26th Annual Conference of the International Speech Communication Association (INTERSPEECH 2025)

Rotterdam, The Netherlands
17-21 August 2025

Volume 1 of 8

ISBN: 979-8-3313-3179-5

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571

CURRAN ASSOCIATES INC.
proceedings
.com

Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by International Speech Communication Association
All rights reserved.

Printed with permission by Curran Associates, Inc. (2026)

For permission requests, please contact International Speech Communication Association
at the address below.

International Speech Communication Association
c/o Mme Emmanuelle FOXONET
4 Rue des Fauvettes - Lous Tourils
F-66390 Baixas, France

Phone: 49 228 735 643
Fax: 33 468 385 827

secretariat@isca-speech.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

TABLE OF CONTENTS

VOLUME 1

SPOKEN MACHINE TRANSLATION 1

Speech Transcription from South Tyrolean Dialect to Standard German with Whisper 1
Luca Ducceschi, Greta H. Franzini

Length Aware Speech Translation for Video Dubbing 6
Aswin Shanmugam Subramanian, Harveen Chadha, Vikas Joshi, Shubham Bansal, Jian Xue, Rupeshkumar Mehta, Jinyu Li

ArticulateX: End-To-End Monolingual Speech Translation in Articulator Space 11
Vishal Kumar, Vinayak Abrol

CMSP-ST: Cross-Modal Mixup with Speech Purification for End-To-End Speech Translation 16
Jiale Ou, Hongying Zan

End-To-End Speech Translation Guided by Robust Translation Capability of Large Language Model 21
Yosuke Higuchi, Tetsuji Ogawa, Tetsunori Kobayashi

Empowering Large Language Models for End-To-End Speech Translation Leveraging Synthetic Data 26
Yu Pu, Xiaoqian Liu, Guangyu Zhang, Zheng Yan, Wei-Qiang Zhang, Xie Chen

Speech-To-Text Translation with Phoneme-Augmented CoT: Enhancing Cross-Lingual Transfer in Low-Resource Scenarios 31
Gerard I. Gállego, Oriol Pareras, Martí Cortada Garcia, Lucas Takanori, Javier Hernando

Scheduled Interleaved Speech-Text Training for Speech-To-Speech Translation with LLMs 36
Hayato Futami, Emiru Tsunoo, Yosuke Kashiwagi, Yuki Ito, Hassan Shahmohammadi, Siddhant Arora, Shinji Watanabe

End-To-End Speech Translation for Low-Resource Languages Using Weakly Labeled Data 41
Aishwarya Pothula, Bhavana Akkiraju, Srihari Bandarupalli, Charan D, Santosh Kesiraju, Anil Kumar Vuppala

Self-Improvement for Audio Large Language Model Using Unlabeled Speech 46
Shaowen Wang, Xinyuan Chen, Yao Xu

REAL-TIME SPEECH ENHANCEMENT

Optimized Real-Time Speech Enhancement with Deep SSMs on Raw Audio 51
Yan Ru Pei, Ritik Shrivastava, Fnu Sidharth

A Two-Stage Hierarchical Deep Filtering Framework for Real-Time Speech Enhancement............................ 56
Shenghui Lu, Hukai Huang, Jinanglong Yao, Kaidi Wang, Qingyang Hong, Lin Li

Real-Time Audio-Visual Speech Enhancement Using Pre-Trained Visual Representations 61
Teng Ma, Sile Yin, Li-Chia Yang, Shuo Zhang

Lightweight Speech Enhancement Model Based on Harmonic Attention and Phase Estimation with Skin-Attachable Accelerometer 66
Yonghun Song, Yeeun Kim, Yoonyoung Chung

TSDT-Net: Ultra-Low-Complexity Two-Stage Model Combining Dual-Path-Transformer and Transform-Average-Concatenate Network for Speech Enhancement 71
Yi Gao, Hangting Chen, Siyu Zhang, Qingshan Yang, Jingcong Chen

Structured Codebook Based Hierarchical Framework for DNN for Computationally Efficient Speech Enhancement 76
Chidambar B, Hanumanth Rao Naidu

MULTILINGUALITY, CROSS-LINGUISTIC STUDIES, L2 SPEECH

Evaluation of Three Automatic Alignment Tools for the Processing of Non-Native French........ 81
Qian Zhou, Mathilde Hutin

CrossPhon: An Auto Phone Mapping Tool to Streamline Cross-Language Modeling for Phone Alignment of Low-Resource Languages 86
Hongchen Wu, Yixin Gu

Multi-Lingual and Zero-Shot Speech Recognition by Incorporating Classification of Language-Independent Articulatory Features........ 91
Ryo Magoshi, Shinsuke Sakai, Jaeyoung Lee, Tatsuya Kawahara

Instantaneous Changes in Acoustic Signals Reflect Syllable Progression and Cross-Linguistic Syllable Variation 96
Haley Hsu, Dani Byrd, Khalil Iskarous, Louis Goldstein

Influence of Proficiency and L2 Experience on Dynamic Spectral Cue Utilization in L2 Vowel Perception and Production........ 101
Linda Bakkouche, Brechtje Post

A Bayesian Approach to L2 Fluency Ratings by Native and Nonnative Listeners........ 106
Kakeru Yazawa, Takayuki Konishi

Are Loan Sequences Different from Foreign Sequences? a Perception Study with Japanese Listeners on Coronal Obstruent – High Front Vowel Sequences........111
Silke Hamann, Andrea Alicehajic

Relative Cue Weighting in Multilingual Stop Voicing Production 116
Le Xuan Chan, Annika Heuser

Variability in Intervocalic /T/ and Community Diversity in Australian English 121
Hannah White, Joshua Penney, Felicity Cox

SPEECH EMOTION RECOGNITION 1

Vector Quantized Cross-Lingual Unsupervised Domain Adaptation for Speech Emotion Recognition 126
Pravin Mote, Donita Robinson, Elizabeth Richerson, Carlos Busso

HYFuse: Aligning Heterogeneous Speech Pre-Trained Representations in Hyperbolic Space for Speech Emotion Recognition 131
Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Pailla Balakrishna Reddy, Arun Balaji Buduru, Rajesh Sharma

Meta-PerSER: Few-Shot Listener Personalized Speech Emotion Recognition Via Meta-Learning 136
Shi-Xin Fang, Liang-Yeh Shen, Yi-Cheng Lin, Huang-Cheng Chou, Hung-Yi Lee

Breaking Resource Barriers in Speech Emotion Recognition Via Data Distillation 141
Yi Chang, Zhao Ren, Zhonghao Zhao, Thanh Tam Nguyen, Kun Qian, Tanja Schultz, Björn W. Schuller

Multi-Teacher Language-Aware Knowledge Distillation for Multilingual Speech Emotion Recognition 146
Mehedi Hasan Bijoy, Dejan Porjazovski, Tamás Grósz, Mikko Kurimo

Learning More with Less: Self-Supervised Approaches forLow-Resource Speech Emotion Recognition 151
Ziwei Gong, Pengyuan Shi, Kaan Donbekci, Lin Ai, Run Chen, David Sasu, Zehui Wu, Julia Hirschberg

MULTIMODAL RESOURCES

Hearing from Silence: Reasoning Audio Descriptions from Silent Videos Via Vision-Language Model 156
Yong Ren, Chenxing Li, Le Xu, Hao Gu, Duzhen Zhang, Yujie Chen, Manjie Xu, Ruibo Fu, Shan Yang, Dong Yu

Mitigating Audiovisual Mismatch in Visual-Guide Audio Captioning.. 161
Le Xu, Chenxing Li, Yong Ren, Yujie Chen, Yu Gu, Ruibo Fu, Shan Yang, Dong Yu

ViCocktail: Automated Multi-Modal Data Collection for Vietnamese Audio-Visual Speech Recognition 166
Thai-Binh Nguyen, Thi Van Nguyen, Quoc Truong Do, Chi Mai Luong

GALAXY: A Large-Scale Open-Domain Dataset for Multimodal Learning.. 171
Yihan Wu, Yichen Lu, Yijing Chen, Jiaqi Song, William Chen, Ruihua Song, Shinji Watanabe

FD-Bench: A Full-Duplex Benchmarking Pipeline Designed for Full Duplex Spoken Dialogue Systems.. 176
Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, Eng Siong Chng

PersonaTAB: Predicting Personality Traits Using Textual, Acoustic, and Behavioral Cues in Fully-Duplex Speech Dialogs 181
Sho Inoue, Shuai Wang, Haizhou Li

FFD: Fine-Finger Diffusion Model for Music to Fine-Grained Finger Dance Generation 186
Boya Dong, Wentao Lei, Li Liu

