed that PDAs still have some impairments on tone production of Standard Chinese.

WED-PM3-SS6.3: Multi-Task Based Mispronunciation Detection of Children Speech Using Multi-Lingual Information

Linxuan Wei, Wenwei Dong, Binghuai Lin and Jinsong Zhang

BLCU, Beijing Language and Culture University

In developing a Computer-Aided Pronunciation Training (CAPT) system for Chinese ESL (English as a Second Language) children, we suffered from insufficient task-specific data. To address this issue, we propose to utilize first language (L1) and second language (L2) knowledge from both adult and children data through multitask-based transfer learning according to Speech Learning Model (SLM). Experimental set-up includes the TDNN acoustic modelling using the following training data: 70 hours of English speech by American Children (AC), 100 hours by American Adults (AA), 5 hours of Chinese speech by Chinese Children (CC), and 89 hours by Chinese Adults (CA). Testing data includes 2 hours of ESL speech by Chinese children. Experimental results showed that the inclusion of AA data brought about 13% relative Detection Error Rate (DER) reduction compared to AC only. Further inclusion of CC and CA data through L1 transfer learning brought about a total of 21% relative improvement in DER. These results suggested the proposed method is effective in mitigating insufficient data problem.

WED-PM3-SS6.4: Sounds of Personality: Inference from Voices by Non-Native Speakers

Bin Li, Yihan Guan and Si Chen

City University of Hong Kong, The Hong Kong Polytechnic University

People listen to get information in oral communication, and may consciously or unconsciously draw inferences on speakers based on their voices. A variety of phonetic cues have been found relevant in perceiving personalities from speech. This correlation has been documented in research involving native speech perception, though perceptual judgements on personality parameters were reported varying across cases. The tendency and sensitivity in voice perception has also attracted attention on non-native

communication, where both language proficiency and language-specific features are considered influential as well. It is thus interesting to examine if similar or different sets of phonetic cues may affect non-native listeners' assessment of voices and personality. This study examined English-as-a-second-language (ESL) speakers' listening comprehension and perception of personality when listening to speech samples in American English. Speech extracts were modified to include variations in temporal and spectral dimensions. Results show that modification to pitch seemed beneficial to improve ESL listeners' comprehension accuracy. The modification also resulted in more favorable judgement of personality traits. Changes to the speaking rate yielded similar positive correlation with comprehension and personality judgement.

WED-PM3-SS6.5: Acquisition and Interpretation of Mandarin Speech Prosody by Native Speakers and Cantonese Learners

Xi Chen and Si Chen

The Hong Kong Polytechnic University

The Interface Hypothesis posited that internal and external interfaces pose difficulties for adult L2 learners (Sorace, 2006). Recent studies showed that using speech prosody to mark information structure can be challenging to L2 learners of English (Lee, Perdomo & Kaan, 2019). However, fewer studies examined the acquisition of speech prosody in tonal languages. This study aims to test the ability to match acoustic cues to different focus types and positions by advanced Cantonese L2 learners of Mandarin under the modalities of auditory-only and visual-auditory. Following the design by Roettger (2019), participants were instructed to make a 5-Likert response to rate their preferences for the conversations they heard. Results show that visual-aids facilitated the perception of prosody; L2 learners showed fewer difficulties in differentiating narrow and contrastive focus than native Mandarin speakers. These findings provide significances for prosodic perception, second language acquisition, and bilingual education.

WED-PM3-SS6.6: Acquisition of L2 Mandarin Rhythm By Russian and Japanese

Yiran Ding, Yanlu Xie and Jinsong Zhang

Beijing Language and Culture University

For Chinese as second language (CSL) leaners with different mother tongues (L1), the developments of their speech rhythm received little attention. Based on the interval-based acoustic rhythm metrics, we compared the speech productions of L2 Mandarin by 15 Japanese and 15 Russian learners with different proficiency level. The data included 103 sentences in read speech by each speaker (3605 sentences in total). Preliminary results showed: a.)During the progress from beginners toward intermediate level, the durational variability decreased in both groups of learners, which indicated acquisition of L2 Mandarin rhythm followed similar developmental paths from more stress-timed toward more syllable-timed; b.)During the progress from intermediate toward advanced level, Russian learners kept kind of stress-timed rhythm, Japanese learners appeared mora-timed rhythm, it indicated the transfer effects were influential at this learning stages.

WED-PM3-SS7
Recent Topics on Signal and Information Processing for Active Control of Sound

Time: Wednesday, Nov 20, 17:00-18:40

Place: A7

Chairs: Yoshinobu Kajikawa, Chuang Shi

WED-PM3-SS7.1: A min-max optimization algorithm for global active acoustic radiation control

Rong Han, Ming Wu, Kexun Chi, Lan Yin, Hongling Sun and Jun Yang

Institute of Acoustics, Chinese Academy of Sciences, Institute of Automation, Beijing Information Science and Technology University

Generally, global active noise control is to minimize the sum of the energy of the residual sound field. But on some occasions, we are more concerned that the energy of the residual noise at every direction does not exceed a certain value. In this paper, an algorithm for global active noise control is introduced to achieve a global active acoustic radiation noise control by minimizing the maximum residual sound energy of the far field after active acoustic radiation control. The proposed algorithm adjusts the weights of the secondary sound sources based on the min-max optimization. Simulation results show that the proposed algorithm can reduce the maximum of the far-field residual sound pressure 2.2 dB more than the maximum of the residual sound pressure based on the traditional global acoustic radiation control algorithm.

WED-PM3-SS7.2: Audio Integrated Active Noise Control System with Auto Gain Controller

Kenta Iwai and Takanobu Nishiura

Ritsumeikan University

This paper proposes an audio integrated active noise control (ANC) system with an auto gain controller. An ANC system is one of the techniques for reducing unwanted noise and used to reduce the factory noise, engine noise, and so forth. In general, the ANC system cannot completely reduce the unwanted noise due to its principle. To solve this problem, an audio integrated ANC (AIANC) system has been proposed. The AIANC system uses the additional audio signal to mask the residual noise called error signal. Also, the AIANC system can be used for telecommunication under noisy environment, in which the voice is treated as the audio signal of the AIANC system. However, in the conventional AIANC system, the power of the audio signal cannot be adjusted to that of the error signal and it causes that the audio signal is too larger or smaller than the error signal. To solve this problem, the AIANC with an auto gain controller is proposed. The proposed AIANC system has the auto gain controller to adjust the power of the audio signal and that of the error signal. Then, the audio signal is emitted with the same power as the error signal. Simulation results shows that the proposed AIANC system can reduce the unwanted noise and adjust the power of the audio signal to that of the error signal.

WED-PM3-SS7.3: Beam Steering of Portable Parametric Array Loudspeaker

Kyosuke Nakagawa, Chuang Shi and Yoshinobu Kajikawa

Kansai University, School of Information and Communication Engineering , University of Electronic Science and Technology of China

Portable devices such as smartphones and tablet PCs have become increasingly sophisticated and explosively spread. Opportunities for outdoor use have been consequently increasing. When the portable devices are used in public areas, personal audio system is required to avoid sound spread in the vicinity. We have already proposed the portable parametric array loudspeaker which can realize personal audio without using earphones and headphones. In this system, parametric array loudspeakers are mounted on two edges of tablet PCs and can radiate highly directional stereo sound to the user. However, the radiated sound beams may not focus on the user' s ears when the user' s head is moving. In this paper, we examine the phased array technique to steer the sound beam based on the user' s head position. We demonstrate that the sound beam angle can be appropriately steered by using the phased array technique through experimental results.

