



Provisional Programme

09/07/24

The programme is subject to changes.

Monday 02/09

Time Slot 1

Oral Session: Audio Event Detection and Classification 1

A5-01

Location: Acesso

246 Scaling up masked audio encoder learning for general audio classification

Heinrich Dinkel (Xiaomi); Zhiyong Yan (Xiaomi); Yongqing Wang (xiaomi); Junbo Zhang (Xiaomi); Yujun Wang (xiaomi); Bin Wang (Xiaomi AI Lab)

591 Low-Complexity Acoustic Scene Classification Using Parallel Attention-Convolution Network

Yanxiong Li (South China University of Technology); Jiaxin Tan (South China University of Technology); Guoqing Chen (South China University of Technology); Jialong Li (South China University of Technology); Yongjie Si (South China University of Technology); Qianhua He (South China University of Technology)

714 MAT-SED: A Masked Audio Transformer with Masked-Reconstruction Based Pre-training for Sound Event Detection

Pengfei Cai (University of Science and Technology of China); Yan Song (USTC); Kang Li (University of Science and Technology of China, National Engineering Research Center of Speech and Language Information Processing.); Haoyu Song (The Australian National University); Ian McLoughlin (Singapore Institute of Technology)

1274 Audio Mamba: Selective State Spaces for Self-Supervised Audio Representations

Sarthak Yadav (Aalborg University); Zheng-Hua Tan (Aalborg University)

1350 Can Synthetic Audio From Generative Foundation Models Assist Audio Recognition and Speech Modeling?

Tiantian Feng (University of Southern California); Dimitrios Dimitriadis (Amazon); Shrikanth Narayanan (USC)

2075 Sound Event Bounding Boxes

Jane Ebberts (Mitsubishi Electric Research Laboratories (MERL)); François G Germain (Mitsubishi Electric Research Laboratories (MERL)); Gordon Wichern (Mitsubishi Electric Research Laboratories (MERL)); Jonathan Le Roux (Mitsubishi Electric Research Laboratories (MERL))

Oral Session: Speech Synthesis: Voice Conversion 1

A7-01

Location: Aegle A

1107 Spatial Voice Conversion: Voice Conversion Preserving Spatial Information and Non-target Signals

Kentaro Seki (The University of Tokyo); Shinnosuke Takamichi (The University of Tokyo); Norihiro Takamune (the University of Tokyo); Yuki Saito ("The University of Tokyo, Japan"); Kanami Imamura (The University of Tokyo); Hiroshi Saruwatari (The University of Tokyo)

1298 Neural Codec Language Models for Disentangled and Textless Voice Conversion

Alan Baade (The University of Texas at Austin); Puyuan Peng (The University of Texas at Austin); David Harwath (The University of Texas at Austin)

1857 DualVC 3: Leveraging Language Model Generated Pseudo Context for End-to-end Low Latency

Streaming Voice Conversion

Ziqian Ning (Northwestern Polytechnical University); Shuai Wang (Shenzhen Research Institute of Big Data, Chinese University of Hong Kong (Shenzhen)); Pengcheng Zhu (Fuxi AI Lab, NetEase Inc.); Zhichao Wang (Northwestern Polytechnical University); Jixun Yao (Northwestern Polytechnical University); Lei Xie (NWPU); Mengxiao Bi (NetEase Fuxi AI Lab)

1941 Towards Realistic Emotional Voice Conversion using Controllable Emotional Intensity

Tianhua Qi (Southeast University); Shiyan Wang (Southeast University); Cheng Lu (Southeast University); Yan Zhao (Southeast University); Yuan Zong (Southeast University); Wenming Zheng (Southeast University)

2351 Fine-Grained and Interpretable Neural Speech Editing

Max Morrison (Northwestern University); Cameron Churchwell (Northwestern University); Nathan Pruyne (Northwestern University); Bryan Pardo (Northwestern University)

2387 FastVoiceGrad: One-step Diffusion-Based Voice Conversion with Adversarial Conditional Diffusion Distillation

Takuhiko Kaneko (NTT Corporation); Hirokazu Kameoka (NTT Communication Science Laboratories, NTT Corporation); Kou Tanaka (NTT Corporation); Yuto Kondo (NTT)

Oral Session: Speaker Diarization 1

A4-01 Location: Aegle B

561 On the Success and Limitations of Auxiliary Network Based Word-Level End-to-End Neural Speaker Diarization

Yiling Huang (Google); Weiran Wang (Google); Guanlong Zhao (Google); Hank Liao (Google); Wei Xia (Google); Quan Wang (Google)

668 EEND-M2F: Masked-attention mask transformers for speaker diarization

Marc Härkönen (Fano Labs); Samuel J Broughton (Fano Labs); Lahiru T Samarakoon (Fano Labs)

764 AFL-Net: Integrating Audio, Facial, and Lip Modalities with a Two-step Cross-attention for Robust Speaker Diarization in the Wild

Yongkang Yin (Peking University); Xu Li (ARC Lab, Tencent); Ying Shan (Tencent); Yuexian Zou (Peking University)

1044 Investigating Confidence Estimation Measures for Speaker Diarization

Anurag Chowdhury (Solventum); Abhinav Misra (Solventum); Mark Fuhs (Solventum); Monika Woszczyna (Solventum)

1146 Exploiting Wavelet Scattering Transform for an Unsupervised Speaker Diarization in Deep Neural Network Framework

Arunav Arya (IIT Jammu); Murtiza Ali (Indian Institute of Technology, Jammu); Karan Nathwani (Indian Institute of Technology, Jammu)

1174 Speakers Unembedded: Embedding-free Approach to Long-form Neural Diarization

Xiang Li (AWS AI Labs); Vivek Govindan (AWS AI Labs); Rohit Paturi (AWS AI Labs); Sundararajan Srinivasan (AWS AI Labs)

Oral Session: Decoding Algorithms

A9-01 Location: Hippocrates

404 Towards Effective and Efficient Non-autoregressive Decoding Using Block-based Attention Mask

Tianzi Wang (The Chinese University of HongKong); Xurong Xie (Institute of Software, Chinese Academy of Sciences); Zhaoqing Li (The Chinese University of Hong Kong); Shoukang Hu (Nanyang Technological University); Zengrui Jin (The Chinese University of Hong Kong); Jiajun Deng (The Chinese University of HongKong); Mingyu Cui (The Chinese University of Hong Kong); Shujie HU (The Chinese University of Hong Kong); Mengzhe GENG (The Chinese University of Hong Kong); Guinan Li (Chinese University of HongKong); Helen Meng (The Chinese University of Hong Kong); Xunying Liu (The Chinese University of Hong Kong)

1349 Contextual Biasing with the Knuth-Morris-Pratt Matching Algorithm

Weiran Wang (Google); Zelin Wu (Google LLC); Diamantino Caseiro (Google); Tsendsuren Munkhdalai (Google LLC); Khe C Sim (Google Inc.); Pat Rondon (Google); Golan Pundak (Google); Gan Song (Google); Rohit Prabhavalkar (Google); Zhong Meng (Google); Ding Zhao (Google); Tara Sainath (Google); Yanzhang He (Google); Pedro J Moreno (Google)

1591 Speed of Light Exact Greedy Decoding for RNN-T Speech Recognition Models on GPU

Daniel Galvez (NVIDIA); Vladimir Bataev (NVIDIA, University of London); Hainan Xu (NVIDIA); Tim Kaldewey (NVIDIA)

1891 E-Paraformer: A Faster and Better Parallel Transformer for Non-autoregressive End-to-End Mandarin Speech Recognition

Kun Zou (PingAn Technology); Fengyun Tan (PingAn Technology); Ziyang Zhuang (Ping An Technology); Chenfeng Miao (Ping An technology); Tao Wei (Ping An Technology); Shaodan Zhai (Coupang); Zijian Li (Georgia Institute Technology); Wei Hu (Ping An Technology); Shaojun Wang (PAII Inc.); Jing Xiao (Ping An Insurance (Group) Company of China)

2222 Text-only Domain Adaptation for CTC-based Speech Recognition through Substitution of Implicit Linguistic Information in the Search Space

TATSUNARI TAKAGI (Toyohashi University of Technology); Yukoh Wakabayashi (Toyohashi University of Technology); Atsunori Ogawa (NTT Corporation); Norihide Kitaoka (Toyohashi University of Technology)

2457 Beam-search SIEVE for low-memory speech recognition

Martino Ciaperoni (Aalto University); Athanasios Katsamanis ("ATHENA R.C., Behavioral Signal Technologies"); Aristides Gionis (KTH Royal Institute of Technology); Panagiotis Karras (University of Copenhagen)

Oral Session: Pronunciation Assessment**A10-01**

Location: Iasso

123 MultiPA: A Multi-task Speech Pronunciation Assessment Model for Open Response Scenarios

Yu-Wen Chen (Columbia university); Zhou Yu (Columbia university); Julia Hirschberg (Columbia University)

459 A Framework for Phoneme-Level Pronunciation Assessment Using CTC

Xinwei Cao (NTNU); Zijian Fan (NTNU); Torbjørn Svendsen (NTNU); Giampiero Salvi (NTNU)

1166 Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task

Nhan Chi Phan (Aalto University); Ekaterina Voskoboini (Aalto University); Tamas Grosz (Aalto University); Mikko Kurimo (Aalto University); Anna Zansen (University of Helsinki); Raili Hilden (University of Helsinki); Maria Kautonen (University of Jyväskylä)

2217 Phonological-Level Mispronunciation Detection and Diagnosis

Mostafa Shahin (UNSW); Beena Ahmed (School of Electrical Engineering and Telecommunications, UNSW Australia)

2297 Pitch-Aware RNN-T for Mandarin Chinese Mispronunciation Detection and Diagnosis

XINTONG WANG (National University of Singapore); Mingqian Shi (National University of Singapore); Ye Wang (National University of Singapore)

2498 Acoustic Feature Mixup for Balanced Multi-aspect Pronunciation Assessment
Heejin Do (POSTECH); Wonjun Lee (POSTECH); Gary Geunbae Lee (Postech)

Oral Session: Acoustic Echo Cancellation

A6-01 Location: Melampus

107 Low Complexity Echo Delay Estimator Based on Binarized Feature Matching

Jacky Gao (Tencent Chengdu Branch); Xiang Su (Tencent Chengdu Branch)

763 SDAEC: Signal Decoupling for Advancing Acoustic Echo Cancellation

Zhao Fei (School of Computer Science, Inner Mongolia University); Jinjiang Liu (College of Computer Science, Inner Mongolia University); Xueliang zhang (Inner Mongolia University)

834 Multi-mic Echo Cancellation Coalesced with Beamforming for Real World Adverse Acoustic Conditions

Premanand Vinayak Nayak (Samsung Research and Development Institute - Bangalore, India); Kamini Sabu (Samsung Research and Development Institute - Bangalore, India); Mahaboob Ali Basha Shaik (Samsung Research and Development Institute - Bangalore, India)

1355 MSA-DPCRN: A Multi-Scale Asymmetric Dual-Path Convolution Recurrent Network with Attentional Feature Fusion for Acoustic Echo Cancellation

Ye Ni (southeast university); Cong Pang (Southeast University); Chengwei Huang (Zhejiang Lab); Cairong Zou (Southeast University)

1414 Interference Aware Training Target for DNN based joint Acoustic Echo Cancellation and Noise Suppression

Vahid Khanagha (Meta); Dimitris Koutsaidis (Meta); Kaustubh Kalgaonkar (Meta); Sriram Srinivasan (Meta)

1957 Efficient Joint Beamforming and Acoustic Echo Cancellation Structure for Conference Call Scenarios
Ofer Schwartz (CEVA Ltd.); Sharon Gannot (Bar-Ilan University)

Oral Session: L2 Speech, Bilingualism and Code-Switching

A1-01 Location: Panacea Amphitheater

Surv-01-2 Second language (L2) speech and perception

Ann Bradlow

418 Towards a better understanding of receptive multilingualism: listening conditions and priming effects

Wei Xue (Saarland University); Ivan Yuen (Saarland University); Bernd Möbius (Saarland University)

551 Characterizing code-switching: Applying Linguistic Principles for Metric Assessment and Development

Jie Chi (University of Edinburgh); Electra Wallington (University of Edinburgh); Peter Bell (University of Edinburgh)

1669 The influence of L2 accent strength and different error types on personality trait ratings

Sarah Wesolek (Leibniz-Centre General Linguistics); Piotr Gulowski (University of Wrocław); Marzena Zygis (Leibniz-

ZAS)

1991 Re-evaluating the word token for bilingual speech processing: The case for Intonation Units
 Rebecca Pattichis (University of California, Los Angeles); Dora L LaCasse (University of MONTana); Rena Torres Cacoullos (Pennsylvania State University)

Poster Session: Neural Network Architectures for ASR 2

A8-P1 Location: Poster Area 1A, Poster Area 1B

243 SEQ-former: A context-enhanced and efficient automatic speech recognition framework
 Qinglin Meng (Mashang Consumer Finance Co.,Ltd.); Min Liu (Mashang Consumer Finance Co.,Ltd.); Kaixun Huang (NWPU); Kun Wei (School of Computer Science, Northwestern Polytechnical University); Lei Xie (NWPU); Zongfeng Quan (Mashang Consumer Finance Co.,Ltd.); Weihong Deng (Mashang Consumer Finance Co.,Ltd.); Quan Lu (Mashang Consumer Finance Co.,Ltd.); Ning Jiang (Mashang Consumer Finance Co., Ltd.); Guoqing Zhao (Mashang Consumer Finance Co., Ltd)

565 Exploring the limits of decoder-only models trained on public speech recognition corpora
 Ankit Gupta (IBM Research); George Saon (IBM); Brian Kingsbury (IBM Research)

619 InterBiasing: Boost Unseen Word Recognition through Biasing Intermediate Predictions
 YU NAKAGOME (LINE WORKS); Michael Hentschel (LINE WORKS Corporation)

768 Lightweight Transducer Based on Frame-Level Criterion
 Genshun Wan (University of Science and Technology of China); Mengzhi Wang (iflytek); Tingzhi Mao (iflytek); Hang Chen (USTC); Zhongfu Ye (University of Science and Technology of China)

870 How Much Context Does My Attention-Based ASR System Need?
 Robert J Flynn (Sheffield University); Anton Ragni (University of Sheffield)

965 Contextual Biasing Speech Recognition in Speech-enhanced Large Language Model
 Xun Gong (Shanghai Jiaotong University); Anqi Lv (AntGroup); Zhiming Wang (AntGroup); Yanmin Qian (Shanghai Jiao Tong University)

994 Exploring the Capability of Mamba in Speech Applications
 Koichi Miyazaki (CyberAgent, Inc.); Yoshiki Masuyama (Tokyo Metropolitan University)

2029 Rich speech signal: exploring and exploiting end-to-end automatic speech recognizers'ability to model hesitation phenomena
 Vincenzo Norman Vitale (University of Naples ""Federico II""; Loredana Schettino (University of Naples ""Federico II""; Francesco Cutugno (University of Naples Federico II)

2374 Transmitted and Aggregated Self-Attention for Automatic Speech Recognition
 Tian-Hao Zhang (University of Science and Technology Beijing); Xinyuan Qian (USTB); Feng Chen (EEasy Technology Co. LTD); Xu-Cheng Yin (University of Science and Technology Beijing)

2384 Multi-Conformer: Extending Conformer with Multiple Convolution Kernels
 Darshan Deepak Prabhu (Indian Institute of Technology, Bombay); Yifan Peng (Carnegie Mellon University); Preethi Jyothi (Indian Institute of Technology Bombay); Shinji Watanabe (Carnegie Mellon University)

2489 Robust Voice Activity Detection using Locality-Sensitive Hashing and Residual Frequency-Temporal Attention
 Shu Li (Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology); Peng Zhang (Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology); Ye Li (Shandong Provincial Key Laboratory of Computer Networks, Shandong Computer Science Center, Qilu University of Technology)

Poster Session: Speech and Audio Analysis and Representations

A5-P1-A

Location: Poster Area 2A

29 M2D-CLAP: Masked Modeling Duo Meets CLAP for Learning General-purpose Audio-Language Representation

Daisuke Niizumi (NTT Corporation); Daiki Takeuchi (NTT Corporation); Yasunori Ohishi (NTT Corporation); Noboru Harada (NTT); Masahiro Yasuda (NTT); Shunsuke Tsubaki (Doshisha University); Keisuke Imoto (Doshisha University)

74 AlignNet: Learning dataset score alignment functions to enable better training of speech quality estimators

Jaden Pieper (Institute for Telecommunication Sciences); Stephen Voran (Institute for Telecommunication Sciences)

245 Audio Fingerprinting with Holographic Reduced Representations

Yusuke Fujita (LY Corporation); Tatsuya Komatsu (LY Corporation)

396 Reduce, Reuse, Recycle: Is Perturbed Data Better than Other Language Augmentation for Low Resource Self-Supervised Speech Models

ASAD ULLAH (University College Dublin); Alessandro Ragano (University College Dublin); Andrew Hines (University College Dublin)

1805 YOLOPitch: A Time-Frequency Dual-Branch YOLO Model for Pitch Estimation

Xuefei Li (Xinjiang University); Hao Huang (Xinjiang University); Ying Hu (Xinjiang University); Liang HE (Tsinghua University); 家宝张 (新疆大学); 玉奕王 (Xinjiang University)

1863 MINT: Boosting Audio-Language Model via Multi-Target Pre-Training and Instruction Tuning

Hang Zhao (ByteDance Inc.); Yifei Xin (Peking University); Zhesong Yu (Bytedance AI Lab); Bilei Zhu (ByteDance AI Lab); Lu Lu (Bytedance); Zejun Ma (Bytedance)

2203 RAST: A Reference-Audio Synchronization Tool for Dubbed Content

David Meyer (ETH Zürich); Eitan Abecassis (Disney Entertainment & ESPN Technology); Clara Fernandez (Disney Research); Christopher Schroers (DisneyResearch|Studios)

Poster Session: Acoustic Event Detection and Classification 2

A5-P1-B

Location: Poster Area 2B

145 MFF-EINV2: Multi-scale Feature Fusion across Spectral-Spatial-Temporal Domains for Sound Event Localization and Detection

Da Mu (Beijing University of Posts and Telecommunications); Zhicheng Zhang (Beijing University of Posts and Telecommunications); Haobo Yue (Beijing University of Posts and Telecommunications)

216 Diversifying and Expanding Frequency-Adaptive Convolution Kernels for Sound Event Detection

Hyeonuk Nam (KAIST); Seong-Hu Kim (KAIST); Deokki Min (Korea Advanced Institute of Science and Technology (KAIST)); Junhyeok Lee (Supertone Inc.); Yong-Hwa Park (Kaist)

573 Stream-based Active Learning for Streaming Anomalous Sound Detection in Machine Condition Monitoring

Ho Tuan Vu (Hitachi, Ltd.); Kota Dohi (Hitachi Ltd.); Yohei Kawaguchi (Hitachi, Ltd.)

1205 Sound of Traffic: A Dataset for Acoustic Traffic Identification and Counting

Shabnam Ghaffarzadegan (BOSCH Research North America); Luca Bondi (Bosch Research); Wei-Cheng Lin (Bosch)

Research); Abinaya Kumar (BOSCH Research North America); Ho-Hsiang Wu (Bosch Research); Hans-Georg Horst (Robert Bosch GmbH); Samarjit Das (Bosch Research)

1703 FakeSound: Deepfake General Audio Dataset

Zeyu Xie (Shanghai Jiao Tong University); Baihan Li (Shanghai Jiao Tong University); Xuenan Xu (Shanghai Jiao Tong University); Zheng Liang (Shanghai Jiao Tong University); Mengyue Wu (Shanghai Jiao Tong University); Kai Yu (Shanghai Jiao Tong University)

1761 AnoPatch: Towards Better Consistency in Machine Anomalous Sound Detection

Anbai Jiang (Tsinghua University); Bing Han (Shanghai Jiao Tong University); zhiqiang lv (HUAKONG AI Plus); Yufeng Deng (Huakong AI Plus); Wei-Qiang Zhang (Tsinghua University); Xie Chen (Shanghai Jiaotong University); Yanmin Qian (Shanghai Jiao Tong University); Jia Liu (Tsinghua University); Pingyi Fan (Tsinghua University)

2285 Improving Audio Classification with Low-Sampled Microphone Input: An Empirical Study Using Model Self-Distillation

Dawei Liang (UT Austin); Alice Zhang (UT Austin); David Harwath (The University of Texas at Austin); Edison Thomaz (The University of Texas at Austin)

Poster Session: Spoken Language Processing

A12-P1-A Location: Poster Area 3A

100 Query-by-Example Keyword Spotting Using Spectral-Temporal Graph Attentive Pooling and Multi-Task Learning

ZHENYU WANG (UTD); Shuyu Kong (Meta); Li Wan (Meta); Biqiao Zhang (Meta); Yiteng Huang (Meta Platforms); Mumin Jin (MIT); Ming Sun (Meta); Xin Lei (Meta); Zhaojun Yang (Meta)

229 Relational Proxy Loss for Audio-Text based Keyword Spotting

Youngmoon Jung (Samsung Research); Seungjin Lee (Samsung Research); Joon-Young Yang (Samsung Research); Jaeyoung Roh (Samsung Research); Chang Woo Han (Samsung Reserch); Hoon-Young Cho (Samsung Research)

706 CTC-aligned Audio-Text Embedding for Streaming Open-vocabulary Keyword Spotting

Sichen Jin (Samsung); Youngmoon Jung (Samsung Research); Seungjin Lee (Samsung Research); Jaeyoung Roh (Samsung Research); Chang Woo Han (Samsung Reserch); Hoon-Young Cho (Samsung Research)

789 Text-aware Speech Separation for Multi-talker Keyword Spotting

Haoyu Li (Shanghai Jiao Tong University); Baochen Yang (Shanghai Jiao Tong University); Yu Xi (Shanghai Jiao Tong University); Linfeng Yu (Shanghai Jiao Tong University); Tian Tan (Shanghai Jiao Tong University); Hao Li (AISpeech Ltd, Suzhou China); Kai Yu (Shanghai Jiao Tong University)

1194 OWSM v3.1: Better and Faster Open Whisper-Style Speech Models based on E-Branchformer

Yifan Peng (Carnegie Mellon University); Jinchuan Tian (Carnegie Mellon University); William Chen (Carnegie Mellon University); Siddhant Arora (Carnegie Mellon University); Brian Yan (Carnegie Mellon University); Yui Sudo (Honda Research Institute Japan); Muhammad Dr. Shakeel (Honda Research Institute Japan); Kwanghee Choi (Carnegie Mellon University); Jiatong Shi (Carnegie Mellon University); Xuankai Chang (Carnegie Mellon University); Jee-weon Jung (Carnegie Mellon University); Shinji Watanabe (Carnegie Mellon University)

1342 Language-Universal Speech Attributes Modeling for Zero-Shot Multilingual Spoken Keyword Recognition

Hao Yen (Georgia Institute of Technology); Pin-Jui Ku (Georgia Institute of Technology); Sabato M Siniscalchi (Università degli Studi di Palermo); Chin-hui Lee (Georgia Institute of Technology)

1664 Adding User Feedback To Enhance CB-Whisper

Raul Monteiro (Priberam Informática S. A.)

Poster Session: Spoken Machine Translation 2

A12-P1-B

Location: Poster Area 3B

490 Towards Speech-to-Pictograms Translation

Cécile Macaire (Université Grenoble Alpes); Chloé Dion (Université Grenoble Alpes); Didier Schwab (Université Grenoble Alpes); Benjamin Lecouteux (University Grenoble Alpes (UGA)); Emmanuelle Esperança-Rodier (Université Grenoble Alpes)

759 Parameter-Efficient Adapter Based on Pre-trained Models for Speech Translation

Nan Chen (Inner Mongolian University); Yonghe Wang (Inner mongolia university); Feilong Bao (Inner Mongolia University)

903 Translating speech with just images

Dan Oneata (Politehnica University of Bucharest); Herman Kamper (Stellenbosch University)

1088 ZeroST: Zero-Shot Speech Translation

Sameer Khurana (Mitsubishi Electric Research Lab); Chiori Hori (Mitsubishi Electric Research Laboratories (MERL)); Antoine Laurent (Le Mans University); Gordon Wichern (Mitsubishi Electric Research Laboratories (MERL)); Jonathan Le Roux (Mitsubishi Electric Research Laboratories (MERL))

1507 Soft Language Identification for Language-Agnostic Many-to-One End-to-End Speech Translation

Peidong Wang (Microsoft); JIAN XUE (Microsoft Corporation); Jinyu Li (Microsoft); Junkun Chen (Microsoft); Aswin Shanmugam Subramanian (Microsoft)

1759 Navigating the Minefield of MT Beam Search in Cascaded Streaming Speech Translation

Rastislav Rabatin (Meta); Ernie Chang (Meta Inc.); Frank Seide (Meta Platforms, Inc.)

2109 Wave to Interlingua: Analyzing Representations of Multilingual Speech Transformers for Spoken Language Translation

Badr M Abdullah (Saarland University); Mohammed Maqsood Shaik (Saarland University); Dietrich Klakow (Saarland University)

2346 Knowledge-Preserving Pluggable Modules for Multilingual Speech Translation Tasks

Nan Chen (Inner Mongolian University); Feilong Bao (Inner Mongolia University); Yonghe Wang (Inner mongolia university)

2548 A Unit-based System and Dataset for Expressive Direct Speech-to-Speech Translation

Anna Min (Tsinghua University); Chenxu Hu (Tsinghua University); Yi Ren (ByteDance); Hang Zhao (Tsinghua University)

Poster Session: Biosignal-enabled Spoken Communication

SS-2

Location: Poster Area 4A

739 Auditory Attention Decoding in Four-Talker Environment with EEG

Yujie Yan (Peking University); Xiran Xu (Peking University); HaoLin Zhu (Peking University); Pei Tian (Peking University); Zhongshu Ge (Peking University); Xihong Wu (Peking University); Jing Chen (Peking University)

753 ASA: An Auditory Spatial Attention Dataset with Multiple Speaking Locations

Zijie Lin (South China University of Technology); Tianyu He (The Chinese University of Hong Kong, Shenzhen); Siqi

Cai (National University of Singapore); Haizhou Li (The Chinese University of Hong Kong, Shenzhen)

1217 Leveraging Graphic and Convolutional Neural Networks for Auditory Attention Detection with EEG
Saurav Pahuja (University of Bremen); Gabriel Ivucic (University of Bremen); Pascal Himmelmann (University of Bremen); Siqi Cai (National University of Singapore); Tanja Schultz (University of Bremen); Haizhou Li (The Chinese University of Hong Kong, Shenzhen)

1289 Using articulated speech EEG signals for imagined speech decoding
Chris S Bras (Delft University of Technology); Tanvina Patel (Multimedia computing, Delft University of Technology); Odette Scharenborg (Multimedia Computing Group, Delft University of Technology)

1568 Towards EMG-to-Speech with Necklace Form Factor
Peter Wu (UC Berkeley); Ryan Kaveh (UC Berkeley); Raghav A Nautiyal (University of California, Berkeley); Christine Zhang (UC Berkeley); Albert Guo (UC Berkeley); Anvitha Kachinthaya (UC Berkeley); Tavish Mishra (UC Berkeley); Bohan Yu (UC Berkeley); Alan Black (CMU); Rikky Muller (UC Berkeley); Gopala Krishna Anumanchipalli (UC Berkeley)

1571 A multimodal approach to study the nature of coordinative patterns underlying speech rhythm
Jinyu Li (Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)); Leonardo Lancia (Laboratoire Parole et Langage (CNRS/Aix-Marseille Université))

1756 Articulatory synthesis using representations learnt through phonetic label-aware contrastive loss
Jesuraj Bandekar (IISc); Sathvik Udupa (Indian Institute of Science); Prasanta Kumar Ghosh (Indian Institute of Science (IISc), Bangalore)

1864 Optical Flow Guided Tongue Trajectory Generation for Diffusion-based Acoustic to Articulatory Inversion
Yudong yang (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences); Rongfeng Su (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Rukiye Ruzi (Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Manwa L Ng (Division of Speech and Hearing Sciences, University of Hong Kong); Shaofeng Zhao (Department of Rehabilitation Medicine, The Eighth Affiliated Hospital of Sun Yat-sen University); Nan Yan (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Lan Wang (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences)

2153 Direct Speech Synthesis from Non-Invasive Neuromagnetic Signals
Jinuk Kwon (The University of Texas at Austin); David Harwath (The University of Texas at Austin); Debadatta Dash (University of Texas at Austin); Paul Ferrari (Helen DeVos Children's Medical Center); Jun Wang (University of Texas at Austin)

2223 Multimodal Segmentation for Vocal Tract Modeling
Rishi Jain (UC Berkeley); Bohan Yu (UC Berkeley); Peter Wu (UC Berkeley); Tejas S Prabhune (Berkeley Speech Lab); Gopala Krishna Anumanchipalli (UC Berkeley)

Time Slot 2

Oral Session: Detection and Classification of Bioacoustic Signals

A5-02

Location: Acesso

143 DB3V: A Dialect Dominated Dataset of Bird Vocalisation for Cross-corpus Bird Species Recognition

Xin Jing (Technische Universität München); Luyang Zhang (Beijing Forestry University); Alexander Gebhard (Technical University of Munich); Jiangjian Xie (Beijing forestry university); Alice Baird (Hume AI); Prof. Dr. Bjoern Schuller (Imperial College London)

283 SimuSOE: A Simulated Snoring Dataset for Obstructive Sleep Apnea-Hypopnea Syndrome Evaluation during Wakefulness

Jie Lin (Wuhan University); Xiuping Yang (Zhongnan Hospital of Wuhan University); Li Xiao (School of Computer Science, Wuhan University); Li Xinhong (Wuhan University); weiyang WY (Wuhan university); Yuhong Yang (Wuhan University); Weiping Tu (Wuhan University); Xiong Chen (Zhongnan Hospital of Wuhan University)

439 Study Selectively: An Adaptive Knowledge Distillation based on a Voting Network for Heart Sound Classification

Xihang Qiu (Beijing Institute of Technology); Lixian Zhu (Beijing Institute of Technology); Zikai Song (Beijing Institute of Technology); Zeyu Chen (Nanjing University of Aeronautics and Astronautics); Haojie Zhang (Beijing Institute of Technology); Kun Qian (Beijing Institute of Technology); Ye Zhang (Shenzhen MSU-BIT University); Bin Hu (Beijing Institute of Technology); Yoshiharu Yamamoto (University of Tokyo); Bjorn W. Schuller (Imperial College London)

1096 Investigating self-supervised speech models ability to classify animal vocalizations: The case of gibbon's vocal signatures

Jules Cauzinille (ILCB); Benoît Favre (Lab. Informatique et Systèmes / Aix-Marseille University / CNRS); Ricard Marxer (Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon); Dena J Clink (Cornell University); Abdul H AHMAD (Universiti Malaysia Sabah); Arnaud Rey (Centre de Recherche en Psychologie et Neurosciences, Aix Marseille Univ, CNRS)

1412 Vision Transformer Segmentation for Visual Bird Sound Denoising

Sahil Kumar (Yeshiva university); Jialu Li (Cornell University); Youshan Zhang (Yeshiva University)

1453 Multi-label Bird Species Classification from Field Recordings using Mel _ Graph-GCN Framework

Noumida A (College Of Engineering Trivandrum); Rajeev Rajan (Government Engineering College, Barton Hill, Trivandrum)

Oral Session: Zero-shot TTS

A7-02

Location: Aegle A

549 DINO-VITS: Data-Efficient Zero-Shot TTS with Self-Supervised Speaker Verification Loss for Noise Robustness

Vikentii Pankov (Huawei Technologies Co. Ltd.); Valeria Pronina (Huawei Technologies Co. Ltd.); Alexander Kuzmin (ITMO University); Maksim Borisov (ITMO University); Nikita Usoltsev (HSE); Xingshan Zeng (Huawei Noah's Ark Lab); Alexander Golubkov (Huawei Technologies Co. Ltd.); Nikolai Ermolenko (Huawei Technologies Co. Ltd.); Aleksandra Shirshova (Saint Petersburg State University); Yulia Matveeva (Huawei Technologies Co. Ltd.)

803 Lightweight Zero-shot Text-to-Speech with Mixture of Adapters

Kenichi Fujita (NTT Corporation); Takanori Ashihara (NTT Corp.); Marc Delcroix (NTT); Yusuke Ijima (NTT Corporation)

1056 Improving Audio Codec-based Zero-Shot Text-to-Speech Synthesis with Multi-Modal Context and Large Language Model

Jinlong Xue (Beijing University of Posts and Telecommunications); Yayue Deng (Beijing University of Posts and Telecommunications); Yichen Han (Beijing University of Posts and Telecommunications); Yingming Gao (Beijing University of Posts and Telecommunications); Ya Li (Beijing University of Posts and Telecommunications)

1336 An Investigation of Noise Robustness for Flow-Matching-Based Zero-Shot TTS

Xiaofei Wang (Microsoft); Sefik Emre Eskimez (Microsoft); Manthan Thakker (Microsoft); Hemin Yang (Microsoft);

Zirun Zhu (Microsoft); Min Tang (Microsoft); Yufei Xia (Microsoft); Jinzhu Li (Microsoft); sheng zhao (microsoft); Jinyu Li (Microsoft); Naoyuki Kanda (Microsoft)

Oral Session: Paralinguistics

A3-01

Location: Aegle B

Surv-03-2 Paralinguistics: Past and Present

Anton Batliner

493 Real-world PTSD Recognition: A Cross-corpus and Cross-linguistic Evaluation

Alexander Kathan (University of Augsburg); Martin Bürger (University of Augsburg); Andreas Triantafyllopoulos (Technical University of Munich); Sabrina Milkus (LMU Munich); Jonas Hohmann (LMU Munich); Pauline Muderlak (LMU Munich); Jürgen Schottdorf (Zentrumspraxis Friedberg); Richard Musil (LMU Munich); Björn Schuller (Technical University of Munich); Shahin Amiriparian (Technical University of Munich)

514 Cross-transfer Knowledge Between Speech and Text Encoders to Evaluate Customer Satisfaction

Luis Felipe Parra Gallego (University of Antioquia); Juan Rafael Orozco-Arroyave (University of Antioquia); Tilak Purohit (Idiap Research Institute); Bogdan Vlasenko (Idiap Research Institute); Mathew Magimai.-Doss (Idiap Research Institute)

1224 Switching Tongues, Sharing Hearts: Identifying the Relationship between Empathy and Code-switching in Speech

Debasmita Bhattacharya (Columbia University); Eleanor M Lin (Columbia University); Run Chen (Columbia University); Julia Hirschberg (Columbia University)

2237 Fine-tuning of Pre-trained Models for Classification of Vocal Intensity Category from Speech Signals

Manila Kodali (Aalto University); Sudarsana Reddy Kadiri (University of Southern California); Paavo Alku (Aalto University)

Oral Session: Noise Robustness, Far-Field, and Multi-Talker ASR

A8-01

Location: Hippocrates

90 LibriheavyMix: A 20,000-Hour Dataset for Single-Channel Reverberant Multi-Talker Speech Separation, ASR and Speaker Diarization

Zengrui Jin (The Chinese University of Hong Kong); Yifan Yang (Shanghai Jiao Tong University); Mohan Shi (University of California, Los Angeles); Wei Kang (Xiaomi Corp., Beijing, China); Xiaoyu Yang (Xiaomi Corp., Beijing); Zengwei Yao (Xiaomi Corp.); Fangjun Kuang (Xiaomi Corp.); Liyong Guo (Xiaomi Corp.); Lingwei Meng (The Chinese University of Hong Kong); Long Lin (Xiaomi Corp.); Yong Xu (Tencent); Shi-Xiong Zhang (Capital One); Daniel Povey (Xiaomi, Inc.)

700 A JOINT NOISE DISENTANGLEMENT AND ADVERSARIAL TRAINING FRAMEWORK FOR ROBUST SPEAKER VERIFICATION

Xujiang Xing (Xinjiang University); Mingxing Xu (Tsinghua University); Thomas Fang Zheng (""CSLT, Tsinghua University"")

1062 Unified Multi-Talker ASR with and without Target-speaker Enrollment

Ryo Masumura (NTT Corporation); Naoki Makishima (NTT); Tomohiro Tanaka (NTT); Mana Ihori (NTT Corporation); Naotaka Kawata (NTT); Shota Orihashi (NTT Corporation); Kazutoshi Shinoda (NTT Corporation); Taiga Yamane (NTT Corporation); Saki Mizuno (NTT Corporation); Keita Suzuki (Nippon Telegraph and Telephone Corporation); Satoshi

Suzuki (NTT Computer and Data Science Laboratories / The University of Electro-Communications); Nobukatsu Hojo (NTT Corporation); Takafumi Moriya (NTT); Atsushi Ando (NTT Corporation)

1137 Neural Blind Source Separation and Diarization for Distant Speech Recognition

Yoshiaki Bando (National Institute of Advanced Industrial Science and Technology); Tomohiko Nakamura (National Institute of Advanced Industrial Science and Technology (AIST)); Shinji Watanabe (Carnegie Mellon University)

1710 Serialized Output Training by Learned Dominance

Ying Shi (Harbin Institute of Technology); Lantian Li (Beijing University of Posts and Telecommunications); Shi Yin (Haiwei Technologies Co., Ltd.); Dong Wang (Tsinghua University); jiqing Han (Harbin Institute of Technology)

2211 SOT Triggered Neural Clustering for Speaker Attributed ASR

Xianrui Zheng (University of Cambridge); Guangzhi Sun (University of Cambridge Department of Engineering); Chao Zhang (Tsinghua University); Phil Woodland (Machine Intelligence Laboratory, Cambridge University Department of Engineering)

Oral Session: Spoken Language Understanding

A11-01

Location: Iasso

103 Using Large Language Model for End-to-End Chinese ASR and NER

Yuang Li (Huawei); Jiawei Yu (Xiamen University); Min Zhang (Huawei); Mengxin Ren (Huawei); Yanqing Zhao (Huawei); Xiaofeng Zhao (Huawei); Shimin Tao (Huawei); Jinsong Su (Xiamen University); Hao Yang (Huawei)

712 Finding Task-specific Subnetworks in Multi-task Spoken Language Understanding Model

Hayato Futami (Sony Group Corporation); Siddhant Arora (Carnegie Mellon University); Yosuke Kashiwagi (Sony); Emiru Tsunoo (Sony Group Corporation); Shinji Watanabe (Carnegie Mellon University)

940 Out-of-distribution generalisation in spoken language understanding

Dejan Porjazovski (Aalto University); Anssi Moisio (Aalto University); Mikko Kurimo (Aalto University)

957 Speech-MASSIVE: A Multilingual Speech Dataset for SLU and Beyond

Beomseok LEE (University of Trento); Ioan Calapodescu (Naver Labs Europe); Marco Gaido (Fondazione Bruno Kessler); Matteo Negri (Fondazione Bruno Kessler); laurent besacier (Naver Labs Europe)

1133 A dual task learning approach to fine-tune a multilingual semantic speech encoder for Spoken Language Understanding

Gaëlle Laperrière (LIA - Avignon University); Sahar Ghannay (LISN); Bassam Jabaian (LIA - Avignon university); Yannick Estève (LIA - Avignon University)

1219 A Contrastive Learning Approach to Mitigate Bias in Speech Models

Alkis Koudounas (Politecnico di Torino); Flavio Giobergia (Politecnico di Torino); Eliana Pastor (Politecnico di Torino); Elena Baralis (Politecnico di Torino)

Oral Session: Spoken Machine Translation 1

A12-01

Location: Melambus

616 Diffusion Synthesizer for Efficient Multilingual Speech to Speech Translation

Nameer M Hirschkind (Roblox); Mahesh Nandwana (Roblox); Xiao Yu (Roblox); Joseph Liu (Roblox); Elois DuBois (Roblox); Dao Le (Roblox); Nicolas Thiebaut (Roblox); Colin Sinclair (Roblox); Kyle Spence (Roblox); Chong Shang (Roblox); Zoe Abrams (Roblox); Morgan McGuire (Roblox)

790 Lightweight Audio Segmentation for Long-form Speech Translation

Jaesong Lee (NAVER Cloud); Soyeon Kim (NAVER Cloud); Hanbyul Kim (NAVER Cloud); Joon Son Chung (KAIST)

1858 Investigating Decoder-only Large Language Models for Speech-to-text Translation

Chao-Wei Huang (National Taiwan University); Hui Lu (The Chinese University of Hong Kong); Hongyu Gong (Meta AI); Hirofumi Inaguma (Meta); Ilia Kulikov (Meta); Ruslan Mavlyutov (Meta); Sravya Popuri (Facebook Inc)

2022 Sign Value Constraint Decomposition for Efficient 1-Bit Quantization of Speech Translation Tasks

Nan Chen (Inner Mongolian University); Feilong Bao (Inner Mongolia University); Yonghe Wang (Inner mongolia university)

2289 Leveraging Multilingual LLMs for Zero-Resource Cross-lingual Transfer in Speech Translation and ASR

Karel Mundnich (Amazon); Xing Niu (Amazon); Prashant Mathur (Amazon); Srikanth Ronanki (Amazon); Brady Houston (AWS AI Labs); Veera Raghavendra Elluru (AWS AI Labs); Nilaksh Das (AWS AI Labs, Amazon); Zejiang Hou (Amazon); Goeric Huybrechts (Amazon); Anshu Bhatia (Amazon); Daniel Garcia-Romero (AWS AI); Kyu Han (Amazon Web Services); Katrin Kirchhoff (Amazon)

2426 Contrastive Feedback Mechanism for Simultaneous Speech Translation

Haotian Tan (Nara Institute of Science and Technology); Sakriani Sakti (Nara Institute of Science and Technology / Japan Advanced Institute of Science and Technology)

Oral Session: Individual and Social Factors in Phonetics

A2-01 Location: Panacea Amphitheater

628 Entrainment Analysis and Prosody Prediction of Subsequent Interlocutor's Backchannels in Dialogue

Keiko Ochi (Kyoto University); Koji Inoue (Kyoto University); Divesh Lala (Kyoto University); Tatsuya Kawahara (Kyoto University)

1154 Exploring the anatomy of articulation rate in spontaneous English speech: relationships between utterance length effects and social factors

James Tanner (University of Glasgow); Morgan Sonderegger (McGill University); Jane Stuart-Smith (University of Glasgow); Tyler Kendall (University of Oregon); Jeff Mielke (North Carolina State University); Robin Dodsworth (North Carolina State University); Erik Thomas (North Carolina State University)

1763 Familiar and Unfamiliar Speaker Identification in Speech and Singing

Katelyn L Taylor (University of York); Amelia Gully (University of York); Helena Daffern (University of York)

2013 Echoes of Implicit Bias. Exploring Aesthetics and Social Meanings of Swiss German Dialect Features

Tillmann Pistor (University of Bern); Adrian Leemann (University of Bern)

2260 Modelled Multivariate Overlap: A method for measuring vowel merger

Irene B. R. Smith (McGill University); Morgan Sonderegger (McGill University); The Spade Consortium (University of Glasgow)

2456 In search of structure and correspondence in intra-speaker trial-to-trial variability

Vivian G. Li (Yale)