Towards Diverse and Efficient Audio Captioning Via Diffusion Models.. 191
Manjie Xu, Chenxing Li, Yong Ren, Xinyi Tu, Ruibo Fu, Wei Liang, Dong Yu

Pull it Together: Reducing the Modality Gap in Contrastive Learning 196
Amit Sofer, Yoav Goldman, Shlomo E. Chazan

INTERPRETABILITY IN AUDIO AND SPEECH TECHNOLOGY

EnvSDD: Benchmarking Environmental Sound Deepfake Detection 201
Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, Haohe Liu, Wenwu Wang, Mark D Plumbley

Echoes of Phonetics: Unveiling Relevant Acoustic Cues for ASR Via Feature Attribution 206
Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, Luisa Bentivogli

Benchmarking Time-Localized Explanations for Audio Classification Models........211
Cecilia Bolaños, Leonardo Pepino, Martin Meza, Luciana Ferrer

Spectrotemporal Modulation: Efficient and Interpretable Feature Representation for Classifying Speech, Music, and Environmental Sounds 216
Andrew Chang, Yike Li, Iran R. Roman, David Poeppel

Discrete Tokens Exhibit Interlanguage Speech Intelligibility Benefit: An Analytical Study Towards Accent-Robust ASR Only with Native Speech Data 221
Kentaro Onda, Keisuke Imoto, Satoru Fukayama, Daisuke Saito, Nobuaki Minematsu

Analysis of Semantic and Acoustic Token Variability Across Speech, Music, and Audio Domains 226
Takanori Ashihara, Marc Delcroix, Tsubasa Ochiai, Kohei Matsuura, Shota Horiguchi

Is Your Model Big Enough? Training and Interpreting Large-Scale Monolingual Speech Foundation Models 231
Yaroslav Getman, Tamás Grósz, Tommi Lehtonen, Mikko Kurimo

Semantic-Aware Interpretable Multimodal Music Auto-Tagging 236
Andreas Patakis, Vassilis Lyberatos, Spyridon Kantarelis, Edmund Dervakos, Giorgos Stamou

From Words to Waves: Analyzing Concept Formation in Speech and Text-Based Foundation Models 241
Asim Ersoy, Basel Ahmad Mousi, Shammur Absar Chowdhury, Firoj Alam, Fahim I Dalvi, Nadir Durrani

Effective Context in Neural Speech Models 246
Yen Meng, Sharon Goldwater, Hao Tang

Word Stress in Self-Supervised Speech Models: A Cross-Linguistic Comparison 251
Martijn Bentum, Louis Ten Bosch, Tomas O. Lentz

What Do Self-Supervised Speech Models Know About Dutch? Analyzing Advantages of Language-Specific Pre-Training 256
Marianne De Heer Kloots, Hosein Mohebbi, Charlotte Pouw, Gaofei Shen, Willem Zuidema, Martijn Bentum

Iterative Refinement, Not Training Objective, Makes HuBERT Behave Differently from Wav2vec 2.0 261
Robin Huo, Ewan Dunbar

On the Reliability of Feature Attribution Methods for Speech Classification 266
Gaofei Shen, Hosein Mohebbi, Arianna Bisazza, Afra Alishahi, Grzegorz Chrupala

An Exploration of Interpretable Deep Learning Models for the Assessment of Mild Cognitive Impairment 271
Emma Cathrine Liisborg Leschly, Oliver Roesler, Michael Neumann, Jackson Liscombe, Abhishek Hosamath, Lakshmi Arbatti, Line H. Clemmensen, Melanie Ganz, Vikram Ramanarayanan

SUMMARIZATION

Towards Multi-Level Transcript Segmentation: LoRA Fine-Tuning for Table-Of-Contents Generation 276
Steffen Freisinger, Philipp Seeberger, Thomas Ranzenberger, Tobias Bocklet, Korbinian Riedhammer

Pick and Summarize: Integrating Extractive and Abstractive Speech Summarization 281
Takatomo Kano, Atsunori Ogawa, Marc Delcroix, Ryo Fukuda, William Chen, Shinji Watanabe

Beyond Similarity Scoring: Detecting Entailment and Contradiction in Multilingual and Multimodal Contexts 286
Othman Istaiteh, Salima Mdhaffar, Yannick Estève

Comparison-Based Automatic Evaluation for Meeting Summarization 291
Ziwei Gong, Lin Ai, Harsh Deshpande, Alexander Johnson, Emmy Phung, Zehui Wu, Ahmad Emami, Julia Hirschberg

SHOW AND TELL 1: ASR / TOOLS

Voxplorer: Voice Data Exploration and Projection in an Interactive Dashboard 296
Alessandro De Luca, Srikanth Madikeri, Volker Dellwo

ASR-FAIRBENCH: Measuring and Benchmarking Equity Across Speech Recognition Systems 298
Anand Rai, Satyam Rahangdale, Utkarsh Anand, Animesh Mukherjee

Transcribing Oral History Recordings Using the Transcription Portal 300
Christoph Draxler, Julian Pömp, Henk Van Den Heuvel, Fabio Ardolino, Arjan Van Hessen

LiRI Corpus Platform: Demonstration of a Web-Based Infrastructure for Multimodal Corpus Analysis 302
Teodora Vukovic, Jeremy Zehr, Jonathan Schaber, Igor Mustac, Nikolina Rajovic, Daniel McDonald, Johannes Graën, Noah Bubenhofer

Speech Annotation for A: Accuracy, Access, and Application 304
Zirong Li, Hongchen Wu, Yixin Gu, Yao Du, Yang Yue

LATE: Open Source Toolkit for Latvian and Latgalian Speech Transcription 306
Arturs Znotins, Didzis Gosko, Normunds Gruzitis

Scalable Offline ASR for Command-Style Dictation in Courtrooms 308
Kumarmanas Nethil, Vaibhav Mishra, Kriti Anandan, Kavya Manohar

MODELS OF SPEECH PRODUCTION

Towards a Dynamical Model of Transitions Between Fluent and Stuttered Speech 310
Yijing Lu, Khalil Iskarous, Louis Goldstein

Study of Vocal Fold Vibration Using M-Mode Ultrasound: A Proof of Concept 315
Juliette Dindart, Agnès Rouxel, Crystal Lin, Trung Kien Bui, Muriel Lefort, Claire Pillot-Loiseau, Christophe Trésallet, Frédérique Frouin

Articulatory Feature Prediction from Surface EMG During Speech Production.. 320
Jihwan Lee, Kevin Huang, Kleanthis Avramidis, Simon Pistrosch, Monica Gonzalez-Machorro, Yoonjeong Lee, Björn W. Schuller, Louis Goldstein, Shrikanth Narayanan

Enhancing Acoustic-To-Articulatory Speech Inversion by Incorporating Nasality... 325
Saba Tabatabaee, Suzanne Boyce, Liran Oren, Mark Tiede, Carol Espy-Wilson

SPEECH AND GRAMMAR/ARTICULATORY ANALYSES

Modeling Probabilistic Reduction Using Information Theory and Naive Discriminative Learning 330
Anna Stein, Kevin Tang

Contextual Predictability Effects on Acoustic Distinctiveness in Read Polish Speech 335
Zofia Malisz, Jan Foremski, Maglorzata Kul

How Do Both Phonological and Syntactic Complexity Influence Speech Planning? 340
Ivan Yuen, Katherine Demuth, Stefanie Shattuck-Hufnagel

When Focus Shapes the Flow: Prosodic Restructuring in Mandarin Complex Nominals............................... 344
Anqi Xu, Yu-Yin Hsu

Investigating the Impact of Word Informativeness on Speech Emotion Recognition 349
Sofoklis Kakouros

Lexical Stress Affects Lenition: The Case of Italian Palato-Alveolar Affricates.. 354
Bowei Shao, Philipp Buech, Anne Hermes, Maria Giavazzi

Evaluation of a Model for Sound Radiation from the Vocal Tract Wall 359
Peter Birkholz, Tianyi Zhang

FROST-EMA: Finnish and Russian Oral Speech Dataset of Electromagnetic Articulography
Measurements with L1, L2 and Imitated L2 Accents 364
Satu Hopponen, Tomi Kinnunen, Alexandre Nikolaev, Rosa González Hautamäki, Lauri Tavi, Einar Meister

SPEAKING STYLES, REGISTER AND CONVERSATIONAL SPEECH

Modeling Formant Dynamics in Mandarin /Ai/: Effects of Speech Style and Speech Rate............................ 369
Yunzhuo Xiang, Jingyi Sun