WED-PM3-SS7.4: A Simulation Investigation of Modified FxLMS Algorithms for Feedforward Active Noise Control

Chuang Shi, Nan Jiang, Rong Xie and Huiyong Li

University of Science and Technology of China, National University of Defense Technology

In this paper, two modified FxLMS algorithms are proposed based on the post-masking-based LMS (PMLMS) algorithm. They are the PMI-FxLMS and the signed PMI-FxLMS (SPMI-FxLMS) algorithms. In both algorithms, I denotes the length of the error signal memory. The control filter coefficients are updated by the maximum absolute value in the error signal memory, instead of the immediate value. The difference between the two modified FxLMS algorithms is that the PMI-FxLMS algorithm keeps the sign of the error sample with the largest absolute value, while the SPMI-FxLMS algorithm uses the sign of the immediate error sample. The simulation results show that the SPMI-FxLMS algorithm converges faster than the standard FxLMS algorithm with the same step-size, and the PMI-FxLMS algorithm may be difficult to converge when I is large.

WED-PM3-SS8
Advanced Signal Processing for 5G Communication

Time: Wednesday, Nov 20, 17:00-18:40

Place: A8

Chairs: Na Chen, Minoru Okada

WED-PM3-SS8.1: Mobile Robot Object Recognition in The Internet of Things based on Fog Computing

Meixia Fu and Songlin Sun

Beijing University of Posts and Communications

Mobile robot object recognition has attracted significant attention in the internet of things recently, in which there are many challenging tasks, such as the objects, the communication networks and the computer system. It still needs a large of computation, communication and storage capability for the whole system. In this paper, we propose a scheme of mobile robot object recognition in IOT and use edge nodes to process the data from robot vision instead of cloud computing. Besides, we adopt YOLOv3 as the main algorithm in the edge nodes to process the video data. The video from the camera on the robot is transmitted to the fog node by a wifi router. The advantages of using edge nodes for computation local robot clusters are reliability, real-time ability and flexibility compared with the cloud. In our experiment, we efficiently train the computer model using COCO database deployed to GPU. The robot could recognize the objects in real-time. We achieve mAP of 31.0% and response time of 52ms that illustrates the state-of-art performance of the proposed framework.

WED-PM3-SS8.2: Joint Sparse Channel Estimation in Downlink NOMA system

Jia Haohui, Na Chen, Minoru Okada and Takasei Higashino

Nara Institute of Science and Technology

Non-orthogonal multiple access(NOMA) is regarded as one of the most important technique for the future 5G systems and it is easy to combine with Massive MIMO technique.In the downlink general NOMA schemes, the received NOMA information will be studied through two parallel channel stat Information after sparse multiple path channel fading .In this paper,by exploiting the inherent sparsity of sparse channel, we proposed a Low-complexity joint channel estimation in a large number of antennas system,based on a structured compressed sensing to detector each layer channel state information.As a comparision, the performance of stuctured compressed sensing (CS)is better than the conventional Method LS and MMSE.

WED-PM3-SS8.3: Time-domain signal recovery for OFDM system in the industrial environment

Chengbo Liu, Na Chen, Yafei Hou and Minoru Okada

Nara Institute of Science and Technology, Okayama University

This paper will propose a method to recover the damaged time-domain signal of the OFDM system over the AWGN and fading channel in the industrial environment. We will show that, for OFDM signal, the whole time-domain signal can be rebuilt using the partial time-domain signal. Therefore, we can recover the damaged partial time-domain signal assuming that the channel is perfectly detected. The simulated results show that the OFDM system using M-algorithm can achieve better BER performance than that of the conventional OFDM system even the half time-domain signal is missing in the AWGN channel. On the other hand, the OFDM system using M -algorithm can achieve better BER performance compared to that of the conventional OFDM system in the AWGN channel when 1/8, 1/4 and 1/2 time- domain signal is missing. Although the BER performance of the OFDM system with M -algorithm using MMSE is better than that of the OFDM system with M-algorithm using ZF in the fading channel when 1/8, 1/4 and 1/2 time-domain signals are missing, the BER performance of the OFDM system with M-algorithm using ZF is better than that of the conventional OFDM system.

WED-PM3-SS8.4: Deep Reinforcement Learning for Resource Allocation in 5G Communications

Mau-Luen Tham, Amjad Iqbal and Yoong Choon Chang

Universiti Tunku Abdul Rahman (UTAR)

The rapid growth of data traffic has pushed the mobile telecommunication industry towards the adoption of fifth generation (5G) communications. Cloud radio access network (CRAN), one of the 5G key enabler, facilitates fine-grained management of network resources by separating the remote radio head (RRH) from the baseband unit (BBU) via a high-speed front-haul link. Classical resource allocation (RA) schemes rely on numerical techniques to optimize various performance metrics. Most of these works can be defined as instantaneous since the optimization decisions are derived from the current network state without considering past network states. While utility theory can incorporate long-term optimization effect into these optimization actions, the growing heterogeneity and complexity of mobile network environments has made the RA issue become intractable. One potential solution is to resort to reinforcement learning (RL), a dynamic programming framework which solves the RA problems optimally over varying network states. Still, such method cannot handle the highly dimensional state/action spaces in the context of CRAN problems. Driven by the success of machine learning, researchers begin to explore the potential of deep reinforcement learning (DRL) to solve the RA problems. In this work, an overview of the major state-of-the-art approaches to DRL in CRAN is provided. We complete this article by pinpointing current challenges and open future directions for research.

WED-PM3-SS8.5: A Survey on Applications of Deep Reinforcement Learning in Resource Management for 5G Heterogeneous Networks

Ying Loong Lee and Donghong Qin

Universiti Tunku Abdul Rahman, Guangxi University for Nationalities

Heterogeneous networks (HetNets) have been regarded as the key technology for fifth generation (5G) communications to support the explosive growth of mobile traffics. By deploying small-cells within the macrocells, the HetNets can boost the network capacity and support more users especially in the hotspot and indoor areas. Nonetheless, resource management for such networks becomes more complex compared to conventional cellular networks due to the interference arise between small-cells and macrocells, which thus making quality of service provisioning more challenging. Recent advances in deep reinforcement learning (DRL) have inspired its applications in resource management for 5G HetNets. In this paper, a survey on the applications of DRL in resource management for 5G HetNets is conducted. In particular, we review the DRLbased resource management schemes for 5G HetNets in various domains including energy harvesting, network slicing, cognitive HetNets, coordinated multipoint transmission, and big data. An insightful comparative summary and analysis on the surveyed studies is provided to shed some light on the shortcomings and research gaps in the current advances in DRL-based resource management for 5G HetNets. Last but not least, several open issues and future directions are presented.