Poster Session: Hearing Disorders

A13-P1-A

Location: Poster Area 1A

111 Signal processing algorithm effective for sound quality of hearing loss simulators*Toshio Irino (Wakayama University); Shintaro Doan (Wakayama University); Minami Ishikawa (Wakayama University)***234 Automatic Assessment of Speech Production Skills for Children with Cochlear Implants Using Wav2Vec2.0 Acoustic Embeddings***Seonwoo LEE (Seoul National University); Sunhee Kim (Seoul National University); Minhwa Chung (Seoul National University)***414 Automatic Detection of Hearing Loss from Children's Speech using Wav2Vec 2.0 Features***Jessica Monaghan (National Acoustic Laboratories); Arun Sebastian (National Acoustic Laboratories); Nicky Chong-White (National Acoustic Laboratories); Vicky W Zhang (National Acoustic Laboratories); Viji Easwar (National Acoustic Laboratories); Padraig Kitterick (National Acoustic Laboratories)***432 SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Crossmodal Audio Token Synchronization***Young Jin Ahn (KAIST); Jungwoo Park (Kwangwoon University); Sangha Park (Ajou University); Jonghyun Choi (Seoul National University); Kee-Eung Kim (KAIST)***871 Production of fricative consonants in French-speaking children with cochlear implants and typically hearing: acoustic and phonological analyses.***Sophie Fagniat (University of Mons, Belgium); Brigitte Charlier (ULB - Centre Comprendre et Parler); Bernard Harmegnies (UMONS - ULB); Anne Huberlant (Centre Comprendre et Parler); Kathy Huet (UMONS); Myriam Piccaluga (UMONS); Véronique Delvaux (FNRS & UMONS)***961 Evaluating a 3-factor listener model for prediction of speech intelligibility to hearing-impaired listeners***Mark Huckvale (University College London); Gaston Hilkhuisen (University College London)***1940 Auditory Spatial Attention Detection Based on Feature Disentanglement and Brain Connectivity-Informed Graph Neural Networks***Yixiang Niu (East China University of Science and Technology); Ning Chen (East China University of Science and Technology); Hongqing Zhu (East China University of Science and Technology); Zhiying Zhu (East China University of Science and Technology); Guangqiang Li (East China University of Science and Technology); Yibo Chen (East China University of Science and Technology)*

Poster Session: Speech Disorders 2

A13-P1-B

Location: Poster Area 1B

188 A Cross-Attention Layer coupled with Multimodal Fusion Methods for Recognizing Depression from Spontaneous Speech*Loukas Ilias (National Technical University of Athens); Dimitris Askounis (National Technical University of Athens)***627 Cascaded Transfer Learning Strategy for Cross-Domain Alzheimer's Disease Recognition through Spontaneous Speech***Guanlin Chen (Jiangsu Normal University); Yun Jin (Jiangsu Normal University)***1009 Automatic Longitudinal Investigation of Multiple Sclerosis Subjects**

Gábor Gosztolya (MTA-SZTE Research Group on AI); Veronika Svindt (Research Center for Linguistics); Judit Bóna (ELTE Eötvös Loránd University); Ildikó Hoffmann (Research Center for Linguistics)

1452 DysArinVox: DYSpHonia & DYSarThria mandARIN speech corpus

Haojie Zhang (Tianjin University); Zhang Tao (Tianjin University); Ganjun Liu (Tianjin University); Dehui Fu (Tianjin Medical University); Xiaohui Hou (Tianjin University); Ying Lv (Tianjin University)

1563 Improving Speech-based Dysarthria Detection using Multi-task Learning with Gradient Projection

Yan Xiong (Arizona State University); Visar Berisha (Arizona State University); Julie Liss (Arizona State University); Chaitali Chakrabarti (Arizona State University)

1855 YOLO-Stutter: End-to-end Region-Wise Speech Dysfluency Detection

Xuanru Zhou (Berkeley Speech Group); Anshul P Kashyap (UC Berkeley); Steve Li (Berkeley Speech Group); Ayati Sharma (University of California, Berkeley); Brittany Morin (University of California San Francisco); David Baquirin (University of California San Francisco); Jet Vonk (University of California San Francisco); Zoe Ezzes (University of California San Francisco); Zachary Miller (University of California San Francisco); Maria Luisa Gorno Tempini (UCSF); Jiachen Lian (University of California Berkeley); Gopala Krishna Anumanchipalli (UC Berkeley)

1966 Multimodal Continuous Fingerspelling Recognition via Visual Alignment Learning

Katerina Papadimitriou (University of Thessaly); Gerasimos Potamianos (ECE, University of Thessaly)

2063 Segmental and Suprasegmental Speech Foundation Models for Classifying Cognitive Risk Factors: Evaluating Out-of-the-Box Performance

Si-loi Ng (Arizona State University); Lingfeng Xu (Arizona State University); Kimberly D. Mueller (University of Wisconsin-Madison); Julie Liss (Arizona State University); Visar Berisha (Arizona State University)

2236 Contrastive Learning Approach for Assessment of Phonological Precision in Patients with Tongue Cancer Using MRI Data

Tomas Arias-Vergara (Friedrich-Alexander-Universitaet Erlangen-Nuernberg); Paula Andrea Pérez-Toro (Friedrich-Alexander-Universität Erlangen-Nürnberg); Xiaofeng Liu (Yale University); Fangxu Xing (Massachusetts General Hospital / Harvard Medical School); Maureen Stone (University of Maryland); Jiachen Zhuo (University of Maryland); Jerry L Prince (Johns Hopkins University); Maria Schuster (Ludwig-Maximilians Universität München); Elmar Noeth (Friedrich Alexander Universität, Erlangen-Nuremberg); Jonghye Woo (Massachusetts General Hospital / Harvard Medical School); Andreas K Maier (Pattern Recognition Lab, FAU Erlangen-Nuremberg)

2293 Whister: Using Whisper's representations for Stuttering detection

Vrushank Changawala (Dalhousie University); Frank Rudzicz (Dalhousie University)

Poster Session: Speaker Recognition: Adversarial and Spoofing Attacks

A4-P1 Location: Poster Area 2A, Poster Area 2B

21 Textual-Driven Adversarial Purification for Speaker Verification

Sizhou Chen (Chengdu University of Information Technology); Yibo Bai (The University of Hong Kong); Jiadi Yao (Northwestern Polytechnical University); Zhang XiaoLei (Northwestern Polytechnical University); Xuelong Li (Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd)

251 Spoofing Speech Detection by Modeling Local Spectro-Temporal and Long-term Dependency

Haochen Wu (University of Science and Technology of China); Wu Guo (University of Science and Technology of China); Zhentao Zhang (China Merchants Bank); Wenting Zhao (China Merchants Bank); Shengyu Peng (University of Science and Technology of China); Jie Zhang (University of Science and Technology of China (USTC))

346 Boosting the Transferability of Adversarial Examples with Gradient-Aligned Ensemble Attack for Speaker Recognition

Zhuhai Li (University of Science and Technology of China); Jie Zhang (University of Science and Technology of China)

China (USTC)); Wu Guo (University of Science and Technology of China); Haochen Wu (University of Science and Technology of China)

367 Improving Copy-Synthesis Anti-Spoofing Training Method with Rhythm and Speaker Perturbation
Jingze Lu (Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China); Yuxiang Zhang (Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences); Zhuo Li (OPPO); Zengqiang Shang (The Institute of Acoustics of the Chinese Academy of Sciences); Wenchao Wang (Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China); pengyuan zhang (Institute of Acoustics, Chinese Academy of Sciences)

372 VoiceDefense: Protecting Automatic Speaker Verification Models Against Black-box Adversarial Attacks

Yip Keng Kan (Huawei International); Ke Xu (Huawei International); hao li (Huawei Technology); Jie Shi (Huawei International)

403 Anti-spoofing Ensembling Model: Dynamic Weight Allocation in Ensemble Models for Improved Voice Biometrics Security

Eros Rosello (University of Granada); Angel M. Gomez (University of Granada); Iván López-Espejo (University of Granada); Antonio M. Peinado (University of Granada); Juan M. Martín-Doñas (Vicomtech)

659 Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection

Duc-Tuan Truong (Nanyang Technological University); Ruijie Tao (National University of Singapore); Tuan V. A. Nguyen (Institute for Infocomm Research, A*STAR); Hieu-Thi Luong (Nanyang Technological University); Kong Aik Lee (The Hong Kong Polytechnic University); Eng Siong Chng (Nanyang Technological University)

1191 Neural Codec-based Adversarial Sample Detection for Speaker Verification

Xuanjun Chen (National Taiwan University); JIAWEI DU (National Taiwan University); Haibin Wu (National Taiwan University); Roger Jang (); Hung-yi Lee (National Taiwan University)

1365 "Spoof Diarization: "What Spoofed When" in Partially Spoofed Audio"

Lin Zhang (National Institute of Informatics); Xin Wang (National Institute of Informatics); Erica Cooper (National Institute of Information and Communications Technology); Mireia Diez (Brno University of Technology); Federico Landini (Brno University of Technology); Nicholas Evans (EURECOM); Junichi Yamagishi (National Institute of Informatics)

Poster Session: Source Separation 2

A5-P2-B

Location: Poster Area 2B

197 TSE-PI: Target Sound Extraction under Reverberant Environments with Pitch Information

Yiwen Wang (Peking University); Xihong Wu (Peking University)

530 Towards Explainable Monaural Speaker Separation with Auditory-based Training

Hassan Taherian (The Ohio State University); Vahid Ahmadi Kalkhorani (The Ohio State University); Ashutosh Pandey (META); Daniel D.E. Wong (Meta Platforms Inc.); Buye Xu (Meta Reality Labs Research); DeLiang Wang (Ohio State University)

787 SA-WavLM: Speaker-Aware Self-Supervised Pre-training for Mixture Speech

Jingru Lin (National University of Singapore); Meng Ge (Tianjin University); Junyi Ao (The Chinese University of Hong Kong (Shenzhen)); Liqun Deng (Huawei Noah's Ark Lab); Haizhou Li (The Chinese University of Hong Kong (Shenzhen))

1066 PARIS: Pseudo-AutoRegressive Siamese Training for Online Speech Separation

Zexu Pan (Mitsubishi Electric Research Laboratories (MERL)); Gordon Wichern (Mitsubishi Electric Research Laboratories (MERL)); François G Germain (Mitsubishi Electric Research Laboratories (MERL)); Kohei Saijo (Waseda

University); Jonathan Le Roux (Mitsubishi Electric Research Laboratories (MERL))

1131 Does the Lombard Effect Matter in Speech Separation? Introducing the Lombard-GRID-2mix Dataset
Iva Ewert (University of Bremen); Marvin Borsdorf (University of Bremen); Haizhou Li (The Chinese University of Hong Kong, Shenzhen); Tanja Schultz (University of Bremen)

1241 Enhanced Reverberation as Supervision for Unsupervised Speech Separation
Kohei Saijo (Waseda University); Gordon Wichern (Mitsubishi Electric Research Laboratories (MERL)); François G Germain (Mitsubishi Electric Research Laboratories (MERL)); Zexu Pan (Mitsubishi Electric Research Laboratories (MERL)); Jonathan Le Roux (Mitsubishi Electric Research Laboratories (MERL))

1265 Multimodal Representation Loss Between Timed Text and Audio for Regularized Speech Separation
Tsun-An Hsieh (Indiana University Bloomington); Heeyoul Choi (Handong Global University); Minje Kim (University of Illinois at Urbana-Champaign)

2322 OR-TSE: An Overlap-Robust Speaker Encoder for Target Speech Extraction
Yiru Zhang (Nanjing University of Aeronautics and Astronautics); Linyu Yao (Nanjing University of Aeronautics and Astronautics); Yang Qun (Nanjing University of Aeronautics and Astronautics)

Poster Session: Contextual Biasing and Adaptation

A9-P1 Location: Poster Area 3A, Poster Area 3B

298 Speculative Speech Recognition by Audio-Prefixed Low-Rank Adaptation of Language Models
Bolaji Yusuf (Bogazici University); Murali Karthick Baskar (Google Inc); Andrew Rosenberg (Google LLC); Bhuvana Ramabhadran (Google)

348 Dual-Pipeline with Low-Rank Adaptation for New Language Integration in Multilingual ASR
Yerbolat Khassanov (ByteDance); Zhipeng Chen (Bytedance); Tianfeng Chen (Bytedance); Tze Yuang Chong (Bytedance); Wei Li (Bytedance); Jun Zhang (Bytedance); Lu Lu (Bytedance); Yuxuan Wang (ByteDance AI Lab)

391 Keyword-Guided Adaptation of Automatic Speech Recognition
Aviv Shamsian (Bar Ilan University); Aviv Navon (Bar-Ilan University); Neta Glazer (Aiola); Gill Hetz (aiola); Joseph Keshet (Technion - Israel Institute of Technology)

633 Factor-Conditioned Speaking-Style Captioning
Atsushi Ando (NTT Corporation); Takafumi Moriya (NTT); Shota Horiguchi (NTT Corporation); Ryo Masumura (NTT Corporation)

653 Incorporating Class-based Language Model for Named Entity Recognition in Factorized Neural Transducer
Peng Wang (Key Lab of Speech Acoustics and Content Understanding, Institute of Acoustics, CAS, China; University of Chinese Academy of Sciences, China); Yifan Yang (Shanghai Jiao Tong University); Zheng Liang (Shanghai Jiao Tong University); Tian Tan (Shanghai Jiao Tong University); Shiliang Zhang (Alibaba Group); Xie Chen (Shanghai Jiaotong University)

675 Modality Translation Learning for Joint Speech-Text Model
Pin-Yen Liu (National Yang Ming Chiao Tung University); Jen-Tzung Chien (National Yang Ming Chiao Tung University)

749 Improving Neural Biasing for Contextual Speech Recognition by Early Context Injection and Text Perturbation
Ruizhe Huang (Johns Hopkins University); Mahsa Yarmohammadi (Johns Hopkins University); Daniel Povey (Xiaomi, Inc.); Sanjeev Khudanpur (Johns Hopkins University)

812 Improved Factorized Neural Transducer Model For text-only Domain Adaptation

Junzhe Liu (Shanghai Jiaotong University); Jianwei Yu (Tencent AI lab); Xie Chen (Shanghai Jiaotong University)

869 Contextual Biasing with Confidence-based Homophone Detector for Mandarin End-to-End Speech Recognition

Chengxu Yang (University of Chinese Academy of Sciences; Institute of Acoustics, Chinese Academy of Sciences); Lin Zheng (University of Chinese Academy of Sciences; Institute of Acoustics, Chinese Academy of Sciences); Gaofeng Cheng (Institute of Acoustics, Chinese Academy of Sciences); Sanli Tian (University of Chinese Academy of Sciences; Institute of Acoustics, Chinese Academy of Sciences); Sujie Xiao (University of Chinese Academy of Sciences; Institute of Acoustics, Chinese Academy of Sciences); ta li (Institute of Acoustics, Chinese Academy of Sciences)

1002 Fast Context-Biasing for CTC and Transducer ASR models with CTC-based Word Spotter

Andrei Andrusenko (NVIDIA); Aleksandr Laptev (NVIDIA, ITMO University); Vladimir Bataev (NVIDIA, University of London); Vitaly Lavrukhin (NVIDIA); Boris Ginsburg (NVIDIA)

1006 SAML: Speaker Adaptive Mixture of LoRA Experts for End-to-End ASR

Qiuming Zhao ("CSLT, Tsinghua University"); Guangzhi Sun (University of Cambridge Department of Engineering); Chao Zhang (Tsinghua University); Mingxing Xu (Tsinghua University); Thomas Fang Zheng ("CSLT, Tsinghua University")

1162 Prompt Tuning for Speech Recognition on Unknown Spoken Name Entities

Xizi Wei (action.ai); Stephen E McGregor (action.ai)

1762 Improving Speech Recognition with Prompt-based Contextualized ASR and LLM-based Re-predictor

Nguyen Manh Tien Anh (VinBigData Joint Stock Company); Hồ Sỹ Thạch (VinBigdata)

2368 Domain-Aware Data Selection for Speech Classification via Meta-Reweighting

Junghun Kim (Seoul National University); Ka Hyun Park (Seoul National University); Hoyoung Yoon (Seoul National University); U Kang (Seoul National University)

Poster Session: Noise Reduction, Reverberation, and Echo Cancellation

A6-P1-A

Location: Poster Area 4A

336 Graph Attention Based Multi-Channel U-Net for Speech Dereverberation With Ad-Hoc Microphone Arrays

Hongmei Guo (Northwestern Polytechnical University); Yijiang Chen (Northwestern Polytechnical University); Zhang XiaoLei (Northwestern Polytechnical University); Xuelong Li (Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd)

662 Deep Echo Path Modeling for Acoustic Echo Cancellation

Zhao Fei (School of Computer Science, Inner Mongolia University); Jinjiang Liu (College of Computer Science, Inner Mongolia University); Chenggang Zhang (Inner Mongolia Minzu University); Shulin He (College of Computer Science, Inner Mongolia University); Xueliang zhang (Inner Mongolia University)

854 Elucidating Clock-drift Using Real-world Audios In Wireless mode For Time-offset Insensitive End-to-End Asynchronous Acoustic Echo Cancellation

Premanand Vinayak Nayak (Samsung Research and Development Institute - Bangalore, India); Mahaboob Ali Basha Shaik (Samsung Research and Development Institute - Bangalore, India)

1151 ANIMAL-CLEAN –A Deep Denoising Toolkit for Animal-Independent Signal Enhancement

Alexander Barnhill (Friedrich-Alexander University); Elmar Noeth (Friedrich Alexander Universität, Erlangen-Nuremberg); Andreas K Maier (Pattern Recognition Lab, FAU Erlangen-Nuremberg); Christian Bergler (Technical University of Applied Sciences Amberg-Weiden, Department of Electrical Engineering, Media and Computer Science)

1173 Speech dereverberation constrained on room impulse response characteristics

Louis Bahrman (Télécom Paris); Mathieu Fontaine (Télécom Paris); Jonathan Le Roux (Mitsubishi Electric Research Laboratories (MERL)); Gaël Richard (LTCl, Telecom Paris, Institut polytechnique de Paris)

2180 DeWinder: Single-Channel Wind Noise Reduction using Ultrasound Sensing

Kuang Yuan (Carnegie Mellon University); Shuo Han (Carnegie Mellon University); Swarun Kumar (Carnegie Mellon University); Bhiksha Raj (Carnegie Mellon University)

Poster Session: Computationally-Efficient Speech Enhancement**A6-P1-B**

Location: Poster Area 4B

785 Streamlining Speech Enhancement DNNs: an Automated Pruning Method Based on Dependency Graph with Advanced Regularized Loss Strategies

zugang zhao (Beijing University of Posts and Telecommunications); Jinghong Zhang (Beijing University of Posts and Telecommunications); yonghui liu (Fanvil Link Technology Co.); jianbing liu (Fanvil Link Technology Co.); kai niu (Beijing University of Posts and Telecommunications); zhiqiang he (Beijing University of Posts and Telecommunications)

958 Dynamic Gated Recurrent Neural Network for Compute-efficient Speech Enhancement

Longbiao Cheng (Institute of Neuroinformatics, University of Zurich and ETH Zurich); Ashutosh Pandey (META); Buye Xu (Meta Reality Labs Research); Tobi Delbruck (Sensors Group, Inst. of Neuroinformatics, UZH-ETH Zurich); Shih-Chii Liu (Institute of Neuroinformatics, University of Zurich and ETH Zurich)

1017 MUSE: Flexible Voiceprint Receptive Fields and Multi-Path Fusion Enhanced Taylor Transformer for U-Net-based Speech Enhancement.

Zizhen Lin (Sichuan University); XiaoTing Chen (YunNan University); Junyu Wang (Sichuan University)

1368 Lightweight Dynamic Sparse Transformer for Monaural Speech Enhancement

Zehua Zhang (Harbin Institute of Technology(Shenzhen)); Xuyi Zhuang (Harbin Institute of Technology (Shenzhen)); yukun qian (Harbin Institute of Technology(Shenzhen)); Mingjiang Wang (Harbin Institute of Technology Shenzhen)

1383 Knowledge Distillation for Tiny Speech Enhancement with Latent Feature Augmentation

Behnam Gholami (Samsung); Mostafa El-Khamy (Samsung Research USA); KeeBong Song (Samsung Semiconductor Inc.)

1444 Speech Boosting: Low-Latency Live Speech Enhancement for TWS Earbuds

Hanbin Bae (Speech Intelligence Team, Samsung Research); Pavel Andreev (Samsung AI Center Moscow); Azat Saginbaev (MIPT); Nicholas Babaev (Samsung AI center); WonJun Lee (Samsung); Hosang Sung (Samsung Research); Hoon-Young Cho (Samsung Research)

1927 Sub-PNWR: Speech Enhancement Based on Signal Sub-Band Splitting and Pseudo Noisy Waveform Reconstruction Loss

Yuewei Zhang (Shanghai Jiao Tong University); Huanbin Zou (Xiaohongshu); jie zhu (Shanghai Jiao Tong University)

Poster Session: TAUADIAL Challenge: Speech-Based Cognitive Assessment in Chinese and English (Special Session)**SS-13**

Location: Yanis Club

914 Cognitive Insights Across Languages: Enhancing Multimodal Interview Analysis

David Ortiz-Perez (University of Alicante); Jose Garcia-Rodriguez (University of Alicante); David Tomás (University of Alicante)

984 Combining Acoustic Feature Sets for Detecting Mild Cognitive Impairment in the Interspeech'24 TAUkADIAL Challenge

Gábor Gosztolya (MTA-SZTE Research Group on AI); Tóth László (University of Szeged)

1352 The Interspeech 2024 TAUkADIAL Challenge: Multilingual Mild Cognitive Impairment Detection with Multimodal Approach

Benjamin Barrera-Altuna (University of South Florida); Daeun Lee (Sungkyunkwan University); Zaima Zarnaz (University of South Florida); Jinyoung Han (Sungkyunkwan University); Seungbae Kim (University of South Florida)

1422 Translingual Language Markers for Cognitive Assessment from Spontaneous Speech

Bao Hoang (Michigan State University); Yijiang Pang (Michigan State University); Hiroko Dodge (Harvard Medical School); Jiayu Zhou (Michigan State University)

1807 Connected Speech-Based Cognitive Assessment in Chinese and English

Saturnino Luz (The University of Edinburgh); Sofia De La Fuente Garcia (The University of Edinburgh); Fasih Haider (The University of Edinburgh); Davida Fromm Fromm (Carnegie Mellon University); Brian MacWhinney (CMU); Alyssa Lanzi (University of Delaware); Ya-Ning Chang (Miin Wu School of Computing, National Cheng Kung University); Chia-Ju Chou (Cardinal Tien Hospital); Yi-Chien Liu (Neurology department, CTH hospital)

2030 Leveraging Universal Speech Representations for Detecting and Assessing the Severity of Mild Cognitive Impairment Across Languages

Anna Favaro (Johns Hopkins University); Tianyu Cao (Johns Hopkins University); Najim Dehak (Johns Hopkins University); Laureano Moro-Velazquez (Johns Hopkins University)

2115 Multilingual Speech and Language Analysis for the Assessment of Mild Cognitive Impairment: Outcomes from the TaukadiAl Challenge

Paula Andrea Pérez-Toro (Friedrich-Alexander-Universität Erlangen-Nürnberg); Tomas Arias-Vergara (Friedrich-Alexander-Universität Erlangen-Nürnberg); Philipp Klumpp (Pattern Recognition Lab, FAU Erlangen-Nuremberg); Tobias Weise (Friedrich-Alexander-Universität Erlangen-Nürnberg); Maria Schuster (Ludwig Maximilian University of Munich); Elmar Noeth (Friedrich Alexander Universität, Erlangen-Nuremberg); Juan Rafael Orozco-Arroyave (University of Antioquia); Andreas K Maier (Pattern Recognition Lab, FAU Erlangen-Nuremberg)

2386 Pre-trained Feature Fusion and Matching for Mild Cognitive Impairment Detection

Junwen Duan (Central South University); Fangyuan Wei (Central South University); Hong-Dong Li (Central South University); Jin Liu (Central South University)

Tuesday 03/09

Time Slot 1

Oral Session: Phonetics and Phonology of Second Language Acquisition

A2-02

Location: Acesso

1014 Automatic Speech Recognition with parallel L1 and L2 acoustic phone models to evaluate /l/ allophony in L2 English speech production

Anisia Popescu (Université Paris Saclay - LISN); Lori Lamel (CNRS LISN); Ioana Vasilescu (LIMSI); Laurence Y. Devillers (LISN-CNRS)

1296 Mmm whatcha say? Uncovering distal and proximal context effects in first and second-language word perception using psychophysical reverse correlation

Paige Tuttösi (Simon Fraser University); Henny Yeung (Simon Fraser University); Yue Wang (Simon Fraser University); Fenqi Wang (Simon Fraser University); Guillaume Denis (Independent Web Developer); Jean-Julien Aucouturier (Institut FEMTO-ST); Angelica Lim (Simon Fraser University)

1627 Exploring Impact of Pausing and Lexical Stress Patterns on L2 English Comprehensibility in Real Time

Sylvain Coulange (Université Grenoble Alpes); Tsuneo Kato (Doshisha University); Solange Rossato (Univ. Grenoble Alpes); Monica Masperi (Université Grenoble Alpes)

1806 Mandarin T3 Production by Chinese and Japanese Native Speakers

Wu Qi (University of Tsukuba)

2175 Analysis of articulatory setting for L1 and L2 English speakers using MRI data

Kevin Y Huang (University of Southern California); Jack Goldberg (University of Southern California); Louis Goldstein (University of Southern California); Shrikanth Narayanan (USC)

2287 Bilingual Rhotic Production Patterns: A Generational Comparison of Spanish-English Bilingual Speakers in Canada

Ioana Colgiu (Western University); Yasaman Rafat (Western University); Laura Spinu (City University of New York - Kingsborough Community College); Rajiv Rao (University of Wisconsin-Madison)

Oral Session: Spoofing and Deepfake Detection

A4-02

Location: Aegle A

422 Revisiting and Improving Scoring Fusion for Spoofing-aware Speaker Verification Using Compositional Data Analysis

Xin Wang (National Institute of Informatics); Tomi H. Kinnunen (University of Eastern Finland); Kong Aik Lee (The Hong Kong Polytechnic University); Paul-Gauthier Noé (Laboratoire Informatique d'Avignon, Avignon Université); Junichi Yamagishi (National Institute of Informatics)

1283 Source Tracing of Audio Deepfake Systems

Nicholas M Klein (Pindrop Security); Tianxiang Chen (Pindrop Security Inc.); Hemlata Tak (Pindrop); Ricardo Casal (Pindrop Security Inc.); Elie Khoury (pindrop)

1724 DGNP: A Dual Graph Prototypical Network for Few-Shot Speech Spoofing Algorithm Recognition
 Zirui Ge (Nanjing University of Posts and Telecommunications); Xinzhou Xu (Nanjing University of Posts and Telecommunications); Haiyan Guo (Nanjing University of Posts and Telecommunications); Tingting Wang (Nanjing University of Posts and Telecommunications); Zhen Yang (Nanjing University of Posts and Telecommunication); Bjorn W. Schuller (Imperial College London)

2009 How Do Neural Spoofing Countermeasures Detect Partially Spoofed Audio?
 Tianchi Liu (National University of Singapore); Lin Zhang (National Institute of Informatics); Rohan Kumar Das (Fortemedia); Yi Ma (National University of Singapore); Ruijie Tao (National University of Singapore); Haizhou Li (The Chinese University of Hong Kong, Shenzhen)

2156 Interpretable Temporal Class Activation Representation for Audio Spoofing Detection
 Menglu Li (Toronto Metropolitan University); Xiao-Ping Zhang (Toronto Metropolitan University)

2349 SecureSpectra: Safeguarding Digital Identity from Deep Fake Threats via Intelligent Signatures
 Oguzhan Baser (The University of Texas at Austin); Kaan Kale (Bogazici University); Sandeep Chinchali (The University of Texas at Austin)

Oral Session: Speech Synthesis: Evaluation

A7-03 Location: Hippocrates

781 Assessing the impact of contextual framing on subjective TTS quality
 Jens Edlund (KTH Royal Institute of Technology); Christina Tännander (KTH Royal Institute of Technology); Sébastien Le Maguer (University of Helsinki); Petra Wagner (Bielefeld University)

937 Uncertainty-Aware Mean Opinion Score Prediction
 Hui Wang (Nankai University); Shiwan Zhao (Nankai University); Jiaming Zhou (Nankai University); Xiguang Zheng (University of Wollongong); Haoqin Sun (Nankai University); Xuechen Wang (Nankai University); Yong Qin (Nankai University)

959 Lifelong Learning MOS Prediction for Synthetic Speech Quality Evaluation
 Félix Saget (LIUM); Meysam Shamsi (LIUM); Marie Tahon (LIUM)

1193 What do people hear? Listeners' Perception of Conversational Speech
 Adaeze Adigwe (University of Edinburgh); Simon King (University of Edinburgh); Sarenne Wallbridge (University of Edinburgh)

1540 SVSNet+: Enhancing Speaker Voice Similarity Assessment Models with Representations from Speech Foundation Models
 Chun Yin (National Yang Ming Chiao Tung University; Academia Sinica); Tai-Shih Chi (National Yang Ming Chiao Tung University); Yu Tsao (Academia Sinica); Hsin-Min Wang (Academia Sinica)

2418 Enhancing Out-of-Distribution Performance of Indian TTS Systems for Practical Applications through Low-Effort Data Strategies
 Srija Anand (AI4Bharat); Praveen S V (Indian Institute of Technology Madras); Ashwin Sankar (AI4Bharat); Giri Raju (AI4Bharat); Mitesh M. Khapra (Indian Institute of Technology Madras)

Oral Session: Multilingual ASR

A8-02 Location: Iasso

- 205** **Continual Learning Optimizations for Auto-regressive Decoder of Multilingual ASR systems**
Chin Yuen Kwok (Nanyang Technological University); Jia Qi Yip (Alibaba Group / Nanyang Technological University); Eng Siong Chng (Nanyang Technological University)
- 734** **Weighted Cross-entropy for Low-Resource Languages in Multilingual Speech Recognition**
Andrés Piñeiro-Martín (Balidea S.L. / Universidade de Vigo); Carmen García-Mateo (Universidade de Vigo); Laura Docio-Fernandez (Universidade de Vigo); María del Carmen López-Pérez (Balidea S.L.); Georg Rehm (DFKI)
- 890** **MSR-86K: An Evolving, Multilingual Corpus with 86,300 Hours of Transcribed Audio for Speech Recognition Research**
Song Li (Meituan); Yongbin You (Meituan); Xuezhi Wang (Meituan); Zhengkun Tian (Meituan); Ke Ding (Meituan); Guanglu Wan (Meituan)
- 1278** **Improving Multilingual ASR Robustness to Errors in Language Input**
Brady Houston (AWS AI Labs); Omid Sadjadi (Amazon); Zejiang Hou (AWS AI Labs); Srikanth Vishnubhotla (AWS AI Labs); Kyu Han (Amazon Web Services)
- 1492** **M²ASR: Multilingual Multi-task Automatic Speech Recognition via Multi-objective Optimization**
A F M Saif (Rensselaer Polytechnic Institute); Lisha Chen (RENSSELAER POLYTECHNIC INST); Xiaodong Cui (IBM T. J. Watson Research Center); Songtao Lu (IBM Thomas J. Watson Research Center); Brian Kingsbury (IBM Research); Tianyi Chen (Rensselaer Polytechnic Institute)
- 2248** **ML-SUPERB 2.0: Benchmarking Multilingual Speech Models Across Modeling Constraints, Languages, and Datasets**
Jiatong Shi (Carnegie Mellon University); SHIH-HENG WANG (National Taiwan University); William Chen (Carnegie Mellon University); Martijn Bartelds (Stanford University); Vanya Bannihatti Kumar (Carnegie Mellon University); Jinchuan Tian (Carnegie Mellon University); Xuankai Chang (Carnegie Mellon University); Dan Jurafsky (Stanford University); Karen Livescu (TTI-Chicago); Hung-yi Lee (National Taiwan University); Shinji Watanabe (Carnegie Mellon University)

Oral Session: Generative Speech Enhancement

A6-02

Location: Melampus

- 138** **Universal Score-based Speech Enhancement with High Content Preservation**
Robin Scheibler (LINE Corporation); Yusuke Fujita (LY Corporation); Yuma Shirahata (LY Corp.); Tatsuya Komatsu (LY Corporation)
- 579** **Schrödinger Bridge for Generative Speech Enhancement**
Ante Jukić (NVIDIA); Roman Korostik (NVIDIA); Jagadeesh Balam (NVIDIA); Boris Ginsburg (NVIDIA)
- 590** **Genhancer: High-Fidelity Speech Enhancement via Generative Modeling on Discrete Codec Tokens**
Haici Yang (Indiana University); Jiaqi Su (Adobe Research); Minje Kim (University of Illinois at Urbana-Champaign); Zeyu Jin (Adobe Research)
- 841** **Thunder : Unified Regression-Diffusion Speech Enhancement with a Single Reverse Step using Brownian Bridge**
Thanapat Trachu (Chulalongkorn University); Chawan Piansaddhayanon (Chulalongkorn University); Ekapol Chuangsuwanich (Chulalongkorn University)
- 1077** **Pre-training Feature Guided Diffusion Model for Speech Enhancement**
Yiyuan Yang (University of Oxford); Niki Trigoni (University of Oxford); Andrew Markham (University of Oxford)
- 1545** **Guided conditioning with predictive network on score-based diffusion model for speech enhancement**

Dail Kim (Hanyang University); Da-Hee Yang (Hanyang University); Donghyun Kim (Hanyang University); Joon-Hyuk Chang (Hanyang University); Jeonghwan Choi (Samsung Electronics); Moa Lee (Samsung Electronics); Jaemo Yang (Samsung Electronics); Han-gil Moon (Samsung Electronics)

Oral Session: Pathological Speech Analysis 1

A13-01

Location: Panacea Amphitheater

96 Sustained Vowels for Pre- vs Post-Treatment COPD Classification

Andreas Triantafyllopoulos (Technical University of Munich); Anton Batliner (University of Munich); Wolfgang Mayr (University Hospital Augsburg); Markus Fendler (University Hospital Augsburg); Florian B Pokorny (Medical University of Graz); Maurice Gerczuk (University of Augsburg); Shahin Amiriparian (Technical University of Munich); Thomas Berghaus (University Hospital Augsburg); Prof. Dr. Bjoern Schuller (Imperial College London)

522 Exploiting Foundation Models and Speech Enhancement for Parkinson's Disease Detection from Speech in Real-World Operative Conditions

Moreno La Quatra (Kore University of Enna); Maria Francesca Turco (Kore University of Enna); Torbjørn Svendsen (NTNU); Giampiero Salvi (NTNU); Juan Rafael Orozco-Aroyave (University of Antioquia); Sabato M Siniscalchi (Università degli Studi di Palermo)

1075 Adversarial Robustness Analysis in Automatic Pathological Speech Detection Approaches

Mahdi Amiri (Idiap Research Institute; École Polytechnique Fédérale de Lausanne); Ina Kodrasi (Idiap Research Institute)

1799 Automatic Children Speech Sound Disorder Detection with Age and Speaker Bias Mitigation

Gahye Kim (Sogang University); Yunjung Eom (Sogang university); Selina(Seim) Sung (University of Wisconsin-Madison, Sogang University); Seunghee Ha (Hallym University); Tae-Jin Yoon (Sungshin Women's University); Jung-min So (Sogang University)

2039 The MARRYS helmet: A new device for researching and training "jaw dancing"

Vidar Freyr Gudmundsson (University of Southern Denmark); Keve Márton Gönczi (University of Southern Denmark); Malin Svensson Lundmark (Lund University); Donna M Erickson (Haskins labs); Oliver Niebuhr (University of Southern Denmark)

Poster Session: Corpora-based Approaches in Automatic Emotion Recognition

A3-P1-A

Location: Poster Area 1A

921 Reinforcement Learning based Data Augmentation for Noise Robust Speech Emotion Recognition

Sumit Ranjan (TCS Research); Rupayan Chakraborty (TCS Research); Sunil Kumar Kopparapu (TCS Research)

1248 Unsupervised Domain Adaptation for Speech Emotion Recognition using K-Nearest Neighbors Voice Conversion

Pravin Mote (The University of Texas at Dallas); Berrak Sisman (The University of Texas at Dallas); Carlos Busso (University of Texas at Dallas)

1374 An Effective Local Prototypical Mapping Network for Speech Emotion Recognition

Yuxuan Xi (National Engineering Research Center of Speech and Language Information Processing); Yan Song (USTC); Lirong Dai (University of Science and Technology of China); Haoyu Song (The Australian National University); Ian McLoughlin (Singapore Institute of Technology)

1956 Confidence-aware Hypothesis Transfer Networks for Source-Free Cross-Corpus Speech Emotion Recognition

Wang jincen (southeast university); Yan Zhao (Southeast University); Cheng Lu (Southeast University); Hailun lian (Southeast University); Hongli Chang (University of Electronic Science and Technology of China); Yuan Zong (South-east University); Wenming Zheng (Southeast University)

2385 Speech Emotion Recognition with Multi-Level Acoustic and Semantic Information Extraction and Interaction

Yuan Gao (Kyoto University); Hao Shi (Kyoto University); Chenhui Chu (Kyoto University); Tatsuya Kawahara (Kyoto University)

Poster Session: Analysis of Speakers States and Traits**A3-P1-B**

Location: Poster Area 1B

277 How rhythm metrics are linked to produced and perceived speaker charisma

Oliver Niebuhr (University of Southern Denmark); Nafiseh Taghva (Shiraz University)

347 Detecting Empathy in Speech

Run Chen (Columbia University); Haozhe Chen (Columbia University); Anushka Kulkarni (Columbia University); Eleanor M Lin (Columbia University); Linda Pang (Columbia University); Divya Tadimeti (UC Berkeley); Jun Shin (Columbia University); Julia Hirschberg (Columbia University)

1097 Exploring Gender-Specific Speech Patterns in Automatic Suicide Risk Assessment

Maurice Gerczuk (University of Augsburg); Shahin Amiriparian (Technical University of Munich); Justina Lutz (District Hospital Augsburg); Wolfgang Strube (District Hospital Augsburg); Irina Papazova (District Hospital Augsburg); Alkomiet Hasan (District Hospital Augsburg); Prof. Dr. Bjoern Schuller (Imperial College London)

1871 Modelling Lexical Characteristics of the Healthy Aging Population: A Corpus-Based Study

Kunmei Han (National University of Singapore)

1962 A Functional Trade-off between Prosodic and Semantic Cues in Conveying Sarcasm

Zhu Li (University of Groningen); Xiyuan Gao (Groningen University); Yuqing Zhang (University of Groningen); Shekhar Nayak (University of Groningen); Matt Coler (University of Groningen)

2103 Multimodal Belief Prediction

John Murzaku (Stony Brook University); Adil Soubki (Stony Brook University); Owen Rambow (Stony Brook University)

2483 Learning Representation of Therapist Empathy in Counseling Conversation Using Siamese Hierarchical Attention Network

Dehua TAO (The Chinese University of Hong Kong); Tan Lee (The Chinese University of Hong Kong); Harold Chui (The Chinese University of Hong Kong); Sarah Luk (The Chinese University of Hong Kong)

Poster Session: Audio Captioning, Tagging, and Audio-Text Retrieval**A5-P2-A**

Location: Poster Area 2A

65 Enhancing Automated Audio Captioning via Large Language Models with Optimized Audio Encoding

Jizhong Liu (Xiaomi); Gang Li (Xiaomi); Junbo Zhang (Xiaomi); Heinrich Dinkel (Xiaomi); Yongqing Wang (xiaomi); Zhiyong Yan (Xiaomi); Yujun Wang (xiaomi); Bin Wang (Xiaomi AI Lab)

242 Streaming Audio Transformers for Online Audio Tagging

Heinrich Dinkel (Xiaomi); Zhiyong Yan (Xiaomi); Yongqing Wang (xiaomi); Junbo Zhang (Xiaomi); Yujun Wang (xiaomi); Bin Wang (Xiaomi AI Lab)

593 Audio-text Retrieval with Transformer-based Hierarchical Alignment and Disentangled Cross-modal Representation

Yifei Xin (Peking University); Zhihong Zhu (Peking University); Xuxin Cheng (Peking University); Xusheng Yang (Peking University); Yuexian Zou (Peking University)

1268 PFCA-Net: Pyramid Feature Fusion and Cross Content Attention Network for Automated Audio Captioning

Jianyuan Sun (University of Sheffield); Wenwu Wang (University of Surrey); Mark D. Plumbley (University of Surrey)

1315 ParaCLAP -- Towards a general language-audio model for computational paralinguistic tasks

Xin Jing (Technische Universität München); Andreas Triantafyllopoulos (Technische Universität München); Prof. Dr. Bjoern Schuller (Imperial College London)

1331 Efficient CNNs with Quaternion Transformations and Pruning for Audio Tagging

Aryan Chaudhary (IIIT Delhi); Arshdeep Singh (University of Surrey); Vinayak Abrol (Indraprastha Institute of Technology Delhi); Mark D. Plumbley (University of Surrey)

1680 Efficient Audio Captioning with Encoder-Level Knowledge Distillation

Xuenan Xu (Shanghai Jiao Tong University); Haohe Liu (University of Surrey); Mengyue Wu (Shanghai Jiao Tong University); Wenwu Wang (University of Surrey); Mark D. Plumbley (University of Surrey)

Poster Session: Speech Enhancement**A6-P2-A**

Location: Poster Area 3A

129 Are Recent Deep Learning-Based Speech Enhancement Methods Ready to Confront Real-World Noisy Environments?

Candy Olivia Mawlim (Japan Advanced Institute of Science and Technology); Masashi Unoki (JAIST); Shogo Okada (Japan Advanced Institute of Science and Technology)

166 Neural Network Augmented Kalman Filter for Robust Acoustic Howling Suppression

Yixuan Zhang (The Ohio State University); Hao Zhang (Tencent AI Lab); Meng Yu (Tencent); Dong Yu (Tencent AI Lab)

620 Reducing Speech Distortion and Artifacts for Speech Enhancement by Loss Function

Haixin Guan (University of Science and Technology of China; Unisound AI Technology Co., Ltd); Wei Dai (Unisound); Guangyong Wang (Unisound AI Technology Co., Ltd); Xiaobin Tan (University of Science and Technology of China); Peng Li (Beijing iplustek Co., Ltd); Jiaen Liang (Unisound)

639 QMixCAT: Unsupervised Speech Enhancement Using Quality-guided Signal Mixing and Competitive Alternating Model Training

Shilin Wang (Shanghai Normal University); Haixin Guan (University of Science and Technology of China; Unisound AI Technology Co., Ltd); yanhua long (Shanghai Normal University)

1069 Improving Speech Enhancement by Integrating Inter-Channel and Band Features with Dual-branch Conformer

Jizhen Li (Wuhan University); Xinmeng Xu (Wuhan University); Weiping Tu (Wuhan University); Yuhong Yang (Wuhan University); Rong Zhu (Wuhan University)

1266 Beyond Performance Plateaus: A Comprehensive Study on Scalability in Speech Enhancement

Wangyou Zhang (Shanghai Jiao Tong University); Kohei Saijo (Waseda University); Jee-weon Jung (Carnegie Mellon University); Chenda Li (Shanghai Jiao Tong University); Shinji Watanabe (Carnegie Mellon University); Yanmin Qian (Shanghai Jiao Tong University)

1488 Improved Remixing Process for Domain Adaptation-Based Speech Enhancement by Mitigating Data Imbalance in Signal-to-Noise Ratio

Li Li (CyberAgent, Inc.); Shogo Seki (CyberAgent, Inc.)