Representation of Perceived Prosodic Similarity of Conversational Feedback 374
Livia Qian, Carol Figueroa, Gabriel Skantze

Prolongation in Romanian 379
Oana Niculescu, Monica Vasileanu

Speech Reduction in French: The Relationship Between Vowel Space and Articulation Dynamics................ 384
Kübra Bodur, Corinne Fredouille, Christine Meunier

Stress in Spoken and Whistled Greek.. 389
Andre Batchelder-Schwab, Vasileios Michos, Jonathan Barnes

EMOTIONAL DISTRESS IN SPEECH

Leveraging Text and Speech Processing for Suicide Risk Classification in Chinese Adolescents 394
Justyna Krzywdziak, Bartlomiej Eljasiak, Joanna Stepien, Michal Swiatek, Agnieszka Pruszek

The 1st SpeechWellness Challenge: Detecting Suicide Risk Among Adolescents 399
Wen Wu, Ziyun Cui, Chang Lei, Yinan Duan, Diyang Qu, Ji Wu, Bowen Zhou, Runsen Chen, Chao Zhang

Leveraging Large Language Models for Spontaneous Speech-Based Suicide Risk Detection 404
Yifan Gao, Jiao Fu, Long Guo, Hong Liu

Predicting Adolescent Suicidal Risk from Multi-Task-Based Speech: An Ensemble Learning Approach 409
Xi Chen, Renzhe Yu, Yanshen Tan, Yiyi Li, Quan Qian, Ying Lin

In-Context Learning Capabilities of Large Language Models to Detect Suicide Risk Among Adolescents from Speech Transcripts 414
Filomene Roquefort, Alexandre Ducorroy, Rachid Riad

Language-Agnostic Suicidal Risk Detection Using Large Language Models 419
June-Woo Kim, Wonkyo Oh, Haram Yoon, Sung-Hoon Yoon, Dae-Jin Kim, Dong-Ho Lee, Sang-Yeol Lee, Chan-Mo Yang

Network of Acoustic Characteristics for the Automatic Detection of Suicide Risk from Speech. Contribution to the 2025 SpeechWellness Challenge by the Semawave Team 424
Vincent P. Martin, Charles Brazier, Maxime Amblard, Michel Musiol, Jean-Luc Rouas

PROSODY IN SPEECH SYNTHESIS

ProMode: A Speech Prosody Model Conditioned on Acoustic and Textual Inputs 429
Eray Eren, Qingju Liu, Hyeongwoo Kim, Pablo Garrido, Abeer Alwan

Counterfactual Activation Editing for Post-Hoc Prosody and Mispronunciation Correction in TTS Models 434
Kyowoon Lee, Artyom Stitsyuk, Gunu Jho, Inchul Hwang, Jaesik Choi

Investigating Stochastic Methods for Prosody Modeling in Speech Synthesis 439
Paul Mayer, Florian Lux, Alejandro Pérez-González-De-Martos, Angelina Elizarova, Lindsey Vanderlyn, Dirk Väth, Ngoc Thang Vu

GST-BERT-TTS: Prosody Prediction Without Accentual Labels for Multi-Speaker TTS Using BERT with Global Style Tokens 444
Tadashi Ogura, Takuma Okamoto, Yamato Ohtani, Erica Cooper, Tomoki Toda, Hisashi Kawai

ExagTTS: An Approach Towards Controllable Word Stress Incorporated TTS for Exaggerated Synthesized Speech Aiding Second Language Learners 449
Anindita Mondal, Monica Surtani, Anil Kumar Vuppala, Parameswari Krishnamurthy, Chiranjeevi Yarra

Synthetic Data Generation for Phrase Break Prediction with Large Language Model.................................... 454
Hoyeon Lee, Sejung Son, Ye-Eun Kang, Jong-Hwan Kim

DEPRESSION DETECTION AND ASSESSMENT 1

Speech Reference Intervals: An Assessment of Feasibility in Depression Symptom Severity Prediction 459
Lauren White, Ewan Carr, Judith Dineley, Catarina Botelho, Pauline Conde, Faith Matcham, Carolin Oetzmann, Amos Folarin, George Fairs, Agnes Norbury, Stefano Goria, Srinivasan Vairavan, Til Wykes, Richard Dobson, Vaibhav Naraya, Matthew Hotopf, Alberto Abad, Isabel Trancoso, Nicholas Cummins

DepressGEN: Synthetic Data Generation Framework for Depression Detection........ 464
Wenrui Liang, Rong Zhang, Xuezhen Zhang, Ying Ma, Wei-Qiang Zhang

Emotion-Guided Graph Attention Networks for Speech-Based Depression Detection Under Emotion-Inducting Tasks........ 469
Yuqiu Zhou, Yongjie Zhou, Yudong Yang, Yang Liu, Jun Huang, Shuzhi Zhao, Rongfeng Su, Lan Wang, Nan Yan

Explainable Depression Detection Using Masked Hard Instance Mining........ 474
Patawee Prakrankamanant, Shinji Watanabe, Ekapol Chuangsuwanich

Test-Time Training for Speech-Based Depression Detection........ 479
Sri Harsha Dumpala, Chandramouli S. Sastry, Rudolf Uher, Sageev Oore

Leveraging Ordinal Information for Speech-Based Depression Classification........ 484
Lishi Zuo, Man-Wai Mak

Zero-Shot Speech-Based Depression and Anxiety Assessment with LLMs........ 489
Erfan Loweimi, Sofia De La Fuente Garcia, Saturnino Luz

Towards the Objective Characterisation of Major Depressive Disorder Using Speech Data from a 12-Week Observational Study with Daily Measurements........ 494
Robert Lewis, Szymon Fedor, Nelson Hidalgo Julia, Joshua Curtiss, Jiyeon Kim, Noah Jones, David Mischoulon, Thomas F Quatieri, Nicholas Cummins, Paola Pedrelli, Rosalind Picard

Can Speech Accurately Detect Depression in Patients with Comorbid Dementia? an Approach for Mitigating Confounding Effects of Depression and Dementia........ 499
Sophie Young, Fuxiang Tao, Bahman Mirheidari, Madhurananda Pahar, Markus Reuber, Heidi Christensen

SPEECH ANALYSIS, DETECTION AND CLASSIFICATION 1

Temporal Convolutional Network with Smoothed and Weighted Losses for Distant Voice Activity and Overlapped Speech Detection........ 504
Shaojie Li, Qintuya Si, De Hu

Attention is Not Always the Answer: Optimizing Voice Activity Detection with Simple Feature Fusion........ 509
Kumud Tripathi, Chowdam Venkata Kumar, Pankaj Wasnik

SpeechMLC: Speech Multi-Label Classification........ 514
Miseul Kim, Seyun Um, Hyeonjin Cha, Hong-Goo Kang

Fully End-To-End Streaming Open-Vocabulary Keyword Spotting with W-CTC Forced Alignment........ 519
Dohyun Kim, Jiwook Hwang

Comparative Evaluation of Acoustic Feature Extraction Tools for Clinical Speech Analysis.......................... 524
Anna Seo Gyeong Choi, Alexander Richardson, Ryan Partlan, Sunny X. Tang, Sunghye Cho

Can We Trust Machine Learning? the Reliability of Features from Open-Source Speech Analysis Tools for Speech Modeling.......................... 529
Tahiya Chowdhury, Veronica Romero

SPEECH-BASED COGNITIVE ASSESSMENT 1

HK-GenSpeech: A Generative AI Scene Creation Framework for Speech Based Cognitive Assessment.......................... 534
Vi Jun Sean Yong, Serkan Kumyol, Pau Le Lisa Low, Suk Wai Winnie Leung, Tristan Braud

Leveraging AM and FM Rhythm Spectrograms for Dementia Classification and Assessment.......................... 539
Parismita Gogoi, Vishwanath Pratap Singh, Seema Khadirnaikar, Soma Siddhartha, Sishir Kalita, Jagabandhu Mishra, Md Sahidullah, Priyankoo Sarmah, S. R. M. Prasanna

Leveraging Cascaded Binary Classification and Multimodal Fusion for Dementia Detection Through Spontaneous Speech.......................... 544
Yin-Long Liu, Yuanchao Li, Rui Feng, Liu He, Jia-Xin Chen, Yi-Ming Wang, Yu-Ang Chen, Yan-Han Peng, Jia-Hong Yuan, Zhen-Hua Ling

Whisper-Based Multilingual Alzheimer's Disease Detection and Improvements for Low-Resource Language.......................... 549
Kaichen Jia, Jinpeng Li, Ke Li, Wei-Qiang Zhang