THU-AM1-SS1
Advanced Signal Processing and Machine Learning for Audio and Speech Applications

Time: Thursday, Nov 21, 10:20-12:00

Place: A2

Chairs: Shoji Makino, Hiroshi Saruwatari

THU-AM1-SS1.1: Griffin--Lim phase reconstruction using short-time Fourier transform with zero-padded frame analysis

Yukoh Wakabayashi and Nobutaka Ono

Tokyo Metropolitan University

In this paper, we present the short-time Fourier transform (STFT) with zero-padded frame analysis to introduce frequency redundancy into a time--frequency representation, and we investigate its application to phase reconstruction by the Griffin--Lim algorithm. Recent studies on phase reconstruction have suggested that the use of a small STFT frame shift improves the performance of phase reconstruction techniques, which implies that increasing the temporal redundancy of the time--frequency representation makes phase reconstruction easier. Motivated by this, we consider the STFT with zero padding to increase the redundancy of the STFT along the frequency axis. The linearity of this transform and its inverse enables the use of the Griffin--Lim algorithm on the time--frequency domain. We evaluate the performance of phase reconstruction using the STFT with and without zero padding using the spectral distance and perceptual evaluation of speech quality as the criteria. The experimental results show that increasing the frequency redundancy with zero padding improves the phase reconstruction performance similarly to using a small frame shift.

THU-AM1-SS1.2: Robust Demixing Filter Update Algorithm Based on Microphone-wise Coordinate Descent for Independent Deeply Learned Matrix Analysis

Naoki Makishima, Norihiro Takamune, Daichi Kitamura, Hiroshi Saruwatari, Yu Takahashi and Kazunobu Kondo

The University of Tokyo, National Institute of Technology, Kagawa College, Yamaha Corporation

In this paper, we propose a robust demixing filter update algorithm for audio source separation, which is the task of recovering source signals from multichannel mixtures observed in a microphone array. Recently, independent deeply learned matrix analysis (IDLMA) has been proposed as a state-of-the-art separation method. IDLMA utilizes the deep neural network (DNN) inference of source models and the blind estimation of demixing filters based on sources' independence. In conventional IDLMA, iterative projection (IP) is exploited to estimate the demixing filters. Although IP is a fast algorithm, when a specific source model is not accurate owing to an unfavorable SNR condition, the subsequent update of filters will fail. This is because IP updates the demixing filters in a sourcewise manner, where only one source model is used for each update. In this paper, we derive a new microphone-wise update algorithm that exploits all information of the source models simultaneously for each update. The microphone-wise update problem cannot be solved by IP, but instead, a new type of vectorwise coordinate descent algorithm is introduced into the proposed algorithm to realize convergence-guaranteed parameter estimation. Experimental results show that the proposed update algorithm achieves better separation performance than IP.

THU-AM1-SS1.3: Evaluation of multichannel hearing aid system using rank-constrained spatial covariance matrix estimation

Masakazu Une, Yuki Kubo, Norihiro Takamune, Daichi Kitamura, Hiroshi Saruwatari and Shoji Makino

University of Tsukuba, The University of Tokyo, National Institute of Technology , Kagawa College

In a noisy environment, speech extraction techniques make hearing aid systems more effective and practical. Blind source separation~(BSS) is suitable for hearing aid because it can be employed without any a priori spatial information. Among many BSS methods, independent low-rank matrix analysis (ILRMA) achieves high-quality separation performance. In the diffuse noise environment, however, ILRMA cannot suppress the noise since the method is based on the determined situation. On the other hand, rank-constrained spacial covariance matrix~(SCM) estimation overcomes the problem. The method utilizes spatial parameters accurately estimated by ILRMA and compensates for the deficiency of spatial basis of diffuse noise. Application of BSS methods to the multichannel binaural hearing aid system with the smartphone have never been studied in detail so far. To clarify the efficacy of the BSS methods in real environments, we record real sounds by constructing a hearing aid system with a dummy head and a smartphone. In this paper, we investigate the efficacy of BSS for multichannel binaural hearing aid system including microphones on a smartphone. Furthermore, we apply ILRMA and the rank-constrained SCM estimation to the recorded data and evaluate them in terms of separation performance.

THU-AM1-SS1.4: Comparative Study of Deep Learning Based and Traditional Single-Channel Noise-Reduction Algorithms

Ningning Pan, Jingdong Chen and Biing-Hwang Juang

Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Georgia Institute of Technology

Deep neural networks (DNN) have been applied to the problem of noise reduction and promising results have been reported widely, leading to the impression that the traditional techniques based on blind noise estimation may no longer be needed. However, there lacks comprehensive and rigorous evaluation and comparison between DNN based and traditional noise reduction algorithms for their pros and cons. In this work, we attempt to evaluate some widely used DNN based noise-reduction algorithms and compare them to a traditional noise-reduction method. We also evaluate a method that straightforwardly combines a DNN based regression method with the optimal filtering technique. Through experiments, it is observed that: 1) DNN based methods have advantages over the traditional methods in scenarios with non-stationary noise and low signal-to-noise ratios (SNRs); 2) generalization remains a challenging issue with DNN based methods and for noise type unseen in the training data, which happens often in practical environments, DNN based methods do not show any advantages over the traditional technique; 3) combining DNN based regression method and the optimal filtering technique shows some potential in improving noise-reduction performance as well as system generalization.

THU-AM1-SS1.5: An Efficient Dilated Convolutional Neural Network for UAV Noise Reduction at Low Input SNR

Zhi-Wei Tan, Anh Nguyen and Andy Khong

Nanyang Technological University – School of Electrical and Electronics Engineering

Acoustic applications on a multi-rotor unmanned aerial vehicle (UAV) have been hindered by its low input signal-to-noise ratio (SNR). Such low SNR condition poses prominent challenges for beamforming algorithms, statistical methods, and existing mask-based deep learning algorithms. We propose the small model on low SNR (SMoLnet), a compact convolutional neural network (CNN) to suppress UAV noise in noisy speech signals recorded off a microphone array mounted on the UAV. The proposed SMoLnet employs a large analysis window to achieve high spectral resolution since the loud UAV noise exhibits a narrow-band harmonic pattern. In the proposed SMoLnet model, exponentially-increasing dilated convolution layers were adopted to capture the global relationship across the frequency dimension. Furthermore, we performed direct spectral mapping between noisy and clean complex spectrogram to cater to the low SNR scenario. Simulation results show that the proposed SMoLnet outperforms existing dilation-based models in terms of speech quality and objective speech intelligibility metrics for UAV noise reduction. In addition, the proposed SMoLnet requires fewer parameters and achieves lower latency than the compared models.

THU-AM1-SS2
Multilingual Speech and Language Processing

Time: Thursday, Nov 21, 10:20-12:00

Place: A3

Chairs: Zhiyuan Tang, Dong Wang, GuanyuLi, Mijiti Ablimit

THU-AM1-SS2.1: Multi-lingual Transformer Training for Khmer Automatic Speech Recognition

Kak Soky, Sheng Li, Tatsuya Kawahara and Sopheap Seng

National Institute of Posts, Telecommunications and ICT (NIPTICT), National Institute of Information & Communications Technology (NICT), Kyoto University

Currently, there are three challenges for constructing reliable ASR systems for the Khmer language: (1) the lack of language resources (text and speech corpora) in digital form, (2) the writing system without explicit word boundary and (3) the pronunciation model has not well studied. In this paper, to avoid the extensive work on selecting proper acoustic units (e.g., phones, syllables) and prepare the frame-level labels on the traditional GMM-HMM/DNN-HMM framework, we directly use words or characters as the label using state-of-the-art transformer-based End-to-End model. Moreover, we use the multi-lingual training framework to tackle the low-resource data problem. All experiments are performed on the Basic Expressions Travel Corpus (BTEC) datasets, which is first time evaluated by the transformer-based models to our best knowledge. The experiments show that the proposed multi-lingual transformer-based End-to-End model can achieve significant improvement compared to the DNN-HMM baseline model.