1831 An Exploration of Length Generalization in Transformer-Based Speech Enhancement

Qiquan Zhang (The University of New South Wales); Hongxu Zhu (National University of Singapore); Xinyuan Qian (USTB); Eliathamby Ambikairajah (The University of New South Wales); Haizhou Li (The Chinese University of Hong Kong (Shenzhen))

2114 RaD-Net 2: A causal two-stage repairing and denoising speech enhancement network with knowledge distillation and complex axial self-attention

Mingshuai Liu (NWPU); Zhuangqi Chen (Bytedance); Xiaopeng Yan (Northwestern Polytechnical University); Yuanjun Lv (Northwestern Polytechnical University); Xianjun Xia (RTC Lab, ByteDance); Chuanzeng Huang (Speech, Audio and Music Intelligence (SAMI) group, ByteDance); Yijian Xiao (ByteDance); Lei Xie (NWPU)

2244 DNN-based monaural speech enhancement using alternate analysis windows for phase and magnitude modification

Xi Liu (Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, TX, USA); John H Hansen (Univ. of Texas at Dallas)

Poster Session: General Topics in ASR

A8-P2 Location: Poster Area 3A, Poster Area 3B

104 A Multitask Training Approach to Enhance Whisper with Open-Vocabulary Keyword Spotting

Yuang Li (Huawei); Min Zhang (Huawei); Chang Su (Huawei); Yinglu Li (HUAWEI TECHNOLOGIES CO., LTD.); Xiaosong Qiao (Huawei); Mengxin Ren (Huawei); Miaomiao Ma (Huawei); Daimeng Wei (Huawei); Shimin Tao (Huawei); Hao Yang (Huawei)

368 Investigating ASR Error Correction with Large Language Model and Multilingual 1-best Hypotheses

Sheng Li (National Institute of Information & Communications Technology (NICT)); Chen Chen (Nanyang Technological University); Chin Yuen Kwok (Nanyang Technological University); Chenhui Chu (Kyoto University); Eng Siong Chng (Nanyang Technological University); Hisashi Kawai (NICT)

377 Improving Domain-Specific ASR with LLM-Generated Contextual Descriptions

Jiwon Suh (Hanyang University); Injae Na (Hanyang University); Woohwan Jung (Hanyang University)

444 DYSARTHIC SPEECH RECOGNITION USING CURRICULUM LEARNING AND ARTICULATORY FEATURE EMBEDDING

I-Ting Hsieh (National Cheng Kung University); Chung-Hsien Wu (National Cheng Kung University)

472 A Comparative Analysis of Bilingual and Trilingual Wav2Vec Models for Automatic Speech Recognition in Multilingual Oral History Archives

Jan Lehečka (University of West Bohemia); Josef V. Psutka (University of West Bohemia); Lubos Smidl (University of West Bohemia); Pavel Ircing (University of West Bohemia); Josef Psutka (University of West Bohemia)

583 An efficient text augmentation approach for contextualized Mandarin speech recognition

Naijun Zheng (Huawei Technologies Co., Ltd.); Xucheng Wan (Huawei Technologies Co., Ltd.); KAI LIU (Huawei Technologies Co., Ltd.); Ziqing Du (Huawei Technologies Co., Ltd.); zhou huan (AARC, Huawei Technologies Co., Ltd.)

731 CrisperWhisper: Accurate Timestamps on Verbatim Speech Transcriptions

Mario Zusag (*myReha GmbH*); Laurin Wagner (*nyra health GmbH*); Bernhad Thallinger (*nyra health GmbH*)

855 DualPure: An Efficient Adversarial Purification Method for Speech Command Recognition

Hao Tan (*Harbin Institute of Technology (Shenzhen)*); Xiaochen Liu (*Pengcheng Lab*); Huan Zhang (*Harbin Institute of Technology*); Zhang Junjian (*Guangzhou University*); Yaguan QIAN (*Zhejiang University of Science and Technology*); Zhaoquan Gu (*Peng Cheng Laboratory*)

1231 Fine-Tuning Strategies for Dutch Dysarthric Speech Recognition: Evaluating the Impact of Healthy, Disease-Specific, and Speaker-Specific Data

Spyretta Leivaditi (*University of Groningen*); Tatsunari Matsushima (*IGSA Inc.*); Matt Coler (*University of Groningen*); Shekhar Nayak (*University of Groningen*); Vass Verkhodanova (*University of Groningen*)

1360 Enhancing Dysarthric Speech Recognition for Unseen Speakers via Prototype-Based Adaptation

shiyao wang (*Nankai University*); Shiwang Zhao (*Nankai University*); Jiaming Zhou (*Nankai University*); Aobo Kong (*Nankai University*); Yong Qin (*Nankai University*)

2157 On Disfluency and Non-lexical Sound Labeling for End-to-end Automatic Speech Recognition

Peter Mihajlik (*BME-TMIT*); Yan Meng (*BME-TMIT*); Mate S Kadar (*Budapest University of Technology and Economics*); Julian Linke (*TU Graz (SPSC)*); Barbara Schuppler (*Graz University of Technology*); Katalin Mády (*Hungarian Research Centre for Linguistics*)

2246 Inclusive ASR for Disfluent Speech: Cascaded Large-Scale Self-Supervised Learning with Targeted Fine-Tuning and Data Augmentation

Dena F Mujtaba (*Michigan State University*); Nihar R Mahapatra (*Michigan State University*); Megan Arney (*Michigan State University*); J Scott Yaruss (*Michigan State University*); Caryn Herring (*Friends: The National Association of Young People Who Stutter*); Jia Bin (*Michigan State University*)

2341 A layer-wise analysis of Mandarin and English suprasegmentals in SSL speech models

Cibran A de la Fuente (*Stanford University*); Dan Jurafsky (*Stanford University*)

Poster Session: Spoken Language Understanding

A12-P2-A

Location: Poster Area 4A

26 Prompting Whisper for QA-driven Zero-shot End-to-end Spoken Language Understanding

Mohan Li (*Toshiba Europe Ltd*); Simon Keizer (*Toshiba Europe Ltd*); Rama S Doddipatla (*Toshiba Europe LTD*)

543 Efficient SQA from Long Audio Contexts: A Policy-driven Approach

Alexander Johnson (*UCLA*); Peter W Plantinga (*JPMorgan Chase & Co.*); Pheobe Sun (*JP Morgan Chase & Co.*); Swaroop Gadiyaram (*JP Morgan Chase & Co.*); Abenezzer Girma (*JP Morgan Chase & Co.*); Ahmad Emami (*JP Morgan*)

783 Textless Dependency Parsing by Labeled Sequence Prediction

Shunsuke Kando (*The University of Tokyo*); Yusuke Miyao (*University of Tokyo*); Jason Naradowsky (*Square-Enix*); Shinnosuke Takamichi (*The University of Tokyo*)

840 Towards Speech Classification from Acoustic and Vocal Tract data in Real-time MRI

yaoyao yue (*The University of Sydney*); Michael I Proctor (*Macquarie University*); Amelia Gully (*University of York*); Kirrie Ballard (*University of Sydney*); Luping Zhou (*University of Sydney*); Rijul Gupta (*The University of Sydney*); Tharinda Piyadasa (*The University of Sydney*); Craig T Jin (*The University of Sydney*)

1292 AR-NLU: A Framework for Enhancing Natural Language Understanding Model Robustness against ASR Errors

Emmy Phung (*JP Morgan Chase & Co.*); Harsh Saiprasad Deshpande (*JP Morgan Chase & Co.*); Ahmad Emami (*JP*

Morgan); Kanishk Singh (Columbia University)

1976 VN-SLU: A Vietnamese Spoken Language Understanding Dataset

Tuyen Tran (Hanoi University of Science and Technology); Khanh Le (Hanoi University of Science and Technology); Ngoc Dang Nguyen (Hanoi University of Science and Technology); Minh Duc Vu (Hanoi University of Science and Technology); Huyen Ngo (Hanoi University of Science and Technology); Woomyoung Park (Naver Corporation); Thi Thu Trang Nguyen (Hanoi University of Science and Technology)

Poster Session: Speech and Multimodal Resources

A12-P2-B Location: Poster Area 4B

42 BESST Dataset: A Multimodal Resource for Speech-based Stress Detection and Analysis

Jan Pešán (Faculty of Information Technology - Brno University of Technology); Vojtěch Juřík (Faculty of Civil Engineering - Brno University of Technology); Martin Karafiat (BUT speech@fit); Jan Honza Cernocky (Brno University of Technology)

70 GLOBE: A High-quality English Corpus with Global Accents for Zero-shot Speaker Adaptive Text-to-Speech

Wenbin Wang (University of New South Wales); Yang Song (University of New South Wales); Sanjay Jha (University of New South Wales)

417 HebDB: a Weakly Supervised Dataset for Hebrew Speech Processing

Arnon Turetzky (Hebrew University of Jerusalem); Or Tal (The Hebrew University of Jerusalem); Yael Segal (Technion); Yehoshua Disen (Technion - Israel Institute of Technology); Ella Zeldes (The Hebrew University of Jerusalem); Amit Roth (The Hebrew University of Jerusalem); Eyal Cohen (Technion); Yosi Shrem (); Bronya Roni Chernyak (Technion - Israel Institute of Technology); Olga Seleznova (Technion); Joseph Keshet (Technion - Israel Institute of Technology); Yossi Adi (The Hebrew University of Jerusalem)

644 STraDa: A Singer Traits Dataset

Yuxuan KONG (Deezer); Viet-Anh Tran (Deezer); Romain Hennequin (Deezer Research)

978 MaViLS, a Benchmark Dataset for Video-to-Slide Alignment, Assessing Baseline Accuracy with a Multimodal Alignment Algorithm Leveraging Speech, OCR, and Visual Features

Katharina Anderer (Karlsruhe Institute of Technology); Anderas Reich (University of Hohenheim); Matthias Wölfel (Karlsruhe University of Applied Sciences)

1794 MultiTalk: Enhancing 3D Talking Head Generation Across Languages with Multilingual Video Dataset

Kim Sung-Bin (POSTECH); Lee Chae-Yeon (POSTECH); Gihun Son (Inha University); Oh Hyun-Bin (POSTECH); Janghoon Ju (KRAFTON); Suekyeong Nam (KRAFTON Inc.); Tae-Hyun Oh (POSTECH)

2273 Towards measuring fairness in speech recognition: Fair-Speech dataset

Irina-Elena Veliche (Meta); Zhuangqun Huang (Meta AI); Vineeth Ayyat Kochaniyan (apex fintech services LLC); Fuchun Peng (Meta); Ozlem Kalinli (Meta); Mike Seltzer (Meta)

2428 Codecfake: An Initial Dataset for Detecting LLM-based Deepfake Audio

Yi Lu (State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences); Yuankun Xie (Communication University of China); Ruibo Fu (National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences); Wen Zhengqi (CASIA); Jianhua Tao (Tsinghua University); Zhiyong Wang (University of Chinese Academy of Sciences); Xin Qi (University of Chinese Academy of Sciences); Xuefei Liu (Qiyuan Lab); Yongwei Li (Chinese Academy of Sciences); Yukun Liu (UCAS); Xiaopeng Wang (UCAS); Shuchen Shi (Shanghai Polytechnic University)

2440 SER Evals: In-domain and Out-of-domain benchmarking for speech emotion recognition

Mohamed Osman (Virginia Commonwealth University); Daniel Z Kaplan (N/A); Tamer Nadeem (Virginia Common-

wealth University)

Poster Session: Speech and Language in Health: from Remote Monitoring to Medical Conversations - 1 (Special Session)

SS-5A

Location: Yanis Club

15 Predicting Acute Pain Levels Implicitly from Vocal Features

Jennifer Williams (University of Southampton); Eike Schneiders (University of Nottingham); Henry Card (University of Southampton); Tina Seabrooke (University of Southampton); Beatrice JH Pakenham-Walsh (University of Southampton); Tayyaba Azim (University of Southampton); Lucy M Valls-Reed (Southampton University); Ganesh Vigneswaran (University of Southampton); John Robert Bautista (University of Missouri-Columbia); Rohan Chandra (UT Austin); Arya Farahi (University of Texas at Austin)

131 Developing an End-to-End Framework for Predicting the Social Communication Severity Scores of Children with Autism Spectrum Disorder

Jihyun Mun (Seoul National University); Sunhee Kim (Seoul National University); Minhwa Chung (Seoul National University)

156 Multimodal Fusion for Vocal Biomarkers Using Vector Cross-Attention

Vladimir Despotovic (Luxembourg institute of Health); Abir Elbeji (Luxembourg Institute of Health); Petr Nazarov (Luxembourg Institute of Health); Guy Fagherazzi (Luxembourg Institute of Health)

975 Towards Intelligent Speech Assistants in Operating Rooms: A Multimodal Model for Surgical Workflow Analysis

Kubilay Can Demir (Friedrich-Alexander Universität Erlangen-Nürnberg); Belén Lojo Rodríguez (Friedrich-Alexander Universität Erlangen-Nürnberg); Tobias Weise (Friedrich-Alexander-Universität Erlangen-Nürnberg); Andreas K Maier (Pattern Recognition Lab, FAU Erlangen-Nuremberg); Seung Hee Yang (Friedrich-Alexander Universität Erlangen-Nürnberg (FAU))

1092 Revealing Confounding Biases: A Novel Benchmarking Approach for Aggregate-Level Performance Metrics in Health Assessments

Stefano Gorla (thymia); Roseline Polle (thymia); Salvatore Fara (thymia); Nicholas Cummins (King's College London)

1484 Comparing ambulatory voice measures during daily life with brief laboratory assessments in speakers with and without vocal hyperfunction

Daryush Mehta (MGH); Jarrad Van Stan (Mass General, Harvard Medical School, MGH IHP); Hamzeh Ghasemzadeh (Massachusetts General Hospital); Robert Hillman (Massachusetts General Hospital)

1907 Reference-Free Estimation of the Quality of Clinical Notes Generated from Doctor-Patient Conversations

Mojtaba Kadhodaie Elyaderani (3M Health Information Systems); John Glover (3M | Health Information Systems / DSC); Thomas Schaaf (Solventum | Health Information Systems / M*Modal)

1926 Developing Multi-Disorder Voice Protocols: A team science approach involving clinical expertise, bio-ethics, standards and DEI.

yael bensoussan (University of South Florida); Satrajit S Ghosh (MIT); Anais Rameau (Weill Cornell Medicine); Micah Boyer (University of South Florida); Ruth Bahr (University of South Florida); stephanie watts (universiyt of south florida); Frank Rudzicz (University of Toronto); Don Bolser (University of South Florida); Jordan Lerner-Ellis (University of South Florida); Shaheen Awan (University of South Florida); Maria E Powell (Vanderbilt University Medical Center); Jean-Christophe Belisle-Pipon (Simon Fraser university); Vardit Ravitsky (Hasting Center); Alistair Johnson (Glowry); Alexandros Sigaras (Weill Cornell); Olivier Elemento (Weill Cornell); David Dorr (OHSU); Philip R Payne

(Washington University in St. Louis)

2224 A Multimodal framework for the assessment of the Schizophrenia spectrum

Gowtham Premananth (University of Maryland); Yashish M. Siriwardena (University of Maryland College Park); Philip Resnik (University of Maryland); Sonia Bansal (University of Maryland School of Medicine); Deanna L. Kelly (University of Maryland School of Medicine); Carol Y Espy-Wilson (University of Maryland)

2344 Self-Supervised Embeddings for Detecting Individual Symptoms of Depression

Sri Harsha Dumpala (Dalhousie University/Vector Institute); Katerina Dikaos (McMaster University); Abraham Nunes (Dalhousie University); Frank Rudzicz (Dalhousie University); Rudolf Uher (Dalhousie University); Sageev Oore (Dalhousie University/Vector Institute)

Time Slot 2

Oral Session: Speech and Brain

A1-02

Location: Acesso

223 Toward Fully-End-to-End Listened Speech Decoding from EEG Signals

Jihwan Lee (University of Southern California); Aditya Kommineni (University of Southern California); Tiantian Feng (University of Southern California); Kleanthis Avramidis (University of Southern California); Xuan Shi (University of Southern California); Sudarsana Reddy Kadiro (University of Southern California); Shrikanth Narayanan (University of Southern California)

382 Towards an End-to-End Framework for Invasive Brain Signal Decoding with Large Language Models

Sheng Feng (Shanghai Jiao Tong University); Heyang Liu (Shanghai Jiao Tong University); Yu Wang (Shanghai Jiao Tong University); Yan-Feng Wang (Cooperative medianet innovation center of Shanghai Jiao Tong University)

604 Refining Self-supervised Learnt Speech Representation using Brain Activations

HengYu Li (University of Science and Technology of China); KangDi Mei (University of Science and Technology of China); Zhaoci Liu (University of Science and Technology of China); Yang Ai (University of Science and Technology of China); Liping chen (University of Science and Technology of China); Jie Zhang (University of Science and Technology of China (USTC)); Zhen-Hua Ling (University of Science and Technology of China)

1616 Large Language Model-based FMRI Encoding of Language Functions for Subjects with Neurocognitive Disorder

Yuejiao Wang (The Chinese University of Hong Kong); Xianmin Gong (The Chinese University of Hong Kong); Lingwei Meng (The Chinese University of Hong Kong); Xixin Wu (The Chinese University of Hong Kong); Helen Meng (The Chinese University of Hong Kong)

2186 Exploring the Complementary Nature of Speech and Eye Movements for Profiling Neurological Disorders

Yuzhe Wang (Johns Hopkins University); Anna Favaro (Johns Hopkins University); Thomas Thebaud (Johns Hopkins University); Jesus Antonio Villalba (Johns Hopkins University); Najim Dehak (Johns Hopkins University); Laureano Moro-Velazquez (Johns Hopkins University)

2531 From sound to meaning in the auditory cortex: A neuronal representation and classification analysis

Kumar Neelabh (IIIT Hyderabad); Vishnu Sreekumar (International Institute of Information Technology, Hyderabad)

Oral Session: Emotion Recognition: Resources and Benchmarks

A3-02

Location: Aegle A

97 INTERSPEECH 2009 Emotion Challenge Revisited: Benchmarking 15 Years of Progress in Speech Emotion Recognition

Andreas Triantafyllopoulos (Technical University of Munich); Anton Batliner (University of Munich); Simon Rampp (University of Augsburg); Manuel Milling (University of Augsburg); Prof. Dr. Bjoern Schuller (Imperial College London)

788 EmoBox: Multilingual Multi-corpus Speech Emotion Recognition Toolkit and Benchmark

Ziyang Ma (Shanghai Jiao Tong University); Mingjie Chen (University of Sheffield); Hezhao Zhang (The University of Sheffield); Zhisheng Zheng (Shanghai Jiao Tong University); Wenxi Chen (Shanghai Jiao Tong University); Xiquan Li (Shanghai Jiao Tong University); JiaXin Ye (Fudan University); Xie Chen (Shanghai Jiaotong University); Thomas Hain (University of Sheffield)

1227 WHiSER: White House Tapes Speech Emotion Recognition Corpus

Abinay Reddy Naini (The University of Texas at Dallas); Lucas Goncalves (The University of Texas at Dallas); Mary Kohler (Laboratory for Analytic Sciences, North Carolina State University); Donita Robinson (Laboratory for Analytic Sciences, North Carolina State University); Elizabeth Richerson (Laboratory for Analytic Sciences, North Carolina State University); Carlos Busso (University of Texas at Dallas)

1827 Evaluating Transformer-Enhanced Deep Reinforcement Learning for Speech Emotion Recognition

Siddique Latif (Queensland University of Technology); Raja Jurdak (Queensland University of Technology); Prof. Dr. Bjoern Schuller (Imperial College London)

1947 Boosting Cross-Corpus Speech Emotion Recognition using CycleGAN with Contrastive Learning

Wang jincen (southeast university); Yan Zhao (Southeast University); Cheng Lu (Southeast University); Chuangao Tang (Nanjing Institute of Technology); Sunan Li (Southeast University); Yuan Zong (Southeast University); Wenming Zheng (Southeast University)

1983 What Does it Take to Generalize SER Model Across Datasets? A Comprehensive Benchmark

Adham Ibrahim Ibrahim (MBZUAI); Shady Shehata (Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)); Ajinkya Kulkarni (ValidSoft MBZUAI); Mukhtar Mohamed (Mohamed bin Zayed University of Artificial Intelligence); Muhammad Abdul-Mageed (iSchool@UBC)

Oral Session: Vision and Speech

A9-02

Location: Aegle B

Surv-09-1 Visually grounded speech models

David Harwath

526 AVCap: Leveraging Audio-Visual Features as Text Tokens for Captioning

Jongsuk Kim (KAIST); Jiwon Shin (KAIST); Junmo Kim (KAIST)

1063 Joint Speaker Features Learning for Audio-visual Multichannel Speech Separation and Recognition

Guinan Li (Chinese University of HongKong); Jiajun Deng (The Chinese University of HongKong); Youjun Chen (The Chinese University of Hong Kong); Mengzhe Geng (National Research Council Canada); Shujie HU (The Chinese University of Hong Kong); Zhe LI (Hong Kong Polytechnic University); Zengrui Jin (The Chinese University of Hong Kong); Tianzi Wang (The Chinese University of HongKong); Xurong Xie (Institute of Software, Chinese Academy of Sciences); Helen Meng (The Chinese University of Hong Kong); Xunying Liu (The Chinese University of Hong Kong)

1918 Visually-Conditioned Generative Error Correction for Robust Automatic Speech Recognition

Sreyan Ghosh (University of Maryland, College Park); Sonal Kumar (University of Maryland, College Park); Ashish Seth (IIT Madras); Purva Chiniya (Cisco Systems); Utkarsh Tyagi (University of Maryland, College Park); Ramani Duraiswami (The University of Maryland); Dinesh Manocha (University of Maryland at College Park)

2509 CNVSR 2023: The First Chinese Continuous Visual Speech Recognition Challenge

Chen Chen (Tsinghua University); Zehua Liu (Beijing University of Posts and Telecommunications); Xiaolou Li (Beijing University of Posts and Telecommunications); Lantian Li (Beijing University of Posts and Telecommunications); Dong Wang (Tsinghua University)

Oral Session: Speech Synthesis: Singing Voice Synthesis

A7-04 Location: Hippocrates

33 Singing Voice Data Scaling-up: An Introduction to ACE-Opencpop and ACE-KiSing

Jiatong Shi (Carnegie Mellon University); Yueqian Lin (Duke University); Xinyi Bai (Cornell University); Keyi Zhang (Imperial College London); Yuning Wu (Renmin University of China); Yuxun Tang (Renmin University of China); Yifeng Yu (Georgia Institute of Technology); Qin Jin (Renmin University of China); Shinji Watanabe (Carnegie Mellon University)

49 X-Singer: Code-Mixed Singing Voice Synthesis via Cross-Lingual Learning

Ji-Sang Hwang (Netmarble AI Center); HyeongRae Noh (Netmarble AI Center); Yoonseok Hong (Netmarble AI Center); Insoo Oh (Netmarble AI Center)

678 MakeSinger: A Semi-Supervised Training Method for Data-Efficient Singing Voice Synthesis via Classifier-free Diffusion Guidance

Semin Kim (Seoul National University); Myeonghun Jeong (Seoul National University); Hyeonseung Lee (Seoul National University); Minchan Kim (Seoul National University); Byoung Jin Choi (Seoul National University); Nam Soo Kim (Seoul National University)

1720 Period Singer: Integrating Periodic and Aperiodic Variational Autoencoders for Natural-Sounding End-to-End Singing Voice Synthesis

Taewoo Kim (Korea Electronics Technology Institute); Choongsang Cho (Korea Electronics Technology Institute); Young Han Lee (Korea Electronics Technology Institute)

1837 An End-to-End Approach for Chord-Conditioned Song Generation

Shuochen Gao (Tsinghua University); Shun Lei (Tsinghua University); Fan Zhuo (Kunlun Skywork Technology Co., Ltd.); Hangyu Liu (Kunlun Skywork Technology Co., Ltd.); Feng Liu (Kunlun Skywork Technology Co., Ltd.); Boshi Tang (Tsinghua University); Qiaochu Huang (Tsinghua University); Shiyin Kang (Kunlun Skywork Technology Co., Ltd.); Zhiyong Wu (Tsinghua University)

2504 Challenge of Singing Voice Synthesis Using Only Text-To-Speech Corpus With FIRNet Source-Filter Neural Vocoder

Takuma Okamoto (National Institute of Information and Communications Technology); Yamato Ohtani (National Institute of Information and Communications Technology); Sota Shimizu (Kobe University); Tomoki Toda (Nagoya University); Hisashi Kawai (NICT)

Oral Session: LLM in ASR

A8-03 Location: Iasso

Surv-08-2 Development of Spoken Language Models

Hung-yi Lee

571 Speech ReaLLM –Real-time Speech Recognition with Multimodal Language Models by Teaching the Flow of Time

Frank Seide (Meta Platforms, Inc.); Yangyang Shi (Meta Platforms, Inc.); Morrie Doulaty (Meta); Yashesh Gaur (Meta Platforms, Inc.); Junteng Jia (Meta Platforms, Inc.); Chunyang Wu (Meta Platforms, Inc.)

968 A Transcription Prompt-based Efficient Audio Large Language Model for Robust Speech Recognition

Yangze Li (Northwestern Polytechnical University); xiong wang (Tencent); Songjun Cao (Tencent); Yike Zhang (Tencent); Long Ma (Tencent); Lei Xie (NWPU)

987 Pinyin Regularization in Error Correction for Chinese Speech Recognition with Large Language Models

Zhiyuan Tang (Tencent); Dong Wang (Tsinghua University); Shen Huang (Tencent Research); Shi-dong Shang (tencent)

1903 Speech Prefix-Tuning with RNNT Loss for Improving LLM Predictions

Murali Karthick Baskar (Google Inc); Andrew Rosenberg (Google LLC); Bhuvana Ramabhadran (Google); Neeraj Gaur (Google); Zhong Meng (Google)

Oral Session: Spoken Document Summarization**A12-02**

Location: Melampus

349 Sentence-wise Speech Summarization: Task, Datasets, and End-to-End Modeling with LM Knowledge Distillation

Kohei Matsuura (NTT); Takanori Ashihara (NTT); Takafumi Moriya (NTT); Masato Mimura (NTT corporation); Takatomo Kano (NTT Corporation); Atsunori Ogawa (NTT Corporation); Marc Delcroix (NTT)

1428 An End-to-End Speech Summarization Using Large Language Model

Hengchao Shang (HW-TSC); ZongYao LI (HW-TSC); Jiaxin GUO (Huawei); Shaojun Li (Huawei TSC); Zhiqiang Rao (Huawei); Yuanchang Luo (HW-TSC); Daimeng Wei (Huawei); Hao Yang (Huawei)

2213 Prompting Large Language Models with Audio for General-Purpose Speech Summarization

Wonjune Kang (Massachusetts Institute of Technology); Deb Roy (Massachusetts Institute of Technology)

2250 Real-time Speech Summarization for Medical Conversations

Khai Le-Duc (University of Toronto); Khai-Nguyen Nguyen (College of William and Mary); Long Vo-Dang (University of Cincinnati); Truong Son Hy (Indiana State University)

2268 Optimizing the role of human evaluation in LLM-based spoken document summarization systems

Kelsey N Kraus (Cisco Systems); Margaret Kroll (Cisco Systems)

2389 Key-Element-Informed sLLM Tuning for Document Summarization

Sangwon Ryu (POSTECH); Heejin Do (POSTECH); Yunsu Kim (aiXplain, Inc.); Gary Geunbae Lee (Postech); Jungseul Ok (POSTECH)

Oral Session: Audio-Text Retrieval**A5-03**

Location: Panacea Amphitheater

41 Domain Adaptation for Contrastive Audio-Language Models

Soham Deshmukh (Microsoft); Rita Singh (Carnegie Mellon University); Bhiksha Raj (Carnegie Mellon University)

193 tinyCLAP: Distilling Contrastive Language-Audio Pretrained Models

Francesco Paissan (Fondazione Bruno Kessler); Elisabetta Farella (Fondazione Bruno Kessler)

405 DiffATR: Diffusion-based Generative Modeling for Audio-Text Retrieval

Yifei Xin (Peking University); Xuxin Cheng (Peking University); Zhihong Zhu (Peking University); Xusheng Yang (Peking University); Yuexian Zou (Peking University)

420 Bridging Language Gaps in Audio-Text Retrieval

Zhiyong Yan (Xiaomi); Heinrich Dinkel (Xiaomi); Yongqing Wang (xiaomi); Jizhong Liu (Xiaomi); Junbo Zhang (Xiaomi); Yujun Wang (xiaomi); Bin Wang (Xiaomi AI Lab)

492 BTS: Bridging Text and Sound Modalities for Metadata-Aided Respiratory Sound Classification

June-Woo Kim (Kyungpook National University); Miika Toikkanen (MODULABS); Yera Choi (NAVER Digital Healthcare LAB, NAVER Cloud); Seong-Eun Moon (NAVER Digital Healthcare LAB); Ho Young Jung (Kyungpook National University)

1433 Enhanced Feature Learning with Normalized Knowledge Distillation for Audio Tagging

Yuwu Tang (Hangzhou Huacheng Network Technology Co., Ltd.); Ziang Ma (Hangzhou Huacheng Network Technology Co., Ltd.); Haitao Zhang (Hangzhou Huacheng Network Technology Co., Ltd.)

Poster Session: Innovative Methods in Phonetics and Phonology

A2-P1

Location: Poster Area 1A

181 K-means and hierarchical clustering of f0 contours

Constantijn Kaland (Institute of Linguistics - University of Cologne); Jeremy Steffman (Linguistics and English Language - University of Edinburgh); Jennifer S Cole (Northwestern University)

318 Phonological Feature Detection for US English using the Phonet Library

Harsha Veena Tadavarthy (University of Georgia); Austin Jones (University of Georgia); Margaret E L Renwick (University of Georgia)

357 The sub-band cepstrum as a tool for locating local spectral regions of phonetic sensitivity: A first attempt with multi-speaker vowel data

Michael Lambropoulos (Australian National University); Frantz Clermont (Australian National University); Shunichi Ishihara (Australian National University)

429 Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment

Rotem Rouso (Technion); Eyal Cohen (Technion - Israel institute of technology); Joseph Keshet (Technion - Israel Institute of Technology); Eleanor Chodroff (University of Zurich)

1155 Using wav2vec 2.0 for phonetic classification tasks: methodological aspects

LILA KIM (Laboratoire de Phonétique et Phonologie); Cédric Gendrot (LPP)

1208 Speaker- and Text-Independent Estimation of Articulatory Movements and Phoneme Alignments from Speech

Tobias Weise (Friedrich-Alexander-Universität Erlangen-Nürnberg); Philipp Klumpp (Pattern Recognition Lab, FAU Erlangen-Nuremberg); Kubilay Can Demir (Friedrich-Alexander Universität Erlangen-Nürnberg); Paula Andrea Pérez-Toro (Friedrich-Alexander-Universität Erlangen-Nürnberg); Maria Schuster (LMU); Elmar Noeth (Friedrich Alexander Universität, Erlangen-Nuremberg); Bjoern Heismann (Friedrich-Alexander-Universität Erlangen-Nürnberg); Andreas K Maier (Pattern Recognition Lab, FAU Erlangen-Nuremberg); Seung Hee Yang (Friedrich-Alexander Universität Erlangen-Nürnberg (FAU))

1269 Speaker-Independent Acoustic-to-Articulatory Inversion through Multi-Channel Attention Discrim-

inator

Woojin Chung (Yonsei University); Hong-Goo Kang (Yonsei University)

1517 Preprocessing for acoustic-to-articulatory inversion using real-time MRI movies of Japanese speech
 Anna Oura (Waseda university); Hideaki Kikuchi (Waseda university); Tetsunori Kobayashi (Waseda University)

1990 Deciphering Assamese Vowel Harmony with Featural InfoWaveGAN
 Sneha Ray Barman (Indian Institute of Technology Guwahati); Shakuntala Mahanta (Indian Institute of Technology Guwahati); Neeraj Sharma (IIT Guwahati)

2286 The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panāra
 Emily P Ahn (University of Washington); Eleanor Chodroff (University of Zurich); Myriam Lapierre (University of Washington); Gina-Anne Levow (University of Washington)

Poster Session: Voice, Tones and F0

A2-P1-B Location: Poster Area 1B

303 Impact of the tonal factor on diphthong realizations in Standard Mandarin with Generalized Additive Mixed Models
 Chenyu Li (Université Paris Cité, CNRS, LLF); Jalal Al-Tamimi (Université Paris Cité, CNRS, LLF)

1169 The Use of Modifiers and f0 in Remote Referential Communication with Human and Computer Partners
 Iona Gessinger (University College Dublin); Bistra Andreeva (Saarland University); Benjamin R. Cowan (University College Dublin)

1393 A Study on the Information Mechanism of the 3rd Tone Sandhi Rule in Mandarin Disyllabic Words
 Liu Xiaowang (Beijing Language and Culture University); Jinsong Zhang (Beijing Language and Culture University)

1592 Gender and age based f0-variation in the German Plapper Corpus
 Melanie Weirich (Friedrich-Schiller-University Jena); Daniel Duran (Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)); Stefanie Jannedy (Leibniz-ZAS Berlin)

2240 Voice quality in telephone speech: Comparing acoustic measures between VoIP telephone and high-quality recordings
 Chenzi Xu (University of Oxford); Vincent Hughes (University of York); Paul Foulkes (University of York); Philip Harrison (University of York); Jessica Wormald (University of York); David L. van der Vloed (Netherlands Forensic Institute); Finnian Kelly (Oxford Wave Research); Poppy Welch (University of York)

Poster Session: Speaker and Language Identification and Diarization

A4-P2 Location: Poster Area 2A, Poster Area 2B

235 Data Cleansing for End-to-End Neural Singer Diarization Using Neural Analysis and Synthesis Framework
 Hokuto Munakata (LY Corporation); Ryo Terashima (LY Corporation); Yusuke Fujita (LY Corporation)

436 SOMSRED: Sequential Output Modeling for Joint Multi-talker Overlapped Speech Recognition and Speaker Diarization
 Naoki Makishima (NTT); Naotaka Kawata (NTT); Mana Ihori (NTT); Tomohiro Tanaka (NTT); Shota Orihashi (NTT)

Corporation); Atsushi Ando (NTT Corporation); Ryo Masumura (NTT Corporation)

802 Exploring Spoken Language Identification Strategies for Automatic Transcription of Multilingual Broadcast and Institutional Speech

Martina Valente (Almawave); Fabio Brugnara (Almawave); Giovanni Morrone (Almawave); Enrico Zovato (Almawave); Leonardo Badino (Almawave)

845 AG-LSEC: Audio Grounded Lexical Speaker Error Correction

Rohit Paturi (AWS AI Labs); Xiang Li (AWS AI Labs); Sundararajan Srinivasan (AWS AI Labs)

917 ASOBO: Attentive Beamformer Selection for Distant Speaker Diarization in Meetings

Théo Mariotte (LTCI, Télécom Paris, Institut Polytechnique de Paris); Anthony Larcher (Université du Mans - LIUM); Silvio Montresor (Le Mans University - LAUM, UMR 6613, IA-GS); Jean-Hugh Thomas (Le Mans University - LAUM, UMR 6613, IA-GS)

923 Multi-latency look-ahead for streaming speaker segmentation

Bilal Rahou (IRIT); Hervé Bredin (CNRS)

1286 Once more Diarization: Improving meeting transcription systems through segment-level speaker reassignment

Christoph Boeddeker (Paderborn University); Tobias Cord-Landwehr (Paderborn University); Reinhold Haeb-Umbach (University of Paderborn)

1287 Hybrid-Diarization System with Overlap Post-Processing for the DISPLACE 2024 Challenge

Gabriel Pirlogeanu (University Politehnica of Bucharest/ Zevo-Tech); Octavian Pascu (University Politehnica of Bucharest/ Zevo-Tech); Alexandru-Lucian Georgescu (University Politehnica of Bucharest/ Zevo-Tech); Horia Cucu (University Politehnica of Bucharest)

1657 Exploring Energy-Based Models for Out-of-Distribution Detection in Dialect Identification

yaqian hao (China Mobile Research Institute); Chenguang Hu (chinamobile); Yingying Gao (China Mobile Research Institute); Shilei Zhang (China Mobile Research Institute); Junlan Feng (China Mobile Research Institute)

1833 The Second DISPLACE Challenge: Diarization of SPEAKER and LANGUAGE in Conversational Environments

Shareef Babu Kalluri (Indian Institute of Science, Bangalore); Prachi Singh (Indian Institute of Science, Bangalore); Pratik Roy chowdhuri (National Institute of Technology Karnataka, Surathkal); Apoorva Kulkarni (Indian Institute of science); Shikha Baghel (National Institute of Technology Karnataka, Surathkal); Pradyoth Hegde (Indian Institute of Information Technology Dharwad); swapnil sontakke (Indian Institute of Information Technology Dharwad); Deepak T (IIT-Dharwad); Mahadeva Prasanna (IIT Dharwad); Deepu Vijayasanen (National Institute of Technology Karnataka, Surathkal); Sriram Ganapathy (Google Research ; Indian Institute of Science, Bangalore, India, 560012)

1885 Speaker Change Detection with Weighted-sum Knowledge Distillation based on Self-supervised Pre-trained Models

Hang SU (Beijing Xiaomi Mobile Software Co., Ltd); Yuxiang Kong (Xiaomi Inc.); Lichun Fan (Beijing Xiaomi Mobile Software Co., Ltd); Peng Gao (Beijing Xiaomi Mobile Software Co., Ltd); Yujun Wang (xiaomi); Zhiyong Wu (Tsinghua University)

2462 TalTech-IRIT-LIS Speaker and Language Diarization Systems for DISPLACE 2024

Joonas Kalda (Tallinn University of Technology); Tanel Alumae (Tallinn University of Technology); Martin Lebourdais (IRIT/CNRS); Hervé Bredin (CNRS); Séverin BAROUDI (LIS); Ricard Marxer (Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon)

Poster Session: **Speech Coding**

A6-P2-B

Location: Poster Area 3B

108 A Low-Bitrate Neural Audio Codec Framework with Bandwidth Reduction and Recovery for High-Sampling-Rate Waveforms

Yang Ai (University of Science and Technology of China); Ye-Xin Lu (University of Science and Technology of China); Xiao-Hang Jiang (University of Science and Technology of China); Zheng-Yan Sheng (University of Science and Technology of China); Rui-Chen Zheng (University of Science and Technology of China); Zhen-Hua Ling (University of Science and Technology of China)

823 TD-PLC: A Semantic-Aware Speech Encoding for Improved Packet Loss Concealment

Jinghong Zhang (Beijing University of Posts and Telecommunications); zugang zhao (Beijing University of Posts and Telecommunications); kai niu (Beijing University of Posts and Telecommunications); zhiqiang he (Beijing University of Posts and Telecommunications); yonghui liu (Fanvil Link Technology Co.); jianbing liu (Fanvil Link Technology Co.)

1061 On Improving Error Resilience of Neural End-to-End Speech Coders

Kishan Gupta (Fraunhofer IIS); Nicola Pia (Fraunhofer IIS); Srikanth Korse (Fraunhofer IIS); Andreas Brendel (Fraunhofer IIS); Guillaume Fuchs (Fraunhofer IIS); Markus Multus (Fraunhofer IIS)

1072 Speech quality evaluation of neural audio codecs

Thomas Muller (Orange Innovation); Stephane Ragot (Orange); Laetitia Gros (Orange Innovation); Pierrick Philippe (Orange, Rennes); Pascal Scalart (IRISA Laboratory)

1764 BS-PLCNet 2: Two-stage Band-split Packet Loss Concealment Network with Intra-model Knowledge Distillation

Zihan Zhang (Northwestern Polytechnical University); Xianjun Xia (RTC Lab, ByteDance); Chuanzeng Huang (Speech, Audio and Music Intelligence (SAMI) group, ByteDance); Yijian Xiao (ByteDance); Lei Xie (NWPU)

2093 CodecFake: Enhancing Anti-Spoofing Models Against Deepfake Audios from Codec-Based Speech Synthesis Systems

Haibin Wu (National Taiwan University); Yuan Tseng (National Taiwan University); Hung-yi Lee (National Taiwan University)

Poster Session: Speech Synthesis: Expressivity and Emotion

A7-P1-A

Location: Poster Area 4A

398 EmoSphere-TTS: Emotional Style and Intensity Modeling via Spherical Emotion Vector for Controllable Emotional Text-to-Speech

Deok-Hyeon Cho (Korea University); Hyung-Seok Oh (Korea university); Seung-Bin Kim (Korea University); Sang-Hoon Lee (Ajou University); Seong-Whan Lee (Korea University)

980 Retrieval Augmented Generation in Prompt-based Text-to-Speech Synthesis with Context-Aware Contrastive Language-Audio Pretraining

Jinlong Xue (Beijing University of Posts and Telecommunications); Yayue Deng (Beijing University of Posts and Telecommunications); Yingming Gao (Beijing University of Posts and Telecommunications); Ya Li (Beijing University of Posts and Telecommunications)

1187 Differentiable Time-Varying Linear Prediction in the Context of End-to-End Analysis-by-Synthesis

Chin-Yun Yu (Queen Mary University of London); George Fazekas (QMUL)

1337 Controlling Emotion in Text-to-Speech with Natural Language Prompts

Thomas Bott (University of Stuttgart); Florian Lux (University of Stuttgart); Ngoc Thang Vu (University of Stuttgart)

1581 Expressive paragraph text-to-speech synthesis with multi-step variational autoencoder

xuyuan li (The Institute of Acoustics of the Chinese Academy of Sciences); Zengqiang Shang (The Institute of Acoustics of the Chinese Academy of Sciences); Peiyang Shi (The Institute of Acoustics of the Chinese Academy of Sciences)

ences); Hua Hua (Key Laboratory of Speech Acoustics and Content Understanding, Institute of Acoustics, China); ta li (Institute of Acoustics, Chinese Academy of Sciences); pengyuan zhang (Institute of Acoustics, Chinese Academy of Sciences)

1734 TSP-TTS: Text-based Style Predictor with Residual Vector Quantization for Expressive Text-to-Speech
Donghyun Seong (Hanyang University); Hoyoung Lee (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1862 Text-aware and Context-aware Expressive Audiobook Speech Synthesis
Dake Guo (Northwestern Polytechnical University); Xinfa Zhu (Northwestern Polytechnical University); Liumeng Xue (Northwestern Polytechnical University); yongmao zhang (Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, China); Wenjie Tian (Northwestern Polytechnical University); Lei Xie (NWPUP)

1989 Spontaneous Style Text-to-Speech Synthesis with Controllable Spontaneous Behaviors Based on Language Models

Weiqin Li (Tsinghua University); Peiji Yang (Tencent Technology (Shenzhen) Co.Ltd); yicheng zhong (Tencent Technology (Shenzhen) Company Li); Yixuan Zhou (Tsinghua University); Zhisheng Wang (Tencent); Zhiyong Wu (Tsinghua University); Xixin Wu (The Chinese University of Hong Kong); Helen Meng (The Chinese University of Hong Kong)

2216 GTR-Voice: Articulatory Phonetics Informed Controllable Expressive Speech Synthesis
Zehua Kcriss Li (University of Rochester); Meiying Chen (University of Rochester); Yi Zhong (Independent Researcher); Pinxin Liu (University of Rochester); Zhiyao Duan (University of Rochester)

2311 Emotion Arithmetic: Emotional Speech Synthesis via Weight Space Interpolation
Pavan Kalyan Tankala (IIT Bombay); Preeti Rao (Indian Institute of Technology Bombay); Preethi Jyothi (Indian Institute of Technology Bombay); Pushpak Bhattacharya (IIT Bombay)

Poster Session: Speech Synthesis: Tools and Data

A7-P1-B Location: Poster Area 4B

266 MSceneSpeech: A Multi-Scene Speech Dataset For Expressive Speech Synthesis
Qian Yang (Zhejiang University); Jialong Zuo (Zhejiang University); Zhe Su (Zhejiang University); Ziyue Jiang (Zhejiang University); Mingze Li (Zhejiang University); Zhou Zhao (Zhejiang University); chen feiyang (huawei); Zhefeng Wang (Huawei Cloud); Baoxing Huai (Huawei Cloud)

388 SRC4VC: Smartphone-Recorded Corpus for Voice Conversion Benchmark
Yuki Saito ("The University of Tokyo, Japan"); Takuto Igarashi (The University of Tokyo); Kentaro Seki (The University of Tokyo); Shinnosuke Takamichi (The University of Tokyo); Ryuichi Yamamoto (LY Corp.); Kentaro Tachibana (LY Corp.); Hiroshi Saruwatari (The University of Tokyo)

692 LibriTTS-P: A Corpus with Speaking Style and Speaker Identity Prompts for Text-to-Speech and Style Captioning

Masaya Kawamura (LY Corp.); Ryuichi Yamamoto (LY Corp.); Yuma Shirahata (LY Corp.); Takuya Hasumi (LY Corporation); Kentaro Tachibana (LINE Corp.)