PPGs-BERT: Leveraging Phoneme Sequence and BERT for Alzheimer’s Disease Detection from Spontaneous Speech.......................... 554
Qi Sun, Ziyue Qiu, Yu Pu, Jinpeng Li, Xuchu Chen, Wei-Qiang Zhang

LARGE LANGUAGE MODELS IN SPEECH RECOGNITION

LLM-Based Phoneme-To-Grapheme for Phoneme-Based Speech Recognition.......................... 559
Te Ma, Min Bi, Saierdaer Yusuyin, Hao Huang, Zhijian Ou

Pinyin-Guided Chinese Speech Recognition with Large Language Model.......................... 564
Jie Zhengjie, Gaofeng Cheng

Text-Enhanced Audio Encoder for Large Language Model Based Speech Recognition Via Cross-Modality Pre-Training with Unpaired Audio-Text Data.......................... 569
Hang Su, Yuxiang Kong, Lichun Fan, Jian Luan

Towards Atypical Speech Transcription Using LLM-Based ASR.......................... 574
Jinda Zhang, Aanchan Mohan

Better Pseudo-Labeling with Multi-ASR Fusion and Error Correction by SpeechLLM.......................... 579
Jeena Prakash, Blessingh Kumar, Kadri Hacioglu, Bidisha Sharma, Sindhuja Gopalan, Malolan Chetlur, Shankar Venkatesan, Andreas Stolcke

Leveraging LLM and Self-Supervised Training Models for Speech Recognition in Chinese Dialects: A Comparative Analysis.......................... 584
Tianyi Xu, Hongjie Chen, Qing Wang, Lv Hang, Jian Kang, Jie Li, Zhennan Lin, Yongxiang Li, Lei Xie

SPEECH CODING AND ECHO CANCELLATION

Synonymity-Based Semantic Coding for Efficient Speech Compression 589
Shanhui Gan, Zijian Liang, Kai Niu, Ping Zhang

Towards an Ultra-Low-Delay Neural Audio Coding with Computational Efficiency..................................... 594
Byeong Hyeon Kim, Hyungseob Lim, Inseon Jang, Hong-Goo Kang

SpecTokenizer: A Lightweight Streaming Codec in the Compressed Spectrum Domain 599
Zixiang Wan, Guochang Zhang, Yifeng He, Jianqiang Wei

TS3-Codec: Transformer-Based Simple Streaming Single Codec 604
Haibin Wu, Naoyuki Kanda, Sefik Emre Eskimez, Jinyu Li

Towards Bitrate-Efficient and Noise-Robust Speech Coding with Variable Bitrate RVQ............................... 609
Yunkee Chae, Kyogu Lee

LSPnet: An Ultra-Low Bitrate Hybrid Neural Codec... 614
Bowen Zhang, Ian McLoughlin, Xiaoxiao Miao, As Madhukumar

Vision-Integrated High-Quality Neural Speech Coding 619
Yao Guo, Yang Ai, Rui-Chen Zheng, Hui-Peng Du, Xiao-Hang Jiang, Zhen-Hua Ling

Neural Spectral Band Generation for Audio Coding 624
Woongjib Choi, Byeong Hyeon Kim, Hyungseob Lim, Inseon Jang, Hong-Goo Kang

Multi-Channel Acoustic Echo Cancellation Based on Direction-Of-Arrival Estimation 629
Fei Zhao, Xueliang Zhang, Zhong-Qiu Wang

VOLUME 2

DECODING ALGORITHMS

Simultaneous Masked and Unmasked Decoding with Speculative Decoding Masking for Fast ASR
Without Accuracy Loss 634
Koji Okabe, Hitoshi Yamamoto

WIND: Accelerated RNN-T Decoding with Windowed Inference for Non-Blank Detection.......................... 639
Hainan Xu, Vladimir Bataev, Lilit Grigoryan, Boris Ginsburg

NGPU-LM: GPU-Accelerated N-Gram Language Model for Context-Biasing in Greedy ASR
Decoding 644
Vladimir Bataev, Andrei Andrusenko, Lilit Grigoryan, Aleksandr Laptev, Vitaly Lavrukhin, Boris Ginsburg

Pushing the Limits of Beam Search Decoding for Transducer-Based ASR Models 649
Lilit Grigoryan, Vladimir Bataev, Andrei Andrusenko, Hainan Xu, Vitaly Lavrukhin, Boris Ginsburg

Skip-Salsa: Skip Synchronous Fusion of ASR LLM Decoders 654
Ashish Mittal, Darshan Prabhu, Sunita Sarawagi, Preethi Jyothi

Efficient Trie-Based Biasing Using K-Step Prediction for Rare Word Recognition 659
Chin Yuen Kwok, Jia Qi Yip

QUEER AND TRANS SPEECH SCIENCE AND TECHNOLOGY

Web-Based Application for Real-Time Biofeedback of Vocal Resonance in Gender-Affirming Voice Training: Design and Usability Evaluation 664
Tara McAllister, Collin Eagen, Yi Shan, Peter Traver, Daphna Harel, Tae Hong Park, Vesna Novak

On the Production and Perception of a Single Speaker's Gender 669
Robin Netzorg, Naomi Carvalho, Andrea Guzman, Lydia Wang, Juliana Francis, Klo Vivienne Garoute, Keith Johnson, Gopala Anumanchipalli

Conveying Gender Through Speech: Insights from Trans Men 674
Alice Ross, Cliodhna Hughes, Eddie L. Ungless, Catherine Lai

Queer Waves: A German Speech Dataset Capturing Gender and Sexual Diversity from Podcasts and YouTube 679
Ingo Siegert, Jan Marquenie, Sven Grawunder

Reddit FlairShare: A Human-Annotated Dataset of Gender-Progressive Online Discourse 684
Carlos Hartmann

Voices of `cyborg Awesomeness': Posthuman Embodiment of Nonbinary Gender Expression in AI Speech Technologies 689
Maxwell Hope, Éva Székely

TONE

Neutral Tone Variation in Beijing Mandarin: Is Neutral Tone Toneless? 694
Xiao Dong, Fengming Liu, Chien-Jer Lin, Monica Nesbitt, Shuju Shi

The Role of Syntactic Structures in Shaping Directionality in Trisyllabic Tone Sandhi: Evidence from Tianjin Mandarin 699
Siqi Lu, Hui Feng, Ziyu Xiong

Acoustic Representation and Realization of Weak Elements Subcategories: In the Case of Tianjin Mandarin 704
Zhijie Li, Hui Feng

Lexical Competition in the Process of Cantonese Tone Merging: Diverse Impact Mechanisms Across Different Individuals and Tone Pairs 709
Lishan Li, Yaolin Zhou, Xiaoying Xu

Tonal Perception in Changde Mandarin 714
Zhenrui Zhang, Fang Hu

Tonal Contrasts in the Malipo Variety of the Mienic Language 719
Changhong Du, Fang Hu

CROSS-LINGUAL AND MULTILINGUAL PROCESSING

Building an Accurate Open-Source Hebrew ASR System Through Crowdsourcing 723
Yanir Marmor, Yair Lifshitz, Yoad Snapir, Kinneret Misgav

A Practitioner's Guide to Building ASR Models for Low-Resource Languages: A Case Study on Scottish Gaelic 728
Ondrej Klejch, William Lamb, Peter Bell

Automatic Speech Recognition for Low-Resourced Middle Eastern Languages 733
Razhan Hameed, Sina Ahmadi, Hanah Hadi, Rico Sennrich

In-Context Language Learning for Endangered Languages in Speech Recognition 738
Zhaolin Li, Jan Niehues

CS-FLEURS: A Massively Multilingual and Code-Switched Speech Dataset 743
Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Sai Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, Shinji Watanabe

Weakly Supervised Data Refinement and Flexible Sequence Compression for Efficient Thai LLM-Based ASR 748
Mingchen Shao, Xinfa Zhu, Chengyou Wang, Bingshen Mu, Hai Li, Ying Yan, Junhui Liu, Danming Xie, Lei Xie

Can We Train ASR Systems on Code-Switch Without Real Code-Switch Data? Case Study for Singapore's Languages 753
Tuan Nguyen, Huy Dat Tran

Swedish Whispers; Leveraging a Massive Speech Corpus for Swedish Speech Recognition 758
Leonora Vesterbacka, Faton Rekathati, Robin Kurtz, Justyna Sikora, Agnes Toftgård

ECHO CANCELLATION, FEEDBACK CONTROL, AND NEAR-END ENHANCEMENT

Room Impulse Response as a Prompt for Acoustic Echo Cancellation 763
Fei Zhao, Shulin He, Xueliang Zhang