THU-AM1-SS2.2: A Study on Low-resource Language Identification

Zhaodi Qi, Yong Ma and Mingliang Gu

School of Physics and Electronic Engineering, Jiangsu Normal University, Kewen College Jiangsu Normal University, School of Physics and Electronic Engineering Jiangsu Normal University

Modern language identification (LID) systems require a large amount of data to train language-discriminative models, either statistical (e.g., i-vector) or neural (e.g., x-vector). Unfortunately, most of languages in the world have very limited accumulation of data resources, which result in limited performance on most languages. In this study, two approaches are investigated to deal with the LID task on low-resource languages. The first approach is data augmentation, which enlarges the data set by incorporating various distortions into the original data; and the second approach is multi-lingual bottleneck feature extraction, which extracts multiple sets of bottleneck features (BNF) based on speech recognition systems of multiple languages. Experiments conducted on both the i-vector and x-vector models demonstrated that the two approach are effective, and can obtain promising results on both in-domain data and out-of-domain data.

THU-AM1-SS2.3: A morpheme sequence and convolutional neural network based Kazakh text classification

Sardar Parhat, Gao Ting, Mijit Ablimit and Askar Hamdulla

College of Information Science and Engineering, Xinjiang University, School of Information Science and Engineering, Xinjiang University

Word embedding techniques can map language units into a sequential vector space based on context. And it is a natural way to extract and predict out-of-vocabulary (OOV) from context information, word-vector based morphological analysis has provided a convenient way for low resource languages processing tasks. In this paper, we discuss Kazakh text classification experiment based on the m2asr morphological analyzer for small agglutinative languages. Morpheme segmentation and stem extraction from noisy data based on stem-vector similarity representation are experimented on Kazakh language. After preparing both word and morpheme-based training text corpora, we apply convolutional neural networks (CNN) as a feature selection and text classification algorithm to perform text classification tasks. Experimental results show that morpheme-based approach outperforms word-based approach.

THU-AM1-SS2.4: Zero-resource language recognition

Jiawei Yu and Jinsong Zhang

Beijing Language and Culture University

Language recognition (LRE) can be categorized into two configurations: in the close-set setting, a test segment is classified into one of several pre-defined languages, and in the open-set setting, a segment that is not in any of the pre-defined languages will be labelled as 'unknown'. In real applications, there is another scenario: we hope to register a new language with several utterances and then this language will be recognized by the system, although this language is not involved when the system is constructed. We call this zero-resource LRE (ZR-LRE). In this paper, we explore the language embedding approach and apply it to tackle the ZR-LRE problem. Specifically, we first train an embedding space of languages based on i-vector or d-vector, and then new languages can be registered and recognized within this space. The experiments were conducted on the AP18-OLR database including 10 languages for training the embedding space and another 8 languages as zero-resource (ZR) languages for registration and recognition. To explore the influence of different test condition to the performance of ZR- LRE, we evaluated various configurations that involve different numbers of enrollment utterances and different duration of test utterances. The results show that embedding based on i-vectors is suitable for ZR-LRE, which achieved an Equal Error Rate (EER) of 8.7%.

THU-AM1-SS2.5: Cross-lingual Automatic Speech Recognition Exploiting Articulatory Features

Qingran Zhan, Petr Motlicek, Shixuan Du, Yahui Shan, Sifan Ma and Xiang Xie

Beijing Institute of Technology, Idiap Research Institute

Articulatory features (AFs) provide language-independent attribute by exploiting the speech production knowledge. This paper proposes a cross-lingual automatic speech recognition (ASR) based on AF methods. Various neural network (NN) architectures are explored to extract cross-lingual AFs and their performance is studied. The architectures include muti-layer perception(MLP), convolutional NN (CNN) and long short-term memory recurrent NN (LSTM). In our cross-lingual setup, only the source language (English, representing a well-resourced language) is used to train the AF extractors. AFs are then generated for the target language (Mandarin, representing an under-resourced language) using the trained extractors. he frame-classification accuracy indicates that the LSTM has an ability to perform a knowledge transfer through the robust cross-lingual AFs from well-resourced to under-resourced language. The final ASR system is built using traditional approaches (e.g. hybrid models), combining AFs with conventional MFCCs. The results demonstrate that the cross-lingual AFs improve the performance in under-resourced ASR task even though the source and target languages come from different language family. Overall, the proposed cross-lingual ASR approach provides slight improvement over the monolingual LF-MMI and cross-lingual (acoustic model adaptation-based) ASR systems.

THU-AM1-SS2.6: AP19-OLR Challenge: Three Tasks and Their Baselines

Zhiyuan Tang, Dong Wang and Liming Song

Tsinghua University, SpeechOcean

This paper introduces the fourth oriental language recognition (OLR) challenge AP19-OLR, including the data profile, the tasks and the evaluation principles. The OLR challenge has been held successfully for three consecutive years, along with APSIPA Annual Summit and Conference (APSIPA ASC). The challenge this year still focuses on practical and challenging tasks, precisely (1) short-utterance LID, (2) cross-channel LID and (3)zero-resource LID. The event this year includes more languages and more real-life data provided by SpeechOcean and the NSFC M2ASR project. All the data is free for participants. Recipes for x-vector system and back-end evaluation are also conducted as baselines for the three tasks. The participants can refer to these online-published recipes to deploy LID systems for convenience. We report the baseline results on the three tasks and demonstrate that the three tasks are worth some efforts on them.

THU-AM1-SS3
Deep Learning Systems for Cloud, Fog, and Edge Computing, and Applications

Time: Thursday, Nov 21, 10:20-12:00

Place: A4

Chair: Jia-Ching Wang

THU-AM1-SS3.1: A Real-time and Online Multiple-Type Object Tracking Method with Deep Features

Yi-Hsuan Hsu and Jiun-In Guo

National Chiao Tung University

Object tracking is one of the most important things in intelligent vision system. Meanwhile, the most challenging issue in object tracking is how to keep the target' s identity unchangeable with limited power consumption. In this paper, we propose a real-time and online tracking method to track multiple types of objects (e.g. pedestrian and car). Furthermore, to handle the ID switching problem, we provide a lightweight deep learning model which can recognize the similarity of objects. It can effectively solve the ID switching problem resulted from occlusion. Finally, we do some experiments to demonstrate that the proposed method achieves the state-of-the-art performance with less power consumption. The proposed method can solve the problem of high computation of tracking and keep the high accuracy of counting results with low ID switching rate. The experimental result shows that the average counting accuracy of the proposed method can reach more than 93% on pedestrian and vehicle counting applications. Also, it shows that the proposed method improves 68.2% on average of ID switching rate than previous works.