1356 FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks
Min Ma (Google DeepMind); Yuma Koizumi (Google DeepMind); Shigeki Karita (Google DeepMind); Heiga Zen (Google); Jason Riesa (Google); Haruko Ishikawa (Google DeepMind); Michiel Bacchiani (Google)

1554 SaSLaW: Dialogue Speech Corpus with Audio-visual Egocentric Information Toward Environment-adaptive Dialogue Speech Synthesis
Osamu Take (The University of Tokyo); Shinnosuke Takamichi (The University of Tokyo); Kentaro Seki (The University of Tokyo)

of Tokyo); Yoshiaki Bando (National Institute of Advanced Industrial Science and Technology); Hiroshi Saruwatari (The University of Tokyo)

2281 1000 African Voices: Advancing inclusive multi-speaker multi-accent speech synthesis

Sewade O Ogun (Inria); Abraham T Owodunni (Intron Health); Tobi Olatunji (Intron Inc); Eniola Alese (amazethu); Babatunde Oladimeji (Amazethu); Tejumade Afonja (CISPA Helmholtz Center for Information Security, Saarland University, and AI Saturdays Lagos); Kayode K Olaleye (University of Pretoria); Naome A. Etori (University of Minnesota-Twin Cities); Tosin Adewumi (EISLAB, Machine Learning Group)

2343 WenetSpeech4TTS: A 12,800-hour Mandarin TTS Corpus for Large Speech Generation Model Benchmark

linhan ma (Northwestern Polytechnical University); Dake Guo (Northwestern Polytechnical University); Kun Song (Northwestern Polytechnical University); Yuepeng Jiang (Northwestern Polytechnical University); Shuai Wang (Shenzhen Research Institute of Big Data, Chinese University of Hong Kong (Shenzhen)); Liumeng Xue (Northwestern Polytechnical University); Weiming Xu (Northwest Polytechnic University); Huan Zhao (Northwestern Polytechnical University); Binbin Zhang (WeNet Open Source Community); Lei Xie (NWPU)

2421 Rasa: Building Expressive Speech Synthesis Systems for Indian Languages in Low-resource Settings

Praveen S V (Indian Institute of Technology Madras); Ashwin Sankar (AI4Bharat); Giri Raju (AI4Bharat); Mitesh M. Khapra (Indian Institute of Technology Madras)

Poster Session: Speech and Language in Health: from Remote Monitoring to Medical Conversations - 2 (Special Sessions)

SS-5B

Location: Yanis Club

158 Automatic Prediction of Amyotrophic Lateral Sclerosis Progression using Longitudinal Speech Transformer

Liming Wang (Massachusetts Institute of Technology); Yuan Gong (Massachusetts Institute of Technology); Nauman Dawalatabad (Massachusetts Institute of Technology); Marco Vilela (Takeda Development Center Americas, Inc.); Katerina Placek (Takeda Development Center Americas, Inc.); Brian Tracey (Takeda Development Center Americas, Inc.); Yishu Gong (Takeda); Alan Premasiri (ALS Therapy Development Institute); Fernando Vieira (ALS Therapy Development Institute); James Glass (Massachusetts Institute of Technology)

267 Towards objective and interpretable speech disorder assessment: a comparative analysis of CNN and transformer-based models

Malo Maisonneuve (Avignon Université - LIA); Corinne Fredouille (Avignon Université - LIA); Muriel Lalain (Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France); Alain Ghio (Aix-Marseille Univ, CNRS, LPL, Aix-en-Provence, France); Virginie Woisard (Hospitals of Toulouse)

517 Automatic recognition and detection of aphasic natural speech

Mara Barberis (KU Leuven); Pieter De Clercq (KU Leuven); Bastiaan Tamm (KU Leuven); Hugo Van hamme (KU Leuven); Maaïke Vandermosten (KU Leuven)

744 "So . . . my child . . ."—How Child ADHD Influences the Way Parents Talk

Anika A. Spiesberger (Technical University of Munich); Andreas Triantafyllopoulos (University of Augsburg); Alexander Kathan (University of Augsburg); Anastasia Semertzidou (Technical University of Munich); Caterina Gawrilow (University of Tuebingen); Tilman Reinelt (University of Zurich); Wolfgang Rauch (Ludwigsburg University of Education); Prof. Dr. Bjoern Schuller (Imperial College London)

852 Perceiver-Prompt: Flexible Speaker Adaptation in Whisper for Chinese Disordered Speech Recognition

Yicong Jiang (University of Chinese Academy of Sciences); Tianzi Wang (The Chinese University of HongKong);

Xurong Xie (Institute of Software, Chinese Academy of Sciences); Juan Liu (Shenzhen Institute of Advanced Technology); Wei Sun (Institute of Software, Chinese Academy of Sciences); Nan Yan (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Hui Chen (Institute of Software, Chinese Academy of Sciences); Lan Wang (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Xunying Liu (The Chinese University of Hong Kong); Feng Tian (Institute of Software, Chinese Academy of Sciences)

1074 Variability of speech timing features across repeated recordings: a comparison of open-source extraction techniques

Judith Dineley (King's College London); Ewan Carr (King's College London); Lauren White (King's College London); Catriona Lucas (King's College London); Zahia Rahman (King's College London); Tian Pan (King's College London); Faith Matcham (University of Sussex); Johnny Downs (King's College London); Richard Dobson (King's College London); Thomas Quatieri (Massachusetts Institute of Technology Lincoln Laboratory); Nicholas Cummins (King's College London)

1255 Macro-descriptors for Alzheimer's disease detection using large language models

Catarina Botelho (INESC-ID/IST, University of Lisbon); John Mendonça (INESC-ID); Anna Maria Pompili (INESC-ID); Tanja Schultz (University of Bremen); Alberto Abad (INESC-ID/IST); Isabel Trancoso (INESC ID)

1344 Zero-Shot End-To-End Spoken Question Answering In Medical Domain

Yanis Labrak (LIA - Avignon University); Adel Moumen (Avignon University); Mickael Rouvier (LIA - Avignon University); Richard Dufour (LS2N - Nantes University)

1390 How Consistent are Speech-Based Biomarkers in Remote Tracking of ALS Disease Progression Across Languages? A Case Study of English and Dutch

Hardik Kothare (Modality.AI); Michael Neumann (Modality.AI, Inc.); Cathy Zhang (Modality.AI); Jackson Liscombe (Modality.AI); Jordi W J van Unnik (University Medical Center Utrecht); Lianne C M Botman (University Medical Center Utrecht); Leonard H van den Berg (University Medical Center Utrecht); Ruben P A van Eijk (University Medical Center Utrecht); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI)

1541 Towards Scalable Remote Assessment of Mild Cognitive Impairment Via Multimodal Dialog

Oliver Roesler (Modality.AI Inc.); Jackson Liscombe (Modality.AI Inc.); Michael Neumann (Modality.AI, Inc.); Hardik Kothare (Modality.AI); Abhishek Hosamath (Modality.AI); Lakshmi Arbatti (Modality.AI, Inc.); Doug Habberstad (Modality.AI, Inc.); Christiane Suendermann-Oeft (Modality.AI, Inc.); Meredith Bartlett (Modality.AI, Inc.); Cathy Zhang (Modality.AI, Inc.); Nikhil Sukhdev (Modality.AI, Inc.); Kolja Wilms (Modality.AI, Inc.); Anusha Badathala (San Francisco Veterans Affairs Medical Center & University of California, San Francisco); Sandrine Iestas (ICON Strategic Solutions, ICON plc.); Steve Ruhmel (Janssen Research and Development, LLC.); Bryan Hansen (Janssen Research and Development, LLC.); Madeline Hannan (Janssen Research and Development, LLC.); David Henley (Janssen Research and Development, LLC.); Arthur Wallace (San Francisco Veterans Affairs Medical Center & University of California, San Francisco); Ira Shoulson (Modality.AI, Inc.); David Suendermann-Oeft (Modality.AI Inc.); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI)

2183 When Whisper Listens to Aphasia: Advancing Robust Post-Stroke Speech Recognition

Giulia Sanguedolce (Imperial College London); Sophie Mei Brook (Imperial College London); Dragos Gruia (Imperial College London); Patrick A. Naylor (Imperial College London); Fatemeh Geranmayeh (Imperial College London)

2205 Time to take action: acoustic modeling of motor verbs to detect Parkinson's disease

Daniel Escobar-Grisales (University of Antioquia); Cristian David Ríos Urrego (University of Antioquia); Ilja Baumann (Technische Hochschule Nürnberg Georg Simon Ohm); Korbinian Riedhammer (Technische Hochschule Nürnberg Georg Simon Ohm); Elmar Noeth (Friedrich Alexander Universität, Erlangen-Nuremberg); Tobias Bocklet (TH Nürnberg); Adolfo Garcia (Global Brain Health Institute, University of California); Juan Rafael Orozco-Arroyave (University of Antioquia)

2496 Infusing Acoustic Pause Context into Text-Based Dementia Assessment

Franziska Braun (Technische Hochschule Nürnberg); Sebastian P Bayerl (University of Applied Sciences Rosen-

heim); Florian Hönig (KST Institut GmbH); Hartmut Lehfeld (Klinikum Nürnberg); Thomas Hillemacher (Klinikum Nürnberg); Tobias Bocklet (TH Nürnberg); Korbinian Riedhammer (Technische Hochschule Nürnberg Georg Simon Ohm)

Time Slot 3

Oral Session: Prosody

A2-03

Location: Acesso

289 The prosody of the verbal prefix ge-: historical and experimental evidence

Chiara Riegger (University of Konstanz); Tina Bögel (University of Konstanz); George Walkden (University of Konstanz)

895 Automatic pitch accent classification through image classification

Na Hu (Radboud University); Hugo Schnack (Utrecht University); Amalia Arvaniti (Radboud University)

1909 Form and Function in Prosodic Representation: In the Case of “ma” in Tianjin Mandarin

Tianqi Geng (Tianjin University); Hui Feng (Tianjin University)

2004 Influences of Morphosyntax and Semantics on the Intonation of Mandarin Chinese Wh-indeterminates

Hongchen Wu (Georgia Institute of Technology); Jiwon Yun (Stony Brook University)

2172 Urdu Alternative Questions: A Hat Pattern

Benazir Mumtaz (University of Konstanz); Miriam Butt (University of Konstanz)

2513 On Comparing Time- and Frequency-Domain Rhythm Measures in Classifying Assamese Dialects

Joyshree Chakraborty (Indian Institute of Technology Guwahati); Leena Dihingia (Gauhati University); Priyankoo Sarmah (Indian Institute of Technology Guwahati); Rohit Sinha (Indian Institute of Technology Guwahati)

Oral Session: Foundational Models for Deepfake and Spoofed Speech Detection

A4-03

Location: Aegle A

24 Balance, Multiple Augmentation, and Re-synthesis: A Triad Training Strategy for Enhanced Audio Deepfake Detection

Thien-Phuc Doan (Soongsil University); Long Nguyen-Vu (Soongsil University); Kihun Hong (Soongsil University); Souhwan Jung (Soongsil University)

253 Adapter Learning from Pre-trained Model for Robust Spoof Speech Detection

Haochen Wu (University of Science and Technology of China); Wu Guo (University of Science and Technology of China); Shengyu Peng (University of Science and Technology of China); Zhuhai Li (University of Science and Technology of China); Jie Zhang (University of Science and Technology of China (USTC))

292 Speech Formants Integration for Generalized Detection of Synthetic Speech Spoofing Attacks

Kexu Liu (Nanjing University of Posts and Telecommunications); Wang Yuanxin (Nanjing University of Posts and Telecommunications); Shengchen Li (Xi'an Jiaotong-Liverpool University); Xi Shao (Nanjing University of Posts and Telecommunications)

481 Spoofed speech detection with a focus on speaker embedding

Hoan My Tran (irisa); David Guennec (Univ Rennes, CNRS, IRISA); Philippe Martin (Univ Rennes, CNRS, IRISA); Aghilas Sini (Le Mans Université); Damien Lolive (Univ Rennes, CNRS, IRISA); Arnaud Delhay (Univ Rennes, IRISA, CNRS); Pierre-François Marteau (Univ Bretagne Sud, CNRS, IRISA)

942 Exploring Self-supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection

Juan M. Martín-Doñas (Vicomtech); Aitor Álvarez (Vicomtech); Eros Rosello (University of Granada); Angel M. Gomez (University of Granada); Antonio M. Peinado (University of Granada)

1472 Attentive Merging of Hidden Embeddings from Pre-trained Speech Model for Anti-spoofing Detection

Zihan Pan (Institute for Infocomm Research (I2R), A*STAR, Singapore); Tianchi Liu (National University of Singapore); Hardik B Sailor (I2R, ASTAR, Singapore); Qiongqiong Wang (A*STAR)

Oral Session: ASR and LLMs

A10-02 Location: Aegle B

10 MM-KWS: Multi-modal Prompts for Multilingual User-defined Keyword Spotting

Zhiqi Ai (Shanghai University); Zhiyong Chen (Shanghai University); Shugong Xu (Shanghai University)

80 HuBERT-EE: Early Exiting HuBERT for Efficient Speech Recognition

Ji Won Yoon (Chung-Ang University); Beom Jun Woo (Seoul National University); Nam Soo Kim (Seoul National University)

322 Whisper-Flamingo: Integrating Visual Features into Whisper for Audio-Visual Speech Recognition and Translation

Andrew Rouditchenko (MIT CSAIL); Yuan Gong (Massachusetts Institute of Technology); Samuel Thomas (IBM Research AI); Leonid Karlinsky (IBM-Research); Hilde Kuehne (University of Bonn); Rogerio Feris (MIT-IBM Watson AI Lab, IBM Research); James Glass (Massachusetts Institute of Technology)

376 Spoken-to-written text conversion with Large Language Model

HyunJung Choi (Korea National University of Science and Technology(UST)); Muyeol Choi (ETRI); Yohan Lim (University of Science and Technology); Minkyu Lee (ETRI); Seonhui Kim (Korea National University of Science and Technology(UST)); Seung Yun (ETRI); Donghyun Kim (ETRI); Sang Hun Kim (ETRI)

488 MaLa-ASR: Multimedia-Assisted LLM-Based ASR

guanrou yang (Shanghai Jiao Tong University); Ziyang Ma (Shanghai Jiao Tong University); Fan Yu (Speech Lab of DAMO Academy, Alibaba Group); zhifu gao (alibaba); Shiliang Zhang (Alibaba Group); Xie Chen (Shanghai Jiaotong University)

2290 Speech Recognition Models are Strong Lip-readers

Prajwal K R (VGG, Oxford); Triantafyllos Afouras (Meta); Andrew Zisserman (University of Oxford)

Oral Session: Speech Synthesis: Prosody

A7-05 Location: Hippocrates

715 Word-level Text Markup for Prosody Control in Speech Synthesis

Yuliya Korotkova (Just AI); Ilya Kalinovskiy (Just AI); Tatiana Vakhrusheva (Just AI)

1327 Total-Duration-Aware Duration Modeling for Text-to-Speech Systems

Sefik Emre Eskimez (Microsoft); Xiaofei Wang (Microsoft); Manthan Thakker (Microsoft Corporation); Chung-Hsien Tsai (Microsoft Corporation); Canrun Li (Microsoft Corporation); Zhen Xiao (Microsoft Corporation); Hemin Yang (Microsoft); Zirun Zhu (Microsoft); Min Tang (Microsoft); Jinyu Li (Microsoft); Sheng Zhao (Microsoft); Naoyuki Kanda (Microsoft)

1582 Should you use a probabilistic duration model in TTS? Probably! Especially for spontaneous speech
Shivam Mehta (KTH Royal Institute of Technology); Harm Lameris (KTH Royal Institute of Technology); Rajiv Punmiya (Self); Jonas Beskow (KTH Royal Institute of Technology); Eva Szekely (KTH Royal Institute of Technology); Gustav Eje Henter (KTH Royal Institute of Technology)

2055 A Human-in-the-Loop Approach to Improving Cross-Text Prosody Transfer
Himanshu Maurya (University of Edinburgh); Atli Thor Sigurgeirsson (University of Edinburgh)

2133 Multi-Modal Automatic Prosody Annotation with Contrastive Pretraining of Speech-Silence and Word-Punctuation
Jinzuomu Zhong (the University of Edinburgh); Yang Li (Department of AI Technology, Transsion); Hui Huang (Department of AI Technology, Transsion); Korin Richmond (University of Edinburgh); Jie Liu (Department of AI Technology, Transsion); zhibasu (Transsion); Jing Guo (Department of AI Technology, Transsion); Benlai Tang (Department of AI Technology, Transsion); Fengjie Zhu (Department of AI Technology, Transsion)

2506 Towards Expressive Zero-Shot Speech Synthesis with Hierarchical Prosody Modeling
Yuepeng Jiang (Northwestern Polytechnical University); Tao Li (School of Computer Science, Northwestern Polytechnical University, Xi'an); Fengyu Yang (xiaomi); Lei Xie (NWPU); Meng Meng (xiaomi); Yujun Wang (xiaomi)

Oral Session: Neural Network Adaptation

A8-04 Location: Iasso

45 AdaRA: Adaptive Rank Allocation of Residual Adapters for Speech Foundation Model
Zhouyuan Huo (Google); Dongseong Hwang (Google); Gan Song (Google); Khe C Sim (Google Inc.); Weiran Wang (Google)

334 Leveraging Adapter for Parameter-Efficient ASR Encoder
Kyuhyong Shim (Qualcomm AI Research); Jinkyu Lee (Qualcomm AI Research); Hyunjae Kim (Qualcomm AI Research)

513 Whisper Multilingual Downstream Task Tuning using Task Vector
Ji-Hun Kang (Hanyang University); Jae-Hong Lee (Hanyang University); Mun-Hak Lee (Hanyang University); Joon-Hyuk Chang (Hanyang University)

930 Qifusion-Net: Layer-adapted Stream/Non-stream Model for End-to-End Multi-Accent Speech Recognition
jinming chen (Qifu Technology); Jingyi Fang (Qifu Technology); Yuanzhong Zheng (Qifu Technology); Yaoyuan Wang (Qifu Technology); Haojun Fei (QIFU technology)

1105 Shared-Adapters: A novel Transformer-based parameter efficient transfer learning approach for children's automatic speech recognition
Thomas Rolland (INESC-ID); Alberto Abad (INESC-ID/IST)

1873 Speaker-Smoothed kNN Speaker Adaptation for End-to-End ASR
Shaojun Li (Huawei TSC); Daimeng Wei (Huawei); Hengchao Shang (HW-TSC); Jiaxin GUO (Huawei); ZongYao LI (HW-TSC); Zhanglin Wu (HW-TSC); Zhiqiang Rao (Huawei); Yuanchang Luo (Huawei); Xianghui He (Huawei); Hao Yang (Huawei)

Oral Session: Speech Processing Using Discrete Speech Units (Special Session)

SS-10

Location: Melampus

330 Codec-ASR: Training Performant Automatic Speech Recognition Systems with Discrete Speech Representations

Kunal Dhawan (NVIDIA); Nithin Rao Koluguri (NVIDIA); Ante Jukić (NVIDIA); Ryan Langman (NVIDIA); Jagadeesh Balam (NVIDIA); Boris Ginsburg (NVIDIA)

1878 The Interspeech 2024 Challenge on Speech Processing Using Discrete Units

Xuankai Chang (Carnegie Mellon University); Jiatong Shi (Carnegie Mellon University); Jinchuan Tian (Carnegie Mellon University); Yuning Wu (Renmin University of China); Yuxun Tang (Renmin University of China); Yihan Wu (Renmin University of China); Shinji Watanabe (Carnegie Mellon University); Yossi Adi (The Hebrew University of Jerusalem); Xie Chen (Shanghai Jiaotong University); Qin Jin (Renmin University of China)

2135 How Should We Extract Discrete Audio Tokens from Self-Supervised Models?

Pooneh Mousavi (Concordia University); jarod duet (LIA); Salah Zaiem (Telecom Paris); Luca Della Libera (Concordia University); Artem Ploujnikov (Université de Montréal); Cem Subakan (Mila); Mirco Ravanelli (Université de Montréal)

2251 MMM: Multi-Layer Multi-Residual Multi-Stream Discrete Speech Representation from Self-supervised Learning Model

Jiatong Shi (Carnegie Mellon University); Xutai Ma (Meta); Hirofumi Inaguma (Meta); Anna Sun (Meta); Shinji Watanabe (Carnegie Mellon University)

2291 SingOMD: Singing Oriented Multi-resolution Discrete Representation Construction from Speech Models

Yuxun Tang (Renmin University of China); Yuning Wu (Renmin University of China); Jiatong Shi (Carnegie Mellon University); Qin Jin (Renmin University of China)

2360 TokSing: Singing Voice Synthesis based on Discrete Tokens

Yuning Wu (Renmin University of China); Chunlei zhang (Tencent AI Lab); Jiatong Shi (Carnegie Mellon University); Yuxun Tang (Renmin University of China); Shan Yang (Tencent AI Lab); Qin Jin (Renmin University of China)

Oral Session: Source Separation 1

A5-04

Location: Panacea Amphitheater

327 Noise-robust Speech Separation with Fast Generative Correction

Helin Wang (Johns Hopkins University); Jesus Antonio Villalba (Johns Hopkins University); Laureano Moro-Velazquez (Johns Hopkins University); Jiarui Hai (Johns Hopkins University); Thomas Thebaud (Johns Hopkins University); Najim Dehak (Johns Hopkins University)

337 Towards Audio Codec-based Speech Separation

Jia Qi Yip (Alibaba Group / Nanyang Technological University); Shengkui Zhao (Alibaba Group); Dianwen Ng (Alibaba Group / Nanyang Technological University); Eng Siong Chng (Nanyang Technological University); Bin Ma (Alibaba)

576 Improving Generalization of Speech Separation in Real-World Scenarios: Strategies in Simulation, Optimization, and Evaluation

Ke Chen (University of California San Diego); Jiaqi Su (Adobe Research); Taylor Berg-Kirkpatrick (UCSD); Shlomo Dubnov (UC San Diego); Zeyu Jin (Adobe Research)

1136 Unsupervised Improved MVDR Beamforming for Sound Enhancement

Jacob Kealey (*Université de Sherbrooke*); John Hershey (*Google*); François Grondin (*Université de Sherbrooke*)

1378 Enhanced Deep Speech Separation in Clustered Ad Hoc Distributed Microphone Environments

Jihyun Kim (*Yonsei University*); Stijn Kindt (*UGent*); Nilesh Madhu (*IDLab, Ghent University - imec*); Hong-Goo Kang (*Yonsei University*)

2537 MSDET: Multitask Speaker Separation and Direction-of-Arrival Estimation Training

Roland Hartanto (*Tokyo Institute of Technology*); Sakriani Sakti (*Nara Institute of Science and Technology / Japan Advanced Institute of Science and Technology*); Koichi SHINODA (*Tokyo Institute of Technology*)

Poster Session: Pathological Speech Analysis 3

A13-P2-A

Location: Poster Area 1A

470 A Cluster-based Personalized Federated Learning Strategy for End-to-End ASR of Dementia Patients

Wei Tung Hsu (*National Tsing Hua University*); Chin-Po Chen (*Department of Electrical Engineering, National Tsing Hua University*); Yun-Shao Lin (*Electrical Engineering Department, National Tsing Hua University*); Chi-Chun Lee (*National Tsing Hua University*)

996 A Comparative Analysis of Federated Learning for Speech-Based Cognitive Decline Detection

Stefan Kalabakov (*Hasso Plattner Institute, University of Potsdam*); Monica Gonzalez Machorro (*audEERING GmbH / Technical University of Munich*); Florian Eyben (*audEERING*); Prof. Dr. Bjoern Schuller (*Imperial College London*); Bert Arnrich (*Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam*)

1018 Clever Hans Effect Found in Automatic Detection of Alzheimer's Disease through Speech

Yin-Long Liu (*University of Science and Technology of China*); Rui Feng (*University of Science and Technology of China*); Jiahong Yuan (*University of Science and Technology of China*); Zhen-Hua Ling (*University of Science and Technology of China*)

1318 Multimodal Digital Biomarkers for Longitudinal Tracking of Speech Impairment Severity in ALS: An Investigation of Clinically Important Differences

Hardik Kothare (*Modality.AI*); Michael Neumann (*Modality.AI, Inc.*); Jackson Liscombe (*Modality.AI*); Emma Cathrine Liisborg Leschly (*Modality.AI*); Oliver Roesler (*Modality.AI Inc.*); Vikram Ramanarayanan (*University of California, San Francisco & Modality.AI*)

1577 Developing vocal system impaired patient-aimed voice quality assessment approach using ASR representation-included multiple features

Shaoliang Dang (*Nagoya University*); Tetsuya Matsumoto (*Nagoya University*); Yoshinori Takeuchi (*Daido University*); Takashi Tsuboi (*Nagoya University Graduate School of Medicine*); Yasuhiro Tanaka (*Aichi Gakuin University*); Daisuke Nakatsubo (*Nagoya University Graduate School of Medicine*); Satoshi Maesawa (*Nagoya University Graduate School of Medicine*); Ryuta Saito (*Nagoya University Graduate School of Medicine*); Masahisa Katsuno (*Nagoya University Graduate School of Medicine*); Hiroaki Kudo (*Nagoya University*)

1887 Leveraging Phonemic Transcription and Whisper toward Clinically Significant Indices for Automatic Child Speech Assessment

Yeh-Sheng Lin (*Institute of Linguistics, Academia Sinica*); Shu-Chuan Tseng (*"Institute of Linguistics, Academia Sinica"*); Jyh-Shing Roger Jang (*National Taiwan University*)

2134 Towards Self-Attention Understanding for Automatic Articulatory Processes Analysis in Cleft Lip and Palate Speech

Ilja Baumann (*Technische Hochschule Nürnberg Georg Simon Ohm*); Dominik Wagner (*Technische Hochschule Nuernberg Georg Simon Ohm*); Maria Schuster (*Ludwig-Maximilians University*); Korbinian Riedhammer (*Technische Hochschule Nürnberg Georg Simon Ohm*); Elmar Noeth (*Friedrich Alexander Universität, Erlangen-Nuremberg*);

Tobias Bocklet (TH Nürnberg)

Poster Session: Speech Disorders 3

A13-P2-B

Location: Poster Area 1B

68 Acoustic changes in speech prosody produced by children with autism after robot-assisted speech training

Si Chen (The Hong Kong Polytechnic University); Bruce Wang (Hong Kong Polytechnic University); Yitian Hong (Hong Kong Polytechnic University); Fang Zhou (Hong Kong Polytechnic University); Angel Chan (Hong Kong Polytechnic University); Po-yi Tang (Hong Kong Polytechnic University); Bin Li (City University of Hong Kong); Chunyi Wen (Hong Kong Polytechnic University); James Cheung (Hong Kong Polytechnic University); Yan Liu (The Hong Kong Polytechnic University); Zhuoming Chen (The First Hospital of Jinan University)

366 PARAN: Variational Autoencoder-based End-to-End Articulation-to-Speech System for Speech Intelligibility

seyun um (yonsei university); Doyeon Kim (Yonsei University); Hong-Goo Kang (Yonsei University)

578 Learnings from curating a trustworthy, well-annotated, and useful dataset of disordered English speech

Pan-Pan Jiang (Google); Jimmy Tobin (Google Research); Katrin Tomanek (Google); Robert MacDonald (Google); Katie Seaver (MGH Institute of Health Professions); Richard Cave (Language and Cognition UCLggle); Marilyn Ladewig (CPUnlimited); Rus Heywood (Google); Jordan Green (MGH Institute of Health Professions)

672 Towards Improving NAM-to-Speech Synthesis Intelligibility using Self-Supervised Speech Models

Neil Shah (IIIT Hyderabad); Shirish Karande (TCS Research); Vineet Gandhi (IIIT Hyderabad)

879 Enhancing Voice Wake-Up for Dysarthria: Mandarin Dysarthria Speech Corpus Release and Customized System Design

Ming Gao (University of Science and Technology of China); Hang Chen (USTC); Jun Du (University of Science and Technology of China); Xin Xu (Beijing AISHELL Technology Co., Ltd.); Hongxiao Guo (Beijing AISHELL Technology Co., Ltd.); Hui Bu (AISHELL); Jianxing Yang (Beijing Polytechnic); Ming Li (Duke Kunshan University); Chin-Hui Lee (Georgia Institute of Technology)

995 Wav2vec 2.0 Embeddings Are No Swiss Army Knife -- A Case Study for Multiple Sclerosis

Gábor Gosztolya (MTA-SZTE Research Group on AI); Mercedes Vetrab (University of Szeged); Veronika Svindt (Research Center for Linguistics); Judit Bóna (ELTE Eötvös Loránd University); Ildikó Hoffmann (Research Center for Linguistics)

1645 Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric Speech Synthesis

Wing-Zin Leung (University of Sheffield); Mattias George Cross (University of Sheffield); Anton Ragni (University of Sheffield); Stefan Goetze (University of Sheffield)

1969 Fine-Tuning Automatic Speech Recognition for People with Parkinson's: An Effective Strategy for Enhancing Speech Technology Accessibility

Xiuwen Zheng (University of Illinois Urbana-Champaign); Bornali Phukon (University of Illinois Urbana Champaign); Mark A Hasegawa-Johnson (University of Illinois)

Poster Session: Accented Speech, Prosodic Features, Dialect, Emotion, Sound Classification

A8-P3

Location: Poster Area 2A, Poster Area 2B

- 44 The Processing of Stress in End-to-End Automatic Speech Recognition Models**
Martijn Bentum (Radboud University); Louis ten Bosch (radboud university); Tomas O Lentz (Tilburg University)
- 1569 LingWav2Vec2: Linguistic-augmented wav2vec 2.0 for Vietnamese Mispronunciation Detection**
*Tuan V. A. Nguyen (Institute for Infocomm Research, A*STAR); Tran Huy Dat (Institute for Infocomm Research (I2R))*
- 1606 Towards End-to-End Unified Recognition for Mandarin and Cantonese**
Meiling Chen (Beijing Institute of Computer Technology and Application); Pengjie Liu (Beijing Institute of Computer Technology and Application); Heng Yang (Beijing Institute of Computer Technology and Application); Haofeng Wang (Beijing Institute of Computer Technology and Application)
- 1628 Learning from memory-based models**
Rhiannon Mogridge (University of Sheffield); Anton Ragni (University of Sheffield)
- 1733 Cross-modal Features Interaction-and-Aggregation Network with Self-consistency Training for Speech Emotion Recognition**
Ying Hu (Xinjiang University); Huamin Yang (Xinjiang University); Hao Huang (Xinjiang University); Liang He (Tsinghua University)
- 1820 Exploring Multilingual Unseen Speaker Emotion Recognition: Leveraging Co-Attention Cues in Multi-Task Learning**
Arnav Goel (IIITD); Medha Hira (IIITD); Anubha Gupta (Indraprastha Institute of Information Technology-Delhi (IIITD))
- 2257 SELM: Enhancing Speech Emotion Recognition for Out-of-Domain Scenarios**
Hazim T Bukhari (Carnegie Mellon university); Soham Deshmukh (Microsoft); Hira Dhamyal (Carnegie Mellon University); Bhiksha Raj (Carnegie Mellon University); Rita Singh (Carnegie Mellon University)
- 2261 Performant ASR Models for Medical Entities in Accented Speech**
Tejumade Afonja (CISPA Helmholtz Center for Information Security, Saarland University, and AI Saturdays Lagos); Tobi Olatunji (Intron Inc); Sewade O Ogun (Inria); Naome A. Etori (university of Minnesota-Twin Cities); Abraham T Owodunni (Intron Health); Moshood O Yekini (African Masters of Machine Intelligence)
- 2376 LAHAJA: A Robust Multi-accent Benchmark for Evaluating Hindi ASR Systems**
Tahir Javed (Indian Institute of Technology Madras); Janki A Nawale (AI4Bharat); Sakshi Joshi (AI4bharat IIT Madras); Eldho Ittan George (AI4Bharat); Kaushal Santosh Bhogale (Indian Institute of Technology, Madras); Deovrat Mehendale (AI4Bharat); Mitesh M. Khapra (Indian Institute of Technology Madras)
- 2392 LearnerVoice: A Dataset of Non-Native English Learners' Spontaneous Speech**
Haechan Kim (KAIST); Junho Myung (KAIST); Seoyoung Kim (KAIST); Sungpah Lee (Ringle); Dongyeop Kang (University of Minnesota); Juho Kim (KAIST)
- 2414 MinSpeech: A Corpus of Southern Min Dialect for Automatic Speech Recognition**
Jiayan Lin (Xiamen University); Shenghui Lu (Xiamen University); Hui Bu (AISHELL); BINBIN XU (XIAMEN UNIVERSITY); Wenhao Guan (Xiamen University); Hukai Huang (Xiamen University); Lin Li (Xiamen University); Qingyang Hong (Xiamen University)
- 2438 Improving Self-supervised Pre-training using Accent-Specific Codebooks**
Abhishek Kumar Gupta (Indian Institute of Technology Bombay); Darshan Deepak Prabhu (Indian Institute of Technology, Bombay); Omkar M Nitsure (Indian Institute of Technology, Bombay); Preethi Jyothi (Indian Institute of Technology Bombay); Sriram Ganapathy (Google Research ; Indian Institute of Science, Bangalore, India, 560012)

Poster Session: Audio-Visual and Generative Speech Enhancement

A6-P3-A

Location: Poster Area 3A

85 An Analysis of the Variance of Diffusion-based Speech Enhancement*Bunlong Lay (Universität Hamburg); Timo Gerkmann (Universität Hamburg)***162 Locally Aligned Rectified Flow Model for Speech Enhancement Towards Single-Step Diffusion***LI ZHENGXIAO (Tokyo Institute Technology); Nakamasa Inoue (Tokyo Institute of Technology)***494 Diffusion Gaussian Mixture Audio Denoise***Pu Wang (College of Science, University of Science and Technology Liaoning); Junhui Li (College of Science, University of Science and Technology Liaoning); Jialu Li (Cornell University); Liangdong Guo (College of Science, University of Science and Technology Liaoning); Youshan Zhang (Yeshiva University)***516 RT-LA-VocE: Real-Time Low-SNR Audio-Visual Speech Enhancement***Honglie Chen (Meta); Rodrigo Mira (Imperial College London); Stavros Petridis (Meta); Maja Pantic (Imperial College London)***611 Complex Image-Generative Diffusion Transformer for Audio Denoising***Junhui Li (University of Science and Technology Liaoning); Pu Wang (College of Science, University of Science and Technology Liaoning); Jialu Li (Cornell University); Youshan Zhang (Yeshiva University)***701 FlowAVSE: Efficient Audio-Visual Speech Enhancement with Conditional Flow Matching***Chaeyoung Jung (KAIST); Suyeon Lee (Korea Advanced Institute of Science and Technology); Ji-Hoon Kim (KAIST); Joon Son Chung (KAIST)***929 Noise-aware Speech Enhancement using Diffusion Probabilistic Model***Yuchen Hu (Nanyang Technological University); Chen Chen (Nanyang Technological University); Ruizhe Li (University of Aberdeen); Qiushi Zhu (University of Science and Technology of China); Eng Siong Chng (Nanyang Technological University)*

Poster Session: Speech Privacy and Bandwidth Expansion

A6-P3-B

Location: Poster Area 3B

11 Multi-Stage Speech Bandwidth Extension with Flexible Sampling Rates Control*Ye-Xin Lu (University of Science and Technology of China); Yang Ai (University of Science and Technology of China); Zheng-Yan Sheng (University of Science and Technology of China); Zhen-Hua Ling (University of Science and Technology of China)***31 A New Approach to Voice Authenticity***Nicolas M Müller (Fraunhofer AISEC); Piotr Kawa (Wrocław University of Science and Technology); Shen Hu (TU Munich); Matthias Neu (BSI); Jennifer Williams (University of Southampton); Philip Sperl (Fraunhofer AISEC); Konstantin Böttinger (Fraunhofer AISEC)***117 Privacy PORCUPINE: Anonymization of Speaker Attributes Using Occurrence Normalization for Space-Filling Vector Quantization***Mohammad Hassan Vali (Aalto University); Tom Bäckström (Aalto University)***174 SilentCipher: Deep Audio Watermarking***Mayank Kumar Singh (Sony Research Japan); Naoya Takahashi (Sony Research); Wei-Hsiang Liao (Sony Group Corporation); Yuki Mitsuji (Sony AI)***463 SWiBE: A Parameterized Stochastic Diffusion Process for Noise-Robust Bandwidth Expansion**

Yin-Tse Lin (Institute of Communication Engineering, National Tsing Hua University, Taiwan); Shreya G. Upadhyay (National Tsing Hua University); Bo-Hao Su (Department of Electrical Engineering, National Tsing Hua University); Chi-Chun Lee (National Tsing Hua University)

534 TraceableSpeech: Towards Proactively Traceable Text-to-Speech with Watermarking

Junzuo Zhou (Institute of Automation, Chinese Academy of Sciences); Jiangyan Yi (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences); Tao Wang (Institute of Automation, Chinese Academy of Sciences); Jianhua Tao (Tsinghua University); Ye Bai (Institute of Automation, Chinese Academy of Sciences); Chu Yuan Zhang (Institute of Automation, Chinese Academy of Sciences); Yong Ren (Institute of Automation, Chinese Academy of Sciences); Zhengqi Wen (Institute of Automation, Chinese Academy of Sciences)

740 Frequency-mix Knowledge Distillation for Fake Speech Detection

Cunhang Fan (Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University); Dong Shunbo (Anhui University); Jun Xue (Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University); Yujie Chen (Anhui University); Jiangyan Yi (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences); zhao lv (anhui university)

1164 HarmoNet: Partial DeepFake Detection Network based on Multi-scale HarmoF0 Feature Fusion

Liwei Liu (The Chinese University of Hong Kong, Shenzhen); huihui wei (shanghai jiaotong university); dongya liu (NWPU); 中华付 (西北工业大学)

1584 MaskSR: Masked Language Model for Full-band Speech Restoration

Xu Li (Dolby Laboratories); Qirui Wang (Southeast University); Xiaoyu Liu (Dolby Laboratories)

1652 Unmasking Neural Codecs: Forensic Identification of AI-compressed Speech

Denise Moussa (Friedrich-Alexander-University Erlangen-Nuremberg/Federal Criminal Police Office of Germany); Sandra Bergmann (Friedrich-Alexander University Erlangen-Nürnberg); Christian Riess (Friedrich-Alexander University Erlangen-Nuremberg)

Poster Session: Speaker Recognition 1

A4-P3

Location: Poster Area 4A, Poster Area 4B

260 Fine-tune Pre-Trained Models with Multi-Level Feature Fusion for Speaker Verification

Shengyu Peng (University of Science and Technology of China); Wu Guo (University of Science and Technology of China); Haochen Wu (University of Science and Technology of China); Zuoliang Li (University of Science and Technology of China); Jie Zhang (University of Science and Technology of China (USTC))

365 Speaker Conditional Sinc-Extractor for Personal VAD

En-Lun Yu (National Taiwan Normal University); Kuan-Hsun Ho (NTNU); Jieh-wei Hung (National Chi Nan University); Shih-Chieh Huang (Realtek Semiconductor Corp.); Berlin Chen (National Taiwan Normal University)

601 Enhancing ECAPA-TDNN with Feature Processing Module and Attention Mechanism for Speaker Verification

Shiu-Hsiang Liou (National Sun Yat-sen University); Po-Cheng Chan (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); Chia-Ping Chen (National Sun Yat-sen University); Tzu-Chieh Lin (National Sun Yat-sen University); Chung-li Lu (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); yu-han cheng (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); Hsiang Feng Chuang (Advanced Technology Laboratory, Chunghwa Telecom Laboratories); Wei-Yu Chen (Advanced Technology Laboratory, Chunghwa Telecom Laboratories)

897 DB-PMMAE: Dual-Branch Prototypical Masked AutoEncoder with locality for domain robust speaker

verification

Wei-Lin Xie (University of Science and Technology of China); Yuxuan Xi (National Engineering Research Center of Speech and Language Information Processing); Yan Song (USTC); Jian-Tao Zhang (University of Science and Technology of China); Hao-Yu Song (The Australian National University); Ian McLoughlin (Singapore Institute of Technology)

939 MR-RawNet: Speaker verification system with multiple temporal resolutions for variable duration utterances using raw waveforms

Seung-bin Kim (University of Seoul); Chan-yeong Lim (University of Seoul); Jungwoo Heo (University of Seoul); Ju-ho Kim (University of Seoul); Hyun-seo Shin (University of Seoul); Kyo-Won Koo (University of Seoul); Ha-Jin Yu (University of Seoul)

1124 Disentangled Representation Learning for Environment-agnostic Speaker Recognition

KiHyun Nam (Korea Advanced Institute of Science and Technology (KAIST)); Hee-Soo Heo (Naver Corp.); Jee-weon Jung (Carnegie Mellon University); Joon Son Chung (KAIST)

1260 Multi-Channel Extension of Pre-trained Models for Speaker Verification

Ladislav Mosner (Brno University of Technology); romain serizel (Université de Lorraine); Lukáš Burget (Brno University of Technology); Plchot Oldřich (Brno University of Technology); Emmanuel Vincent (Inria); Junyi Peng (Brno University of Technology); Jan Honza Cernocky (Brno University of Technology)

1889 Efficient Integrated Features Based on Pre-trained Models for Speaker Verification

Yishuang Li (Xiamen University); Wenhao Guan (Xiamen University); Hukai Huang (Xiamen University); Shiyu Miao (Xiamen University); Qi Su (Xiamen University); Lin Li (Xiamen University); Qingyang Hong (Xiamen University)

2036 Attention-augmented x-vectors for the evaluation of mimicked speech using Sparse Autoencoder-LSTM framework

BHASI K C (GOVERNMENT ENGINEERING COLLEGE BARTON HILL); Rajeev Rajan (Government Engineering College, Barton Hill, Trivandrum); Noumida A (College Of Engineering Trivandrum)

2119 Evaluating the Santa Barbara Corpus: Challenges of the Breadth of Conversational Spoken Language

Matthew Maciejewski (Johns Hopkins University); Dominik Klement (Brno University of Technology); Ruizhe Huang (Johns Hopkins University); Matthew S Wiesner (Johns Hopkins University); Sanjeev Khudanpur (Johns Hopkins University)

2476 SE/BN Adapter: Parametric Efficient Domain Adaptation for Speaker Recognition

Tianhao Wang (Beijing University of Posts and Telecommunications); Lantian Li (Beijing University of Posts and Telecommunications); Dong Wang (Tsinghua University)

2478 A Comprehensive Investigation on Speaker Augmentation for Speaker Recognition

Zhenyu Zhou (Beijing University Of Posts and Telecommunications); Shibiao Xu (School of Artificial Intelligence, Beijing University of Posts and Telecommunications); Shi Yin (Huawei); Lantian Li (Beijing University of Posts and Telecommunications); Dong Wang (Tsinghua University)

Poster Session: Speech Recognition with Large Pretrained Speech Models for Under-represented Languages (Special Session)

SS-9 Location: Yanis Club

326 Interface Design for Self-Supervised Speech Models

Yi-Jen Shih (The University of Texas at Austin); David Harwath (The University of Texas at Austin)