CAGCRN: Real-Time Speech Enhancement with a Lightweight Model for Joint Acoustic Echo Cancellation and Noise Suppression 768
Yuyang Wang, Yonghui Liu, Jianbing Liu, Kai Niu, Zhiqiang He

Exploiting Echo Path Priors for Enhanced Stereo Acoustic Echo Cancellation 773
Jinfu Wang, Ziteng Wang, Xin Liu, Yang Liu, Qing Shi, Zhengqiang Luo, Feiran Yang

Extended Loss: Incorporating Long Context into Training Models When Using Short Audio Frames 778
Quang Minh Dinh, Hoda Rezaee Kaviani, Mehrdad Hosseinzadeh, Yuanhao Yu

Analysis and Extension of a Near-End Listening Enhancement Method Based on Long-Term Fractile Noise Statistics 783
Filippo Villani, Wai-Yip Chan, Zheng-Hua Tan, Jan Østergaard, Jesper Jensen

A Novel Deep Learning Framework for Efficient Multichannel Acoustic Feedback Control 788
Yuan-Kuei Wu, Juan Azcarreta Ortiz, Kashyap Patel, Buye Xu, Jung-Suk Lee, Sanha Lee, Ashutosh Pandey

Diffusion Buffer: Online Diffusion-Based Speech Enhancement with Sub-Second Latency 793
Bunlong Lay, Rostilav Makarov, Timo Gerkmann

PATHOLOGICAL SPEECH ANALYSIS 1

Addressing Task Conflicts in Stuttering Detection Via MMoE-Based Multi-Task Learning 798
Xiaokang Liu, Xingfeng Li, Yudong Yang, Lan Wang, Nan Yan

Comparison of Acoustic and Textual Features for Dysarthria Severity Classification in Amyotrophic Lateral Sclerosis 803
Upendra Vishwanath Y. S., Tanuka Bhattacharjee, Deekshitha G, Sathvik Udupa, Kumar Chowdam Venkata Thirumala, Madassu Keerthipriya, Darshan Chikktimmegowda, Dipti Baskar, Yamini Belur, Seena Vengalil, Atchayaram Nalini, Prasanta Kumar Ghosh

StutterCut: Uncertainty-Guided Normalised Cut for Dysfluency Segmentation 808
Suhita Ghosh, Melanie Jouaiti, Jan-Ole Perschewski, Sebastian Stober

Physiologically-Informed Feature Analysis of Acquired Speech Disorders for Stroke Assessment 813
Giulia Sanguedolce, Jón Guðnason, Dragos C. Gruia, Emilie D'Olne, Fatemeh Geranmayeh, Patrick A. Naylor

HEARING DISORDERS

Robot-Assisted Recognition of Vocal Emotions in Pseudospeech for Cochlear Implanted Adolescents 818
Gloria Araiza-Illan, Luke Meyer, Bert Maat, Deniz Baskent

Using Neurogram Similarity Index Measure (NSIM) to Model Hearing Loss and Cochlear Neural Degeneration 823
Ahsan Cheema, Sunil Puria

Contrastive Learning-Based Syllable-Level Mispronunciation Detection and Diagnosis for Speech Audiometry 828
Longbin Jin, Donghun Min, Jung Eun Shin, Eun Yi Kim

A Deformable Convolution GAN Approach for Speech Dereverberation in Cochlear Implant Users 833
Hsin-Tien Chiang, John H. L. Hansen

L3C-DeepMFC: Low-Latency Low-Complexity Deep Marginal Feedback Cancellation with Closed-Loop Fine Tuning for Hearing Aids 838
Fengyuan Hao, Brian C. J. Moore, Huiyong Zhang, Xiaodong Li, Chengshi Zheng

Semantic Processing During Spoken Word Production by Children with Cochlear Implants 843
Man Wang, Yixin Ding, Niels Schiller

Linguistic Masking and Its Release in Simulated Electric-Acoustic Hearing 848
Yuting Ding, Xuefei Wang, Fei Chen

INTERSPEECH 2025 URGENT CHALLENGE

Lessons Learned from the URGENT 2024 Speech Enhancement Challenge 853
Wangyou Zhang, Kohei Saijo, Samuele Cornell, Robin Scheibler, Chenda Li, Zhaoheng Ni, Anurag Kumar, Marvin Sach, Wei Wang, Yihui Fu, Shinji Watanabe, Tim Fingscheidt, Yanmin Qian

Interspeech 2025 URGENT Speech Enhancement Challenge 858
Kohei Saijo, Wangyou Zhang, Samuele Cornell, Robin Scheibler, Chenda Li, Zhaoheng Ni, Anurag Kumar, Marvin Sach, Yihui Fu, Wei Wang, Tim Fingscheidt, Shinji Watanabe

TS-URGENet: A Three-Stage Universal Robust and Generalizable Speech Enhancement Network 863
Xiaobin Rong, Dahan Wang, Qinwen Hu, Yushi Wang, Yuxiang Hu, Jing Lu

Multistage Universal Speech Enhancement System for URGENT Challenge 868
Xiaohuai Le, Zhuangqi Chen, Siyu Sun, Xianjun Xia, Chuanzeng Huang

Scaling Beyond Denoising: Submitted System and Findings in URGENT Challenge 2025 873
Zhihang Sun, Andong Li, Tong Lei, Rilin Chen, Meng Yu, Chengshi Zheng, Yi Zhou, Dong Yu

DeepFilterGAN: A Full-Band Real-Time Speech Enhancement System with GAN-Based Stochastic Regeneration 878
Sanberk Serbest, Tijana Stojkovic, Milos Cernak, Andrew Harper

FUSE: Universal Speech Enhancement Using Multi-Stage Fusion of Sparse Compression and Token Generation Models for the URGENT 2025 Challenge 883
Nabarun Goswami, Tatsuya Harada

Universal Speech Enhancement with Regression and Generative Mamba 888
Rong Chao, Rauf Nasretdinov, Yu-Chiang Frank Wang, Ante Jukic, Szu-Wei Fu, Yu Tsao

SPOKEN MACHINE TRANSLATION 2

Structured Pruning for Efficient Systolic Array Accelerated Cascade Speech-To-Text Translation 893
Jean-Luc Rouas, Charles Brazier, Leila Ben Letaifa, Rafael Medina, Pedro Palacios, David Atienza, Giovanni Ansaloni

Scaling Pseudo-Labeling Data for End-To-End Low-Resource Speech Translation (the Case of Kurdish Language) 898
Mohammad Mohammadamini, Aghilas Sini, Marie Tahon, Antoine Laurent

Multilingual Query-By-Example KWS for Indian Languages Using Transliteration 903
Kirandevraj R, Vinod K Kurmi, Vinay Namboodiri, Cv Jawahar

Novel Parasitic Dual-Scale Modeling for Efficient and Accurate Multilingual Speech Translation 908
Chenyang Le, Yinfeng Xia, Huiyan Li, Manhong Wang, Yutao Sun, Xingyang Ma, Yanmin Qian

A Multi-Dialectal Dataset for German Dialect ASR and Dialect-To-Standard Speech Translation 913
Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, Barbara Plank

NIRANTAR: Continual Learning with New Languages and Domains on Real-World Speech Data 918
Tahir Javed, Kaushal Bhogale, Mitesh M. Khapra

SPATIAL AUDIO AND ACOUSTICS 1

Temporal Modeling of Room Impulse Response Generation Via Multi-Scale Autoregressive Learning 923
Sheng Lyu, Yuemin Yu, Chenshu Wu

Effect of Noise Floor in Room Impulse Response on Speech Perception Under Spherical Harmonics-Based Spatial Sound Reproduction 928
Yunqi C. Zhang, Dhruv Jagmohan, Hong Kit Li, C. T. Justine Hui, Yusuke Hioka

Direction-Aware Neural Acoustic Fields for Few-Shot Interpolation of Ambisonic Impulse Responses 933
Christopher Ick, Gordon Wichern, Yoshiki Masuyama, François G. Germain, Jonathan Le Roux

AuralNet: Hierarchical Attention-Based 3D Binaural Localization of Overlapping Speakers 938
Linya Fu, Yu Liu, Zhijie Liu, Zedong Yang, Zhong-Qiu Wang, Youfu Li, He Kong

SoundSculpt: Direction and Semantics Driven Ambisonic Target Sound Extraction 943
Tuochao Chen, D Shin, Hakan Erdogan, Sinan Hersek

TF-Mamba: A Time-Frequency Network for Sound Source Localization 948
Yang Xiao, Rohan Kumar Das