THU-AM1-SS3.2: Convolutional Attention Model for Retinal Edema Segmentation

Phuong Le Thi, Tuan Pham and Jia Ching Wang

National Central University, University of Technology and Education - The University of Danang

Deep learning and computer vision that become popular in recent years are advantageous techniques in medical diagnosis. A large database of Optical Coherence Tomography (OCT) images can be used to train a deep learning model which can support and suggest accurately illnesses and status of a patient. Therefore, semantic image segmentation is used to detect and categorize anomaly regions in OCT images. However, numerous existing approaches ignored spatial structure as well as contextual information in a given image. To overcome remaining problems, this work proposes a novel method which takes advantage task of the deep convolutional neural network, attention block, pyramid pooling module and auxiliary connections between layers. Attention block help to detect the spatial structure of a given image. Pyramid pooling module has a responsibility to identify the shape and margin of the anomaly region. Auxiliary connections help to enrich useful information pass through one layer as well as reduce vanishing gradient problem. Our work produces higher accuracy than state-of-the-art methods with 78.19% comparing to Deeplabv3 76.19% and Bisenet 76.85% in term of dice similarity coefficient. Additionally, a number of parameters in our work is smaller than the previous approaches.

THU-AM1-SS3.3: Parallel Capsule Neural Networks for Sound Event Detection

Kai-Wen Liang, Yu Hao Tseng and Pao-Chi Chang

National Central University

In this work, we propose a sound event detection system based on a parallel capsule neural network. The system takes advantage of the capability of capsule neural networks in the detection of overlapping objects. It further develops a parallel architecture and uses the kernel design of different shapes and sizes to effectively utilize the feature information to increase the detection accuracy. The experimental results show that the performance of the proposed system is as low as 52.34% measured by the error rate, which is even lower than the rank 1 system in DCASE2017 challenge.

THU-AM1-SS3.4: Age and Gender Recognition Using Multi-task CNN

Duc-Quang Vu, Thi-Thu-Trang Phung, Chien-Yao Wang and Jia-Ching Wang

National Central University, Thai Nguyen University, Institute of Information Science, Academia Sinica

The investigation into age and gender identification has been receiving more attention from researchers since social and multimedia networks are becoming more popular nowadays. Recently published methods have yielded quite good results in terms of accuracy but have also proven to be ineffective in real-time applications because the models were too complicated. In this paper, we propose a lightweight model that can classify both age and gender. The number of parameters used in this model is 5 times less than existing models. Experiment results show that the accuracy of the proposed method is equivalent to state-of-the-art methods, while the speed of age and gender recognition decreases by 4 times on the Audience benchmark.

THU-AM1-SS3.5: IoT-based Predictive Maintenance for Smart Manufacturing Systems

Chee Him Leong, Yong Poh Yu and Wah Pheng Lee

Tunku Abdul Rahman University College

Manufacturers have been practicing traditional preventive maintenance for many years. However, it is not cost-effective. To avoid ineffective maintenance routine and costs, manufacturers can leverage Industrial IoT and data science. This paper presents a method to optimize the manufacturing processes by using IoT-based predictive maintenance. It illustrates how an IIoT solution can be used to predict a manufacturing defect. The data is collected from multiple smart sensors stored on this welding machine. It is monitored using Statistical process control methods. Machine learning algorithms are applied to reveal hidden correlations in the data sets and detect abnormal data patterns. The recognized data patterns are then reflected in predictive models, classification approaches are used to identify the type of manufacturing processes, namely normal and welding problem. The variables that contribute the most to the failure are identified.

THU-AM1-SS4
Technologies for A Maximized Experience of Multi-dimensional Content: from 2D, 3D Modeling to An Objective Assessment

Time: Thursday, Nov 21, 10:20-12:00

Place: A5

Chairs: Sanghoon Lee, Chia-Hung Yeh

THU-AM1-SS4.1: Physical parameter prediction by embedding human perceptual parameter for 3D garment modeling

Seongmin Lee, Woojae Kim, Sewoong Ahn, Jaekyung Kim and Sanghoon Lee

Yonsei University

To model garments into a virtual environment, it is crucial topredicting the physical parameters of the simulated model.However, it is troublesome for a user or technical directorto intuitively reflect their aesthetic intention using physicalparameters. In this paper, we propose a framework that pre-dicts various physical parameters (e.g.,stretch resistance,bend resistance, ...) by embedding human perceptual param-eters (e.g.,wrinkly, stretchy, ...) in multi-task learning (MTL)perspective. By predicting both physical and perceptual pa-rameters, we can effectively solve this problem, and can givean important cue to model a 3D garment maximizing users' visual presence. Furthermore, by taking a class activationmapping method, our model seeks the intermediate visual un-derstanding of physical and perceptual parameters. Throughthe rigorous experiments, we demonstrate that the predictedphysical and perceptual parameters agree with subjectivevalues

THU-AM1-SS4.2: Generic Video-Based Motion Capture Data Retrieval

Zifei Jiang, Zhen Li, Wei Li, Xueqing Li and Jingliang Peng

Shandong University

In this work we propose a novel and generic scheme for retrieval of motion capture (MoCap) data given a video query. We reconstruct skeleton animations from video clips by a convolutional neural network for 3-dimensional human pose estimation to narrow the gap between videos and MoCap data. A statistical motion signature is computed to extract both morphological and kinematic characteristics from the skeleton animations and the MoCap sequences. This as well ensures that the proposed scheme works on MoCap data with arbitrary skeleton structures. The retrieval is achieved by computing and sorting the distances between the motion signature of the query and those of the MoCap sequences which are pre-computed and stored in the MoCap database. For experimental evaluation, we respectively record a video dataset and capture a MoCap dataset with different performers, and conduct video-based MoCap data retrieval on them. Experimental results demonstrate the effectiveness of the proposed scheme.

THU-AM1-SS4.3: A Lightweight and Robust Face Recognition Network on Noisy Condition

Lulu Guo, Huihui Bai and Yao Zhao

Institute of information science Beijing jiaotong University

Recently, deep learning has a significant breakthrough in face recognition research. Using the state-of-art convolutional neural network (CNN) model is continually improving the accuracy of recognition. However, it is difficult that the large CNN models deploy on mobile phones or embedded devices with limited computation resources and memory. At the same time, these face recognition networks show low performance in the complex environment, such as noise, shadow, illumination and so on. To address these problems, we propose a lightweight and robust face recognition network (LD-MobileFaceNet) to improve the traditional MobileFaceNet in noisy environment. In this paper, an efficient and flexible denoising block is proposed, which is an independent module to apply in MobileFaceNet. The proposed denoising block uses non-local means algorithm to denoise features that are extracted by convolutional layers. With the residual connection and the 1×1 convolution, it can remain more information and be combined with any layers in MobileFaceNet. Furthermore, we set fewer bottleneck layers, replace PReLU with swish nonlinearity to compensate for the loss accuracy. The experimental results demonstrate that LD-MobileFaceNet with swish is 21.35% more accurate on noisy LFW dataset while reducing parameters by 25% compared to MobileFaceNet.