479 Exploring adaptation techniques of large speech foundation models for low-resource ASR: a case study on Northern Sámi

Yaroslav Getman (Aalto University); Tamas Grosz (Aalto University); Katri Hiovain-Asikainen (UiT The Arctic University of Norway); Mikko Kurimo (Aalto University)

503 Interleaved Audio/Audiovisual Transfer Learning for AV-ASR in Low-Resourced Languages

Zhengyang Li (Technische Universität Carolo-Wilhelmina Braunschweig); Patrick Blumenberg (Technische Universität Braunschweig); Jing Liu (Amazon.com); Thomas Graave (Technische Universität Braunschweig); Timo Lohrenz (Technische Universität Braunschweig); Siegfried Kunzmann (Amazon.com); Tim Fingscheidt (Technische Universität Braunschweig)

1045 Learn and Don't Forget: Adding a New Language to ASR Foundation Models

Mengjie Qian (Cambridge University); Siyuan Tang (University of Cambridge); Rao Ma (University of Cambridge); Katherine M Knill (University of Cambridge); Mark Gales (University of Cambridge)

1533 Comparing Discrete and Continuous Space LLMs for Speech Recognition

Yaoxun Xu (Tsinghua University); Shi-Xiong Zhang (Capital One); Jianwei Yu (Tencent AI lab); Zhiyong Wu (Tsinghua University); Dong Yu (Tencent AI Lab)

1587 Adapter pre-training for improved speech recognition in unseen domains using low resource adapter tuning of self-supervised models

Sathvik Udupa (Indian Institute of Science); Jesuraj Bandekar (IISc); Saurabh Kumar (IISc Bengaluru); Deekshitha G (IISc); Sandhya B (IISc); Abhayjeet Singh (Indian Institute of Sciences, Bangalore, India); Savitha S Murthy (IISc); Priyanka Pai (NavanaTech); Srinvasa Raghavan K M (Navana Tech); Raoul Nanavati (Navana Tech India Pvt. Ltd.); Prasanta Kumar Ghosh (Indian Institute of Science (IISc), Bangalore)

1790 Improving Whisper's Recognition Performance for Under-Represented Language Kazakh Leveraging Unpaired Speech and Text

Jinpeng Li (Tsinghua University); Yu Pu (Tsinghua University); Qi Sun (Tsinghua University); Wei-Qiang Zhang (Tsinghua University)

1953 Towards Rehearsal-Free Multilingual ASR: A LoRA-based Case Study on Whisper

Tianyi Xu (NWPU); Pengcheng Guo (Northwestern Polytechnical University); Kaixun Huang (NWPU); Lei Xie (NWPU); Longtao Huang (Alibaba Group); Yu Zhou (Alibaba Group); Hui Xue (Alibaba Group)

2396 Empowering Low-Resource Language ASR via Large-Scale Pseudo Labeling

Kaushal Santosh Bhogale (Indian Institute of Technology, Madras); Deovrat Mehendale (AI4Bharat); Niharika Sri Parasa (AI4BHARAT); Sathish Kumar Reddy G (Indian Institute of Technology, Madras); Tahir Javed (Indian Institute of Technology Madras); Pratyush Kumar (Indian Institute of Technology Madras); Mitesh M. Khapra (Indian Institute of Technology Madras)

Wednesday 04/09

Time Slot 1

Oral Session: Self-Supervised Models in Speaker Recognition

A4-04

Location: Acesso

295 Parameter-efficient Fine-tuning of Speaker-Aware Dynamic Prompts for Speaker Verification

Zhe LI (Hong Kong Polytechnic University); Man-Wai MAK (The Hong Kong Polytechnic University); Hung-yi Lee (National Taiwan University); Helen Meng (The Chinese University of Hong Kong)

362 Self-supervised speaker verification with relational mask prediction

Ju-ho Kim (University of Seoul); Hee-Soo Heo (Naver Corp.); Bong-Jin Lee (Naver Corporation); Youngki Kwon (Naver Corporation); Minjae Lee (NAVER Cloud Corp.); Ha-Jin Yu (University of Seoul)

486 Towards Supervised Performance on Speaker Verification with Self-Supervised Learning by Leveraging Large-Scale ASR Models

Victor Miara (LRE-EPITA); Theo Lepage (LRE-EPITA); Reda DEHAK (ESLR - EPITA)

1515 Whisper-PMFA: Partial Multi-Scale Feature Aggregation for Speaker Verification using Whisper Models

Yiyang Zhao (Tsinghua University); Shuai Wang (Shenzhen Research Institute of Big Data, Chinese University of Hong Kong (Shenzhen)); Guangzhi Sun (University of Cambridge Department of Engineering); Zehua Chen (Tsinghua University); Chao Zhang (Tsinghua University); Mingxing Xu (Tsinghua University); Thomas Fang Zheng ("CSLT, Tsinghua University")

1630 Improving Noise Robustness in Self-supervised Pre-trained Model for Speaker Verification

Chan-yeong Lim (University of Seoul); Hyun-seo Shin (University of Seoul); Ju-ho Kim (University of Seoul); Jung-woo Heo (University of Seoul); Kyo-Won Koo (University of Seoul); Seung-bin Kim (University of Seoul); Ha-Jin Yu (University of Seoul)

2206 On the impact of several regularization techniques on label noise robustness of self-supervised speaker verification systems

Abderrahim Fathan (Computer Research Institute of Montreal (CRIM), Montreal, Quebec, Canada); Xiaolin Zhu (Computer Research Institute of Montreal); Jahangir Alam (Computer Research Institute of Montreal (CRIM), Montreal (Quebec) Canada)

Oral Session: Privacy and Security in Speech Communication 1

A6-03

Location: Aegle A

81 Prompt Tuning for Audio Deepfake Detection: Computationally Efficient Test-time Domain Adaptation with Limited Target Dataset

Hideyuki Oiso (University of Tsukuba); Yuto Matsunaga (NEC corporation); Kazuya Kakizaki (NEC Corporation / University of Tsukuba); Taiki Miyagawa (NEC Corporation)

247 Harder or Different? Understanding Generalization of Audio Deepfake Detection

Nicolas M Müller (Fraunhofer AISEC); Nicholas Evans (EURECOM); Hemlata Tak (Pindrop); Philip Sperl (Fraunhofer

AISEC); Konstantin Böttinger (Fraunhofer AISEC)

458 RW-VoiceShield: Raw Waveform-based Adversarial Attack on One-shot Voice Conversion

Ching-Yu Yang (National Tsing Hua University); Shreya G. Upadhyay (National Tsing Hua University); Ya-Tse Wu (Department of Electrical Engineering, National Tsing Hua University); Bo-Hao Su (Department of Electrical Engineering, National Tsing Hua University); Chi-Chun Lee (National Tsing Hua University)

698 RawBMamba: End-to-End Bidirectional State Space Model for Audio Deepfake Detection

Yujie Chen (Anhui University); Jiangyan Yi (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences); Jun Xue (Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University); chenglong wang (CASIA); XiaoHui Zhang (School of Computer and Information Technology, Beijing Jiaotong University); Dong Shunbo (Anhui University); siding zeng (Institute of Automation, Chinese Academy of Sciences); Jianhua Tao (Tsinghua University); zhao lv (anhui university); Cunhang Fan (Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University)

1258 How Private is Low-Frequency Speech Audio in the Wild? An Analysis of Verbal Intelligibility by Humans and Machines

Ailin Liu (RWTH Aachen University); Pepijn Vunderink (TU Delft); Jose Vargas Quiros (Delft University of Technology); Chirag A Raman (Delft University of Technology); Hayley Hung (TU Delft)

2165 Robust spread spectrum speech watermarking using linear prediction and deep spectral shaping

Nikolay D Gaubitch (Pindrop); David Looney (Pindrop)

Oral Session: Speech Quality Assessment

A5-05 Location: Aegle B

Surv-05-3 Neural Speech Assessment Metrics

Yu Tsao

1243 Embedding Learning for Preference-based Speech Quality Assessment

ChengHung Hu (Nagoya University); Yusuke Yasuda (Nagoya university); Tomoki Toda (Nagoya University)

1778 Experimental evaluation of MOS, AB and BWS listening test designs

Dan Wells (University of Edinburgh); Andrea L Aldana (Edinburgh University); Cassia Valentini (University of Edinburgh); Erica Cooper (National Institute of Informatics); Aidan Pine (National Research Council Canada); Junichi Yamagishi (National Institute of Informatics); Korin Richmond (University of Edinburgh)

1967 IndicMOS: Multilingual MOS Prediction for 7 Indian languages

Sathvik Udupa (Indian Institute of Science); Soumi Maiti (CMU); Prasanta Kumar Ghosh (Indian Institute of Science (IISc), Bangalore)

2527 Enhancing No-Reference Speech Quality Assessment with Pairwise, Triplet Ranking Losses, and ASR Pretraining

Ta Bao Thang (Viettel Cyberspace Center); Tu Minh Le (Viettel AI, Viettel Group); Van Hai Do (Thuyloi University); Huynh Thi Thanh Binh (Hanoi University of Science and Technology)

Oral Session: Novel Architectures for ASR

A9-03 Location: Hippocrates

Surv-09-2 Novel architectures for ASR

Rohit Prabhavalkar

258 Efficient and Robust Long-Form Speech Recognition with Hybrid H3-Conformer

Tomoki Honda (Kyoto University); Shinsuke Sakai (Kyoto University); Tatsuya Kawahara (Kyoto University)

569 SWAN: SubWord Alignment Network for HMM-free word timing estimation in end-to-end automatic speech recognition

Woohyun Kang (Amazon Web Services); Srikanth Vishnubhotla (Amazon Web Services); Rudolf Braun (Amazon Web Services); Yogesh Virkar (AWS AI Labs); Raghuveer Peri (Amazon Web Services Inc.); Kyu Han (Amazon Web Services (AWS))

574 Quantifying Unintended Memorization in BEST-RQ ASR Encoders

Virat Shejwalkar (Google); Om Thakkar (Google); Arun Narayanan (Google Inc.)

702 Rapid Language Adaptation for Multilingual E2E Speech Recognition Using Encoder Prompting

Yosuke Kashiwagi (Sony); Hayato Futami (Sony Group Corporation); Emiru Tsunoo (Sony Group Corporation); Sidhant Arora (Carnegie Mellon University); Shinji Watanabe (Carnegie Mellon University)

Oral Session: Speech Emotion Recognition**A3-03**

Location: Iasso

Surv-03-1 Automatic Emotion Detection/Recognition

Carlos Busso

280 ExHuBERT: Enhancing HuBERT Through Block Extension and Fine-Tuning on 37 Emotion Datasets

Shahin Amiriparian (Technical University of Munich); Filip Packań (University of Augsburg); Maurice Gerczuk (University of Augsburg); Prof. Dr. Bjoern Schuller (Imperial College London)

651 DropFormer: A Dynamic Noise-Dropping Transformer for Speech Emotion Recognition

Jialong Mai (South China University of Technology); Xiaofen Xing (South China University of Technology); Weidong Chen (South China University of Technology); Xiangmin Xu (South China University of Technology)

1430 Dataset-Distillation Generative Model for Speech Emotion Recognition

Fabian Alejandro Ritter-Gutierrez (Nanyang Technological University); Kuan-Po Huang (National Taiwan University); Jeremy H. M. Wong (Institute for Infocomm Research); Dianwen Ng (Alibaba Group / Nanyang Technological University); Hung-yi Lee (National Taiwan University); Nancy Chen (Institute for Infocomm Research); Eng Siong Chng (Nanyang Technological University)

2282 From Text to Emotion: Unveiling the Emotion Annotation Capabilities of LLMs

Minxue Niu (University of Michigan); Mimansa Jaiswal (Norm AI); Emily K Mower Provost (University of Michigan)

Oral Session: Spoken Dialogue Systems and Conversational Analysis 1**A11-02**

Location: Melampus

1008 Contextual Interactive Evaluation of TTS Models in Dialogue Systems

Siyang Wang (KTH Royal Institute of Technology, Stockholm); Eva Szekely (KTH Royal Institute of Technology); Joakim Gustafson (KTH Royal Institute of Technology)

1082 Joint Learning of Context and Feedback Embeddings in Spoken Dialogue

Livia Qian (KTH Royal Institute of Technology); Gabriel Skantze (KTH)

1514 GSQA: An End-to-End Model for Generative Spoken Question Answering

Min-Han Shih (National Taiwan University); ho lam Chung (National Taiwan University); Yu-Chi Pai (National Taiwan University); Ming-Hao Hsu (National Taiwan University); Guan-Ting Lin (National Taiwan University); Shang-Wen Li (FAIR); Hung-yi Lee (National Taiwan University)

1700 DubWise: Video-Guided Speech Duration Control in Multimodal LLM-based Text-to-Speech for Dubbing

Neha Sahipjohn (Sony Research India); Ashishkumar Prabhakar Gudmalwar (Sony Research India); Nirmesh J Shah (Sony Research India); Pankaj Wasnik (Sony Research India); Rajiv Ratn Shah (IIIT Delhi)

2352 Autoregressive cross-interlocutor attention scores meaningfully capture conversational dynamics

Matthew McNeill (CUNY Graduate Center); Rivka Levitan (City university of new york)

2402 ConvoCache: Smart Re-Use of Chatbot Responses

Conor Atkins (Macquarie University); Ian D Wood (Macquarie University); Mohamed Ali Kaafar (Macquarie University); Hassan Jameel Asghar (Macquarie University); Nardine Basta (Macquarie University); Michal Kepkowski (Macquarie University)

Oral Session: Pathological Speech Analysis 2

A13-02 Location: Panacea Amphitheater

1122 Voice Disorder Analysis: a Transformer-based Approach

Alkis Koudounas (Politecnico di Torino); Gabriele Ciravegna (Politecnico di Torino); Marco Fantini (San Feliciano Hospital); Erika Crosetti (San Giovanni Bosco Hospital); Giovanni Succo (San Giovanni Bosco Hospital); Tania Cerquitelli (Dipartimento di Automatica e Informatica Politecnico di Torino); Elena Baralis (Politecnico di Torino)

1400 Quantifying the effect of speech pathology on automatic and human speaker verification

Bence M Halpern (Nagoya University); Thomas B Tienkamp (University of Groningen); Wen-Chin Huang (Nagoya University); Lester Phillip G Violeta (Nagoya University); Teja Rebernik (University of Groningen); Sebastiaan de Visscher (University Medical Center Groningen); Max J. H. Witjes (University Medical Center Groningen); Martijn Wieling (University of Groningen); Defne Abur (University of Groningen); Tomoki Toda (Nagoya University)

1446 Detection of Cognitive Impairment And Alzheimer's Disease Using a Speech- and Language-Based Protocol

Tanya Talkar (Aural Analytics); Sherman Charles (Aural Analytics); Chelsea Krantsevich (Aural Analytics); Kan Kawabata (Aural Analytics)

1737 Investigation of Layer-Wise Speech Representations in Self-Supervised Learning Models: A Cross-Lingual Study in Detecting Depression

Bubai Maji (Indian Institute of Technology Kharagpur); Rajlakshmi Guha (IIT Kharagpur); Aurobinda Routray (IIT Kharagpur); Shazia Nasreen (IIT Kharagpur); Debabrata Majumdar (IIT Kharagpur)

2137 Prosody-Driven Privacy-Preserving Dementia Detection

Dominika C Woszczyk (Imperial College London); Ranya Aloufi (Imperial College London); Soteris Demetriou (Imperial College London)

2288 Analyzing Multimodal Features of Spontaneous Voice Assistant Commands for Mild Cognitive Impairment Detection

Nana Lin (University of Massachusetts Boston); Youxiang Zhu (University of Massachusetts Boston); Xiaohui Liang (University of Massachusetts Boston); John Batsis (University of North Carolina); Caroline Summerour (University of North Carolina)

Poster Session: Databases and Progress in Methodology

A1-P1-A

Location: Poster Area 1A

135 UY/CH-CHILD -- A Public Chinese L2 Speech Database of Uyghur Children

Mewlude Nijat (Xinjiang University); Chen Chen (Tsinghua University); Dong Wang (Tsinghua University); Askar Hamdulla (Xinjiang University)

646 VoxSim: A perceptual voice similarity dataset

Junseok Ahn (KAIST); Youkyum Kim (KAIST); Yeunju Choi (Samsung Research); Doyeop Kwak (KAIST); Ji-Hoon Kim (KAIST); Seongkyu Mun (Samsung Research); Joon Son Chung (KAIST)

1005 Leveraging Large Language Models to Refine Automatic Feedback Generation at Articulatory Level in Computer Aided Pronunciation Training

钟辉航 (北京语言大学); Yanlu Xie (Beijing Language and Culture University); ZiJin Yao (Beijing Language and Culture University)

1404 Decoding Human Language Acquisition: EEG Evidence for Predictive Probabilistic Statistics in Word Segmentation

Bin Zhao (Institute of Linguistics, Chinese Academy of Social Sciences); Mingxuan Huang (Department of Chinese Language and Literature, Fudan University); Chenlu Ma (Department of Chinese Language and Literature, Fudan University); Jinyi Xue (College of Foreign Languages and Literature, Fudan University); Aijun Li ("Institute of Linguistics, CASS"); Kunyu Xu (Institute of Modern Languages and Linguistics, Fudan University)

1505 DBD-CI: Doubling the Band Density for Bilateral Cochlear Implants

Mingyue Shi (Shenzhen University); Huali Zhou (Shenzhen University); Qinglin Meng (South China University of Technology); Nengheng Zheng (Shenzhen University)

2263 State-of-the-art speech production MRI protocol for new 0.55 Tesla scanners

Prakash Kumar (University of Southern California); Ye Tian (University of Southern California); Yongwan Lim (University of Southern California); Sophia Cui (Siemens Healthineers); Christina Hagedorn (CUNY); Uttam Sinha (Keck School of Medicine, University of Southern California); Dani Byrd (University of Southern California); Shrikanth Narayanan (USC); Krishna Nayak (University of Southern California)

Poster Session: Articulation, Convergence and Perception

A1-P1-B

Location: Poster Area 1B

746 Magnitude and timing of acceleration peaks in stressed and unstressed syllables

Malin Svensson Lundmark (Lund University)

1001 Behavioral evidence for higher speech rate convergence following natural than artificial time altered speech

Jérémy Giroud (MRC Cognition and Brain Science Unit, University of Cambridge); Jessica Lei (MRC Cognition and Brain Science Unit, University of Cambridge); Kirsty Phillips (MRC Cognition and Brain Science Unit, University of Cambridge); Matthew Davis (MRC Cognition and Brain Science Unit, University of Cambridge)

1083 What if HAL breathed? Enhancing Empathy in Human-AI Interactions with Breathing Speech Synthesis

Nicolò Loddò (Utrecht University); Francisca Pessanha (Utrecht University); Almila Akdag (Utrecht University)

1598 A novel experimental design for the study of listener-to-listener convergence in phoneme categorization

Noel Nguyen (Aix-Marseille University); Qingye Shen (Aix-Marseille University); Leonardo Lancia (Laboratoire Parole et Langage (CNRS/Aix-Marseille Université))

1662 Cross-Attention-Guided WaveNet for EEG-to-MEL Spectrogram Reconstruction

Hao Li (South University of Science and Technology of China); Yuan Fang (Inner Mongolia university); Xueliang zhang (Inner Mongolia University); Fei Chen (Southern University of Science and Technology); Guanglai Gao (Inner Mongolia University)

Poster Session: Training Methods, Self-Supervised Learning, Adaptation

A8-P4

Location: Poster Area 2A, Poster Area 2B

136 Unsupervised Online Continual Learning for Automatic Speech Recognition

Steven Vander Eeck (KU Leuven); Hugo Van hamme (KU LEUVEN)

163 Online Knowledge Distillation of Decoder-Only Large Language Models for Efficient Speech Recognition

JEEHYE LEE (KakaoBrain); hyeji seo (kakaobrain)

270 Dual-path Adaptation of Pretrained Feature Extraction Module for Robust Automatic Speech Recognition

Hao Shi (Kyoto University); Tatsuya Kawahara (Kyoto University)

536 Efficiently Train ASR Models that Memorize Less and Perform Better with Per-core Clipping

Lun Wang (Google); Om Thakkar (Google); Zhong Meng (Google); Nicole Rafidi (Google); Rohit Prabhavalkar (Google); Arun Narayanan (Google Inc.)

858 Self-Train Before You Transcribe

Robert J Flynn (Sheffield University); Anton Ragni (University of Sheffield)

1153 MSRS: Training Multimodal Speech Recognition Models from Scratch with Sparse Mask Optimization

Adriana Fernandez-Lopez (Meta); Honglie Chen (Meta); Pingchuan Ma (Meta); Lu Yin (University of Aberdeen); Qiao Xiao (Eindhoven University of Technology); Stavros Petridis (Meta); Shiwei Liu (UT Austin); Maja Pantic (Meta)

1403 ROAR: Reinforcing Original to Augmented Data Ratio Dynamics for Wav2vec2.0 Based ASR

Vishwanath Pratap Singh (University of Eastern Finland); Federico Malato (University of Eastern Finland); Ville Hautamäki (University of Eastern Finland); Md Sahidullah (Institute for Advancing Intelligence, TCG CREST); Tomi H. Kinnunen (University of Eastern Finland)

1434 Speaker personalization for automatic speech recognition using Weight-Decomposed Low-Rank Adaptation

George Joseph (Samsung); Arun Baby (Samsung Research, Bangalore)

1542 Hierarchical Multi-Task Learning with CTC and Recursive Operation

Nahomi Kusunoki (Waseda University); Yosuke Higuchi (Waseda University); Tetsuji Ogawa (Waseda University); Tetsunori Kobayashi (Waseda University)

1621 Personality-memory Gated Adaptation: An Efficient Speaker Adaptation for Personalized End-to-end Automatic Speech Recognition

Yue Gu (Harbin Institute of Technology); Zhihao Du (Speech Lab of DAMO Academy, Alibaba Group); Shiliang Zhang (Alibaba Group); jiqing Han (Harbin Institute of Technology); Yongjun He (Harbin Institute of Technology)

1776 Boosting CTC-based ASR using inter-layer attention-based CTC loss

Keigo Hojo (Toyohashi University of Technology); Yukoh Wakabayashi (Toyohashi University of Technology); Kengo Ohta (National Institute of Technology, Anan College, Japan); Atsunori Ogawa (NTT Corporation); Norihide Kitaoka

(Toyohashi University of Technology)

1813 Online Subloop Search via Uncertainty Quantization for Efficient Test-Time Adaptation

Jae-Hong Lee (Hanyang University); Sang-Eon Lee (Hanyang University); Dong-Hyun Kim (Hanyang University); Dohee Kim (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1824 LASER: Learning by Aligning Self-supervised Representations of Speech for Improving Content-related Tasks

Amit Meghanani (University of Sheffield); Thomas Hain (University of Sheffield)

2187 Speech and Language Recognition with Low-rank Adaptation of Pretrained Models

Amrutha Prasad (Idiap Research Institute); Srikanth Madikeri (Idiap); driss khalil (idiap); Petr Motlicek (Idiap); Christof Schuepbach (Armasuisse Science and Technology)

2188 Convolution-Augmented Parameter-Efficient Fine-Tuning for Speech Recognition

Kwangyoun Kim (ASAPP); Suwon Shon (ASAPP); Yi-Te Hsu (ASAPP); Prashant Sridhar (ASAPP); Karen Livescu (TTI-Chicago); Shinji Watanabe (Carnegie Mellon University)

2323 Self-training ASR Guided by Unsupervised ASR Teacher

Hyung Yong Kim (42dot); Byeong-Yeol Kim (42dot); Yunkyu Lim (42dot); Jihwan Park (42dot Inc.); Shukjae Choi (42dot); Yooncheol Ju (42dot, Seoul, Republic of Korea); Jinseok Park (42dot); Youshin Lim (42dot); Seung Woo Yu (42dot); Hanbin Lee (42dot); Shinji Watanabe (Carnegie Mellon University)

Poster Session: Multimodality and Foundation Models

A10-P1

Location: Poster Area 3A

212 SAMSEMO: New dataset for multilingual and multimodal emotion recognition

Pawel Bujnowski (Samsung Research Poland); Joanna Marhula (Samsung Research Poland); Zuzanna Bordzicka (Samsung Research Poland); Bartlomiej Kuzma (Samsung Research Poland); Bartlomiej Paziewski (Samsung Research Poland); Jacek Rutkowski (Samsung Research Poland); Piotr Andruszkiewicz (Samsung Research Institute Poland)

1595 Enhancing Speech-Driven 3D Facial Animation with Audio-Visual Guidance from Lip Reading Expert

Han EunGi (POSTECH); Oh Hyun-Bin (POSTECH); Kim Sung-Bin (POSTECH); Corentin Nivelet Etcheberry (Bordeaux INP); Suekyeong Nam (KRAFTON Inc.); Janghoon Ju (KRAFTON); Tae-Hyun Oh (POSTECH)

1895 Spontaneous Speech-Based Suicide Risk Detection Using Whisper and Large Language Models

Ziyun Cui (Tsinghua University); Chang Lei (Tsinghua University); Wen Wu (University of Cambridge); Yinan Duan (Tsinghua University); Diyang Qu (Tsinghua University); Ji Wu (Tsinghua University); Runsen Chen (Tsinghua University); Chao Zhang (Tsinghua University)

2181 Spoken Word2Vec: Learning Skipgram Embeddings from Speech

Mohammad Amaan Sayeed (MBZUAI); Hanan Aldarmaki (MBZUAI)

2497 Zero-shot Fake Video Detection by Audio-visual Consistency

Xiaolou Li (Beijing University of Posts and Telecommunications); Zehua Liu (Beijing University of Posts and Telecommunications); Chen Chen (Tsinghua University); Lantian Li (Beijing University of Posts and Telecommunications); Li Guo (Beijing University of Posts and Telecommunications); Dong Wang (Tsinghua University)

2550 LLM-Driven Multimodal Opinion Expression Identification

Bonian Jia (tianjin university); Huiyao Chen (Harbin Institute of Technology (Shenzhen)); Yueheng Sun (Tianjin University); Meishan Zhang (Harbin Institute of Technology (Shenzhen)); Min Zhang (Soochow University)

Poster Session: Speech Technology

A12-P4

Location: Poster Area 3B

777 Transferable speech-to-text large language model alignment module*Wu Boyong (Cloudwalk Technology); Yan Chao (Cloudwalk); Haoran Pu (Cloudwalk Technology)***989 Towards interfacing large language models with ASR systems using confidence measures and prompting***Maryam Naderi (Idiap); Enno Hermann (Idiap Research Institute); Alexandre Nanchen (Idiap Research Institute); Seveda Hovsepyan (Idiap Research Institute); Mathew Magimai.-Doss (Idiap Research Institute)***1471 Text Injection for Neural Contextual Biasing***Zhong Meng (Google); Zelin Wu (Google LLC); Rohit Prabhavalkar (Google); Charles C Peyser (Google Inc.); Weiran Wang (Google); Nanxin Chen (Google); Tara Sainath (Google); Bhuvana Ramabhadran (Google)***1583 Acceleration of Posteriorgram-based DTW by Distilling the Class-to-class Distances Encoded in the Classifier Used to Calculate Posteriors***HAITONG SUN (The University of Tokyo); Choi Jaehyun (The University of Tokyo); Nobuaki Minematsu (The University of Tokyo); Daisuke Saito (The University of Tokyo)***1672 VECL-TTS: Voice identity and Emotional style controllable Cross-Lingual Text-to-Speech***Ashishkumar Prabhakar Gudmalwar (Sony Research India); Nirmesh J Shah (Sony Research India); Sai Akarsh C (SRI); Pankaj Wasnik (Sony Research India); Rajiv Ratn Shah (IIIT Delhi)***1772 Prompting Large Language Models with Mispronunciation Detection and Diagnosis Abilities***Minglin Wu (The Chinese University of Hong Kong); Jing Xu (The Chinese University of Hong Kong); Xixin Wu (The Chinese University of Hong Kong); Helen Meng (The Chinese University of Hong Kong)***1779 Resource-Efficient Speech Quality Prediction through Quantization Aware Training and Binary Activation Maps***Mattias Nilsson (Friedrich Miescher Institute); Riccardo Miccini (Technical University of Denmark); Clement Laroche (GN Audio); Tobias Piechowiak (Jabra); Friedemann Zenke (Friedrich Miescher Institute)*

Poster Session: Speech Synthesis: Voice Conversion 2

A7-P2-A

Location: Poster Area 4A

207 Disentangling prosody and timbre embeddings via voice conversion*Olivier Le Blouch (Orange Labs); Nicolas Gengembre (Orange Labs); Cédric Gendrot (LPP)***250 LDM-SVC: Latent Diffusion Model Based Zero-Shot Any-to-Any Singing Voice Conversion with Singer Guidance***shihao chen (University of Science and Technology of China); Yu Gu (Tencent AILab); Jie Zhang (University of Science and Technology of China (USTC)); Na Li (Tencent); Rilin Chen (tencent); Liping chen (University of Science and Technology of China); Lirong Dai (University of Science and Technology of China)***539 Utilizing Adaptive Global Response Normalization and Cluster-Based Pseudo Labels for Zero-Shot Voice Conversion***Ji Sub Um (KAIST); Hoirin Kim (KAIST)***710 Residual Speaker Representation for One-Shot Voice Conversion***Le Xu (NLPR); Jiangyan Yi (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences); Tao Wang (Institute of Automation, Chinese Academy of Sciences); Yong Ren (Institute of Automation,*

Chinese Academy of Sciences); Rongxiu Zhong (China Mobile Research Institute); Zhengqi Wen (Qiyuan Lab); Jianhua Tao ("National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences")

771 Pre-training Neural Transducer-based Streaming Voice Conversion for Faster Convergence and Alignment-free Training

Hiroki Kanagawa (NTT Corporation); Takafumi Moriya (NTT); Yusuke Ijima (NTT Corporation)

972 Noise-Robust Voice Conversion by Conditional Denoising Training Using Latent Variables of Recording Quality and Environment

Takuto Igarashi (The University of Tokyo); Yuki Saito ("The University of Tokyo, Japan"); Kentaro Seki (The University of Tokyo); Shinnosuke Takamichi (The University of Tokyo); Ryuichi Yamamoto (LY Corp.); Kentaro Tachibana (LY Corp.); Hiroshi Saruwatari (The University of Tokyo)

2091 Improvement Speaker Similarity for Zero-Shot Any-to-Any Voice Conversion of Whispered and Regular Speech

Aleksei Gusev (FluentaAI); Anastasia Avdeeva (FluentaAI)

2362 Vec-Tok-VC+: Residual-enhanced Robust Zero-shot Voice Conversion with Progressive Constraints in a Dual-mode Training Strategy

linhan ma (Northwestern Polytechnical University); Xinfu Zhu (Northwestern Polytechnical University); Yuanjun Lv (Northwestern Polytechnical University); Zhichao Wang (Northwestern Polytechnical University); Ziqian Wang (Northwestern Polytechnical University); WENDI HE (xmly); Hongbin Zhou (Ximalaya Inc.); Lei Xie (NWPU)

Poster Session: Speech Synthesis: Text Processing

A7-P2-B

Location: Poster Area 4B

19 G2PA: G2P for Training with Aligned Audio for Mandarin Chinese

Xingxing Yang (HKUST)

66 Learning Pronunciation from Other Accents via Pronunciation Knowledge Transfer

Siqi Sun (The University of Edinburgh); Korin Richmond (University of Edinburgh)

313 A Language Modeling Approach to Diacritic-Free Hebrew TTS

Amit Roth (The Hebrew University of Jerusalem); Arnon Turetzky (Hebrew University of Jerusalem); Yossi Adi (The Hebrew University of Jerusalem)

342 Audio-conditioned phonetic/prosodic annotation for building text-to-speech models from unlabeled speech data

Yuma Shirahata (LY Corp.); Byeongseon Park (LY Corp.); Ryuichi Yamamoto (LY Corp.); Kentaro Tachibana (LINE Corp.)

533 Exploring the Benefits of Tokenization of Discrete Acoustic Units

Avihu Dekel (IBM Research); Raul Fernandez (IBM Research)

708 Enhancing Japanese Text-to-Speech Accuracy with a Novel Combination Transformer-BERT-based G2P: Integrating Pronunciation Dictionaries and Accent Sandhi

Kiyoshi Kurihara (NHK (Japan Broadcasting Corporation)); Masanori Sano (NHK)

949 Homograph Disambiguation with Text-to-Text Transfer Transformer

Markéta Řezáčková (University of West Bohemia); Daniel Tihelka (University of West Bohemia); Jindrich Matousek (University of West Bohemia, Pilsen, Czech Republic)

1237 Positional Description for Numerical Normalization

Deepanshu Gupta (Apple); Javier Latorre (Apple)

1565 Beyond graphemes and phonemes: continuous phonological features in neural text-to-speech synthesis

Christina Tännander (KTH Royal Institute of Technology); Shivam Mehta (KTH Royal Institute of Technology); Jonas Beskow (KTH Royal Institute of Technology); Jens Edlund (KTH Royal Institute of Technology)

Poster Session: Speech Science, Speech Technology, and Gender (Special Session)

SS-1 Location: Yanis Club

544 Voice Quality Variation in AAE: An Additional Challenge for Addressing Bias in ASR Models?

Li-Fang Lai (Pomona College); Nicole Holliday (Pomona College)

1132 Acoustic Effects of Facial Feminisation Surgery on Speech and Singing: A Case Study

Clíodhna Hughes (University of Sheffield); Ning Ma (University of Sheffield); Nicola Dibben (Department of Music, University of Sheffield, UK); Guy J Brown (University of Sheffield)

1177 Articulatory Configurations across Genders and Periods in French Radio and TV archives

Benjamin Elie (University of Edinburgh); David Doukhan (Institut National de l'Audiovisuel (INA)); Remi Uro (Institut National de l'Audiovisuel); Lucas ONDEL YANG (LISN, CNRS); Albert Rilliard (Université Paris Saclay, CNRS, LISN); Simon Devauchelle (Université Paris Saclay)

1543 An Inclusive Approach to Creating a Palette of Synthetic Voices for Gender Diversity

Eva Szekely (KTH Royal Institute of Technology); Maxwell Hope (University of Delaware)

1717 Challenges of German Speech Recognition: A Study on Multi-ethnolectal Speech Among Adolescents

Martha Schubert (Otto von Guericke University); Ingo Siegert (OvG University Magdeburg); Daniel Duran (Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS))

1854 Automatic Classification of News Subjects in Broadcast News: Application to a Gender Bias Representation Analysis

Valentin Pelloin (INA); Léna Dodson (ARCOM); Émile Chapuis (INA); Nicolas Hervé (INA); David Doukhan (Institut National de l'Audiovisuel (INA))

1921 Gender Representation in TV and Radio: Automatic Information Extraction methods versus Manual Analyses

David Doukhan (Institut National de l'Audiovisuel (INA)); Léna Dodson (ARCOM); Manon Conan (ARCOM); Valentin Pelloin (INA); Aurélien Clamouse (ARCOM); Mélina Lepape (ARCOM); Géraldine Van Hille (ARCOM); Cécile Méadel (CARISM); Marlène Coulomb-Gully (LERASS)

1982 Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices.

Atli Thor Sigurgeirsson (University of Edinburgh); Eddie L. Ungless (University of Edinburgh)

2209 On the Encoding of Gender in Transformer-based ASR Representations

Aravind Krishnan (Saarland University); Badr M Abdullah (Saarland University); Dietrich Klakow (Saarland University)

2269 Speech After Gender: A Trans-Feminine Perspective on Next Steps for Speech Science and Technology

Robin Netzorg (UC Berkeley); Alyssa Cote (Alyssa's Voice Training); Sumi Koshin (SumianVoice); Klo Vivienne Garoute (TransVoiceLessons); Gopala Krishna Anumanchipalli (UC Berkeley)

Time Slot 2

Oral Session: Phonetics and Phonology: Segmentals and Suprasegmentals

A2-04

Location: Acesso

525 Aerodynamics of Sakata labial-velar oral stops

Lorenzo Maselli (*Universiteit Gent, Université de Mons*); Véronique Delvaux (*FNRS & UMONS*)

666 Key Acoustic Cues for the Realization of Metrical Prominence in Tone Languages: A Cross-Dialect Study

Yiyang Hu (*Tianjin University*); Hui Feng (*Tianjin University*)

1184 Revisiting Pitch Jumps: F0 Ratio in Seoul Korean

Michaela Watkins (*University of Amsterdam*); Paul Boersma (*University of Amsterdam*); Silke Hamann (*University of Amsterdam*)

1216 Collecting Mandible Movement in Brazilian Portuguese

Donna M Erickson (*Haskins labs*); Albert Rilliard (*Université Paris Saclay, CNRS, LISN*); Malin Svensson Lundmark (*Lund University*); Adelaide Silva (*Universidade Federal do Paraná*); Leticia Rebollo-Couto (*Universidade Federal do Rio de Janeiro (UFRJ)*); Oliver Niebuhr (*University of Southern Denmark*); João Moraes (*Universidad Federal do Rio de Janeiro*)

1731 Frication noise features of Polish voiceless dental fricative and affricate produced by children with and without speech disorder

Zuzanna Miodonska (*Silesian University of Technology*); Michał Marek Kręcichwost (*Silesian University of Technology*); Ewa Kwaśniok (*Medical Facility ""Therapy"" - Center for Health and Human Development*); Agata Sage (*Silesian University of Technology*); Paweł Badura (*Silesian University of Technology*)

2247 Pitch-driven adjustments in tongue positions: Insights from ultrasound imaging

May Pik Yu Chan (*University of Pennsylvania*); Jianjing Kuang (*University of Pennsylvania*)

Oral Session: Multi-Channel Speech Enhancement

A6-04

Location: Aegle A

4 SA-MF: A Novel Self-Attention Mechanism for Multifeature Fusion in Speech Enhancement Networks

Ruizhe Wang (*Harbin institute of technology*)

545 Audio Enhancement from Multiple Crowdsourced Recordings: A Simple and Effective Baseline

Shiran Aziz (*The Hebrew University*); Yossi Adi (*The Hebrew University of Jerusalem*); Shmuel Peleg (*The Hebrew University*)

669 DeFTAN-AA: Array Geometry Agnostic Multichannel Speech Enhancement

Dongheon Lee (*KAIST*); Jung-Woo Choi (*KAIST*)

801 PLDNet: PLD-Guided Lightweight Deep Network Boosted by Efficient Attention for Handheld Dual-Microphone Speech Enhancement

Nan Zhou (*Shenzhen Transsion Holdings Co., Ltd, Shanghai Branch, China*); Youhai Jiang (*Shenzhen Transsion Holdings Co., Ltd, Shanghai Branch, China*); Jialin Tan (*Shenzhen Transsion Holdings Co., Ltd, Shanghai Branch, China*); chongmin Qi (*Shenzhen Transsion Holdings Co., Ltd, Shanghai Branch, China*)

2124 FoVNet: Configurable Field-of-View Speech Enhancement with Low Computation and Distortion

for Smart Glasses

Zhongweiyang Xu (University of Illinois Urbana-Champaign); Ali Aroudi (META); Ke Tan (Meta Platforms, Inc.); Ashutosh Pandey (META); Jung-Suk Lee (META); Buye Xu (Meta Reality Labs Research); Francesco Nesta (META)

2427 Array Geometry-Robust Attention-Based Neural Beamformer for Moving Speakers

Marvin Tammen (University of Oldenburg); Tsubasa Ochiai (NTT); Marc Delcroix (NTT Communication Science Laboratories); Tomohiro Nakatani (NTT Communication Science Laboratories, NTT Corporation); Shoko Araki (NTT Corporation); Simon Doclo (University of Oldenburg)

Oral Session: Error Correction and Rescoring**A9-04**

Location: Aegle B

148 Retrieval Augmented Speech Understanding through Generative Modeling

Hao Yang (Huawei); minghan wang (monash); Jiaxin GUO (Huawei); Min Zhang (Huawei)

385 HypR: A comprehensive study for ASR hypothesis revising with a reference corpus

Yi-Wei Wang (National Taiwan University of Science and Technology); Ke-Han Lu (National Taiwan University); Kuan-Yu CHEN (National Taiwan University of Science and Technology)

1449 TRANSFORMER-BASED MODEL FOR ASR N-BEST RESCORING AND REWRITING

Iwen E Kang (Apple); Christophe Van Gysel (Apple); Manhung Siu (Apple)

1605 Error Correction by Paying Attention to Both Acoustic and Confidence References for Automatic Speech Recognition

Yuchun Shu (Tianjin University); Bo Hu (BAIDU); Yifeng He (Baidu); Hao Shi (Kyoto University); Longbiao Wang (Tianjin University); Jianwu Dang (Tianjin University)

1829 LI-TTA: Language Informed Test-Time Adaptation for Automatic Speech Recognition

Eunseop Yoon (KAIST); Hee Suk Yoon (KAIST); John Harvill (University of Illinois at Urbana-Champaign); Mark A Hasegawa-Johnson (University of Illinois); Chang D. Yoo (KAIST)

2499 SALSA: Speedy ASR-LLM Synchronous Aggregation

Ashish Mittal (IBM Research, IIT Bombay); Darshan Deepak Prabhu (Indian Institute of Technology, Bombay); Sunita Sarawagi (IIT Bombay); Preethi Jyothi (Indian Institute of Technology Bombay)

Oral Session: Neural Network Architectures for ASR 1**A8-05**

Location: Hippocrates

40 SummaryMixing: A Linear-Complexity Alternative to Self-Attention for Speech Recognition and Understanding

Titouan Parcollet (Samsung AI Cambridge / University of Cambridge); Rogier C. van Dalen (Samsung AI Center, Cambridge); Shucong Zhang (Samsung); Sourav Bhattacharya (Samsung)

442 Boosting Hybrid Autoregressive Transducer-based ASR with Internal Acoustic Model Training and Dual Blank Thresholding

Takafumi Moriya (NTT Corporation); Takanori Ashihara (NTT); Masato Mimura (NTT corporation); Hiroshi Sato (NTT); Kohei Matsuura (NTT); Ryo Masumura (NTT Corporation); Taichi Asami (NTT)

445 Conformer without Convolutions

Matthijs Van keirsbilck (NVIDIA Research); Alexander Keller (NVIDIA)

500 Linear-Complexity Self-Supervised Learning for Speech Processing

Shucong Zhang (Samsung); Titouan Parcollet (Samsung AI Cambridge / University of Cambridge); Rogier C. van Dalen (Samsung AI Center, Cambridge); Sourav Bhattacharya (Samsung)

505 RepCNN: Micro-sized, Mighty Models for Wakeword Detection

Arnav Kundu (Apple); Prateeth Nayak (Apple Inc.); Priyanka Padmanabhan (Apple); Devang Naik (Apple)

577 Contemplative Mechanism for Speech Recognition: Speech Encoders can Think

Tien-Ju Yang (Google); Andrew Rosenberg (Google LLC); Bhuvana Ramabhadran (Google)

Oral Session: Speaker Verification**A4-05**

Location: Iasso

742 ERes2NetV2: Boosting Short-Duration Speaker Verification Performance with Computational Efficiency

Yafeng Chen (Speech Lab, Alibaba Group); Siqi Zheng (Alibaba Group); Hui Wang (Speech Lab, Alibaba Group); Luyao Cheng (Alibaba Group); Qian Chen (Speech Lab, DAMO Academy, Alibaba Group); Shiliang Zhang (Alibaba Group); Junjie Li (University of Science and Technology of China)

1011 Extraction of interpretable and shared speaker-specific speech attributes through binary auto-encoder

Imen Ben-Amor (Université d'Avignon); Jean-Francois Bonastre (Université d'Avignon); Salima Mdhaftar (LIA - University of Avignon)

1354 To what extent can ASV systems naturally defend against spoofing attacks?

Jee-weon Jung (Carnegie Mellon University); Xin Wang (National Institute of Informatics); Nicholas Evans (EU-RECOM); Shinji Watanabe (Carnegie Mellon University); Hye-jin Shim (Carnegie Mellon University); Hemlata Tak (Pindrop); Siddhant Arora (Carnegie Mellon University); Junichi Yamagishi (National Institute of Informatics); Joon Son Chung (KAIST)

1800 Collaborative Contrastive Learning for Hypothesis Domain Adaptation

Jen-Tzung Chien (National Yang Ming Chiao Tung University); I-Pin Yeh (National Yang Ming Chiao Tung University); Man-Wai MAK (The Hong Kong Polytechnic University)

2058 Challenging margin-based speaker embedding extractors by using the variational information bottleneck

Themis Stafylakis (Omilia - Conversational Intelligence); Anna Silnova (Brno University of Technology); Johan Rohdin (Brno University of Technology); Plchot Oldřich (Brno University of Technology); Lukáš Burget (Brno University of Technology)

2116 Reshape Dimensions Network for Speaker Recognition

Ivan Yakovlev (ID R & D); Rostislav Makarov (ID R & D Inc.); Andrew Balykin (ID R & D Inc.); Pavel Malov (ID R & D Inc.); Anton Okhotnikov (ID R & D Inc.); Nikita Torgashov (ID R & D Inc.)