ARTICULATORY AND VOCAL TRACT MODELLING

Articulatory Modeling of the S-Shaped F2 Trajectories Observed in Öhman's Spectrographic Analysis of VCV Syllables 953
Frédéric Berthommier

The Role of Voiced Consonant Duration in Sung Vowel-Consonant and Consonant-Vowel Recognition 958
Allan Vurma, Einar Meister, Lya Meister, Jaan Ross, Marju Raju, Veeda Kala, Tuuri Dede

How Sibilant Spectra Shape Gender Perception in Prepubertal Children: A Voice Morphing Study 963
Riccarda Funk, Melanie Weirich, Adrian Simpson

Constrained LDDMM for Dynamic Vocal Tract Morphing: Integrating Volumetric and Real-Time MRI 968
Tharinda Piyadasa, Joan Glaunès, Amelia Gully, Michael Proctor, Kirrie Ballard, Tünde Szalay, Naeim Sanaei, Sheryl Foster, David Waddington, Craig Jin

2D Immersed Boundary Method in Vocal Tract Acoustics: An Eulerian–Lagrangian Model for Simulation of Diphthongs 973
Rongshuai Wu, Debasish Ray Mohapatra, Sidney Fels

Reconstruction of the Complete Vocal Tract Contour Through Acoustic to Articulatory Inversion Using Real-Time MRI Data 978
Sofiane Azzouz, Pierre-André Vuissoz, Yves Laprie

Co-Registration of Real-Time MRI and Respiration for Speech Research 983
Yubin Zhang, Prakash Kumar, Ye Tian, Ziwei Zhao, Xuan Shi, Kevin Huang, Kevin Lee, Haley Hsu, Shrikanth Narayanan, Krishna Nayak, Louis Goldstein

ACOUSTIC ASSESSMENT OF RESPIRATORY HEALTH

SPEAKtoCOPD: A Flashmob Study to Collect COPD Speech 988
Loes Van Bemmel, Lauren G Reinders, Folkert Brijker, Bas Holverda, Frits M. E. Franssen, Hanneke Van Helvoort, Visara Urovi, Marieke Spreeuwenberg, Sami O. Simons

Developing a LeFF Transformer Model for Exacerbated Speech Detection in COPD and Asthma 993
Yuyang Yan, Sami O. Simons, Visara Urovi

Towards Pre-Training an Effective Respiratory Audio Foundation Model 998
Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, Noboru Harada

Effect of Physical Exercise on Voice in People Living with COPD 1003
Lauren G Reinders, Loes Van Bemmel, Alexander Mackay, David Nobbs, Frits M. E. Franssen, Hester Gietema, Simona Schäfer, Sami O. Simons

Adaptive Differential Denoising for Respiratory Sounds Classification 1008
Gaoyang Dong, Zhicheng Zhang, Ping Sun, Minghui Zhang

Disentangling Dual-Encoder Masked Autoencoder for Respiratory Sound Classification 1013
Peidong Wei, Shiyu Miao, Lin Li

Patient-Aware Feature Alignment for Robust Lung Sound Classification: Cohesion-Separation and Global Alignment Losses 1018
Seung Gyu Jeong, Seong Eun Kim

Improving Respiratory Sound Classification with Architecture-Agnostic Knowledge Distillation from Ensembles 1023
Miika Toikkanen, June-Woo Kim

ADVANCES IN MODELLING AND IMAGING

Theoretical Proposal for a Unified Bayesian Model of Adaptation in Non-Interactive and Interactive Speech Production 1028
Mélen Guillaume, Anahita Basirat, Julien Diard

Self-Supervised Optimality-Guided Learning of Speech Articulation 1033
Juraj Šimko, Benjamin Elie, Alice Turk

Extended High-Frequency Cues to Phoneme Recognition: Insights from ASR 1038
Zhe-Chen Guo, Bharath Chandrasekaran

Decoding Speaker-Normalized Pitch from EEG for Mandarin Perception 1043
Jia-Xin Chen, Yi-Ming Wang, Ziyu Zhang, Jiayang Han, Yin-Long Liu, Rui Feng, Xiuyuan Liang, Zhen-Hua Ling, Jia-Hong Yuan

SSF-DST: A Spectro-Spatial Features Enhanced Deep Spatiotemporal Network for EEG-Based Auditory Attention Detection 1048
Tong Zhu, Xiaoke Yang, Jian Zhou, Lu Li, Zhao Lv, Cunhang Fan

Overestimated Performance of Auditory Attention Decoding Caused by Experimental Design in EEG Recordings 1053
Yujie Yan, Xiran Xu, Haolin Zhu, Songyi Li, Bo Wang, Xihong Wu, Jing Chen

A Real-Time MRI Study on Asymmetry in Velum Dynamics During VCV Production with Nasal Sounds 1058
Chetan Sharma, Vaishnavi Chandwanshi, Shreya Shrikant Karkun, Aditya Anand Gupta, Prasanta Kumar Ghosh

Exploratory Analysis of Brainstem fMRI Data During Sustained Phonation 1063
Carey Smith, Hu Cheng, Pertti Palo, Daniel Aalto, Steven M. Lulich

CONVERSATION, COMMUNICATION AND INTERACTION 1

Gaze-Enhanced Multimodal Turn-Taking Prediction in Triadic Conversations 1068
Seongsil Heo, Christi Miller, Calvin Murdock, Michael Proulx

Visual Cues Support Robust Turn-Taking Prediction in Noise... 1073
Sam O'Connor Russell, Naomi Harte

Backchannel Prediction for Natural Spoken Dialog Systems Using General Speaker and Listener Information 1078
Yoshinori Fukunaga, Ryota Nishimura, Kengo Ohta, Norihide Kitaoka

Rapport-Building Dialogue Strategies for Deeper Connection: Integrating Proactive Behavior, Personalization, and Aizuchi Backchannels 1083
Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, Koichiro Yoshino

Does Effortful Speech Production Indicate Communication Difficulty Caused by Noise and Hearing Aid Support? 1088
Lena-Marie Huttner, Jeppe H. Christensen, Gitte Keidser, Tobias May, Torsten Dau, Sergi Rotger-Griful

``Dyadosyncrasy'', Idiosyncrasy and Demographic Factors in Turn-Taking 1093
Julio Cesar Cavalcanti, Gabriel Skantze

ROBUST SPEAKER VERIFICATION

SSPS: Self-Supervised Positive Sampling for Robust Self-Supervised Speaker Verification 1098
Theo Lepage, Reda Dehak

ParaNoise-SV: Integrated Approach for Noise-Robust Speaker Verification with Parallel Joint Learning of Speech Enhancement and Noise Extraction..1103
Minu Kim, Kangwook Jang, Hoirin Kim

Disentangling Speaker and Content in Pre-Trained Speech Models with Latent Diffusion for Robust Speaker Verification 1108
Zhe Li, Man-Wai Mak, Jen-Tzung Chien, Mert Pilanci, Zezhong Jin, Helen Meng

Evaluating Deep Speaker Embedding Robustness to Domain, Sampling Rate, and Codec Variations 1113
Alexandre Ferro Filho, Diogo Fernandes Costa Silva, Pedro Elias Engelberg Silva Borges, Arlindo Rodrigues Galvão Filho

Towards Robust Speaker Recognition Against Intrinsic Variation with Foundation Model Few-Shot Tuning and Effective Speech Synthesis..1118
Zhiyong Chen, Shuhang Wu, Xinnuo Li, Zhiqi Ai, Shugong Xu

Bayesian Learning for Domain-Invariant Speaker Verification and Anti-Spoofing 1123
Jin Li, Man-Wai Mak, Johan Rohdin, Kong Aik Lee, Hynek Hermansky

MULTILINGUAL ASR

Switch Conformer with Universal Phonetic Experts for Multilingual ASR 1128
Masato Mimura, Jaeyoung Lee, Tatsuya Kawahara

Language-Aware Prompt Tuning for Parameter-Efficient Seamless Language Expansion in Multilingual ASR 1133
Hongli Yang, Sheng Li, Hao Huang, Ayiduosi Tuohan, Yizhou Peng

Efficient Multilingual ASR Finetuning Via LoRA Language Experts....................1138
Jiahong Li, Yiwen Shao, Jianheng Zhuo, Chenda Li, Liliang Tang, Dong Yu, Yanmin Qian

Mixture of LoRA Experts for Low-Resourced Multi-Accent Automatic Speech Recognition 1143
Raphaël Bagat, Irina Illina, Emmanuel Vincent