THU-AM1-SS4.4: Deep Learning Approach to Video Frame Rate Up-Conversion Using Bilateral Motion Estimation

Junheum Park, Chul Lee and Chang-Su Kim

Korea University, Dongguk University

We propose a deep learning-based frame rate up-conversion algorithm using bilateral motion estimation, which aims at generating an intermediate frame. We first estimate the bilateral motion fields using the convolutional neural network (CNN). Also, we approximate the intermediate bi-directional motion fields, assuming the linear motions between the successive frames. Finally, we develop the synthesis network to produce the intermediate frame by merging the warped frames obtained using the estimated motion fields. Experimental results demonstrate that the proposed algorithm generates high-quality intermediate frames with less visual artifacts on the challenging sequences with large motions and occlusion, and outperforms the state-of-the-art algorithms.

THU-AM1-SS4.5: 3D Reconstruction using HDR-based SLAM

Chia-Hung Yeh, Min-Hui Lin and Wei-Chieh Lu

National Sun Yat-sen University, National Taiwan Normal University

3D reconstruction is the key technology to emerging technologies such as smart robotics, VR/AR/XR and autonomous driving. To enhance the robustness of our proposed 3D reconstruction system, the HDR-based SLAM is adopted in the camera pose estimation step to improve the qualitative result of geometric reconstruction. The proposed HDR-based SLAM uses the pre-calibrated inverse camera response function (CRF) to map a single RGB image into a radiance map. To exclude the influence of exposure time, normalized radiance maps independent of exposure time are used during tracking. Since ORB feature matching is the basic element of tracking and mapping in our system, the ORB descriptor patch is re-trained especially for normalized radiance maps. Experimental results have shown good performance of our system under challenging low-light environment, which helps expand the applicability of 3D reconstruction system.

THU-AM1-SS5
Recent Trends in Computational Intelligence

Time: Thursday, Nov 21, 10:20-12:00

Place: A6

Chairs: Chern Hong Lim, Mei Kuan Lim

THU-AM1-SS5.1: Using Machine Learning Applied to Radiomic Image Features for Segmenting Tumour Structures

Henry Clifton, Alanna Vial, Andrew Miller, Christian Ritz, Matthew Field, Lois Holloway, Montserrat Ros, Martin Carolan and David Stirling

School of Electrical, Telecommunications and Computer Engineering, University of Wollongong, Centre for Medical Radiation Physics, University of Wollongong, South West Clinical School, University of New South Wales

Lung cancer (LC) was the predicted leading cause of Australian cancer fatalities in 2018 (9,198 deaths). Non-Small Cell Lung Cancer (NSCLC) tumours with larger amounts of heterogeneity have been linked to a worse outcome. Medical imaging is widely used in oncology and non-invasively collects data about the whole tumour. The field of radiomics uses these medical images to extract quantitative image features and promises further understanding of the disease at the time of diagnosis, during treatment and in follow up. It is well known that manual and semi-automatic tumour segmentation methods are subject to inter-observer variability which reduces confidence in the treatment region and extent of disease. This leads to tumour under- and over-estimation which can impact on treatment outcome and treatment-induced morbidity. This research aims to use radiomic features centred at each pixel to segment the location of the lung tumour on Computed Tomography (CT) scans. To achieve this objective, a Decision Tree (DT) model was trained using sampled CT data from eight patients. The data consisted of 25 pixel-based texture features calculated from four Gray Level Matrices (GLMs) describing the region around each pixel. The model was assessed using an unseen patient through both a confusion matrix and interpretation of the segment. The findings showed that the model accurately (AUROC = 83.9%) predicts tumour location within the test data, concluding that pixel based textural features likely contribute to segmenting the lung tumour. The prediction displayed a strong representation of the manually segmented Region of Interest (ROI), which is considered the ground truth for the purpose of this research.

THU-AM1-SS5.2: Computational Intelligence-based Real-time Lane Departure Warning System Using Gabor Features

Ricky Sutopo, Ting Yau Teo, Joanne Mun-Yee Lim and Koksheik Wong

Monash University

Lane detection and lane departure warning are crucial parts of Advanced Driver Assistance Systems (ADAS), which is designed to increase general safety on the road. This paper proposes a novel approach for lane detection, which boosts the accuracy of lane departure warning system, specifically on the highway and the urban road under sunny condition, using Gabor Filter and other image processing algorithms. Gabor Filter is implemented to enhance the directionality of lane marking patterns. This filter automatically eliminates shadows, road marker and other non-related objects due to lack of directionality. Canny edge detection is applied to extract the edges of lane marking and enhance the lane marking pattern. Lastly, Probabilistic Hough Transform is applied to identify the correct left and right lane candidates on the road. We have also included lane departure warning to alert the driver when the vehicle is veered out of the lane. We showed that our framework is capable of real-time implementation using a Raspberry pi 3B to achieve 93% for lane detection and 95% for lane departure warning with 20 frames per second (fps) and only 75% Central Processing Unit (CPU) utilization.

THU-AM1-SS5.3: Optimising Search Operations with Swarm Intelligence

Chung Hou Ng, Wern Han Lim and Mei Kuan Lim

Monash University

The challenge in search and rescue is to identify the most optimal paths when searching the entire location. That difficulty is further intensified by the passing of time and an unknown complex terrain environment. Many of the existing algorithms such as Depth First Search are focused on having only a single agent to sweep through the location. Drawing inspiration from the self-organisation mechanism and the emergence of global behaviour through local interactions between agents in swarm intelligence; this study utilises the information exchange between agents in the swarm to navigate a search area effectively. We demonstrate the proposed swarm-based search method and compare its performance against the existing path finding algorithm breadth first search (BFS) on terrains with different complexity. The simulation results show that the proposed swarm intelligence inspired algorithm is able to reach upwards of 95% the effectiveness of BFS with at most, approximately one-fifth the cost of BFS. This study also provides evidence and analysis on proving the misconception based on logical sense, that an increase in the number of agents in swarm based algorithm in the context of search operations would always result in an increase of efficiency of the traversal.

THU-AM1-SS5.4: Gun Detection in Surveillance Videos using Deep Neural Networks

Jun Yi Lim, Md Istiaque Al Jobayer, Vishnu Monn Baskaran, Joanne Mun Yee Lim, Kok Sheik Wong and John See

Monash University, Monash Universtiy Malaysia, Multimedia University

The ongoing epidemic of gun violence worldwide has compelled various agencies, businesses and consumers to deploy closed-circuit television (CCTV) surveillance cameras in attempt to combat this epidemic. An active-based CCTV system extends this platform to autonomously detect potential firearms within a video surveillance perspective. However, autonomously detecting a firearm across varying CCTV camera angles, depth and illumination represents an arduous task which has seen limited success using existing deep neural networks models. This challenge is in part due to the lack of available contextual hand gun information from CCTV images, which remains unresolved. As such, this paper introduces a novel large scale dataset of hand guns which were captured using a CCTV camera. This dataset serves to substantially improve the state-of-the-art in representation learning of hand guns within a surveillance perspective. The proposed dataset consist of 250 recorded CCTV videos with a total of 5500 images. Each annotated CCTV image realistically captures the presence of a hand gun under 1) varying outdoor and indoor conditions, and 2) different resolutions representing variable scales and depth of a gun relative to a camera' s sensor. The proposed dataset is used to train a single-stage object detector using a multi-level feature pyramid network (\ie\ M2Det). The trained network is then validated using images from the UCF crime video dataset which contains real-world gun violence. Experimental results indicate that the proposed dataset increases the average precision of gun detection at different scales by as much as 18% when compared to existing approaches in firearms detection.