Oral Session: Speech and Audio Modelling**A5-06**

Location: Melampus

Surv-05-1 Toward speech and audio foundation models

Shinji Watanabe

51 GenDistiller: Distilling Pre-trained Language Models based on an Autoregressive Generative Model
 Yingying Gao (China Mobile Research); Shilei Zhang (China Mobile Research); Chao Deng (China Mobile Research Institute); Junlan Feng (China Mobile Research)

953 Gender and Language Identification in Multilingual Models of Speech: Exploring the Genericity and Robustness of Speech Representations
 Séverine Guillaume (LACITO-CNRS); Maxime Fily (LLF); Alexis Michaud (LACITO-CNRS); Guillaume Wisniewski (Université Paris Cité and LLF)

1156 Neural Compression Augmentation for Contrastive Audio Representation Learning
 Zhaoyu Wang (Imperial College London); Haohe Liu (University of Surrey); Harry Coppock (Imperial College London); Bjoern W. Schuller (Imperial College London); Mark D. Plumbley (University of Surrey)

2400 Post-Net: A linguistically inspired sequence-dependent transformed neural architecture for automatic syllable stress detection
 Sai Harshitha Aluru (Vidya Jyothi Institute Of Technology); Jhansi Mallela (Internantional Institute of Information Technology, Hyderabad); Chiranjeevi Yarra (International Institute of Information Technology Hyderabad)

Oral Session: Speech Production and Perception

A1-03 Location: Panacea Amphitheater

1299 Temporal Co-Registration of Simultaneous Electromagnetic Articulography and Electroencephalography for Precise Articulatory and Neural Data Alignment
 Daniel Friedrichs (University of Zurich); Monica Lancheros (University of Geneva); Sam Kirkham (Lancaster University); Lei He (University of Zurich); Andrew Clark (University of Zurich); Clemens Lutz (Univesity of Zurich); Volker Dellwo (University of Zurich); Steven Moran (University of Neuchatel)

1637 ON THE PERFORMANCE OF EMA-SYNCHRONIZED SPEECH AND STAND-ALONE SPEECH IN ACOUSTIC-TO-ARTICULATORY INVERSION
 Qiang Fang (Chinese Academy of Social Sciences)

1835 Glottal inverse filtering and vocal tract tuning for the numerical simulation of vowel /a/ with different levels of vocal effort
 Marc Freixes (La Salle - Universitat Ramon Llull); Marc Arnela (La Salle - Universitat Ramon Llull); Joan Claudi Socoró (La Salle - Universitat Ramon Llull); Luis Joglar-Ongay (La Salle - Universitat Ramon Llull); Oriol Guasch (La Salle - Universitat Ramon Llull); Francesc Alías-Pujol (La Salle - Universitat Ramon Llull)

1970 Measurement and simulation of pressure losses due to airflow in vocal tract models
 Peter Birkholz (TU Dresden); Patrick Häsner (TU Dresden)

2191 Towards a Quantitative Analysis of Coarticulation with a Phoneme-to-Articulatory Model
 Chaofei Fan (Stanford); Jaimie Henderson (Stanford); Christopher D Manning (Stanford University); Francis Willett (Stanford)

2415 A comparative study of the impact of voiceless alveolar and palato-alveolar sibilants in English on lip aperture and protrusion during VCV production
 Chetan Sharma (Indian Institute of Science); Vaishnavi Chandwanshi (Indian Institute of Science); Prasanta Kumar Ghosh (Indian Institute of Science (IISc), Bangalore)

Poster Session: Topics in Paralinguistics

A3-P2

Location: Poster Area 1A

121 Speaking of Health: Leveraging Large Language Models to assess Exercise Motivation and Behavior of Rehabilitation Patients*Suhas BN (Pennsylvania State University); Amanda Rebar (Central Queensland University); Saeed Abdullah (Penn State University, USA)***173 Who Finds This Voice Attractive? A Large-Scale Experiment Using In-the-Wild Data***Hitoshi Suda (National Institute of Advanced Industrial Science and Technology (AIST)); Aya Watanabe (The University of Tokyo); Shinnosuke Takamichi (The University of Tokyo)***268 Acoustical analysis of the initial phones in speech-laugh***RYO SETOBUCHI (Chiba institute of technology); Yoshiko Arimoto (Chiba Institute of Technology)***546 Confidence estimation for automatic detection of depression and Alzheimer's disease based on clinical interviews***Wen Wu (University of Cambridge); Chao Zhang (Tsinghua University); Phil Woodland (Machine Intelligence Laboratory, Cambridge University Department of Engineering)***1643 On Calibration of Speech Classification Models: Insights from Energy-Based Model Investigations***yaqian hao (China Mobile Research Institute); Chenguang Hu (chinamobile); Yingying Gao (China Mobile Research Institute); Shilei Zhang (China Mobile Research Institute); Junlan Feng (China Mobile Research Institute)***2518 Emotion-Aware Speech Self-Supervised Representation Learning with Intensity Knowledge***Rui Liu (Inner Mongolia University); zening ma (Inner Mongolia University)*

Poster Session: Emotion Recognition: Fairness, Variability, Uncertainty

A3-P2-B

Location: Poster Area 1B

119 Dual-Constrained Dynamical Neural ODEs for Ambiguity-aware Continuous Emotion Prediction*Jingyao Wu (University of New South Wales); Ting Dang (University of Melbourne); Vidhyasaharan Sethu (University of New South Wales); Eliathamby Ambikairajah (The University of New South Wales)***461 An Investigation of Group versus Individual Fairness in Perceptually Fair Speech Emotion Recognition***Woan-Shiuan Chien (Department of Electrical Engineering, National Tsing Hua University); Chi-Chun Lee (National Tsing Hua University)***471 An Inter-Speaker Fairness-Aware Speech Emotion Regression Framework***Hsing-Hang Chou (National Tsing Hua University); Woan-Shiuan Chien (Department of Electrical Engineering, National Tsing Hua University); Ya-Tse Wu (Department of Electrical Engineering, National Tsing Hua University); Chi-Chun Lee (National Tsing Hua University)***977 Are you sure? Analysing Uncertainty Quantification Approaches for Real-world Speech Emotion Recognition***Oliver W Schröfer (audEERING); Manuel Milling (Technical University of Munich); Felix Burkhardt (audEERING GmbH); Florian Eyben (audEERING); Bjoern Schuller (audEERING)***1273 Speech emotion recognition with deep learning beamforming on a distant human-robot interaction scenario***Ricardo García (Universidad de Chile); Rodrigo M Mahu (Universidad de Chile); Nicolás Grágeda (Universidad de Chile); Alejandro Luzanto (University of Chile); Nicolas Bohmer (University of Chile); Carlos Busso (University of Chile)*

Texas at Dallas); Néstor Becerra Yoma (*University of Chile*)

2307 The Whole Is Bigger Than the Sum of Its Parts: Modeling Individual Annotators to Capture Emotional Variability

James A. C. Tavernor (*University of Michigan*); Yara El-Tawil (*University of Michigan*); Emily K Mower Provost (*University of Michigan*)

2525 Iterative Prototype Refinement for Ambiguous Speech Emotion Recognition

Haoqin Sun (*Nankai University*); Shiwan Zhao (*Nankai University*); Xiangyu Kong (*University of Leicester*); Xuechen Wang (*Nankai University*); Hui Wang (*Nankai University*); Jiaming Zhou (*Nankai University*); Yong Qin (*Nankai University*)

Poster Session: Spatial Audio and Acoustics

A5-P3-A

Location: Poster Area 2A

141 Spatial Acoustic Enhancement Using Unbiased Relative Harmonic Coefficients

Liang Tao (*Beijing University of Technology*); Maoshen Jia (*Beijing University of Technology*); Yonggang Hu (*Southwest Institute of Electronics and Telecommunication Technology of China*); Changchun Bao (*Beijing University of Technology*)

1295 Design of Feedback Active Noise Cancellation Filter Using Nested Recurrent Neural Networks

Alireza Bayestehtashk (*Skyworks Inc*); Amit Kumar (*Skyworks Inc*); Michael J Wurtz (*Skyworks Inc.*)

1396 Novel-view Acoustic Synthesis From 3D Reconstructed Rooms

Byeongjoo Ahn (*Apple*); Karren D Yang (*Apple*); Brian Hamilton (*Apple*); Jonathan Sheaffer (*Apple*); Anurag Ranjan (*Apple*); Miguel Sarabia (*Apple*); Oncel Tuzel (*Apple*); Jen-Hao Rick Chang (*Apple*)

1951 RevRIR: Joint Reverberant Speech and Room Impulse Response Embedding using Contrastive Learning with Application to Room Shape Classification

Jacob Bitterman (*OrCam Technologies LTD*); Daniel Levi (*Bar-Ilan University*); Hilel Hagai Diamandi (*Yale University*); Sharon Gannot (*Bar-Ilan University*); Tal Rosenwein (*OrCam Technologies LTD*)

2053 RIR-in-a-Box: Estimating Room Acoustics from 3D Mesh Data through Shoebox Approximation

Liam O Kelley (*AIP RIKEN, Télécom Paris*); Diego Di Carlo (*RIKEN*); Mathieu Fontaine (*Télécom Paris*); Aditya Arie Nugraha (*RIKEN*); Yoshiaki Bando (*National Institute of Advanced Industrial Science and Technology*); Kazuyoshi Yoshii (*Kyoto University*)

2081 Classification of Room Impulse Responses and its application for channel verification and diarization

Yuri Khokhlov (*Speech Technology Center*); Tatiana Prisyach (*Speech Technology Center*); Anton Mitrofanov (*Speech Technology Center*); Dmitry Dutov (*Speech Technology Center*); Igor Agafonov (*Speech Technology Center*); Tatiana Timofeeva (*Speech Technology Center*); Aleksei Romanenko (*Speech Technology Center*); Maxim Korenevsky (*Speech Technology Center*)

2185 Neuromorphic Keyword Spotting with Pulse Density Modulation MEMS Microphones

Sidi Yaya Arnaud Yarga (*Université de Sherbrooke*); Sean Wood (*Université de Sherbrooke*)

Poster Session: Generative Models for Speech and Audio

A5-P3-B

Location: Poster Area 2B

38 Efficient Fine-tuning of Audio Spectrogram Transformers via Soft Mixture of Adapters

Umberto Cappellazzo (University of Trento); Daniele Falavigna (FBK); Alessio Brutti (FBK)

325 PAM: Prompting Audio-Language Models for Audio Quality Assessment

Soham Deshmukh (Microsoft); Dareen Alharthi (Carnegie Mellon University); Benjamin Elizalde (Microsoft); Hannes Gamper (Microsoft); Mahmoud Al Ismail (Microsoft); Rita Singh (Carnegie Mellon University); Bhiksha Raj (Carnegie Mellon University); Huaming Wang (Microsoft)

632 Phoneme Discretized Saliency Maps for Explainable Detection of AI Generated Voice

shubham Gupta (Mila); Mirco Ravanelli (Université de Montréal); Pascal Germain (Université Laval); Cem Subakan (Mila)

636 Audio Editing with Non-Rigid Text Prompts

Francesco Paissan (FBK); Luca Della Libera (Concordia University); Zhepei Wang (University of Illinois at Urbana-Champaign); Paris Smaragdis (University of Illinois at Urbana-Champaign); Mirco Ravanelli (Université de Montréal); Cem Subakan (Mila)

1071 Exploring compressibility of transformer based text-to-music (TTM) models

Vasileios Moschopoulos (Information Technologies Institute, Centre for Research and Technology - Hellas, Thessaloniki, Greece); Thanasis Kotsiopoulou (Information Technologies Institute, Centre for Research and Technology - Hellas, Thessaloniki, Greece); Pablo Peso Parada (Samsung Research UK); Konstantinos Nikiforidis (Information Technologies Institute, Centre for Research and Technology - Hellas, Thessaloniki, Greece); Alexandros Stergiadis (Pragma-IoT); Gerasimos Papakostas (Information Technologies Institute, Centre for Research and Technology - Hellas, Thessaloniki, Greece); Md Asif Jalal (Samsung Research UK); Jisi Zhang (Samsung Research UK); Anastasios Drosou (CERTH-ITI); KARTHIKEYAN SARAVANAN (Samsung Research, UK)

1333 ConsistencyTTA: Accelerating Diffusion-Based Text-to-Audio Generation with Consistency Distillation

Yatong Bai (University of California, Berkeley); Trung V Dang (Microsoft); Dung N Tran (Microsoft Corporation); Kazuhito Koishida (Microsoft); Somayeh Sojoudi (UC Berkeley)

1451 Sound of Vision: Audio Generation from Visual Text Embedding through Training Domain Discriminator

Jaewon Kim (Hanyang University); Won-Gook Choi (Hanyang University); Seyun Ahn (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1456 Retrieval-Augmented Classifier Guidance for Audio Generation

Ho-Young Choi (Hanyang University); Won-Gook Choi (Hanyang University); Joon-Hyuk Chang (Hanyang University)

Poster Session: Spoken Language Understanding

A11-P1-A Location: Poster Area 3A

87 This Paper Had the Smartest Reviewers - Flattery Detection Utilising an Audio-Textual Transformer-Based Approach

Lukas Christ (University of Augsburg); Shahin Amiriparian (Technical University of Munich); Friederike Hawighorst (University of Passau); Ann-Kathrin Schill (University of Passau); Angelo Boutalikakis (University of Passau); Lorenz Graf-Vlachy (TU Dortmund); Andreas König (University of Passau); Prof. Dr. Bjoern Schuller (Imperial College London)

882 Applying Reinforcement Learning and Multi-Generators for Stage Transition in an Emotional Support Dialogue System

Jeremy Chang (National Cheng Kung University); Kuan-Yu Chen (National Cheng Kung University); Chung-Hsien Wu (National Cheng Kung University)

1351 Automated Human-Readable Label Generation in Open Intent Discovery

Grant Anderson (Edinburgh Napier University); Emma Hart (Edinburgh Napier University); DIMITRA GKATZIA (Edinburgh Napier University); Ian Beaver (Verint Systems Inc)

1780 Convolutional gated MLP and attention improve end-to-end spoken language understanding

Zheng Bei da (Xinjiang University); Mijit Ablimit (Xinjiang University); Hankiz Yilahun (Xinjiang University); Askar Hamdulla (Xinjiang University)

1900 An Uyghur Extension to the MASSIVE Multi-lingual Spoken Language Understanding Corpus with Comprehensive Evaluations

Ainikaerjiang Aimaiti (Xinjiang University); Di Wu (Xinjiang University); liting Jiang (Xinjiang University); Gulinigeer Abudouwaili (Xinjiang University); Hao Huang (Xinjiang University); Wushour Slamu (xinjiang university)

1911 Towards Unified Evaluation of Continual Learning in Spoken Language Understanding

Muqiao Yang (Carnegie Mellon University); Xiang Li (Carnegie Mellon University); Umberto Cappellazzo (University of Trento); Shinji Watanabe (Carnegie Mellon University); Bhiksha Raj (Carnegie Mellon University)

2276 Unified Framework for Spoken Language Understanding and Summarization in Task-Based Human Dialog processing

Eunice Akani (Aix Marseille University, Enedis); FREDERIC BECHET (Aix Marseille University); Benoît Favre (Lab. Informatique et Systèmes / Aix-Marseille University / CNRS); Romain Gemignani (ENEDIS)

Poster Session: Spoken Dialogue Systems and Conversational Analysis 2**A11-P1-B**

Location: Poster Area 3B

67 Investigation of look ahead techniques to improve response time in spoken dialogue system

Masaya Ohagi (SB Intuitions); Yoshikawa Katsumasa (SB Intuitions); Tomoya Mizumoto (SB Intuitions)

225 Target conversation extraction: Source separation using turn-taking dynamics

Tuochao Chen (University of Washington); Qirui Wang (university of washington); Bohan Wu (University of Washington); Malek Itani (University of Washington); Emre Sefik Eskimez (Microsoft); Takuya Yoshioka (AssemblyAI Inc.); Shyamnath Gollakota (University of Washington)

1163 Detecting the terminality of speech-turn boundary for spoken interactions in French TV and Radio content

Remi Uro (Institut Nation de l'Audiovisuel); Marie Tahon (LIUM); David Doukhan (Institut National de l'Audiovisuel (INA)); Antoine Laurent (Le Mans University); Albert Rilliard (Université Paris Saclay, CNRS, LISN)

1168 Uh, um and mh: Are filled pauses prone to conversational converge?

Mathilde Hutin (FNRS, Université Catholique de Louvain); Junfei Hu (UCLouvain / FNRS); Liesbeth Degand (UCLouvain / FNRS)

1892 Utilization of Text Data for Response Timing Detection in Attentive Listening

Yu Watanabe (Graduate School of Informatics, Nagoya University); Koichiro Ito (Graduate School of Informatics, Nagoya University); Shigeki Matsubara (Information Technology Center, Nagoya University)

2381 Investigating the Influence of Stance-Taking on Conversational Timing of Task-Oriented Speech

Sara B Ng (University of Washington); Gina-Anne Levow (University of Washington); Mari Ostendorf (University of Washington); Richard Wright (University of Washington)

2523 Backchannel prediction, based on who, when and what

YO-HAN PARK (Chungnam National University); Wencke Liermann (Chungnam National University); Yong-Seok Choi (Chungnam National University); Seung H. Kim (ETRI); Jeong-Uk Bang (ETRI); Seung Yun (ETRI); Kong Joo Lee (Chungnam National Univ.)

Poster Session: Speech Synthesis: Paradigms and Methods 1

A7-P3-A

Location: Poster Area 4A

208 An Attribute Interpolation Method in Speech Synthesis by Model Merging*Masato Murata (CyberAgent, Inc.); Koichi Miyazaki (CyberAgent, Inc.); Tomoki Koriyama (CyberAgent, Inc.)***562 Learning Fine-Grained Controllability on Speech Generation via Efficient Fine-Tuning***Chung-Ming Chien (Toyota Technological Institute at Chicago); Andros Tjandra (Meta Platforms, Inc); Apoorv Vyas (Meta); Matt Le (Meta Platforms, Inc); Bowen Shi (Meta Platforms, Inc); Wei-Ning Hsu (Meta)***608 ClariTTS: Feature-ratio Normalization and Duration Stabilization for Code-mixed Multi-speaker Speech Synthesis***Changhwan Kim (Hyundai Motor Group)***1160 Highly Intelligent Speaker-Independent Articulatory Synthesis***Charles G McGhee (University of Cambridge); Katherine M Knill (University of Cambridge); Mark Gales (University of Cambridge)***1192 LiveSpeech: Low-Latency Zero-shot Text-to-Speech via Autoregressive Modeling of Audio Discrete Codes***Trung V Dang (Microsoft); David Aponte (Microsoft); Dung N Tran (Microsoft Corporation); Kazuhito Koishida (Microsoft)***1313 Multi-modal Adversarial Training for Zero-Shot Voice Cloning***John Janiczek (Zoom Video Communications); Arlo Faria (Zoom Video Communications); Yuzong Liu (Zoom Video Communications); Dading Chong (Zoom Video Communication); Bruce Dai (Zoom); George Wang (Zoom Video Communications); Tao Wang (ZOOM)***1559 Single-Codec: Single-Codebook Speech Codec towards High-Performance Speech Generation***Hanzhao Li (Northwestern Polytechnical University); Liumeng Xue (Northwestern Polytechnical University); Haohan Guo (The Chinese University of Hong Kong); Xinfu Zhu (Northwestern Polytechnical University); Yuanjun Lv (Northwestern Polytechnical University); Lei Xie (NWPU); Yunlin Chen (mobvoi); Hao Yin (mobvoi); Zhifei Li (Mobvoi)***2074 Low-dimensional Style Token Control for Hyperarticulated Speech Synthesis***Miku Nishihara (Nagoya Institute of Technology); Dan Wells (University of Edinburgh); Korin Richmond (University of Edinburgh); Aidan Pine (National Research Council Canada)***2087 Modeling Vocal Tract Like Acoustic Tubes Using the Immersed Boundary Method***Rongshuai Wu (University of British Columbia); Debasish Mohapatra (University of British Columbia); Sidney Fels (University of British Columbia)*

Poster Session: Speech Synthesis: Paradigms and Methods 2

A7-P3-B

Location: Poster Area 4B

335 Improving Robustness of LLM-based Speech Synthesis Models by Learning Monotonic Alignment*Paarth Neekhara (NVIDIA); Shehzeen S Hussain (UCSD); Subhankar Ghosh (NVIDIA); Jason Li (NVIDIA); Boris Ginsburg (NVIDIA)***359 Phonetic Enhanced Language Modeling for Text-to-Speech Synthesis***Kun Zhou (Alibaba Group); Shengkui Zhao (Alibaba Group); Yukun Ma (Alibaba Group); Chong Zhang (Alibaba Group); HAO WANG (Alibaba); Dianwen Ng (Alibaba Group / Nanyang Technological University); Chongjia Ni (Alibaba); Trung Hieu Nguyen (Alibaba Group); Jia Qi Yip (Nanyang Technological University); Bin Ma (Alibaba)*

462 FastLips: an End-to-End Audiovisual Text-to-Speech System with Lip Features Prediction for Virtual Avatars

Martin Lenglet (Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab); Olivier Perrotin (Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab); gerard bailly (GIPSA-Lab/CNRS)

465 High Fidelity Text-to-Speech Via Discrete Tokens Using Token Transducer and Group Masked Language Model

Joun Yeop Lee (Samsung Research); Myeonghun Jeong (Seoul National University); Minchan Kim (Seoul National University); Ji-Hyun Lee (Samsung Research); Hoon-Young Cho (Samsung Research); Nam Soo Kim (Seoul National University)

508 Small-E: Small Language Model with Linear Attention for Efficient Speech Synthesis

Théodor Lemerle (IRCAM); Nicolas Obin (STMS (Ircam, CNRS, Sorbonne Université)); Axel Roebel (Ircam)

1913 Synthesizing Long-Form Speech merely from Sentence-Level Corpus with Content Extrapolation and LLM Contextual Enrichment

Shijie Lai (School of Information Science and Engineering, Xinjiang University); Minglu He (Xinjiang University); Zijing Zhao (Xinjiang University); Kai Wang (Xinjiang University); Hao Huang (Xinjiang University); Jichen Yang (GPNU)

2320 FluentEditor: Text-based Speech Editing by Considering Acoustic and Prosody Consistency

Rui Liu (Inner Mongolia University); Jiatian Xi (Inner Mongolia University); Ziyue Jiang (Zhejiang University); Haizhou Li (The Chinese University of Hong Kong, Shenzhen)

Poster Session: Computational Models of Human Language Acquisition, Perception, and Production (Special Session)

SS-3 Location: Yanis Club

394 Dirichlet process mixture model based on topologically augmented signal representation for clustering infant vocalizations

Guillem Bonafofos (Université Jean Monnet); Clara Bourot (Université Aix-Marseille); Pierre Pudlo (Université Aix-Marseille); Jean-Marc Freyermuth (Université Aix-Marseille); Laurence Reboul (Université Aix-Marseille); Samuel Tronçon (Résurgences R & D); Arnaud Rey (Université Aix-Marseille)

822 A data-driven model of acoustic speech intelligibility for optimization-based models of speech production

Benjamin Elie (University of Edinburgh); Juraj Simko (University of Helsinki); Alice Turk (The University of Edinburgh)

1041 Neurocomputational model of speech recognition for pathological speech detection: a case study on Parkinson's disease speech detection

Sevada Hovsepyan (Idiap Research Institute); Mathew Magimai.-Doss (Idiap Research Institute)

1051 Spoken-Term Discovery using Discrete Speech Units

Benjamin van Niekerk (Stellenbosch University); Julian Zaïdi (Ubisoft); Marc-André Carbonneau (Ubisoft); Herman Kamper (Stellenbosch University)

1054 Orthogonality and isotropy of speaker and phonetic information in self-supervised speech representations

Mukhtar Mohamed (The University Of Edinburgh); Oli D Liu (University of Edinburgh); Hao Tang (The University of Edinburgh); Sharon Goldwater (University of Edinburgh)

1473 A Pilot Study of GSLM-based Simulation of Foreign Accentuation Only Using Native Speech Corpora

Kentaro Onda (The University of Tokyo); Joonyong Park (The University of Tokyo); Nobuaki Minematsu (The University of Tokyo); Daisuke Saito (The University of Tokyo)

2102 The Difficulty and Importance of Estimating the Lower and Upper Bounds of Infant Speech Exposure
Joseph R Coffey (ECOLE NORMALE SUPERIEURE); Okko Räsänen (Tampere University); Camila Scaff (University of Zurich, Institute of Evolutionary Medicine (IEM)); Alejandrina Cristia (Exelang, CNRS, LSCP)

2192 Simulating articulatory trajectories with phonological feature interpolation
Angelo Gerardo Ortiz Tandazo (Ecole normale supérieure, Grenoble Alpes Univ.); Thomas Schatz (Aix-Marseille Univ., CNRS, LIS); Thomas Hueber (CNRS/Grenoble Alpes Univ.); Emmanuel Dupoux (EHESS, ENS, PSL University, CNRS, INRIA, META)

2340 Information-theoretic hypothesis generation of relative cue weighting for the voicing contrast
Annika L Heuser (University of Pennsylvania); Jianjing Kuang (University of Pennsylvania)

Time Slot 3

Oral Session: Speech and Audio Analysis

A5-07

Location: Aceso

709 Motion Based Audio-Visual Segmentation

Jiahao Li (Beijing Institute of Technology); Miao Liu (Beijing Institute of Technology); Shu Yang (Tsinghua University); Jing Wang (Beijing Institute of Technology); Xiang Xie (Beijing Institute of Technology)

825 Predefined Prototypes for Intra-Class Separation and Disentanglement

Antonio Almudévar (University of Zaragoza); Théo Mariotte (LTCl, Télécom Paris, Institut Polytechnique de Paris); Alfonso Ortega (Universidad de Zaragoza); Marie Tahon (LIUM); Luis Vicente (University of Zaragoza); Antonio Miguel (University of Zaragoza); Eduardo Lleida Solano (University of Zaragoza)

1019 A Transformer-Based Voice Activity Detector

BISWAJIT KARAN (Stellenbosch University); Joshua Miles Jansen Van Vuren (Stellenbosch University); Febe de Wet (Stellenbosch University); Thomas Niesler (University of Stellenbosch)

1127 VAE-based Phoneme Alignment Using Gradient Annealing and SSL Acoustic Features

Tomoki Koriyama (CyberAgent, Inc.)

1229 XANE: eXplainable Acoustic Neural Embeddings

Sri Harsha Dumpala (Dalhousie University/Vector Institute); Dushyant Sharma (Microsoft Inc); Chandramouli Shama Sastry (Dalhousie University/Vector Institute); Stanislav Kruchinin (Microsoft Inc); James Fosburgh (Microsoft Inc); Patrick A. Naylor (Imperial College London)

2404 A comparative analysis of sequential models that integrate syllable dependency for automatic syllable stress detection

Jhansi Mallela (Internantional Institute of Information Technology, Hyderabad); Sai Harshitha Aluru (Vidya Jyothi Institute Of Technology); Chiranjeevi Yarra (International Institute of Information Technology Hyderabad)

Oral Session: Speech Quality and Intelligibility: Prediction and Enhancement

A6-05

Location: Aegle A

691 Exploring Sentence Type Effects on the Lombard Effect and Intelligibility Enhancement: A Comparative Study of Natural and Grid Sentences

Hongyang Chen (Wuhan university); Yuhong Yang (Wuhan University); Zhongyuan Wang (Wuhan University); Weiping Tu (Wuhan University); Haojun Ai (Wuhan University); Cedar Lin (OPPO)

716 Non-Intrusive Speech Intelligibility Prediction for Hearing Aids using Whisper and Metadata

Ryandhimas E Zezario (Academia Sinica); Fei Chen (Southern University of Science and Technology); Chiou-Shann Fuh (National Taiwan University); Hsin-Min Wang (Academia Sinica); Yu Tsao (Academia Sinica)

1421 No-Reference Speech Intelligibility Prediction Leveraging a Noisy-Speech ASR Pre-Trained Model

Haolan Wang (Queen's University); Amin Edraki (Queen's University); Wai-Yip Geoffrey Chan (Queen's University); Iván López-Espejo (University of Granada); Jesper Jensen (Demant A/S)

2051 The PESQetarian: On the Relevance of Goodhart's Law for Speech Enhancement

Danilo Oliveira (Universität Hamburg); Julius Richter (Universität Hamburg); Simon Welker (Universität Hamburg); Timo Gerkmann (Universität Hamburg)

2258 Transfer Learning from Whisper for Microscopic Intelligibility Prediction

Paul Best (LIS); Santiago Cuervo (LIS); Ricard Marxer (Université de Toulon, Aix Marseille Univ, CNRS, LIS, Toulon)

2538 Enhancing Non-Matching Reference Speech Quality Assessment through Dynamic Weight Adaptation

Ta Bao Thang (Viettel Cyberspace Center); Van Hai Do (Thuyloi University); Huynh Thi Thanh Binh (Hanoi University of Science and Technology)

Oral Session: Speech Assessment**A10-03**

Location: Aegle B

552 Quantifying the Role of Textual Predictability in Automatic Speech Recognition

Sean D Robertson (University of Toronto Computer Science Department); Gerald Penn (University of Toronto); Ewan Dunbar (University of Toronto)

760 Oversampling, Augmentation and Curriculum Learning for Speaking Assessment with Limited Training Data

Tin Mei Lun (Aalto University); Ekaterina Voskoboini (Aalto University); Ragheb Al-Ghezi (Aalto University); Tamas Grosz (Aalto University); Mikko Kurimo (Aalto University)

1078 Optimizing Automatic Speech Assessment: W-RankSim Regularization and Hybrid Feature Fusion Strategies

Chung-Wen Wu (National Taiwan Normal University); Berlin Chen (National Taiwan Normal University)

1142 Context-Aware Speech Recognition Using Prompts for Language Learners

Jian Cheng (Google)

1373 Analysis and Visualization of Directional Diversity in Listening Fluency of World Englishes Speakers in the Framework of Mutual Shadowing

Yu Tomita (The University of Tokyo); Yingxiang Gao (The University of Tokyo); Nobuaki Minematsu (The University of Tokyo); Noriko Nakanishi (Kobe Gakuin University); Daisuke Saito (The University of Tokyo)

1639 A Dataset and Two-Pass System for Reading Miscue Detection

Raj Gothi (Indian Institute of Technology Bombay); Rahul Kumar (Indian Institute of Technology Bombay); Mildred Pereira (Indian Institute of Technology, Bombay); Nagesh Satish Nayak (Indian Institute of Technology Bombay); Preeti Rao (Indian Institute of Technology Bombay)

Oral Session: ASR Model Training Methods

A8-06

Location: Hippocrates

830 Investigating the Effect of Label Topology and Training Criterion on ASR Performance and Alignment Quality

Tina Raissi (RWTH Aachen University); Christoph M. Lüscher (Informatik 6, RWTH Aachen University); Simon Berger (RWTH Aachen University); Ralf Schlüter (RWTH Aachen University); Hermann Ney (RWTH Aachen University)

849 Learnable Layer Selection and Model Fusion for Speech Self-Supervised Learning Models

Sheng-Chieh Chiu (National Tsing Hua University); Chia-Hua Wu (Academia Sinica); Jih-Kang Hsieh (National Tsing Hua University); Yu Tsao (Academia Sinica); Hsin-Min Wang (Academia Sinica)

1257 Contextualized End-to-end Automatic Speech Recognition with Intermediate Biasing Loss

Muhammad Dr. Shakeel (Honda Research Institute Japan); Yui Sudo (Honda Research Institute Japan); Yifan Peng (Carnegie Mellon University); Shinji Watanabe (Carnegie Mellon University)

1898 Cross-Modality Diffusion Modeling and Sampling for Speech Recognition

Chia-Kai Yeh (National Yang Ming Chiao Tung University); Chih-Chun Chen (National Yang Ming Chiao Tung University); Ching-Hsien Hsu (National Yang Ming Chiao Tung University); Jen-Tzung Chien (National Yang Ming Chiao Tung University)

1924 Sampling free text injection

Neeraj Gaur (Google); Rohan Agrawal (Google); Yuan Wang (Google); Parisa Haghani (Google); Andrew Rosenberg (Google LLC); Bhuvana Ramabhadran (Google)

2027 Sequential Editing for Lifelong Training of Speech Recognition Models

Devang Kulshreshtha (Amazon); Nikolaos Pappas (University of Washington); Brady Houston (AWS AI Labs); Saket Dingliwal (Amazon); Srikanth Ronanki (Amazon)

Oral Session: New Avenues in Emotion Recognition

A3-04

Location: Iasso

98 Enrolment-based personalisation for improving individual-level fairness in speech emotion recognition

Andreas Triantafyllopoulos (Technical University of Munich); Prof. Dr. Bjoern Schuller (Imperial College London)

482 Can Modelling Inter-Rater Ambiguity Lead To Noise-Robust Continuous Emotion Predictions?

Ya-Tse Wu (Department of Electrical Engineering, National Tsing Hua University); Jingyao Wu (University of New South Wales); Vidhyasaharan Sethu (University of New South Wales); Chi-Chun Lee (National Tsing Hua University)

1218 Keep, Delete, or Substitute: Frame Selection Strategy for Noise-Robust Speech Emotion Recognition

Seong-Gyun Leem (University of Texas at Dallas); Daniel Fulford (Boston University); JP Onnela (T.H. Chan School of Public Health Harvard University); David Gard (San Francisco State University); Carlos Busso (University of Texas at Dallas)

1735 MFDR: Multiple-stage Fusion and Dynamically Refined Network for multimodal emotion recognition

Ziping Zhao (Tianjin Normal University); Tian Gao (Tianjin Normal University); Haishuai Wang (Zhejiang University); Prof. Dr. Bjoern Schuller (Imperial College London)

1948 Hierarchical Distribution Adaptation for Unsupervised Cross-corpus Speech Emotion Recognition

Cheng Lu (Southeast University); Yuan Zong (Southeast University); Yan Zhao (Southeast University); Hailun Lian (Southeast University); Tianhua Qi (Southeast University); Bjoern W. Schuller (Imperial College London); Wenming Zheng (Southeast University)

2350 Multimodal Fusion of Music Theory-Inspired and Self-Supervised Representations for Improved Emotion Recognition

Xiaohan Shi (Nagoya University); Xingfeng LI (Hainan University); Tomoki Toda (Nagoya University)

Oral Session: Speech Synthesis: Vocoders

A7-06 Location: Melampus

255 BiVocoder: A Bidirectional Neural Vocoder Integrating Feature Extraction and Waveform Generation
Hui-Peng Du (University of Science and Technology of China); Ye-Xin Lu (University of Science and Technology of China); Yang Ai (University of Science and Technology of China); Zhen-Hua Ling (University of Science and Technology of China)

380 FA-GAN: Artifacts-free and Phase-aware High-fidelity GAN-based Vocoder
Rubing Shen (Wuhan University); Yanzhen Ren (Computer School of Wuhan University); Zongkun Sun (Wuhan University)

1447 JenGAN: Stacked Shifted Filters in GAN-Based Speech Synthesis
Hyunjae Cho (Seoul National University); Junhyeok Lee (Supertone Inc.); Wonbin Jung (Korea Advanced Institute of Science and Technology)

2014 QGAN: Low Footprint Quaternion Neural Vocoder for Speech Synthesis
Aryan Chaudhary (IIIT Delhi); Vinayak Abrol (Indraprastha Institute of Technology Delhi)

2371 QHM-GAN: Neural Vocoder based on Quasi-Harmonic Modeling
Shaowen Chen (Nagoya University); Tomoki Toda (Nagoya University)

2407 FreeV: Free Lunch For Vocoders Through Pseudo Inversed Mel Filter
Yuanjun Lv (Northwestern Polytechnical University); Hai Li (iQIYI Inc); Ying Yan (Iqiyi); Junhui Liu (iQIYI Inc); Danming Xie (iQIYI); Lei Xie (NWPU)

Oral Session: Dysarthric Speech Assessment

A13-03 Location: Panacea Amphitheater

1144 Electroglottography for the assessment of dysphonia in Parkinson's disease and multiple system atrophy
Khalid Daoudi (INRIA); Solange Milhé de Saint Victor (University Hospital of Bordeaux); Alexandra Foubert-Samier (University Hospital of Bordeaux); Margherita Fabbri (University Hospital of Toulouse); Anne Pavy-Le Traon (University Hospital of Toulouse); Olivier Rascol (University Hospital of Toulouse); Virginie Woisard (Hospitals of Toulouse); Wassilios Meissner (University Hospital of Bordeaux)

1175 Exploring Syllable Discriminability during Diadochokinetic Task with Increasing Dysarthria Severity for Patients with Amyotrophic Lateral Sclerosis
Neelesh Samptur (PES University, Bengaluru); Tanuka Bhattacharjee (Indian Institute of Science); Anirudh Chakravarty K (PES University, Bengaluru); Seena Vengalil (National Institute of Mental Health and Neurosciences); Yamini BK (NIMHANS); Nalini Atchayaram (NIMHANS); Prasanta Dr Ghosh (Indian Institute of Science (IISc), Bangalore)

1281 Beyond Binary: Multiclass Paraphasia Detection with Generative Pretrained Transformers and End-to-End Models

Matthew Perez (University of Michigan); Aneesha Sampath (University of Michigan); Minxue Niu (University of Michigan); Emily K Mower Provost (University of Michigan)

1597 CDSD: Chinese Dysarthria Speech Database

Yan Wang (CAS Key Laboratory of Behavioral Science, Institute of Psychology); Mengyi Sun (CAS Key Laboratory of Behavioral Science, Institute of Psychology); Xinchun Kang (Beijing Key Laboratory of Information Service Engineering, Beijing Union University); Jingting Li (Chinese Academy of Sciences); Pengfei Guo (Jiangsu University of Science and Technology); Ming Gao (University of Science and Technology of China); Su-Jing Wang (Chinese Academy of Sciences)

1852 CoLM-DSR: Leveraging Neural Codec Language Modeling for Multi-Modal Dysarthric Speech Reconstruction

Xueyuan Chen (The Chinese University of Hong Kong); Dongchao Yang (The Chinese University of Hong Kong); Dingdong Wang (The Chinese University of Hong Kong); Xixin Wu (The Chinese University of Hong Kong); Zhiyong Wu (Tsinghua University); Helen Meng (The Chinese University of Hong Kong)

1939 Automatic Assessment of Dysarthria using Speech and synthetically generated Electroglottograph signal

Fathima Zaheera (National Institute of Technology Patna); Supriya M Shetty (Indian Institute of Information Technology, Dharwad); Gayadhar Pradhan (NIT Patna); Deepak T (IIIT-Dharwad)

Poster Session: Speaker Diarization 2**A4-P4-A**

Location: Poster Area 1A

209 DiarizationLM: Speaker Diarization Post-Processing with Large Language Models

Quan Wang (Google); Yiling Huang (Google); Guanlong Zhao (Google); Evan Clark (Google); Wei Xia (Google); Hank Liao (Google)

399 Variable Segment Length and Domain-Adapted Feature Optimization for Speaker Diarization

Chenyuan Zhang (Department of Automation, Xiamen University); Linkai Luo (Department of Automation, Xiamen University); Hong Peng (Department of Automation, Xiamen University); Wei Wen (Xiamen University)

962 Specializing Self-Supervised Speech Representations for Speaker Segmentation

Séverin BAROUDI (LIS); Thomas Pellegrini (IRIT); Hervé Bredin (CNRS)

1060 On the calibration of powerset speaker diarization models

Alexis Plaquet (IRIT); Hervé Bredin (CNRS)

1874 Efficient Speaker Embedding Extraction Using a Twofold Sliding Window Algorithm for Speaker Diarization

Jeong-Hwan Choi (Hanyang University); Ye-Rin Jeoung (Hanyang University); Ilseok Kim (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1987 Joint vs Sequential Speaker-Role Detection and Automatic Speech Recognition for Air-traffic Control

Alexander Blatt (Saarland University); Aravind Krishnan (Saarland University); Dietrich Klakow (Saarland University)

Poster Session: Speaker Recognition 2**A4-P4-B**

Location: Poster Area 1B

161 On the Usefulness of Speaker Embeddings for Speaker Retrieval in the Wild: A Comparative Study of x-vector and ECAPA-TDNN Models

Erfan Loweimi (University of Cambridge); Mengjie Qian (Cambridge University); Katherine M Knill (University of Cambridge); Mark Gales (University of Cambridge)

354 W-GVKT: Within-Global-View Knowledge Transfer for Speaker Verification

Jin Zezhong (The Hong Kong Polytechnic University); Youzhi TU (The Hong Kong Polytechnic University); Man-Wai MAK (The Hong Kong Polytechnic University)

364 CEC: A Noisy Label Detection Method for Speaker Recognition

Yao Shen (China Mobile Research Institute); Yingying Gao (China Mobile Research Institute); yaqian hao (China Mobile Research Institute); Chenguang Hu (chinamobile); Fulin Zhang (China Mobile Research Institute); Junlan Feng (China Mobile Research Institute); Shilei Zhang (China Mobile Research Institute)

434 Disentangling Age and Identity with Mutual Information Minimization for Cross-Age Speaker Verification

Fengrun Zhang (Kuaishou Technology Company); Wangjin Zhou (Kyoto University); Yiming Liu (Kuaishou Technology, Beijing, China); wang geng (北京达佳互联信息技术有限公司); Yahui Shan (北京达佳互联信息技术有限公司); Chen Zhang (北京达佳互联信息技术有限公司)

973 Contrastive Learning and Inter-Speaker Distribution Alignment Based Unsupervised Domain Adaptation for Robust Speaker Verification

Zuoliang Li (University of Science and Technology of China); Wu Guo (University of Science and Technology of China); Bin Gu (中国科学技术大学); Shengyu Peng (University of Science and Technology of China); Jie Zhang (University of Science and Technology of China (USTC))

1280 Identifying Speakers in Dialogue Transcripts: A Text-based Approach Using Pretrained Language Models

Van Minh Nguyen (University of Oregon); Franck Dernoncourt (MIT); Seunghyun Yoon (Adobe Research); Hanieh Deilamsalehy (Adobe Research); Hao Tan (Adobe Research); Ryan A. Rossi (Adobe Research); Quan Hung Tran (Adobe Research); Trung Bui (Adobe Research); Thien Huu Nguyen (University of Oregon)