Effects of Speaker Count, Duration, and Accent Diversity on Zero-Shot Accent Robustness in Low-Resource ASR....................1148
Zheng Xin Yong, Vineel Pratap, Michael Auli, Jean Maillard

Leveraging Geographic Metadata for Dialect-Aware Speech Recognition 1153
Pouya Mehralian, Hugo Van Hamme

Overcoming Data Scarcity in Multi-Dialectal Arabic ASR Via Whisper Fine-Tuning 1158
Ömer Tarik Özyilmaz, Matt Coler, Matias Valdenegro-Toro

VietASR: Achieving Industry-Level Vietnamese ASR with 50-Hour Labeled Data and Large-Scale Speech Pretraining....................1163
Jianheng Zhuo, Yifan Yang, Yiwen Shao, Yong Xu, Dong Yu, Kai Yu, Xie Chen

Open Universal Arabic ASR Leaderboard....................1168
Yingzhi Wang, Anas Alhmoud, Muhammad Alqurishi

MULTI-CHANNEL SPEECH ENHANCEMENT

Mel-McNet: A Mel-Scale Framework for Online Multichannel Speech Enhancement 1173
Yujie Yang, Bing Yang, Xiaofei Li

A Lightweight Hybrid Dual Channel Speech Enhancement System Under Low-SNR Conditions 1178
Zheng Wang, Xiaobin Rong, Yu Sun, Tianchi Sun, Zhibin Lin, Jing Lu

ARiSE: Auto-Regressive Multi-Channel Speech Enhancement....................1183
Pengjie Shen, Xueliang Zhang, Zhong-Qiu Wang

WTFormer: A Wavelet Conformer Network for MIMO Speech Enhancement with Spatial Cues Peservation 1188
Lu Han, Junqi Zhao, Renhua Peng

A Three-Stage Beamforming with Harmonic Guidance for Multi-Channel Speech Enhancement....................1193
Nurali Alip, Tianrui Wang, Rui Cao, Meng Ge, Jingru Lin, Longbiao Wang, Jianwu Dang

Speech Enhancement with Dual-Path Multi-Channel Linear Prediction Filter and Multi-Norm Beamforming....................1198
Chengyuan Qin, Wenmeng Xiong, Jing Zhou, Maoshen Jia, Changchun Bao

SELF-SUPERVISED LEARNING

Exploring SSL Discrete Speech Features for Zipformer-Based Contextual ASR 1203
Mingyu Cui, Yifan Yang, Jiajun Deng, Jiawen Kang, Shujie Hu, Tianzi Wang, Zhaoqing Li, Shiliang Zhang, Xie Chen, Xunying Liu

Self-Supervised Learning of Speech Representations with Dutch Archival Data 1208
Nik Vaessen, Roeland Ordelman, David A. Van Leeuwen

GigaAM: Efficient Self-Supervised Learner for Speech Recognition.. 1213
Aleksandr Kutsakov, Alexandr Maximenko, Georgii Gospodinov, Pavel Bogomolov, Fyodor Minkin

DiceHuBERT: Distilling HuBERT with a Self-Supervised Learning Objective 1218
Hyung-Gun Chi, Zakaria Aldeneh, Tatiana Likhomanenko, Oggi Rudovic, Takuya Higuchi, Li-Wei Chen, Shinji Watanabe, Ahmed Hussen Abdelaziz

Differentiable K-Means for Fully-Optimized Discrete Token-Based ASR 1223
Kentaro Onda, Yosuke Kashiwagi, Emiru Tsunoo, Hayato Futami, Shinji Watanabe

Towards Early Prediction of Self-Supervised Speech Model Performance.. 1228
Ryan Whetten, Lucas Maison, Titouan Parcollet, Marco Dinarelli, Yannick Estève

SINGING VOICE AND AUDIO SYNTHESIS

VibE-SVC: Vibrato Extraction with High-Frequency F0 Contour for Singing Voice Conversion................ 1233
Joon-Seung Choi, Dong-Min Byun, Hyung-Seok Oh, Seong-Whan Lee

TVC-MusicGen: Time-Varying Structure Control for Background Music Generation Via Self-Supervised Training... 1238
Chenyu Yang, Hangting Chen, Shuai Wang, Haina Zhu, Haizhou Li

Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-To-Audio with Retrieval-Augmented Generation.. 1243
Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, Andros Tjandra

Bridging Speech and Singing: Multi-Stage Speech-Prompted Singing Voice Conversion with Speaker Embedding Adaptation 1248
Mingda Liu, Jiatong Shi

Neurodyne: Neural Pitch Manipulation with Representation Learning and Cycle-Consistency GAN 1253
Yicheng Gu, Chaoren Wang, Zhizheng Wu, Lauri Juvela

VS-Singer: Vision-Guided Stereo Singing Voice Synthesis with Consistency Schrödinger Bridge 1258
Zijing Zhao, Kai Wang, Hao Huang, Ying Hu, Liang He, Jichen Yang

DAFMSVC: One-Shot Singing Voice Conversion with Dual Attention Mechanism and Flow Matching... 1263
Wei Chen, Binzhu Sha, Dan Luo, Jing Yang, Zhuo Wang, Fan Fan, Zhiyong Wu

Simple and Effective Content Encoder for Singing Voice Conversion Via SSL-Embedding Dimension Reduction 1268
Wangjin Zhou, Tianjiao Du, Chenglin Xu, Sheng Li, Yi Zhao, Tatsuya Kawahara

Song Form-Aware Full-Song Text-To-Lyrics Generation with Multi-Level Granularity Syllable Count Control 1273
Yunkee Chae, Eunsik Shin, Suntae Hwang, Seungryeol Paik, Kyogu Lee

ACOUSTIC AND ARTICULATORY CUES IN SPEECH PERCEPTION

Multitalker Babble in English Vowel Perception Training: A Comparison Between Humans and Neural Models 1278
Wenwei Dong, Alif Silpachai, Catia Cucchiarini, Helmer Strik

Speech Stimulus Design to Study the Neural Coding of Speech and the Impact of Cochlear Synaptopathy 1283
Etienne Gaudrain, Sarah Verhulst, Deniz Baskent

Prediction of Listening Effort Ratings for Habitual and clear-Lombard Speech Presented in Noise 1288
Esther Janse, Chen Shen, Martin Cooke

Language and Accent Familiarity Effects on the Use of Acoustic Cues in Talker Identification 1293
Shengyue Xiong, Zhe-Chen Guo, Bharath Chandrasekaran

Characterization of Voice Cue Sensitivity and Vocal Emotion Recognition Across the Adult Lifespan 1298
Laura Rachman, Deniz Baskent

Creaky Voice Facilitates More Efficient Phonological Processing of Mandarin Tone 3 1303
Zixia Fan, Ronny Ibrahim, Joshua Penney, Felicity Cox

AUDIO EVENT DETECTION AND CLASSIFICATION

Training Onset-And-Offset-Aware Sound Event Detection on a Heterogeneous Dataset Via Probabilistic Sequential Modeling........ 1308
Tomoya Yoshinaga, Yoshiaki Bando, Keitaro Tanaka, Keisuke Imoto, Masaki Onishi, Shigeo Morishima

Multi-View Fusion and Parameter Perturbation for Few-Shot Class-Incremental Audio Classification 1313
Yulu Fang, Mingyue He, Qisheng Xu, Jianqiao Zhao, Cheng Yang, Kele Xu, Yong Dou

Fully Few-Shot Class-Incremental Audio Classification Using Multi-Level Embedding Extractor and Ridge Regression Classifier........ 1318
Yongjie Si, Yanxiong Li, Jiaxin Tan, Qianhua He, Il-Youp Kwak

Beyond Conventional Metrics: Using Entropic Triangles to Explain Balancing Methods in Acoustic Scene Classification........ 1323
Claudia Montero-Ramírez, Alba Martínez-Serrano, Jorge Garcelán-Gómez, Francisco J. Valverde-Albacete, Carmen Peláez-Moreno

Domain Adaptation Method and Modality Gap Impact in Audio-Text Models for Prototypical Sound Classification 1328
Emiliano Acevedo, Martín Rocamora, Magdalena Fuentes

Unified Microphone Conversion: Many-To-Many Device Mapping Via Feature-Wise Linear Modulation 1333
Myeonghoon Ryu, Hongseok Oh, Suji Lee, Han Park

INCLUSIVITY

The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages.......... 1338
Chris Emezue, The Naijavoices Community, Busayo Awobade, Abraham Toluwase Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, Chris Pal