THU-AM1-SS5.5: Interpreting Abnormality of a Complex Static Scene using Generative Adversarial Network

Mahamat Moussa and Chern Hong Lim

Monash University

Anomaly detection remains a difficult task in the field of computer vision and image processing. Although several studies have been done to address this challenge, most of these studies focused on analyzing the temporal features to determine abnormality. Examples of temporal features include behavioral changes and new object appears in the target scene. In this paper, we are interpreting abnormality from a new perspective, which is static and complex image scene that focused on the same subject using generative adversarial networks (GANs). Our interpretation of abnormality in such image intended to test two research hypotheses: 1) whether GANs can capture the cognitive features of abnormality from within a complex scene. 2) whether GANs can be used to generate more reliable datasets of abnormal scene. In this work, we selected airplane as our case study, where we defined abnormality as an airplane that involved in an accident, which could be either crashed or falling. And normal otherwise, which could be flying or landed airplane. A custom dataset consist of two classes; normal and abnormal has been collected for this work. We augmented each class to double of its size using GANs, and created three different combinations of datasets (DS1, DS2, DS3) to test our hypotheses. We conducted classification using different supervised machine learning algorithms on each dataset with three different interpretations 1) with original images; 2) with applying PCA, and 3) with applying Local Binary Pattern. The overall results showed that GANs possess the capability of generating images that capture the abnormality features from the static complex scene.

THU-AM1-SS6
Multi-source Data Processing and Analysis: Models, Methods and Applications

Time: Thursday, Nov 21, 10:20-12:00

Place: A7

Chairs: Ping Han, Qiuping Jiang, Runmin Cong, Chongyi Li

THU-AM1-SS6.1: Median based Multi-label Prediction by Inflating Emotions with Dyads for Visual Sentiment Analysis

Tetsuya Asakawa and Masaki Aono

Computer Science and Engineering, Toyohashi University of Technology

Visual sentiment analysis investigates sentiment estimation from images and has been an interesting and challenging research problem. Most studies have focused on estimating a few specific sentiments and their intensities. Multi-label sentiment estimation from the images has not been sufficiently investigated. The purpose of this research is to accurately estimate the sentiments as a multi-label multi-class problem from given images that evoke multiple different emotions simultaneously. We first introduce the emotion inflation method from 6 emotions defined by Emotion6 dataset into 13 emotions (which we call 'Transf13') by means of emotional dyads. We then perform multi-label sentiment analysis using the emotion-inflated dataset, where we propose a combined deep neural network model which enables inputs to come from both hand-crafted features (e.g. BoVW (Bag of Visual Words) features) and CNN features. We also introduce a median-based multi-label prediction algorithm,in which we assume that each emotion has a probability distribution. In other words, after training of our deep neural network, we predict the existence of an evoked emotion for a given unknown image if the intensity of the emotion is larger than the median of the corresponding emotion. Experimental results demonstrate that our model outperforms existing state-of-the-art algorithms in terms of subset accuracy.

THU-AM1-SS6.2: Action Recognition using Convolutional Neural Networks with Joint Supervision

Yupeng Li, Yuxiao Wang, Yongfeng Jiang and Liang Zhang

Civil Aviation University of China, Public Security Bureau in Wenzhou

Mapping the depth video into an optimally representation in two-dimensional space are of vital importance for depth video based human action understanding. Meanwhile, such representation will lost some useful information inevitably, a feature learning approach not only separable but also discriminative are essential for action recognition task from such representation. This paper presents a new method for action recognition base on convolutional neural networks with joint supervision which shares the merits of both representation as mentioned above and convolutional neural networks. The advantages of our method come from (i) The whole procedure of our method is done automatically no matter the generation of representation or deeply feature learned; (ii) The deeply feature using the proposed deep architectures to learned has high discriminative capacity to improve the accuracy of action recognition effectively compared with handcrafted features. We conduct experiments on two challenging datasets: MSRDailyActivity3D and SYSU 3D HOI. Experimental results show that our method outperform previous methods based on hand-crafted features. Our method also achieves superior performance to the state-of-the-art on these datasets.

THU-AM1-SS6.3: A Study of Perceptual Quality Assessment for Stereoscopic Image Retargeting

Zhenqi Fu, Yan Yang, Feng Shao and Xinghao Ding

Ningbo University, Xiamen University

Subjective and objective perceptual quality assessment of stereoscopic retargeted images is a fundamentally important issue in stereoscopic image retargeting (SIR) which has not been deeply investigated. Here, a stereoscopic image retargeting quality assessment (SIRQA) database is proposed to study the perceptual quality of different stereoscopic retargeted images. To construct the database, we collect 720 stereoscopic retargeted images generated by eight representative SIR methods. The perceptual quality (mean opinion scores, MOS) of each stereoscopic retargeted image is subjectively rated by 30 viewers. For objective assessment, several publicly available quality evaluation metrics are tested on the database. Experimental results show that there is a large room for improving the accuracy of objective quality assessment in SIRQA by comprehensively considering geometric distortion, content loss and stereoscopic perceptual quality.

THU-AM1-SS6.4: Infrared Pedestrian Detection with Converted Temperature Map

Yifan Zhao, Jingchun Cheng, Wei Zhou, Chunxi Zhang and Xiong Pan

Beihang University, Tsinghua University, University of Science and Technology of China

Infrared pedestrian detection aims to detect persons in outdoor thermal images. It shows a unique advantage in dark environment or bad weather compared to daytime visible images (the RGB image). Most current methods treat infrared detection the same way as with visible images, e.g. regarding the infrared image as a special gray-scale visible image. In this paper, we tackle this problem with more emphasis on the underlying temperature information in infrared images. We build an image-temperature transformation formula based upon infrared image formation theory, which can convert infrared image into temperature map with the prior of pedestrian pixel-temperature value. The whole detection process follows a two-stage manner. In the first stage, we use a common detector which treats the infrared image as the gray-scale visible image to provide primary detection results and a pedestrian position prior (the highest-confidence pedestrian detection box in each image). In the second stage, we convert infrared images into corresponding temperature maps and train a temperature net for detection. The final results consist of both the primary detection and the temperature net outputs, detecting pedestrians with characteristics in both image and temperature domain. We show that the converted temperature image is less affected by environmental factors, and that its detector shows amazing complementary ability with the primary detector. We carry out extensive experiments and analysis on two public infrared datasets, the OTCBVS dataset and the FLIR dataset; and demonstrate the effectiveness of incorporating temperature maps.