Poster Session: Cross-Lingual and Multilingual Processing

A9-P2 Location: Poster Area 2A, Poster Area 2B

259 SC-MoE: Switch Conformer Mixture of Experts for Unified Streaming and Non-streaming Code-Switching ASR

Shuaishuai Ye (HiThink RoyalFlush AI Research Institute); Shunfei Chen (HiThink RoyalFlush AI Research Institute); Xinhui Hu (HiThink RoyalFlush AI Research Institute); Xinkang Xu (HiThink RoyalFlush AI Research Institute)

892 LoRA-Whisper: Parameter-Efficient and Extensible Multilingual ASR

Zheshu Song (Shanghai Jiao Tong University); Jianheng Zhuo (Shanghai Jiaotong university); Yifan Yang (Shanghai Jiao Tong University); Ziyang Ma (Shanghai Jiao Tong University); Shi-Xiong Zhang (Capital One); Xie Chen (Shanghai Jiaotong University)

907 Whispering in Norwegian: Navigating Orthographic and Dialectic Challenges

Per E Kummervold (National Library of Norway); Javier De la Rosa (National Library of Norway); Freddy Wetjen (National Library of Norway); Rolv-Arild Braaten (National Library of Norway); Per Erik Solberg (National Library of Norway)

938 mHuBERT-147: A Compact Multilingual HuBERT Model

Marcely Zanon Boito (NAVER LABS Europe); Vivek Iyer (The University of Edinburgh); Nikolaos Lagos (Naver Labs)

Europe); laurent besacier (Naver Labs Europe); Ioan Calapodescu (Naver Labs Europe)

1418 Enhancing Neural Transducer for Multilingual ASR with Synchronized Language Diarization

Amir Hussein (Johns Hopkins University); Desh Raj (Johns Hopkins University); Matthew S Wiesner (Johns Hopkins University); Daniel Povey (Xiaomi, Inc.); Paola Garcia (Johns Hopkins University); Sanjeev Khudanpur (Johns Hopkins University)

1714 LUPET: Incorporating Hierarchical Information Path into Multilingual ASR

Wei Liu (The Chinese University of Hong Kong); Jingyong Hou (Tencent); Dong Yang (Tencent); Muyong Cao (Tencent); Tan Lee (The Chinese University of Hong Kong)

1745 A Parameter-efficient Language Extension Framework for Multilingual ASR

Wei Liu (The Chinese University of Hong Kong); Jingyong Hou (Tencent); Dong Yang (Tencent); Muyong Cao (Tencent); Tan Lee (The Chinese University of Hong Kong)

1760 Integrating Speech Self-Supervised Learning Models and Large Language Models for ASR

Ling Dong (Kunming University of Science and Technology); Zhengtao Yu (Kunming University of Science and Technology); Wenjun Wang (Kunming University of Science and Technology); Yuxin Huang (Kunming University of science and technology); Shengxiang Gao (Kunming University of science and technology); Guojiang Zhou (Kunming University of science and technology)

1830 The Greek podcast corpus: Competitive speech models for low-resourced languages with weakly supervised data

Georgios Paraskevopoulos (Athena Research Center); Chara Tsoukala (Athena Research Center); Athanasios Katsamanis ("ATHENA R.C., Behavioral Signal Technologies"); Vassilis Katsouros (Athena Research Center)

1938 On the Effects of Heterogeneous Data Sources on Speech-to-Text Foundation Models

Jinchuan Tian (Carnegie Mellon University); Yifan Peng (Carnegie Mellon University); William Chen (Carnegie Mellon University); Kwanghee Choi (Carnegie Mellon University); Karen Livescu (TTI-Chicago); Shinji Watanabe (Carnegie Mellon University)

2043 A Unified Approach to Multilingual Automatic Speech Recognition with Improved Language Identification for Indic Languages

Nikhil Jakhar (Samsung Research Institute); Sudhanshu Srivastava (SRIB); Arun Baby (Samsung Research, Bangalore)

2199 EFFUSE: Efficient Self-Supervised Feature Fusion for E2E ASR in Low Resource and Multilingual Scenarios

Tejes Srivastava (University of Chicago); Jiatong Shi (Carnegie Mellon University); William Chen (Carnegie Mellon University); Shinji Watanabe (Carnegie Mellon University)

2294 Less is More: Accurate Speech Recognition & Translation without Web-Scale Data

Krishna C Puvvada (NVIDIA); Piotr Żelasko (NVIDIA); He Huang (Nvidia); Oleksii Hrinchuk (NVIDIA); Nithin Rao Koluguri (NVIDIA); Kunal Dhawan (NVIDIA); Somshubra Majumdar (NVIDIA); Elena Rastorgueva (NVIDIA Corporation); Zhehuai Chen (NVIDIA); Vitaly Lavrukhin (NVIDIA); Jagadeesh Balam (NVIDIA); Boris Ginsburg (NVIDIA)

2443 Speech Recognition for Greek Dialects: A Challenging Benchmark

Socrates Vakirtzian (Athena Research Center); Chara Tsoukala (Athena Research Center); Stavros Bompolas (ARCHIMEDES Unit | Athena Research Center); Katerina Mouzou (Athena Research Center); Vivian Stamou (ILSP/Athena R.C.); Georgios Paraskevopoulos (Athena Research Center); Antonios Dimakis (National and Kapodistrian University of Athens); STELLA MARKANTONATOU (ATHENA RC); Angela Ralli (University of Patras); Antonios Anastasopoulos (George Mason University)

2520 All Ears: Building Self-Supervised Learning based ASR models for Indian Languages at scale

Vasista Sai Lodagala (Kanari AI); Abhishek Biswas (Speech Lab, IIT Madras); Shoutrik Das (Indian Institute of Technology, Madras); Jordan F (IIT Madras); S Umesh (IIT Chennai)

Poster Session: Question Answering from Speech and Spoken Dialogue Systems

A11-P2-A

Location: Poster Area 3A

- 745 Cross-Modal Denoising: A Novel Training Paradigm for Enhancing Speech-Image Retrieval**
Lifeng Zhou (NetEase Yidun AI Lab); Yuke Li (NetEase Yidun AI Lab); Rui Deng (NetEase Yidun AI Lab); Yuting Yang (NetEase Yidun AI Lab); Haoqi Zhu (NetEase Yidun AI Lab)
- 839 On the Use of Plausible Arguments in Explainable Conversational AI**
Martina Di Bratto (University of Naples Federico II); Maria Di Maro (University of Naples Federico II); Antonio Origlia (University of Naples Federico II)
- 1036 TM-PATHVQA: 90000+ Textless Multilingual Questions for Medical Visual Question Answering**
Tonmoy Rajkhowa (Indian Institute of Technology, Dharwad); Amartya Roy Chowdhury (IIT Dharwad); Sankalp Nagaonkar (IIT Dharwad); Achyut Mani Tripathi (Indian Institute of Technology Dharwad, Karnataka); Mahadeva Prasanna (IIT Dharwad)
- 1406 Rapport-Driven Virtual Agent: Rapport Building Dialogue Strategy for Improving User Experience at First Meeting**
Muhammad Yeza Baihaqi (Nara Institute of Science and Technology; Guardian Robot Project, RIKEN); Angel F Garcia Contreras (RIKEN); Seiya Kawano (RIKEN); Koichiro Yoshino (RIKEN)
- 1575 Instruction Data Generation and Unsupervised Adaptation for Speech Language Models**
Vahid Noroozi (NVIDIA); Zhehuai Chen (NVIDIA); Somshubra Majumdar (NVIDIA); He Huang (Nvidia); Jagadeesh Balam (NVIDIA); Boris Ginsburg (NVIDIA)
- 1993 Towards Multilingual Audio-Visual Question Answering**
Orchid Chetia Phukan (IIIT Delhi); Priyabrata Mallick (Reliance Jio AICoE); Swarup Ranjan Behera (Reliance Jio AICoE); Aalekhya Satya Narayani (Reliance Jio AICoE); Arun Balaji Buduru (IIIT Delhi); Rajesh Sharma (University of Tartu, Estonia)
- 2147 Reinforcement Learning from Answer Reranking Feedback for Retrieval-Augmented Answer Generation**
Van Minh Nguyen (University of Oregon); Quoc Toan Nguyen (Zoom Video Communications); Kishan KC (Amazon); Zeyu Zhang (University of Arizona); Thuy Vu (Amazon)

Poster Session: Spoken Dialogue Systems and Conversational Analysis 3

A11-P2-B

Location: Poster Area 3B

- 538 MM-NodeFormer: Node Transformer Multimodal Fusion for Emotion Recognition in Conversation**
Zilong Huang (The Hong Kong Polytechnic University); Man-Wai MAK (The Hong Kong Polytechnic University); Kong Aik Lee (The Hong Kong Polytechnic University)
- 635 Evaluating Speech Recognition Performance Towards Large Language Model Based Voice Assistants**
Zhe Liu (Meta); Suyoun Kim (Meta); Ozlem Kalinli (Meta)
- 819 Non-Linear Inference Time Intervention: Improving LLM Truthfulness**
Jan Chojnacki (Samsung Research Poland); Jakub Hoscilowicz (Samsung R & D); Adam Wiacek (Samsung Research Poland); Leszek Michon (Samsung Research Poland); Adam Cieslak (Samsung Research Poland); Artur Janicki (Warsaw University of Technology)
- 1197 Learning from Multiple Annotator Biased Labels in Multimodal Conversation**
Kazutoshi Shinoda (NTT Corporation); Nobukatsu Hojo (NTT); Saki Mizuno (NTT Corporation); Keita Suzuki (Nippon

Telegraph and Telephone Corporation); Satoshi Kobashikawa (NTT); Ryo Masumura (NTT Corporation)

1329 **Participant-Pair-Wise Bottleneck Transformer for Engagement Estimation from Video Conversation**
Keita Suzuki (Nippon Telegraph and Telephone Corporation); Nobukatsu Hojo (Nippon Telegraph and Telephone Corporation); Kazutoshi Shinoda (NTT Corporation); Saki Mizuno (NTT Corporation); Ryo Masumura (NTT Corporation)

1688 **Emotional Cues Extraction and Fusion for Multi-modal Emotion Prediction and Recognition in Conversation**

Haoxiang Shi (University of Science and Technology of China); Ziqi Liang (University of Science and Technology of China); Jun Yu (University of Science and Technology of China)

2122 **""Well"", what can you do with messy data? Exploring the prosody and pragmatic function of the discourse marker ""well"" with found data and speech synthesis"**

Johannah O'Mahony (University of Edinburgh); Catherine Lai (University of Edinburgh); Eva Szekely (KTH Royal Institute of Technology)

Poster Session: Phonetics, Phonology and Prosody

A2-P2-A

Location: Poster Area 4A

310 **Do Speaker-dependent Vowel Characteristics depend on Speech Style?**

Nicolas Audibert (Laboratoire de Phonétique et Phonologie); Cecile FOUGERON (Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne-Nouvelle); Christine Meunier (Aix Marseille University)

317 **Evaluating Italian Vowel Variation with the Recurrent Neural Network Phonet**

Austin Jones (University of Georgia); Margaret E L Renwick (University of Georgia)

833 **Prosodic marking of syntactic boundaries in Khoekhoe**

Kira Tulchynska (Hebrew University of Jerusalem); Sylvanus Job (University of Namibia); Alena Witzlack-Makarevich (Hebrew University of Jerusalem); Margaret Zellers (Kiel University)

1704 **Speaker Detection by the Individual Listener and the Crowd: Parametric Models Applicable to Bonafide and Deepfake Speech**

Tomi H. Kinnunen (University of Eastern Finland); Rosa E Gonzalez Hautamäki (University of Oulu); Xin Wang (National Institute of Informatics); Junichi Yamagishi (National Institute of Informatics)

2060 **Preservation, conservation and phonetic study of the voices of Italian poets: A study on the seven years of the VIP archive**

Federico Lo Iacono (Università di Torino); Valentina Colonna (University of Granada); Antonio Romano (Università di Torino)

2082 **NumberLie: a game-based experiment to understand the acoustics of deception and truthfulness**

Alessandro De Luca (University of Zurich); Andrew Clark (LiRI, University of Zurich); Volker Dellwo (University of Zurich)

2331 **A comparison of voice similarity through acoustics, human perception and deep neural network (DNN) speaker verification systems**

Suyuan Liu (University of British Columbia); Molly Babel (University of British Columbia); Jian Zhu (University of British Columbia)

Poster Session: Segmentals

A2-P2-B

Location: Poster Area 4B

1607 Affricates in Lushootseed*Ted Kye (University of Washington)***1674 Nasal Air Flow During Speech Production In Korebaju***Jenifer A Vega Rodriguez (Gipsa-lab); Nathalie Vallée (GIPSA-lab CNRS / Université Grenoble Alpes / Grenoble-INP); Christophe Savariaux (GIPSA-lab); Silvain Gerber (CNRS)***1841 Intrusive schwa within French stop-liquid clusters: An acoustic analysis***MINMIN YANG (Laboratoire de Phonétique et Phonologie (CNRS & Sorbonne Nouvelle)); Rachid Ridouane (LPP (CNRS & Sorbonne Nouvelle))***2090 Crosslinguistic Comparison of Acoustic Variation in the Vowel Sequences /ia/ and /io/ in Four Romance Languages***Johanna Cronenberg (Université Paris Cité); Ioana Chitoran (Université Paris Cité); Lori Lamel (CNRS LISN); Ioana Vasilescu (LIMS)***2361 Voiced and voiceless laterals in Angami***Viyazonuo Terhija (Indian Institute of Technology Guwahati); Priyankoo Sarmah (Indian Institute of Technology Guwahati)*

Poster Session: Spoken Language Models for Universal Speech Processing (Special Session)

SS-6

Location: Yanis Club

288 NAST: Noise Aware Speech Tokenization for Speech Language Models*Shoval Messica (Hebrew University of Jerusalem (Huji)); Yossi Adi (The Hebrew University of Jerusalem)***457 DeSTA: Enhancing Speech Language Models through Descriptive Speech-Text Alignment***Ke-Han Lu (National Taiwan University); Zhehuai Chen (NVIDIA); Szu-Wei Fu (NVIDIA); He Huang (Nvidia); Boris Ginsburg (NVIDIA); Yu-Chiang Frank Wang (NVIDIA); Hung-yi Lee (National Taiwan University)***1076 Understanding Sounds, Missing the Questions: The Challenge of Object Hallucination in Large Audio-Language Models***Chun-Yi Kuan (National Taiwan University); Wei-Ping Huang (National Taiwan University); Hung-yi Lee (National Taiwan University)***1306 DiscreteSLU: A Large Language Model with Self-Supervised Discrete Speech Units for Spoken Language Understanding***Suwon Shon (ASAPP); Kwangyoung Kim (ASAPP); Yi-Te Hsu (ASAPP); Prashant Sridhar (ASAPP); Shinji Watanabe (Carnegie Mellon University); Karen Livescu (TTI-Chicago)***1346 COSMIC: Data Efficient Instruction-tuning For Speech In-Context Learning***Jing Pan (Microsoft); Jian Wu (Microsoft); Yashesh Gaur (Microsoft.com); Sunit Sivasankaran (Microsoft); Zhuo Chen (Microsoft); Shujie Liu (Microsoft Research Asia); Jinyu Li (Microsoft)***1932 Exploring In-Context Learning of Textless Speech Language Model for Speech Classification Tasks***Kai-Wei Chang (National Taiwan University); Ming-Hao Hsu (National Taiwan University); Shang-Wen Li (FAIR); Hung-yi Lee (National Taiwan University)***2366 Low Bitrate High-Quality RVQGAN-based Discrete Speech Tokenizer**

Slava Shechtman (IBM Research); Avihu Dekel (IBM Research)

2375 On the Effectiveness of Acoustic BPE in Decoder-only TTS

Bohan Li (MoE Key Lab of Artificial Intelligence, AI Institute, X-LANCE Lab Department of Computer Science and Engineering, Shanghai Jiao Tong University); Feiyu Shen (Shanghai Jiao Tong University); Yiwei Guo (Shanghai Jiao Tong University); Shuai Wang (Shenzhen Research Institute of Big Data, Chinese University of Hong Kong (Shenzhen)); Xie Chen (Shanghai Jiaotong University); Kai Yu (Shanghai Jiao Tong University)

2419 Can Large Language Models Understand Spatial Audio?

Changli Tang (Tsinghua University); Wenyi Yu (Tsinghua); Guangzhi Sun (University of Cambridge Department of Engineering); Xianzhao Chen (Bytedance); Tian Tan (Bytedance); Wei Li (Bytedance); Jun Zhang (Bytedance); Lu Lu (Bytedance); Zejun Ma (Bytedance); Yuxuan Wang (ByteDance AI Lab); Chao Zhang (Tsinghua University)

Thursday 05/09

Time Slot 1

Oral Session: Leveraging Large Language Models and Contextual Features for Phonetic Analysis (Special Session)

SS-4

Location: Acesso

665 Exploring Pre-trained Speech Model for Articulatory Feature Extraction in Dysarthric Speech Using ASR

Yuqin Lin (Tianjin University); Longbiao Wang (Tianjin University); Jianwu Dang (Tianjin University); Nobuaki Mine-matsu (The University of Tokyo)

1039 Are Articulatory Feature Overlaps Shrouded in Speech Embeddings?

Erfan Amirzadeh Shams (University College Dublin); Iona Gessinger (University College Dublin); Patrick Cormac English (UCD); Julie Carson-Berndsen (University College Dublin)

1740 Exploring Self-Supervised Speech Representations for Cross-lingual Acoustic-to-Articulatory Inversion

Yun Hao (University of Groningen); Reihaneh Amooie (University of Groningen); Wietse de Vries (University of Groningen); Thomas B Tienkamp (University of Groningen); Rik van Noord (University of Groningen); Martijn Wieling (University of Groningen)

2047 Searching for Structure: Appraising Organisation of Speech Features in wav2vec2.0 embeddings

Patrick Cormac English (UCD); John Kelleher (TCD); Julie Carson-Berndsen (University College Dublin)

2490 Human-like Linguistic Biases in Neural Speech Models: Phonetic Categorization and Phonotactic Constraints in Wav2Vec2.0

Marianne LS de Heer Kloots (University of Amsterdam); Willem Zuidema (ILLC, UvA)

Oral Session: Privacy and Security in Speech Communication 2

A7-07

Location: Aegle A

Surv-07-1 Fake Voice and Voice Privacy (tentative)

Junichi Yamagishi and Xin Wang

328 Anonymising Elderly and Pathological Speech: Voice Conversion Using DDSP and Query-by-Example

Suhita Ghosh (Otto von Guericke University); Melani Jouaiti (University of Birmingham); Arnab Das (Deutsches Forschungszentrum für Künstliche Intelligenz); Yamini Sinha (OvGU Magdeburg); Tim Polzehl (German Research Center for Artificial Intelligence); Ingo Siegert (OvG University Magdeburg); Sebastian Stober (Otto von Guericke University)

502 DiffVC+: Improving Diffusion-based Voice Conversion for Speaker Anonymization

Fan Huang (Sun Yat-sen University); Kun Zeng (Sun Yat-sen University); Wei Zhu (China Mobile Internet Co., Ltd.)

1615 Probing the Feasibility of Multilingual Speaker Anonymization

Sarina Meyer (University of Stuttgart); Florian Lux (University of Stuttgart); Ngoc Thang Vu (University of Stuttgart)

1888 Asynchronous Voice Anonymization Using Adversarial Perturbation On Speaker Embedding

Rui Wang (University of Science and Technology of China); Liping chen (University of Science and Technology of China); Kong Aik Lee (The Hong Kong Polytechnic University); Zhen-Hua Ling (University of Science and Technology of China)

Oral Session: Streaming ASR**A8-07**

Location: Aegle B

705 Decoder-only Architecture for Streaming End-to-end Speech Recognition

Emiru Tsunoo (Sony Group Corporation); Hayato Futami (Sony Group Corporation); Yosuke Kashiwagi (Sony); Sidhant Arora (Carnegie Mellon University); Shinji Watanabe (Carnegie Mellon University)

1527 Learning from Back Chunks: Acquiring More Future Knowledge for Streaming ASR Models via Self Distillation

Yuting Yang (NetEase Yidun AI Lab); Guodong Ma (NetEase Yidun AI Lab); Yuke Li (NetEase Yidun AI Lab); Binbin Du (NetEase Yidun AI Lab); Haoqi Zhu (NetEase Yidun AI Lab); Liang Ruan (NetEase Yidun AI Lab)

1814 Simul-Whisper: Attention-Guided Streaming Whisper with Truncation Detection

Haoyu Wang (Tsinghua University); Guoqiang Hu (Jinan University); Guodong Lin (Tsinghua University); Wei-Qiang Zhang (Tsinghua University); Jian Li (Beijing Sinovoice Technology)

1853 Streaming Decoder-Only Automatic Speech Recognition with Discrete Speech Units: A Pilot Study

Peikun Chen (Northwestern Polytechnical University); Sining Sun (Du Xiaoman); Changhao Shan (Du Xiaoman Financial); qing yang (Du Xiaoman Financial); Lei Xie (NWPU)

1971 Improving Streaming Speech Recognition with Time-Shifted Contextual Attention And Dynamic Right Context Masking

Khanh Duy Le (Zalo AI); Duc Chau (Ho Chi Minh University of Science)

2378 TfCleanformer: A streaming, array-agnostic, full- and sub-band modeling front-end for robust ASR

Jens Heitkaemper (Google); Joseph P Caroselli (Google); Arun Narayanan (Google Inc.); Nathan Howard (Google LLC)

Oral Session: Experimental Phonetics and Laboratory Phonology**A2-05**

Location: Hippocrates

182 Quantity-sensitivity affects recall performance of word stress

Constantijn Kaland (Institute of Linguistics - University of Cologne); Maria Lialiou (Department of German Language and Literature I - University of Cologne)

1118 Modeling probabilistic reduction across domains with Naive Discriminative Learning

Anna S Stein (Heinrich Heine University); Kevin Tang (Heinrich Heine University)

1915 Age-related Differences in Acoustic Cues for the Perception of Checked Syllables in Shengzhou Wu

Bingliang Zhao (Peking University); Jiangping Kong (Department of Chinese Language and Literature of Peking University); Xiyu Wu (Peking University)

2190 Do we EXPECT TO find phonetic traces for syntactic traces?

Jonathan Him Nok Lee (University of Pennsylvania); Mark Liberman (University of Pennsylvania); Martin Salzmann (University of Potsdam)

2212 Perceptual Learning in Lexical Tone: Phonetic Similarity vs. Phonological Categories

Ariëlle Reitsema (Leiden University); Chenxin Li (Leiden University); Leanne van Lambalgen (Leiden University); Laura Preining (Leiden University); Saskia Galindo Jong (Leiden University); Qing Yang (Leiden University); Xinyi Wen (Leiden University); yiya chen (Leiden University)

2334 Phonological Symmetry Does Not Predict Generalization of Perceptual Adaptation to Vowels

Zuheyra Tokac (Northwestern University); Jennifer S Cole (Northwestern University)

Oral Session: Speaker Processing: Evaluation and Resources**A4-06**

Location: Iasso

1158 As Biased as You Measure: Methodological Pitfalls of Bias Evaluations in Speaker Verification Research

Wiebke Hutiri (Sony AI); Tanvina Patel (Multimedia computing, Delft University of Technology); Aaron Yi Ding (TU Delft); Odette Scharenborg (Multimedia Computing Group, Delft University of Technology)

1345 ESPnet-SPK: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models

Jee-weon Jung (Carnegie Mellon University); Wangyou Zhang (Shanghai Jiao Tong University); Jiatong Shi (Carnegie Mellon University); Zakaria Aldeneh (Apple); Takuya Higuchi (Apple); Alex N Gichamba (Carnegie Mellon University Africa); Barry-John Theobald (Apple); Ahmed Hussien Abdelaziz (Apple); Shinji Watanabe (Carnegie Mellon University)

1402 Active Speaker Detection in Fisheye Meeting Scenes with Scene Spatial Spectrums

Xinghao Huang (Huazhong University of Science and Technology); Weiwei Jiang (Huazhong University of Science and Technology); Long Rao (Huazhong University of Science and Technology); Wei Xu (Huazhong University of Science and Technology); Wenqing Cheng (School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China)

1490 VoxBlink2: A 100K+ Speaker Recognition Corpus and the Open-Set Speaker-Identification Benchmark

Yuke Lin (Wuhan University); Ming Cheng (Duke Kunshan University); Fulin Zhang (China Mobile Research Institute); Yingying Gao (China Mobile Research Institute); Shilei Zhang (China Mobile Research Institute); Ming Li (Duke Kunshan University)

1840 WeSep: A Scalable and Flexible Toolkit Towards Generalizable Target Speaker Extraction

Shuai Wang (Shenzhen Research Institute of Big Data, Chinese University of Hong Kong (Shenzhen)); Ke Zhang (Northeastern University); Shaoxiong Lin (Shanghai Jiao Tong University); Junjie Li (Tianjin University); xuefei wang (Shanghai Normal University); Meng Ge (National University of Singapore); Jianwei Yu (Tencent AI lab); Yanmin Qian (Shanghai Jiao Tong University); Haizhou Li (The Chinese University of Hong Kong, Shenzhen)

1972 VSASV: a Vietnamese Dataset for Spoofing-Aware Speaker Verification

Vu Long Hoang (Hanoi University of Science and Technology); Viet Thanh Pham (Hanoi University of Science and Technology); Hoa Thai Nguyen Xuan (HUST); Nhi Thao Pham (Hanoi University of Science and Technology); Phuong Tuan Dat (Hanoi University of Science and Technology); Thi Thu Trang Nguyen (Hanoi University of Science and Technology)

Oral Session: Target Speaker Extraction**A6-06**

Location: Melampus

331 Knowledge boosting during low-latency inference

Vidya Srinivas (University of Washington); Malek Itani (University of Washington); Tuochao Chen (University of Washington); Emre Sefik Eskimez (Microsoft); Takuya Yoshioka (AssemblyAI Inc.); Shyamnath Gollakota (University of Washington)

683 Binaural Selective Attention Model for Target Speaker Extraction

Hanyu Meng (The University of New South Wales); Qiquan Zhang (The University of New South Wales); Xiangyu Zhang (University of New South Wales); Vidhyasaharan Sethu (University of New South Wales); Eliathamby Ambikairajah (The University of New South Wales)

934 Unified Audio Visual Cues for Target Speaker Extraction

Tianci Wu (Inner Mongolia University); Shulin He (College of Computer Science, Inner Mongolia University); Jiahui Pan (Inner Mongolia University); Haifeng Huang (Lenovo); Zhijian Mo (Lenovo); Xueliang zhang (Inner Mongolia University)

1375 Target Speaker Extraction with Curriculum Learning

Yun Liu (National Institute of informatics); Xuechen Liu (National Institute of Informatics); Xiaoxiao Miao (Singapore Institute of Technology); Junichi Yamagishi (National Institute of Informatics)

1560 Centroid Estimation with Transformer-Based Speaker Embedder for Robust Target Speaker Extraction

Woon-Haeng Heo (NCSOFT); Joongyu Maeng (NCSOFT); Yoseb Kang (NCSOFT); Namhyun Cho (NCSOFT)

2168 All Neural Low-latency Directional Speech Extraction

Ashutosh Pandey (META); Sanha Lee (META); Juan Azcarreta Ortiz (Meta Platforms Inc.); Daniel D.E. Wong (Meta Platforms Inc.); Buye Xu (Meta Reality Labs Research)

Oral Session: Speech Type Classification**A5-08**

Location: Panacea Amphitheater

685 E-ODN: An Emotion Open Deep Network for Generalised and Adaptive Speech Emotion Recognition

Liuxian Ma (Beijing Institute of Technology); Lin Shen (Beijing Institute of Technology); Ruobing Li (Beijing Institute of Technology); Haojie Zhang (Beijing Institute of Technology); Kun Qian (Beijing Institute of Technology); Bin Hu (Beijing Institute of Technology); Bjorn W. Schuller (Imperial College London); Yoshiharu Yamamoto (The University of Tokyo)

1228 Enhancing Multilingual Voice Toxicity Detection with Speech-Text Alignment

Joseph Liu (Roblox); Mahesh Kumar Nandwana (Roblox); Janne Pytköinen (Roblox); Hannes Heikinheimo (Roblox); Morgan McGuire (Roblox)

1596 Enhancing Speech and Music Discrimination Through the Integration of Static and Dynamic Features

Liangwei Chen (School of Data Science, University of Science and Technology of China); Xiren Zhou (University of Science and Technology of China); Qiang Tu (Anhui Provincial Hospital); Huanhuan Chen (School of Computer Science and Technology, University of Science and Technology of China)

1934 Speech Topic Classification Based on Multi-Scale and Graph Attention Networks

fangjing niu (Xinjiang University); Xiaozhe Qi (Xinjiang University); Xinya Chen (Xinjiang University); Liang HE (Tsinghua University)

2077 ARAOFFENSE: Detecting Offensive Speech Across Dialects in Arabic Media

Youssef M Nafea (Mohamed bin Zayed University of Artificial Intelligence); Shady Shehata (Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)); Zeerak Talat (Mohamed Bin Zayed University of Artificial Intelligence);

Ahmed Hesham Aboeitta (Mohamed bin Zayed University of Artificial Intelligence: MBZUAI); Ahmed Sharshar (Mohamed bin Zayed University of Artificial Intelligence); Preslav Nakov (MBZUAI)

2370 CogniVoice: Multimodal and Multilingual Fusion Networks for Mild Cognitive Impairment Assessment from Spontaneous Speech

Jiali Cheng (UMass Lowell); Mohamed Elgaar (University of Massachusetts Lowell); Nidhi Vakil (University of Massachusetts, Lowell); Hadi Amiri (UMass Lowell)

Poster Session: L1/L2 Acquisition and Cross-Linguistic Factors

A1-P2-A

Location: Poster Area 1A

274 Ethnolinguistic Identification of Vietnamese-German Heritage Speech

Thanh Lan Truong (University of Tuebingen); Andrea Weber (University of Tuebingen)

416 Cross-Linguistic Intelligibility of Non-Compositional Expressions in Spoken Context

Iuliia Zaitova (Saarland University); Irina Stenger (Saarland University); Wei Xue (Saarland University); Tania Avgustina (Saarland University); Bernd Möbius (Saarland University); Dietrich Klakow (Saarland University)

428 Acquisition of high vowel devoicing in Japanese: A production experiment with three and four year olds

Hyun Kyung Hwang (University of Tsukuba); Manami Hirayama (Seikei University)

614 Effect of Complex Boundary Tones on Tone Identification: An Experimental Study with Mandarin-speaking Preschool Children

Aijun Li ("Institute of Linguistics, CASS"); Jun Gao (Institute of Linguistics, Chinese Academy of Social Sciences); Zhiwei Wang (Institute of Linguistics, Chinese Academy of Social Sciences)

1000 Towards Classifying Mother Tongue from Infant Cries - Findings Substantiating Prenatal Learning Theory

Tim Polzehl (German Research Center for Artificial Intelligence); Tim Herzig (DFKI); Friedrich Wicke (TU Berlin); Kathleen Wermke (University Hospital Würzburg); Razieh Khamsehashari (TUB); Michiko Dahlem (University Hospital Würzburg); Sebastian Möller (TU Berlin)

1611 The Production of Contrastive Focus by 7 to 13-year-olds Learning Mandarin Chinese

Zimeng Li (Nanjing University of Science and Technology); Zhongxuan Mao (Nanjing University of Science and Technology, Nanjing); Shengting Shen (Nanjing University of Science and Technology); Ivan Yuen (Macquarie University); PING TANG (Nanjing University of Science and Technology)

2377 On the relationship between speech production and vocabulary size in 3-5 year olds

Alexis M DeMaere (University of Lethbridge); Nicole A van Rootselaar (University of Lethbridge); Fangfang Li (University of Lethbridge); Robbin Gibb (University of Lethbridge); Claudia Gonzalez (University of Lethbridge)

Poster Session: Speaker Stance, Emotion and Language-External Factors

A1-P2-B

Location: Poster Area 1B

14 Listeners' F0 preferences in quiet and stationary noise

Olympia Simantiraki (Institute of Applied and Computational Mathematics, FORTH); Martin Cooke (Ikerbasque)

473 Joint prediction of subjective listening effort and speech intelligibility based on end-to-end learning

Dirk E Hoffner (Carl von Ossietzky Universität Oldenburg); Jana Roßbach (Carl von Ossietzky Universität Oldenburg)

burg); Bernd Meyer (Carl von Ossietzky Universität Oldenburg)

1562 Depression Enhances Internal Inconsistency between Spoken and Semantic Emotion: Evidence from the Analysis of Emotion Expression in Conversation

Xinyi Wu (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. University of Chinese Academy of Sciences); Changqing Xu (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences); Nan Li (Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences); Rongfeng Su (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Lan Wang (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences); Nan Yan (Key Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences)

1721 Effects of talker and playback rate of reverberation-induced speech on speech intelligibility of older adults

Nao Hodoshima (Tokai University)

Poster Session: Computational Resource Constrained ASR

A9-P3 Location: Poster Area 2A, Poster Area 2B

47 USM RNN-T model weights binarization

Oleg Rybakov (Google); Dmitriy Serdyuk (Google); Chengjian Zheng (Google)

233 RepTor: Re-parameterizable Temporal Convolution for Keyword Spotting via Differentiable Kernel Search

Eunik Park (SqueezeBits Inc.); Daehyun Ahn (SqueezeBits Inc.); Hyungjun Kim (SqueezeBits Inc.)

263 A Small and Fast BERT for Chinese Medical Punctuation Restoration

Tongtao Ling (Guangdong University of Technology); Yutao Lai (Guangdong University of Technology); Lei Chen (Guangdong University of Technology); Shilei Huang (Shenzhen Raisound Technologies, Co., Ltd); Yi Liu (IMSL Shenzhen Key Lab, PKU-HKUST Shenzhen Hong Kong Institute)

607 SparseWAV: Fast and Accurate One-Shot Unstructured Pruning for Large Speech Foundation Models

Tianteng Gu (Shanghai Jiao Tong University); Bei Liu (Shanghai Jiao Tong University); Hang Shao (Shanghai Jiao Tong University); Yanmin Qian (Shanghai Jiao Tong University)

626 DAISY: Data Adaptive Self-Supervised Early Exit for Speech Representation Models

Tzu-Quan Lin (National Taiwan University); Hung-yi Lee (National Taiwan University); Hao Tang (The University of Edinburgh)

642 Global-Local Convolution with Spiking Neural Networks for Energy-efficient Keyword Spotting

shuai Wang (University of Electronic Science and Technology of China); Dehao Zhang (University of Electronic Science and Technology of China); Kexin Shi (University of Electronic Science and Technology of China); Wenjie Wei (University of Electronic Science and Technology of China); Yuchen Wang (University of Electronic Science and Technology of China); Jibin Wu (The Hong Kong Polytechnic University); Malu Zhang (University of Electronic Science and Technology of China)

703 One-pass Multiple Conformer and Foundation Speech Systems Compression and Quantization Using An All-in-one Neural Model

Zhaoqing Li (The Chinese University of Hong Kong); Haoning XU (CUHK); Tianzi Wang (The Chinese University of Hong Kong); Shoukang Hu (Nanyang Technological University); Zengrui Jin (The Chinese University of Hong Kong); Shujie HU (The Chinese University of Hong Kong); Jiajun Deng (The Chinese University of Hong Kong); Mingyu Cui (The Chinese University of Hong Kong); Mengzhe GENG (The Chinese University of Hong Kong); Xunying Liu (The Chinese University of Hong Kong)

Chinese University of Hong Kong)

976 Mitigating Overfitting in Structured Pruning of ASR Models with Gradient-Guided Parameter Regularization

Dong-Hyun Kim (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1330 Dynamic Data Pruning for Automatic Speech Recognition

Qiao Xiao (Eindhoven University of Technology); Pingchuan Ma (Meta); Adriana Fernandez-Lopez (Meta); Boqian Wu (University of Twente); Lu Yin (University of Aberdeen); Stavros Petridis (Meta); Mykola Pechenizkiy (TU Eindhoven); Maja Pantic (Meta); Decebal Constantin Mocanu (University of Luxembourg); Shiwei Liu (UT Austin)

1609 ED-sKWS: Early-Decision Spiking Neural Networks for Rapid, and Energy-Efficient Keyword Spotting

Zeyang Song (National University of Singapore); Qianhui Liu (National University of Singapore); Qu Yang (National University of Singapore); Yizhou Peng (National University of Singapore); Haizhou Li (The Chinese University of Hong Kong (Shenzhen))

Poster Session: Evaluation of Speech Technology Systems

A12-P3-A

Location: Poster Area 3A

32 Beyond Levenshtein: Leveraging Multiple Algorithms for Robust Word Error Rate Computations And Granular Error Classifications

Korbinian Kuhn (Stuttgart Media University); Verena Kersken (Stuttgart Media University); Gottfried Zimmermann (Stuttgart Media University)

1270 Comparing ASR Systems in the Context of Speech Disfluencies

Maria Teleki (Texas A & M University); Xiangjue Dong (Texas A & M University); Soohwan Kim (Texas A & M University); James Caverlee (Texas A & M University)

1869 Deep Prosodic Features in Tandem with Perceptual Judgments of Word Reduction for Tone Recognition in Conversated Speech

Xiang-Li Lu (Feng-Chia University); Yi-Fen Liu (IECS, Feng Chia University)

2033 SeMaScore: A new evaluation metric for automatic speech recognition tasks

Zitha Sasindran (Indian Institute of Science); Harsha Yelchuri (Information Science Engineering RV College of Engineering Bengaluru, India); Prabhakar Venkata Tamma (Electronics Systems Engg)

2107 Leveraging Speech Data Diversity to Document Indigenous Heritage and Culture

Allahsera Auguste Tapo (Rochester Institute of Technology); Éric Le Ferrand (Boston College); Zoey Liu (University of Florida); Christopher M. Homan (Rochester Institute of Technology); Emily Prud'hommeaux (Boston College)

2300 Quantification of stylistic differences in human- and ASR-produced transcripts of African American English

Annika L Heuser (University of Pennsylvania); Tyler Kendall (University of Oregon); Miguel A del Rio Fernandez (Rev); Quinten McNamara (Rev.com); Nishchal Bhandari (Rev); Corey Miller (Rev); Miguel Jetté (Rev)

Poster Session: Neural Network Training for Speech Recognition

A12-P3-B

Location: Poster Area 3B

343 Dynamic Encoder Size Based on Data-Driven Layer-wise Pruning for Speech Recognition

Jingjing Xu (Machine Learning and Human Language Technology Group, RWTH Aachen University); Wei Zhou

(Meta); Zijian Yang (RWTH Aachen University); Eugen Beck (Apptek); Ralf Schlüter (RWTH Aachen University)

363 Guiding Frame-Level CTC Alignments Using Self-Knowledge Distillation

Eungbeom Kim (Seoul National University); Hantae Kim (Naver Cloud); Kyogu Lee (Seoul National University)

555 Enhancing CTC-based speech recognition with diverse modeling units

Shiyi Han (Apple); Mingbin Xu (Apple); Zhihong Lei (Apple); Zhen Huang (Apple); Xingyu Na (Apple)

558 Optimizing Large-Scale Context Retrieval for End-to-end ASR

Zhiqi Huang (University of Massachusetts Amherst); Diamantino Caseiro (Google Inc.); Kandarp Joshi (Google Inc.); Christopher Li (Google); Pat Rondon (Google Inc.); Zelin Wu (Google LLC); Petr Zadrazil (Google Inc.); Lillian Zhou (Google)

588 Revisiting Convolution-free Transformer for Speech Recognition

Zejiang Hou (Amazon); Goeric Huybrechts (Amazon); Anshu Bhatia (Amazon); Daniel Garcia-Romero (AWS AI); Kyu Han (Amazon Web Services (AWS)); Katrin Kirchhoff (Amazon)

1157 Self-Supervised Speech Representations are More Phonetic than Semantic

Kwanghee Choi (Carnegie Mellon University); Ankita Pasad (Toyota Technological Institute at Chicago); Tomohiko Nakamura (National Institute of Advanced Industrial Science and Technology (AIST)); Satoru Fukayama (National Institute of Advanced Industrial Science and Technology (AIST)); Karen Livescu (TTI-Chicago); Shinji Watanabe (Carnegie Mellon University)

Poster Session: Speech Synthesis: Voice Conversion 3

A7-P4-A

Location: Poster Area 4A

46 HybridVC: Efficient Voice Style Conversion with Text and Audio Prompts

Xinlei Niu (Australian National University); Jing Zhang (Australian National University); Charles Patrick Martin (Australian National University)

232 Hear Your Face: Face-based voice conversion with F0 estimation

Jaejun Lee (Seoul National University); Yoori Oh (Seoul National University); Injune Hwang (Seoul National University); Kyogu Lee (Seoul National University)

664 USD-AC: Unsupervised Speech Disentanglement for Accent Conversion

Jen-Hung Huang (Department of Computer Science and Information Engineering, National Cheng Kung University); Wei-Tsung Lee (National Cheng Kung University); Chung-Hsien Wu (National Cheng Kung University)

888 SPA-SVC: Self-supervised Pitch Augmentation for Singing Voice Conversion

Bingsong Bai (Beijing University of Posts and Telecommunications); Fengping Wang (Beijing University of Posts and Telecommunications); Yingming Gao (Beijing University of Posts and Telecommunications); Ya Li (Beijing University of Posts and Telecommunications)

924 Knowledge Distillation from Self-Supervised Representation Learning Model with Discrete Speech Units for Any-to-Any Streaming Voice Conversion

Hiroki Kanagawa (NTT Corporation); Yusuke Ijima (NTT Corporation)

947 PRVAE-VC2: Non-Parallel Voice Conversion by Distillation of Speech Representations

Kou Tanaka (NTT Corporation); Hirokazu Kameoka (NTT Communication Science Laboratories, NTT Corporation); Takuhiro Kaneko (NTT Corporation); Yuto Kondo (NTT)

1256 Towards Naturalistic Voice Conversion: NaturalVoices Dataset with an Automatic Processing Pipeline

Ali N Salman (University of Texas at Dallas); Zongyang Du (The University of Texas at Dallas); Shreeram Suresh Chandra (The University of Texas at Dallas); İsmail Rasim Ülgen (The University of Texas at Dallas); Carlos Busso (University of Texas at Dallas); Berrak Sisman (The University of Texas at Dallas)

1416 Accent Conversion with Articulatory Representations

Yashish M. Siriwardena (University of Maryland College Park); Nathan Swedlow (Dolby Laboratories); Audrey Howard (Dolby Laboratories); Evan Gitterman (Dolby Laboratories); Dan Darcy (Dolby Laboratories); Carol Y Espy-Wilson (University of Maryland); Andrea Fanelli (Dolby Laboratories)

1432 DreamVoice: Text-Guided Voice Conversion

Jiarui Hai (Johns Hopkins University); Karan Thakkar (Johns Hopkins University); Helin Wang (Johns Hopkins University); Zengyi Qin (MIT); Mounya Elhilali (Johns Hopkins University)

Poster Session: Speech Synthesis: Paradigms and Methods 3**A7-P4-B**

Location: Poster Area 4B

63 VoiceTailor: Lightweight Plug-In Adapter for Diffusion-Based Personalized Text-to-Speech

Heeseung Kim (Seoul National University); Sang-gil Lee (NVIDIA); Jiheum Yeom (Seoul National University); Che Hyun Lee (Seoul National University); Sungwon Kim (NVIDIA); Sungroh Yoon (Seoul National University)

1023 PitchFlow: adding pitch control to a Flow-matching based TTS model

Taslima Sadekova (Huawei Noah's Ark Lab); Mikhail Kudinov (Huawei Noah's Ark Lab); Vadim Popov (Huawei Noah's Ark Lab); Assel Yermekova (Huawei Noah's Ark Lab); Artem Khrapov (Huawei Noah's Ark Lab)

1392 SimpleSpeech: Towards Simple and Efficient Text-to-Speech with Scalar Latent Transformer Diffusion Models

Dongchao Yang (The Chinese University of Hong Kong); Dingdong Wang (The Chinese University of Hong Kong); Haohan Guo (The Chinese University of Hong Kong); Xueyuan Chen (The Chinese University of Hong Kong); Xixin Wu (The Chinese University of Hong Kong); Helen Meng (The Chinese University of Hong Kong)

1465 Generating Speakers by Prompting Listener Impressions for Pre-trained Multi-Speaker Text-to-Speech Systems

Zhengyang Chen (Shanghai Jiao Tong University); Xuechen Liu (National Institute of Informatics); Erica Cooper (National Institute of Informatics); Junichi Yamagishi (National Institute of Informatics); Yanmin Qian (Shanghai Jiao Tong University)

1531 TacoLM: GaTed Attention Equipped Codec Language Model are Efficient Zero-Shot Text to Speech Synthesizers

Yakun Song (Shanghai Jiao Tong University); Zhuo Chen (Microsoft); Xiaofei Wang (Microsoft Corp.); Ziyang Ma (Shanghai Jiao Tong University); guanrou yang (Shanghai Jiao Tong University); Xie Chen (Shanghai Jiaotong University)

2005 DualSpeech: Enhancing Speaker-Fidelity and Text-Intelligibility Through Dual Classifier-Free Guidance

Jinhyeok Yang (Supertone, Inc.); Junhyeok Lee (Supertone Inc.); Hyeong-Seok Choi (Seoul National University); Seunghun Ji (Supertone Inc.); Hyeongju Kim (Supertone, Inc.); Juheon Lee (Seoul National University)

2235 Sample-Efficient Diffusion for Text-To-Speech Synthesis

Justin Lovelace (Cornell University); Soham Ray (Asapp); Kwangyoum Kim (ASAPP); Kilian Q Weinberger (Cornell University); Felix Wu (ASAPP)

2337 Exploring the Robustness of Text-to-Speech Synthesis Based on Diffusion Probabilistic Models to Heavily Noisy Transcriptions

Jingyi Feng (Nagoya University); Yusuke Yasuda (Nagoya university); Tomoki Toda (Nagoya University)

Poster Session: Responsible Speech Foundation Models (Special Session)

SS-7

Location: Yanis Club

102 Speech foundation models in healthcare: Effect of layer selection on pathological speech feature prediction

Daniela Wiepert (Mayo Clinic); Rene L Utianski (Mayo Clinic); Joseph Duffy (Mayo Clinic); John Stricker (Mayo Clinic); Leland Barnard (Mayo Clinic); David Jones (Mayo Clinic); Hugo Botha (Mayo Clinic)

454 On the social bias of speech self-supervised models

Yi-Cheng Lin (National Taiwan University); Tzu-Quan Lin (National Taiwan University); Hsi-Che Lin (National Taiwan University); Andy T. Liu (National Taiwan University); Hung-yi Lee (National Taiwan University)

971 Empowering Whisper as a Joint Multi-Talker and Target-Talker Speech Recognition System

Lingwei Meng (The Chinese University of Hong Kong); Jiawen Kang (The Chinese University of Hong Kong); Yuejiao Wang (The Chinese University of Hong Kong); Zengrui Jin (The Chinese University of Hong Kong); Xixin Wu (The Chinese University of Hong Kong); Xunying Liu (The Chinese University of Hong Kong); Helen Meng (The Chinese University of Hong Kong)

1073 Emo-bias: A Large Scale Evaluation of Social Bias on Speech Emotion Recognition

Yi-Cheng Lin (National Taiwan University); Haibin Wu (National Taiwan University); Huang-Cheng Chou (Department of Electrical Engineering at National Tsing Hua University (NTHU)); Chi-Chun Lee (National Tsing Hua University); Hung-yi Lee (National Taiwan University)

1212 Can you Remove the Downstream Model for Speaker Recognition with Self-Supervised Speech Features?

Zakaria Aldeneh (Apple); Takuya Higuchi (Apple); Jee-weon Jung (Carnegie Mellon University); Skyler Seto (Apple); Tatiana Likhomanenko (Apple); Stephen Shum (Apple); Ahmed Hussen Abdelaziz (Apple); Shinji Watanabe (Carnegie Mellon University); Barry-John Theobald (Apple)

2105 Outlier Reduction with Gated Attention for Improved Post-training Quantization in Large Sequence-to-sequence Speech Foundation Models

Dominik Wagner (Technische Hochschule Nuernberg Georg Simon Ohm); Ilja Baumann (Technische Hochschule Nürnberg Georg Simon Ohm); Korbinian Riedhammer (Technische Hochschule Nürnberg Georg Simon Ohm); Tobias Bocklet (TH Nürnberg)

2394 Self-supervised Speech Representations Still Struggle with African American Vernacular English

Kalvin Chang (Carnegie Mellon University); Yi-Hui Chou (Carnegie Mellon University); Jiatong Shi (Carnegie Mellon University); Hsuan-Ming Chen (Carnegie Mellon University); Nicole Holliday (Pomona College); Odette Scharenborg (Multimedia Computing Group, Delft University of Technology); David R. Mortensen (Language Technologies Institute, Carnegie Mellon University)

2494 Unveiling Biases while Embracing Sustainability: Assessing the Dual Challenges of Automatic Speech Recognition Systems

Ajinkya Kulkarni (ValidSoft MBZUAI); Atharva Kulkarni (Erisha); Miguel Couceiro (LORIA); Isabel Trancoso (INESC ID)

Time Slot 2

Oral Session: Spoken Term Detection and Speech Retrieval

A12-03

Location: Acesso

389 Sparse Binarization for Fast Keyword Spotting

Jonathan Svirsky (Bar Ilan University); Uri Shaham (Yale University); Ofir Lindenbaum (Yale)

1633 2DP-2MRC: 2-Dimensional Pointer-based Machine Reading Comprehension Method for Multimodal Moment Retrieval

Jiajun He (Nagoya University); Tomoki Toda (Nagoya University)

1642 GPA: Global and Prototype Alignment for Audio-Text Retrieval

Yuxin Xie (Peking University); Zhihong Zhu (Peking University); Xianwei Zhuang (Peking University); Liming Liang (Peking University); Zhichang Wang (Peking University); Yuexian Zou (Peking University)

1713 Pretraining End-to-End Keyword Search with Automatically Discovered Acoustic Units

Bolaji Yusuf (Bogazici University); Jan Honza Cernocky (Brno University of Technology); Murat Saraçlar (Bogazici University)

1823 Few-Shot Keyword-Incremental Learning with Total Calibration

Ilseok Kim (Hanyang University); Ju-seok Seong (Hanyang University); Joon-Hyuk Chang (Hanyang University)

2296 Few-Shot Keyword Spotting from Mixed Speech

Junming Yuan (Xinjiang University); Ying Shi (Harbin Institute of Technology); Lantian Li (Beijing University of Posts and Telecommunications); Dong Wang (Tsinghua University); Askar Hamdulla (Xinjiang University)

Oral Session: Speech synthesis: Cross-lingual and multilingual aspects

A7-08

Location: Aegle A

589 X-E-Speech: Joint Training Framework of Non-Autoregressive Cross-lingual Emotional Text-to-Speech and Voice Conversion

Houjian Guo (Osaka University, Riken Guardian Robot Group); Chaoran Liu (Riken); Carlos T Ishi (RIKEN); Hiroshi Ishiguro (Osaka University)

775 Improving Multilingual Text-to-Speech with Mixture-of-Language-Experts and Accent Disentanglement

Jing Wu (Fudan University); Ting Chen (Ping An Technology); Minchuan Chen (Ping An Technology); Wei Hu (Ping An Technology); Shaojun Wang (PAII Inc.); Jing Xiao (Ping An Insurance (Group) Company of China)

969 An Initial Investigation of Language Adaptation for TTS Systems under Low-resource Scenarios

Cheng Gong (Tianjin University); Erica Cooper (National Institute of Informatics); Xin Wang (National Institute of Informatics); Chunyu Qiang (Tianjin University); Mengzhe Geng (National Research Council Canada); Dan Wells (University of Edinburgh); Longbiao Wang (Tianjin University); Jianwu Dang (Tianjin University); Marc Tessier (National Research Council Canada); Aidan Pine (National Research Council Canada); Korin Richmond (University of Edinburgh); Junichi Yamagishi (National Institute of Informatics)

1335 Meta Learning Text-to-Speech Synthesis in over 7000 Languages

Florian Lux (University of Stuttgart); Sarina Meyer (University of Stuttgart); Lyonel Behringer (Fraunhofer IIS); Frank Zalkow (Fraunhofer IIS); Phat Do (University of Groningen); Matt Coler (University of Groningen); Emanuel Habets (AudioLabs Erlangen); Ngoc Thang Vu (University of Stuttgart)

1716 Seamless Language Expansion: Enhancing Multilingual Mastery in Self-Supervised Models

Jing Xu (The Chinese University of Hong Kong); Minglin Wu (The Chinese University of Hong Kong); Xixin Wu (The

Chinese University of Hong Kong); Helen Meng (The Chinese University of Hong Kong)

2016 XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model

Edresson Casanova (Nvidia); Kelly Davis (Coqui); Eren Gölge (Coqui.ai); Görkem Gökner (Coqui); Iulian Gulea (Coqui); Logan Hart (Cantina.ai); Aya Jafari (Nvidia); Joshua Meyer (Rabbit); Reuben Morais (Voize); Samuel Olayemi (Coqui); Julian Weber (Cantina.ai)

Oral Session: Self-Supervised Learning for ASR

A8-08 Location: Aegle B

Surv-08-1 Multilingual Multimodal ASR

Bhuvana Ramabhadran

476 What happens in continued pre-training? Analysis of self-supervised speech models with continued pre-training for colloquial Finnish ASR

Yaroslav Getman (Aalto University); Tamas Grosz (Aalto University); Mikko Kurimo (Aalto University)

1875 Balanced-Wav2Vec: Enhancing Stability and Robustness of Representation Learning Through Sample Reweighting Techniques

MUNHAK LEE (Hanyang University); Jae-Hong Lee (Hanyang University); Dohee Kim (Hanyang University); Ye-Eun Ko (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1933 Self-Supervised Learning for ASR Pre-Training with Uniquely Determined Target Labels and Controlling Cepstrum Truncation for Speech Augmentation

Akihiro Kato (RICOH); Hiroyuki Nagano (RICOH); Kohei Chike (RICOH); Masaki Nose (RICOH)

1978 MS-HuBERT: Mitigating Pre-training and Inference Mismatch in Masked Language Modelling methods for learning Speech Representations

Hemant Yadav (IIIT Delhi); Sunayana Sitaram (Microsoft Research); Rajiv Ratn Shah (IIIT Delhi)

Oral Session: Speech Disorders 1

A13-04 Location: Hippocrates

524 Prosody of speech production in latent post-stroke aphasia

Cong Zhang (Newcastle University); Tong Li (Tianjin University); Christos Salis (Newcastle University); Gayle Dede (Temple University)

918 AS-70: A Mandarin stuttered speech dataset for automatic speech recognition and stuttering event detection

Rong Gong (StammerTalk); Hongfei Xue (NWPU); Lezhi Wang (StammerTalk); Xin Xu (AISHELL); Qisheng Li (Almpower.org); Lei Xie (NWPU); Hui Bu (AISHELL); Shaomei Wu (Almpower.org); Jiaming Zhou (Nankai University); Yong Qin (Nankai University); Binbin Zhang (WeNet Open Source Community); Jun Du (University of Science and Technology of China); Jia Bin (StammerTalk); Ming Li (Duke Kunshan University)

1309 Analyzing Speech Motor Movement using Surface Electromyography in Minimally Verbal Adults with Autism Spectrum Disorder

Wazeer Zulfikar (Massachusetts Institute of Technology (MIT)); Nishat Fahmida Protyasha (Massachusetts Institute of Technology); Camila Canales (Massachusetts General Hospital); Heli Patel (Massachusetts General Hospital); James Williamson (MIT Lincoln Laboratory); Laurie Sarnie (Massachusetts General Hospital); Lisa Nowinski (Massachusetts

General Hospital); Nataliya Kosmyrna (MIT); Paige Townsend (Massachusetts General Hospital); Sophia Yuditskaya (MIT Lincoln Laboratory); Tanya Talkar (MIT Lincoln Laboratory); Utkarsh Sarawgi (Massachusetts Institute of Technology (MIT)); Christopher McDougale (Massachusetts General Hospital); Thomas Quatieri (Massachusetts Institute of Technology Lincoln Laboratory); Pattie Maes (MIT Media Lab); Maria Mody (Massachusetts General Hospital)

1458 Missingness-resilient Video-enhanced Multimodal Disfluency Detection

Payal Mohapatra (Northwestern University); Shamika Likhite (Northwestern University); Subrata Biswas (Worcester Polytechnic Institute); Bashima Islam (Worcester Polytechnic Institute); Zhu Qi (Northwestern University)

1497 MMSD-Net: Towards Multi-modal Stuttering Detection

Liangyu Nie (UT Dallas); Sudarsana Reddy Kadiri (University of Southern California); Ruchit Agrawal (University of Oxford)

2120 Large Language Models for Dysfluency Detection in Stuttered Speech

Dominik Wagner (Technische Hochschule Nuernberg Georg Simon Ohm); Sebastian P Bayerl (University of Applied Sciences Rosenheim); Ilja Baumann (Technische Hochschule Nürnberg Georg Simon Ohm); Elmar Noeth (Friedrich Alexander Universität, Erlangen-Nuremberg); Korbinian Riedhammer (Technische Hochschule Nürnberg Georg Simon Ohm); Tobias Bocklet (TH Nürnberg)

Oral Session: Self and Weakly-Labelled Speaker Verification

A4-07

Location: IASSO

Surv-04-4 Self-supervised based speaker recognition

Themos Stafylakis

360 Self-Supervised Learning with Multi-Head Multi-Mode Knowledge Distillation for Speaker Verification

Jin Zezhong (The Hong Kong Polytechnic University); Youzhi TU (The Hong Kong Polytechnic University); Man-Wai MAK (The Hong Kong Polytechnic University)

531 Getting More for Less: Using Weak Labels and AV-Mixup for Robust Audio-Visual Speaker Verification

Anith Selvakumar (LG Electronics); Homa Fashandi (LG Electronics)

752 SCDNet: Self-supervised Learning Feature-based Speaker Change Detection

Yue Li (Northwestern Polytechnical University, China); Xinsheng Wang (Northwestern Polytechnical University); Li Zhang (Northwestern Polytechnical University); Lei Xie (NWPU)

1466 Self-Supervised Speaker Verification with Mini-Batch Prediction Correction

Junxu Wang (Xinjiang university); Zhihua Fang (Xinjiang University); Liang HE (Tsinghua University)

Oral Session: Deep Learning-Based Speech Enhancement: Approaches, Scalability, and Evaluation

A6-07

Location: Melampus

Surv-06-1 Deep learning based speech enhancement

Timo Gerkmann

53 VoiCor: A Residual Iterative Voice Correction Framework for Monaural Speech Enhancement

Rui Cao (Tianjin University); Tianrui Wang (Tianjin University); Meng Ge (National University of Singapore); Andong Li (Tencent AI Lab); Longbiao Wang (Tianjin University); Jianwu Dang (Tianjin University); Yungang Jia (Tianjin Branch of National Computer Network Emergency Response Technical Team/Coordination Center of China)

153 EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation

Julius Richter (Universität Hamburg); Yi-Chiao Wu (META); Steven Krenn (Meta); Alexander Richard (Facebook Reality Labs); Simon Welker (Universität Hamburg); Bunlong Lay (Universität Hamburg); Shinji Watanabe (Carnegie Mellon University); Timo Gerkmann (Universität Hamburg)

1234 Personalized Speech Enhancement Without a Separate Speaker Embedding Model

Tanel Parnamaa (Microsoft); Ando Saabas (Microsoft)

1239 URGENT Challenge: Universality, Robustness, and Generalizability for speech EnhancemeNT

Wangyou Zhang (Shanghai Jiao Tong University); Robin Scheibler (LINE Corporation); Kohei Saijo (Waseda University); Samuele Cornell (Carnegie Mellon University); Chenda Li (Shanghai Jiao Tong University); Zhaoheng Ni (Meta AI); Jan Pirklbauer (Technische Universität Braunschweig); Marvin Sach (Technische Universität Braunschweig); Shinji Watanabe (Carnegie Mellon University); Tim Fingscheidt (Technische Universität Braunschweig); Yanmin Qian (Shanghai Jiao Tong University)

Oral Session: Fake Audio Detection

A5-09 Location: Panacea Amphitheater

177 One-class learning with adaptive centroid shift for audio deepfake detection

Hyun Myung Kim (KAIST); Kangwook Jang (KAIST); Hoirin Kim (KAIST)

254 Generalized Source Tracing: Detecting Novel Audio Deepfake Algorithm with Real Emphasis and Fake Dispersion Strategy

Yuankun Xie (Communication University of China); Ruibo Fu (National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences); Wen Zhengqi (CASIA); Zhiyong Wang (University of Chinese Academy of Sciences); Xiaopeng Wang (UCAS); Haonan Cheng (Communication University of China); Long Ye (Communication University of China); Jianhua Tao (Tsinghua University)

587 Enhancing Partially Spoofed Audio Localization with Boundary-aware Attention Mechanism

Jiafeng Zhong (Shenzhen University); Bin Li (Shenzhen University); Jiangyan Yi (National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences)

1185 Singing Voice Graph Modeling for SingFake Detection

Xuanjun Chen (National Taiwan University); Haibin Wu (National Taiwan University); Roger Jang (); Hung-yi Lee (National Taiwan University)

1302 Towards generalisable and calibrated audio deepfake detection with self-supervised representations

Octavian Pascu (University POLITEHNICA of Bucharest); Adriana Stan (Technical University of Cluj-Napoca); Dan Oneata (Politehnica University of Bucharest); Elisabeta Oneata (Bitdefender); Horia Cucu (University Politehnica of Bucharest)

1851 Genuine-Focused Learning using Mask AutoEncoder for Generalized Fake Audio Detection

Xiaopeng Wang (UCAS); Ruibo Fu (National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences); Wen Zhengqi (CASIA); Zhiyong Wang (University of Chinese Academy of Sciences); Yuankun Xie (Communication University of China); Yukun Liu (UCAS); Jianhua Tao (Tsinghua University); Xuefei Liu (Qiyuan Lab); Yongwei Li (Chinese Academy of Sciences); Xin Qi (University of Chinese Academy of Sciences); Yi Lu (State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences); Shuchen Shi (Shanghai Polytechnic University)

Poster Session: Multimodal Paralinguistics

A3-P3-A

Location: Poster Area 1A

535 Tackling Missing Modalities in Audio-Visual Representation Learning Using Masked Autoencoders
 Georgios Chochlakis (University of Southern California); Chandrashekhara Lavania (AWS AI Labs); Prashant Mathur (Amazon); Kyu Han (Amazon Web Services (AWS))

782 A multimodal analysis of different types of laughter expression in conversational dialogues
 Kexin Wang (Kobe University, ATR); Carlos Toshinori Ishi (Advanced Telecommunications Research Institute International); Ryoko Hayashi (Kobe University)

1226 Bridging Emotions Across Languages: Low Rank Adaptation for Multilingual Speech Emotion Recognition
 Lucas Goncalves (The University of Texas at Dallas); Donita Robinson (Laboratory for Analytic Sciences, North Carolina State University); Elizabeth Richerson (Laboratory for Analytic Sciences, North Carolina State University); Carlos Busso (University of Texas at Dallas)

1462 Enhancing Modal Fusion by Alignment and Label Matching for Multimodal Emotion Recognition
 Qifei Li (Beijing University of Posts and Telecommunications); Yingming Gao (Beijing University of Posts and Telecommunications); Yuhua Wen (Beijing University of Posts and Telecommunications); Cong Wang (Beijing University of Posts and Telecommunications); Ya Li (Beijing University of Posts and Telecommunications)

1512 Prompt Link Multimodal Fusion in Multimodal Sentiment Analysis
 Kang Zhu (Anhui University); Cunhang Fan (Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University); Jianhua Tao (Tsinghua University); Zhao Lv (Anhui University)

1793 LoRA-MER: Low-Rank Adaptation of Pre-Trained Speech Models for Multimodal Emotion Recognition Using Mutual Information
 Yunrui Cai (Tsinghua University); Zhiyong Wu (Tsinghua University); Jia Jia (Tsinghua University); Helen Meng (The Chinese University of Hong Kong)

2364 Enhancing Multimodal Emotion Recognition through ASR Error Compensation and LLM Fine-Tuning
 Jehyun Kyung (Hanyang University); Serin Heo (Hanyang University); Joon-Hyuk Chang (Hanyang University)

Poster Session: Automatic Emotion Recognition

A3-P3-B

Location: Poster Area 1B

427 MFSN: Multi-perspective Fusion Search Network For Pre-training Knowledge in Speech Emotion Recognition
 Haiyang Sun (Institute of Automation, Chinese Academy of Sciences); Fulin Zhang (China Mobile Research Institute); Yingying Gao (CMRI); Shilei Zhang (CMRI); Zheng Lian (Institute of Automation, Chinese Academy of Sciences); Junlan Feng (China Mobile Research)

469 A Layer-Anchoring Strategy for Enhancing Cross-Lingual Speech Emotion Recognition
 Shreya G. Upadhyay (National Tsing Hua University); Carlos Busso (University of Texas at Dallas); Chi-Chun Lee (National Tsing Hua University)

860 Exploring Self-Supervised Multi-view Contrastive Learning for Speech Emotion Recognition with Limited Annotations
 Bulat Khaertdinov (Maastricht University); Pedro Jeruis (Maastricht University); Annanda D F Sousa (Maastricht University)

versity); Enrique Hortal (Maastricht University)

1059 Time-frequency masking-based mask autoencoder for speech emotion recognition

Zhi-Kun Peng (China University of Geosciences); Zhen-Tao Liu (China University of Geosciences); Yu-Jie Zou (China University of Geosciences); Dan Chen (China University of Geosciences)

2233 Are Paralinguistic Representations all that is needed for Speech Emotion Recognition?

Orchid Chetia Phukan (IIIT Delhi); Gautam Siddharth Kashyap (Jamia Hamdard); Arun Balaji Buduru (IIIT Delhi); Rajesh Sharma (University of Tartu, Estonia)

Poster Session: Acoustic Event Detection, Segmentation and Classification

A5-P4-A

Location: Poster Area 2A

106 LungAdapter: Efficient Adapting Audio Spectrogram Transformer for Lung Sound Classification

Li Xiao (School of Computer Science, Wuhan University); Lucheng Fang (wuhan university); Yuhong Yang (Wuhan University); Weiping Tu (Wuhan University)

791 Explainable by-design Audio Segmentation through Non-Negative Matrix Factorization and Probing

Martin Lebourdais (IRIT/CNRS); Théo Mariotte (LTCL, Télécom Paris, Institut Polytechnique de Paris); Antonio Al-mudévar (University of Zaragoza); Marie Tahon (LIUM); Alfonso Ortega (Universidad de Zaragoza)

1361 Multimodal Large Language Models with Fusion Low Rank Adaptation for Device Directed Speech Detection

Shruti Palaskar (Apple); Ognjen Rudovic (Apple); Sameer Dharur (Apple); Florian Pesce (Apple); Gautam Krishna (Apple); Aswin Sivaraman (Apple); John Berkowitz (Apple); Ahmed Hussen Abdelaziz (Apple); Saurabh Adya (Apple); Ahmed Tewfik (Apple)

1454 Comparative Analysis of Personalized Voice Activity Detection Systems: Assessing Real-World Effectiveness

Sai Srujana Buddi (Apple); Satyam Kumar (The University of Texas at Austin); Utkarsh (Oggy) Sarawgi (Apple); Vi-neet Garg (Apple); Shivesh Ranjan (Apple); Ognjen Rudovic (Apple); Ahmed Hussen Abdelaziz (Apple); Saurabh Adya (Apple)

1644 Robust Laughter Segmentation with Automatic Diverse Data Synthesis

Taisei Omine (Kyushu University); Kenta Akita (Kyushu University); Reiji Tsuruno (Kyushu University)

1686 Generalized Fake Audio Detection via Deep Stable Learning

Zhiyong Wang (University of Chinese Academy of Sciences); Ruibo Fu (National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences); Wen Zhengqi (CASIA); Yuankun Xie (Communication University of China); Yukun Liu (UCAS); Xiaopeng Wang (UCAS); Xuefei Liu (Qiyuan Lab); Yongwei Li (Chinese Academy of Sciences); Jianhua Tao (Tsinghua University); Xin Qi (University of Chinese Academy of Sciences); Yi Lu (State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences); Shuchen Shi (Shanghai Polytechnic University)

1723 FastAST: Accelerating Audio Spectrogram Transformer via Token Merging and Cross-Model Knowledge Distillation

Swarup Ranjan Behera (Reliance Jio AICoE); Abhishek Dhiman (Reliance Jio AICoE); Karthik Gowda (Reliance Jio AICoE); Aalekhya Satya Narayani (Reliance Jio AICoE)

1758 Fully Few-shot Class-incremental Audio Classification Using Expandable Dual-embedding Extractor

Yongjie Si (South China University of Technology); Yanxiong Li (South China University of Technology); Jialong Li (South China University of Technology); Jiaxin Tan (South China University of Technology); Qianhua He (SOUTH CHINA UNIVERSITY OF TECHNOLOGY)

1890 ElasticAST: An Audio Spectrogram Transformer for All Length and Resolutions

Jiu Feng (Korea Advanced Institute of Science and Technology); Mehmet Hamza Erol (KAIST); Joon Son Chung (KAIST); Arda Senocak (KAIST)

2150 PREDICTING HEART ACTIVITY FROM SPEECH USING DATA-DRIVEN AND KNOWLEDGE-BASED FEATURES

Gasser Elbanna (Harvard); Zohreh Mostaani (Idiap Research Institute); Mathew Magimai.-Doss (Idiap Research Institute)

2204 Measuring acoustic dissimilarity of hierarchical markers in task-oriented dialogue with MFCC-based dynamic time warping

Natalia Morozova (University of Zurich); Guanghao You (University of Zurich); Sabine Stoll (University of Zurich, Department of Comparative Language Science, Center for the Interdisciplinary Study of Language Evolution); Adrian Bangerter (University of Neuchatel)

2242 CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection

Yongyi Zang (University of Rochester); Jiatong Shi (Carnegie Mellon University); You Zhang (University of Rochester); Ryuichi Yamamoto (Nagoya University); Jionghao Han (Carnegie Mellon University); Yuxun Tang (Renmin University of China); Shengyuan Xu (Timedomain.ai); Wenxiao Zhao (Timedomain.ai); Jing Guo (Timedomain.ai); Tomoki Toda (Nagoya University); Zhiyao Duan (University of Rochester)

Poster Session: Speech and Audio Modelling

A5-P4-B

Location: Poster Area 2B

112 Leveraging Language Model Capabilities for Sound Event Detection

hualai wang (Institute of Computing Technology, Chinese Academy of Sciences); Jianguo Mao (Institute of Computing Technology, Chinese Academy of Sciences); Zhifang Guo (Institute of Computing Technology, Chinese Academy of Sciences); Jiarui Wan (Beijing Jiaotong University); Hong Liu (Institute of Computing Technology, Chinese Academy of Sciences); Xiangdong Wang (Institute of Computing Technology, Chinese Academy of Sciences)

314 Blind Zero-Shot Audio Restoration: A Variational Autoencoder Approach for Denoising and Inpainting

Veranika Boukun (Carl von Ossietzky Universität Oldenburg); Jakob Drefs (Carl von Ossietzky Universität Oldenburg); Jörg Lücke (Universität Oldenburg)

478 DNSMOS Pro: A Reduced-Size DNN for Probabilistic MOS of Speech

Fredrik Cumlin (Codemill AB); Xinyu Liang (Codemill AB); Victor Ungureanu (Google); Chandan K. A. Reddy (Google); Christian Schüldt (Google); Saikat Chatterjee (KTH Royal Institute of Technology)

1692 Enhancing Zero-shot Audio Classification using Sound Attribute Knowledge from Large Language Models

Xuenan Xu (Shanghai Jiao Tong University); Pingyue Zhang (Shanghai Jiao Tong University); Mengyue Wu (Shanghai Jiao Tong University); Ming Yan (Alibaba Group); Ji Zhang (Alibaba Inc.)

1726 DiveSound: LLM-Assisted Automatic Taxonomy Construction for Diverse Audio Generation

Baihan Li (Shanghai Jiao Tong University); Zeyu Xie (Shanghai Jiao Tong University); Xuenan Xu (Shanghai Jiao Tong University); Mengyue Wu (Shanghai Jiao Tong University); Kai Yu (Shanghai Jiao Tong University); Yiwei Guo (Shanghai Jiao Tong University); Ming Yan (Alibaba Group); Ji Zhang (Alibaba Inc.)

1848 LAFMA: A Latent Flow Matching Model for Text-to-Audio Generation

Wenhao Guan (Xiamen University); Kaidi Wang (Xiamen University); Wangjin Zhou (Kyoto University); Yang Wang

(Kuaishou Technology); Feng Deng (Kuaishou); Hui Wang (Nankai University); Lin Li (Xiamen University); Qingyang Hong (Xiamen University); Yong Qin (Nankai University)

Poster Session: Speech Synthesis: Other Topics 1

A7-P5-A

Location: Poster Area 3A

302 Towards realtime co-speech gestures synthesis using STARGATE

Louis ABEL (Université de Lorraine); Vincent Colotte (LORIA); Slim Ouni (LORIA)

339 Towards a General-Purpose Model of Perceived Pragmatic Similarity

Nigel Ward (University of Texas at El Paso, USA); Andres Segura (University of Texas at El Paso); Alejandro Ceballos (University of Texas at El Paso); Divette Marco (University of Texas at El Paso)

400 Zero-shot Out-of-domain is No Joke: Lessons Learned in the VoiceMOS 2023 MOS Prediction Challenge

Marie Kunešová (University of West Bohemia); Jan Lehečka (University of West Bohemia); Josef Michálek (University of West Bohemia); Jindrich Matousek (University of West Bohemia, Pilsen, Czech Republic); Jan Švec (University of West Bohemia)

425 Speak in the Scene: Diffusion-based Acoustic Scene Transfer toward Immersive Speech Generation

Miseul Kim (Yonsei University); Soo-Whan Chung (Naver Corporation); Youna Ji (NAVER Corporation); Hong-Goo Kang (Yonsei University); Min-Seok Choi (NAVER)

1429 PL-TTS: A Generalizable Prompt-based Diffusion TTS Augmented by Large Language Model

Shuhua Li (Jiangsu University); Qirong Mao (Jiangsu University); Jiatong Shi (Carnegie Mellon University)

1435 FLY-TTS: Fast, Lightweight and High-Quality End-to-End Text-to-Speech Synthesis

Yinlin Guo (Zhejiang University); Yening Lv (Zhejiang University); Jinqiao Dou (Zhejiang University); Yan Zhang (Zhejiang University); Yuehai Wang (Zhejiang University)

1771 PPPR: Portable Plug-in Prompt Refiner for Text to Audio Generation

Shuchen Shi (Shanghai Polytechnic University); Ruibo Fu (National Laboratory of Pattern Recognition, Institute of Automation of the Chinese Academy of Sciences); Wen Zhengqi (CASIA); Jianhua Tao (Tsinghua University); Tao Wang (Institute of Automation, Chinese Academy of Sciences); Chunyu Qiang (Tianjin University); Yi Lu (State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences); Xin Qi (University of Chinese Academy of Sciences); Xuefei Liu (Qiyuan Lab); 育坤刘 (CASIA); Yongwei Li (Chinese Academy of Sciences); Zhiyong Wang (University of Chinese Academy of Sciences); Xiaopeng Wang (UCAS)

2380 Neural ATSM: Fully Neural Network-based Adaptive Time-Scale Modification Using Sentence-Specific Dynamic Control

Jaeuk Lee (Hanyang University); Sohee Jang (Hanyang University); Joon-Hyuk Chang (Hanyang University)

2467 LiteFocus: Accelerated Diffusion Inference for Long Audio Synthesis

Zhenxiong Tan (National University of Singapore); Xinyin Ma (National University of Singapore); Gongfan Fang (National University of Singapore); Xinchao Wang (National University of Singapore)

Poster Session: Speech Synthesis: Other Topics 2

A7-P5-B

Location: Poster Area 3B

140 FVTTS : Face Based Voice Synthesis for Text-to-Speech

Minyoung Lee (SUNGKYUNKWAN UNIVERSITY); Eunil Park (Sungkyunkwan University); Sungeun Hong (Sungkyunkwan University)

168 Frame-Wise Breath Detection with Self-Training: An Exploration of Enhancing Breath Naturalness in Text-to-Speech

Dong Yang (The University of Tokyo); Tomoki Koriyama (CyberAgent, Inc., Japan); Yuki Saito ("The University of Tokyo, Japan")

600 Bilingual and Code-switching TTS Enhanced with Denoising Diffusion Model and GAN

Huai-Zhe Yang (National Sun Yat-sen University); Chia-Ping Chen (National Sun Yat-sen University); Shan-Yun He (National Sun Yat-sen University); Cheng-Ruei Li (National Sun Yat-sen University)

1480 H4C-TTS: Leveraging Multi-Modal Historical Context for Conversational Text-to-Speech

Donghyun Seong (Hanyang University); Joon-Hyuk Chang (Hanyang University)

1508 SpeechBERTScore: Reference-Aware Automatic Evaluation of Speech Generation Leveraging NLP Evaluation Metrics

Takaaki Saeki (The University of Tokyo); Soumi Maiti (CMU); Shinnosuke Takamichi (The University of Tokyo); Shinji Watanabe (Carnegie Mellon University); Hiroshi Saruwatari (The University of Tokyo)

1675 UNIQUE : Unsupervised Network for Integrated Speech Quality Evaluation

Juhwan Yoon (Yonsei University); WooSeok Ko (Yonsei University); seyun um (yonsei university); Sungwoong Hwang (Hyundai Motor Company); Soojoong Hwang (Hyundai Motor Company); Changhwan Kim (Hyundai Motor Group); Hong-Goo Kang (Yonsei University)

1749 2.5D Vocal Tract Modeling: Bridging Low-Dimensional Efficiency with 3D Accuracy

Debasish Mohapatra (University of British Columbia); Victor Zappi (Northeastern University); Sidney Fels (University of British Columbia)

2403 Enabling Conversational Speech Synthesis using Noisy Spontaneous Data

Liisa Rätsep (University of Tartu); Rasmus Lellep (University of Tartu); Mark Fishel (University of Tartu)

Poster Session: Noise, Far-Field, Multi-Talker, Enhancement, Audio Classification

A8-P5

Location: Poster Area 4A, Poster Area 4B

306 Enhanced ASR Robustness to Packet Loss with a Front-End Adaptation Network

Yehoshua Dissen (Technion - Israel Institute of Technology); Shiry Yonash (Technion - Israel Institute of Technology); Israel Cohen (Technion - Israel Institute of Technology); Joseph Keshet (Technion - Israel Institute of Technology)

312 Towards Robust Few-shot Class Incremental Learning in Audio Classification using Contrastive Representation

Riyansha Singh (IIT Kanpur); Parinita Nema (Indian Institute of Science Education and Research Bhopal); Vinod Kumar Kurmi (IISER Bhopal)

413 SpeakerBeam-SS: Real-time Target Speaker Extraction with Lightweight Conv-TasNet and State Space Modeling

Hiroshi Sato (NTT); Takafumi Moriya (NTT); Masato Mimura (NTT corporation); Shota Horiguchi (NTT); Tsubasa Ochiai (NTT); Takanori Ashihara (NTT); Atsushi Ando (NTT Corporation); Kentaro Shinayama (NTT); Marc Delcroix (NTT)

1172 wTIMIT2mix: A Cocktail Party Mixtures Database to Study Target Speaker Extraction for Normal and Whispered Speech

Marvin Borsdorf (University of Bremen); Zexu Pan (Mitsubishi Electric Research Laboratories (MERL)); Haizhou Li (The Chinese University of Hong Kong, Shenzhen); Tanja Schultz (University of Bremen)

1264 Transcription-Free Fine-Tuning of Speech Separation Models for Noisy and Reverberant Multi-Speaker Automatic Speech Recognition

William Ravenscroft (The University of Sheffield); George L Close (University of Sheffield); Thomas Hain (University of Sheffield); Stefan Goetze (University of Sheffield); Mohammad Soleymanpour (Solventum); Anurag Chowdhury (Solventum); Mark Fuhs (Solventum)

1622 Hold Me Tight: Stable Encoder-Decoder Design for Speech Enhancement

Daniel Haider (Acoustics Research Institute); Felix Perfler (Acoustics Research Institute); Vincent Lostanlen (LS2N, CNRS); Martin Ehler (University of Vienna); Peter Balazs (Acoustics Research Institute, Austrian Academy of Sciences)

1623 Bird Whisperer: Leveraging Large Pre-trained Acoustic Model for Bird Call Classification

Muhammad Umer Sheikh (Mohamed Bin Zayed University of Artificial Intelligence); Hassan Abid (Mohamed Bin Zayed University of Artificial Intelligence); Bhuiyan Sanjid Shafique (Mohamed Bin Zayed University of Artificial Intelligence); Asif Hanif (Mohamed Bin Zayed University of Artificial Intelligence); Muhammad Haris Khan (Mohamed Bin Zayed University of Artificial Intelligence)

1788 NOTSOFAR-1 Challenge: New Datasets, Baseline, and Tasks for Distant Meeting Transcription

Alon Vinnikov (Microsoft); Amir Ivry Mark (Microsoft); Aviv Hurvitz (Microsoft); Igor Abramovski (Microsoft); Sharon Koubi (Microsoft); Ilya Gurvich (Microsoft Corporation); Shai Peer (Microsoft); Xiong Xiao (Microsoft); Benjamin Elizalde (Microsoft); Naoyuki Kanda (Microsoft); Xiaofei Wang (Microsoft); Shalev Shaer (Technion); Stav Yagev (Microsoft); Yossi Asher (Microsoft); Sunit Sivasankaran (Microsoft); Yifan Gong (Microsoft); Min Tang (Microsoft); Huaming Wang (Microsoft); Eyal Krupka (Microsoft Research)

1796 DGSRN: Noise-Robust Speech Recognition Method with Dual-Path Gated Spectral Refinement Network

Wenjun Wang (Kunming University of Science and Technology); mo shangbin (昆明理工大学); Ling Dong (Kunming University of Science and Technology); Zhengtao Yu (Kunming University of Science and Technology); Junjun Guo (Kunming University of science and technology); Yuxin Huang (Kunming University of science and technology)

2264 RIR-SF: Room Impulse Response Based Spatial Feature for Target Speech Recognition in Multi-Channel Multi-Speaker Scenarios

Yiwen Shao (Johns Hopkins University); Shi-Xiong Zhang (Capital One); Dong Yu (Tencent AI Lab)

2274 Multi-Channel Multi-Speaker ASR Using Target Speaker Solo Segments

Yiwen Shao (Johns Hopkins University); Shi-Xiong Zhang (Capital One); Yong Xu (Tencent); Meng Yu (Tencent); Dong Yu (Tencent AI Lab); Daniel Povey (Xiaomi, Inc.); Sanjeev Khudanpur (Johns Hopkins University)

Poster Session: Connecting Speech-science and Speech-technology for Children's Speech (Special Session)

SS-8

Location: Yanis Club

159 Bridging Child-Centered Speech Language Identification and Language Diarization via Phonetics

Yujia Wang (Johns Hopkins University); Paola Garcia (Johns Hopkins University); Hexin Liu (Nanyang Technological University)

485 Improving child speech recognition with augmented child-like speech

Yuanyuan Zhang (Technische Universiteit Delft); Zhengjun Yue (Technische Universiteit Delft); Tanvina Patel (Multimedia computing, Delft University of Technology); Odette Scharenborg (Multimedia Computing Group, Delft University of Technology)

499 Mixed Children/Adult/Childrenized Fine-Tuning for Children's ASR: How to Reduce Age Mismatch and Speaking Style Mismatch

Thomas Graave (Technische Universität Braunschweig); Zhengyang Li (Technische Universität Carolo-Wilhelmina Braunschweig); Timo Lohrenz (Technische Universität Braunschweig); Tim Fingscheidt (Technische Universität Braunschweig)

540 Enhancing Child Vocalization Classification with Phonetically-Tuned Embeddings for Assisting Autism Diagnosis

Jialu Li (UIUC); Mark A Hasegawa-Johnson (University of Illinois); Karrie Karahalios (University of Illinois at Urbana-Champaign)

717 Exploring Speech Foundation Models for Speaker Diarization in Child-Adult Dyadic Interactions

Anfeng Xu (University of Southern California); Kevin Y Huang (University of Southern California); Tiantian Feng (University of Southern California); Shen Lue (Boston University); Helen Tager-Flusberg (Boston University); Shrikanth Narayanan (USC)

992 Training speech-breathing coordination in computer-assisted reading

Delphine Charuau (GIPSA Lab); Andrea Briglia (GIPSA-lab Université de Grenoble Alpes); Erika Godde (LEAD - Université de Bourgogne); gerard baily (GIPSA-Lab/CNRS)

1095 Self-Supervised Models for Phoneme Recognition: Applications in Children's Speech for Reading Learning

Lucas Block Medin (Lalilo by Renaissance Learning); Thomas Pellegrini (IRIT); Lucile Gelin (Lalilo)

1102 Introduction to Partial fine-tuning: A comprehensive evaluation of end-to-end children's automatic speech recognition adaptation

Thomas Rolland (INESC-ID); Alberto Abad (INESC-ID/IST)

1114 Examining Vocal Tract Coordination in Childhood Apraxia of Speech with Acoustic-to-Articulatory Speech Inversion Feature Sets

Nina R Benway (University of Maryland); Jonathan L Preston (Syracuse University); Carol Y Espy-Wilson (University of Maryland)

1180 Reading Miscue Detection in Primary School through Automatic Speech Recognition

Lingyun Gao (Radboud University Nijmegen); Cristian Tejedor-Garcia (Radboud University Nijmegen); Helmer Strik (Radboud Universiteit Nijmegen); Catia Cucchiari (Radboud Universiteit Nijmegen)

1353 Benchmarking Children's ASR with Supervised and Self-supervised Speech Foundation Models

Ruchao Fan (University of California, Los Angeles); Natarajan Balaji Shankar (University of California Los Angeles); Abeer Alwan (UCLA)

1359 Preliminary Investigation of Psychometric Properties of a Novel Multimodal Dialog Based Affect Production Task in Children and Adolescents with Autism

Carly Demopoulos (UCSF); Linnea Lampinen (UCSF); Cristian Preciado (UCSF); Hardik Kothare (Modality.AI); Vikram Ramanarayanan (University of California, San Francisco & Modality.AI)

2125 Automatic Evaluation of a Sentence Memory Test for Preschool Children

Ilja Baumann (Technische Hochschule Nürnberg Georg Simon Ohm); Nicole Unger (Technische Hochschule Nürnberg Georg Simon Ohm); Dominik Wagner (Technische Hochschule Nuernberg Georg Simon Ohm); Korbinian Riedhammer (Technische Hochschule Nürnberg Georg Simon Ohm); Tobias Bocklet (TH Nürnberg)

2239 How Does Alignment Error Affect Automated Pronunciation Scoring in Children's Speech?

Prad Kadambi (Arizona State University); Tristan J Mahr (University of Wisconsin - Madison); Lucas Annear (University of Wisconsin-Madison); Henry Nomeland (University of Wisconsin-Madison); Julie Liss (Arizona State University); Katherine Hustad (University of Wisconsin - Madison); Visar Berisha (Arizona State University)

2481 Children's Speech Recognition through Discrete Token Enhancement

Vrunda N Sukhadia (QCRI); Shammur Chowdhury (QCRI)