THU-AM1-SS6.5: A Fast and Accurate Cluster Center Initialization Algorithm for PolSAR Superpixel Segmentation

Binbin Han, Ping Han and Zheng Cheng

Civil Aviation University of China

Locate iterative clustering (LIC) gets good performance in superpixel segmentation for PolSAR images. However, there are also two problems in its cluster center initialization. One is that there is a lot of computational redundancy in calculating edge map, which slows down the segmentation speed. The other is that the calculated edge is much wider than the real one, which makes the fine-tuning of center seeds in 33 window not suitable. To solve these problems, a fast and accurate cluster center initialization algorithm for PolSAR superpixel segmentation is proposed in this paper. For the first problem, integral graph is introduced to eliminate the redundancy. For the second problem, a more reasonable window size is selected to conduct fine-tuning. Experiments with measured AIRSAR data sets demonstrate the speed of the proposed edge map calculation method increases 28 times and the center seed is farther away from the edge than before. Furthermore, LIC gets better superpixel segmentation results with the new initialization algorithm.

THU-AM1-SS7
Second Language Speech Perception and Production[2]

Time: Thursday, Nov 21, 10:20-12:00

Place: A8

Chairs: Ying Chen, Jian Gong

THU-AM1-SS7.1: Comparing native Chinese listeners' Speech Reception Thresholds for Mandarin and English consonants

Jian Gong, Yameng Yu, William Bellamy, Feng Wang, Xiaoli Ji and Zhenzhen Yang

Jiangsu University of Science and Technology

The presence of noise can greatly affect listeners' speech perception. Previous studies demonstrated that non-native listeners' speech perception performance reduce more than natives' in noise conditions. Most of previous studies focused on the effects of different noise types on non-native speech perception, and using fixed signal to noise level in different perception tasks. However, the masking effect of noise may be different on individual speech sounds, therefore leave an incomplete picture of non-native speech perception in noise. The current study applied an adaptive procedure to dynamically adjust the signal to noise level to measure listeners' Speech Reception Threshold (SRT) in noise conditions. More specifically, a group of native Chinese listeners' SRTs for Mandarin and English consonants in Speech Shaped Noise were measured and compared. The results showed that Chinese listeners' mean SRT for Mandarin consonants was 3.6dB lower than that for English consonants, indicating a general native language advantage. However, detailed analysis revealed an unexpected result, that is, the mean SRT for the 5 most noise tolerable consonants in Mandarin was 2.6dB higher than that in English. This result suggests that, non-native speech perception in noise may not always be more difficult than the native one. The acoustic features of different sounds could affect their intelligibility in noise conditions.

THU-AM1-SS7.2: Prosodic Realization of Focus in English by Bidialectal Mandarin Speakers

Jiajing Zhang, Ying Chen and Jie Cui

Nanjing University of Science and Technology

This study was designed to explore the prosodic patterns of focus in English by bidialectal Mandarin speakers. One learner group speaks Nanjing Mandarin as first dialect (D1) and standard Mandarin as second dialect (D2), and the other group speaks Changchun Mandarin as D1 and standard Mandarin as D2. This paper compares their prosodic outcome of focus realization in English in a production experiment. Results indicate that both Changchun and Nanjing bidialectal speakers produced clear in-focus expansion of duration, pitch and intensity and post-focus compression (PFC) of pitch and intensity; however, their PFC in English was not found in a nativelike manner. The two learner groups produced statistically similar patterns of prosodic focus in L2 English though their D1s are different dialects of Mandarin. These findings provide further support for the claim that PFC cannot be easily transferred cross-linguistically despite its existence in both dialects of learners' L1 and their L2.

THU-AM1-SS7.3: World Englishes and Prosody: Evidence from the Successful Public Speakers

Yating Cao and Hua Chen

Nanjing University

As English has become the most common vehicle in global communication, disentangling intelligibility becomes an urgent issue in English pronunciation teaching and learning. Previous studies put more emphasis on segmentals and suggest some core features for maintaining intelligibility, but there is little concern on prosody. Based on the Intonation Theory proposed by Halliday, the present study examined the shared prosodic features of 15 successful public speakers under the World Englishes paradigm. Results showed that appropriate pause position, pause duration, tone choices, tonicity and pre-tonic stress, speech rate, and clear enunciation has worked together to contribute to the effective delivery of information. The findings not only provide a better understanding towards the role of prosodic features in intelligibility and effective communication, but also have pedagogical implications for English teaching and learning.

THU-AM1-SS7.4: An Experimental Study on English Majors Weak Form Productions of Prepositions

Jiangbo Zhang, Aijun Li and Na Zhi

Chinese Academy of Social Science, Institute of Linguistics, CASS, Capital Normal University

This study compares the acoustic cues of preposition production between Chinese English learners and American speakers. Six prepositions "at, for, from, in, of and to" with high frequencies in COCA corpus were selected as target words, and then were put into 3 common verb phrases (VP). The target sentences had a consistent structure of Subject+VP+Object+AP, and each was recorded into 6 information structures. The vowel formant, duration, pitch movement of prepositions were analyzed within and across two speaker groups, from which we concluded: Learners are able perceive the differences between weak forms and their normal counterparts in terms of duration, but due to the absence of lax-tense contrast in their mother tongue Chinese, their reduced vowels are still longer than that of the native speakers. As to the vowel quality, not like the native speakers who employ centralization to realize the weak forms of vowels in prepositions, learners tend to borrow from and reduce their L1 vowels, which makes their weak form production either more advanced or backward with a larger mouth openness than the native speakers. There is an exception that the weak form of "of" in well acquired by learners, as the vowel /e/ is present their L1 vowel system. In addition, the pitch movements of learners' weak form production are also deviant from the native speakers. Although they follow the same H-L pattern within the duration of the prepositions, still it is observed that learners' pitch contour on syllables of the reduced prepositions in post-nuclear position is realized by transferring the neutral tone features due to the obvious pitch reset and the sharply falling trend. For those positioned before the nuclear of the sentences, a continuous pitch movement from the preceding vowel to the reduced one in the prepositions is observed from learners' production, which is similar to what the manner native speakers follow to do weak form realization. It might be because of the fact that only the latter syllable in dissyllabic words are able to go with a neutral tone in Chinese, and reduced vowels in pre-nuclear position are regarded as new phonetic environment to which learners never have never been exposed. For these reasons, it would be easier for them to notice the difference between L1 neutral tone features and the L2 weak form representations, which further leads to the relatively good acquisition. In a nutshell, duration and pitch variation are the only phonetic cues learners use to distinguish the weak forms from the normal ones of the English prepositions, and their productions so far are still in a deviant manner.

THU-AM1-SS7.5： Oral Motor Exercises For CSL Learners to Master Productions of Retroflex And Non-Retroflex Consonants

Yixin Zhang and Jinsong Zhang

Beijing Language and Culture University

Retroflex and non-retroflex consonants inMandarin (z[ts]、c[ts']、s[s]、zh[ts]、ch [ts']、sh[s]) are difficult to produce for many Chinese as a second language (CSL) learners with different mother tongues. Aiming at developing an efficient way to help them to master the sounds, this study adopts the idea of Oral Motor Exercise (OME) to devise a speech production training method. The method consists of two steps: the first is non-speech OME, in which muscle groups of jaw, lips and tongue, essential to the production of target consonants, are intentionally exercised; the second is speech OME, in which monosyllables with target consonants are practiced. 30 participants took part in the experiment. The results showed that both non-speech and speech OMEs could effectively reduce subjects' consonant errors in a short time. When the two steps were accomplished, the trainee achieved the best performances. These indicated the effectiveness of the proposal.

Author Index