FaiST: A Benchmark Dataset for Fairness in Speech Technology.......... 1343
Maliha Jahan, Yinglun Sun, Priyam Mazumdar, Zsuzsanna Fagyal, Thomas Thebaud, Jesus Villalba, Mark Hasegawa-Johnson, Najim Dehak, Laureano Moro Velazquez

On the Language and Gender Biases in PSTN, VoIP and Neural Audio Codecs 1348
Kemal Altwlkany, Amar Kuric, Emanuel Lacic

Evaluating Speech Enhancement Performance Across Demographics and Language 1353
Jose Giraldo, Alex Peiró-Lilja, Carme Armentano-Oller, Rodolfo Zevallos, Cristina España-Bonet

VOICE CONVERSION 1

Towards Better Disentanglement in Non-Autoregressive Zero-Shot Expressive Voice Conversion.......... 1358
Seymanur Akti, Tuan-Nam Nguyen, Alexander Waibel

Voice Conversion for Likability Control Via Automated Rating of Speech Synthesis Corpora.......... 1363
Hitoshi Suda, Shinnosuke Takamichi, Satoru Fukayama

REWIND: Speech Time Reversal for Enhancing Speaker Representations in Diffusion-Based Voice Conversion.......... 1368
Ishan D. Biyani, Nirmesh J. Shah, Ashishkumar P. Gudmalwar, Pankaj Wasnik, Rajiv R. Shah

VOLUME 3

Training-Free Voice Conversion with Factorized Optimal Transport.......... 1373
Alexander Lobashev, Assel Yermekova, Maria Larchenko

E2E-BPVC: End-To-End Background-Preserving Voice Conversion Via In-Context Learning.......... 1378
Yihan Liu, Zhengyang Chen, Leying Zhang, Yanmin Qian

Discl-VC: Disentangled Discrete Tokens and In-Context Learning for Controllable Zero-Shot Voice Conversion.......... 1383
Kaidi Wang, Wenhao Guan, Ziyue Jiang, Hukai Huang, Peijie Chen, Weijie Wu, Qingyang Hong, Lin Li

ReFlow-VC: Zero-Shot Voice Conversion Based on Rectified Flow and Speaker Feature Optimization.......... 1388
Pengyu Ren, Wenhao Guan, Kaidi Wang, Peijie Chen, Qingyang Hong, Lin Li

In This Environment, as that Speaker: A Text-Driven Framework for Multi-Attribute Speech Conversion.......... 1393
Jiawei Jin, Zhihan Yang, Yixuan Zhou, Zhiyong Wu

LinearVC: Linear Transformations of Self-Supervised Features Through the Lens of Voice Conversion 1398
Herman Kamper, Benjamin Van Niekerk, Julian Zaïdi, Marc-André Carbonneau

Speaker Normalization and Content Restoration for Zero-Shot Voice Conversion with Attention-Enhanced Discriminator 1403
Desheng Hu, Yang Xiang, Jian Lu, Xinhui Hu, Xinkang Xu

SPEECH-BASED COGNITIVE ASSESSMENT 2

Optimizing Pause Context in Fine-Tuning Pre-Trained Large Language Models for Dementia Detection 1408
Xiaoquan Ke, Man-Wai Mak, Helen Meng

WhisperD: Dementia Speech Recognition and Filler Word Detection with Whisper 1413
Emmanuel Akinrintoyo, Nadine Abdelhalim, Nicole Salomons

Acoustic and Linguistic Biomarkers for Cognitive Impairment Detection from Speech 1418
Catarina Botelho, David Gimeno-Gómez, Francisco Teixeira, John Mendonça, Patrícia Pereira, Diogo A. P. Nunes, Thomas Rolland, Anna Pompili, Rubén Solera-Ureña, Maria Ponte, David Martins De Matos, Carlos-D. Martínez-Hinarejos, Isabel Trancoso, Alberto Abad

Alzheimer’s Dementia Detection Using Perplexity from Paired Large Language Models 1423
Yao Xiao, Heidi Christensen, Stefan Goetze

Understanding Dementia Speech Alignment with Diffusion-Based Image Generation 1428
Mansi, Anastasios Lepipas, Dominika C Woszczyk, Yiying Guan, Soteris Demetriou

ClaritySpeech: Dementia Obfuscation in Speech 1433
Dominika C Woszczyk, Ranya Aloufi, Soteris Demetriou

SOURCE SEPARATION 1

Quadruple Path Modeling with Latent Feature Transfer for Permutation-Free Continuous Speech Separation 1438
Jihyun Kim, Doyeon Kim, Hyewon Han, Jinyoung Lee, Jonguk Yoo, Chang Woo Han, Jeongook Song, Hoon-Young Cho, Hong-Goo Kang

End-To-End DOA-Guided Speech Extraction in Noisy Multi-Talker Scenarios 1443
Kangqi Jing, Wenbin Zhang, Yu Gao

Speaker Separation for an Unknown Number of Speakers with Encoder-Decoder-Based Contextual Information Module 1448
Xue Yang, Guiru Shen, Yu Yang

Attractor-Based Speech Separation of Multiple Utterances by Unknown Number of Speakers 1453
Yuzhu Wang, Archontis Politis, Konstantinos Drossos, Tuomas Virtanen

ReSepNet: A Unified-Light Model for Recursive Speech Separation with Unknown Speaker Count 1458
Hadi Alizadeh, Rahil Mahdian Toroghi, Hassan Zareian

Deep-Simplex Multichannel Speech Separation 1463
Tzlil Avidan, Bracha Laufer-Goldshtein

FLASepformer: Efficient Speech Separation with Gated Focused Linear Attention Transformer 1468
Haoxu Wang, Yiheng Jiang, Gang Qiao, Pengteng Shi, Biao Tian

Power Spectral Density Estimation for Acoustic Source Separation Using a Spherical Microphone Array 1473
Liang Tao, Maoshen Jia, Yonggang Hu

Exploring Efficient Directional and Distance Cues for Regional Speech Separation 1478
Yiheng Jiang, Haoxu Wang, Yafeng Chen, Gang Qiao, Biao Tian

LANGUAGE AND ACCENT IDENTIFICATION AND SPEAKER PRIVACY

Teacher-Free Knowledge Distillation for Improving Short-Utterance Spoken Language Identification 1483
Spandan Dey, Hirak Mondal, Sanjay Kumar Kurmi

LID Models Are Actually Accent Classifiers: Implications and Solutions for LID on Accented Speech 1488
Niyati Bafna, Matthew Wiesner

Analyzing the Impact of Accent on English Speech: Acoustic and Articulatory Perspectives 1493
Gowtham Premananth, Vinith Kugathasan, Carol Espy-Wilson

A Study of Speech Embedding Similarities Between Australian Aboriginal and High-Resource Languages 1498
Eliathamby Ambikairajah, Jingyao Wu, Ting Dang, Vidhyasaharan Sethu

An Investigative Study on Recent Sharpness- And Flatness-Based Optimizers for Enhanced Self-Supervised Speaker Verification 1503
Abderrahim Fathan, Jahangir Alam, Xiaolin Zhu

Privacy-Preserving Speaker Verification Via End-To-End Secure Representation Learning 1508
Chenguang Hu, Yaqian Hao, Fulin Zhang, Xiaoxue Luo, Yao Shen, Yingying Gao, Chao Deng, Shilei Zhang, Junlan Feng

Novel Loss-Enhanced Universal Adversarial Patches for Sustainable Speaker Privacy 1513
Elvir Karimov, Alexander Varlamov, Danil Ivanov, Dmitrii Korzh, Oleg Rogov

Federated Learning with Feature Space Separation for Speaker Recognition 1518
Ying Meng, Zhihua Fang, Liang He

SOURCE TRACING: THE ORIGINS OF SYNTHETIC OR MANIPULATED SPEECH

Audio Deepfake Source Tracing Using Multi-Attribute Open-Set Identification and Verification 1523
Pierre Falez, Tony Marteau, Damien Lolive, Arnaud Delhay

Unveiling Audio Deepfake Origins: A Deep Metric Learning and Conformer Network Approach with Ensemble Fusion 1528
Ajinkya Kulkarni, Sandipana Dowerah, Tanel Alumäe, Mathew Magimai Doss

Codec-Based Deepfake Source Tracing Via Neural Audio Codec Taxonomy 1533
Xuanjun Chen, I-Ming Lin, Lin Zhang, Jiawei Du, Haibin Wu, Hung-Yi Lee, Jyh-Shing Roger Jang

TADA: Training-Free Attribution and Out-Of-Domain Detection of Audio Deepfakes 1538
Adriana Stan, David Combei, Dan Oneata, Horia Cucu

Source Verification for Speech Deepfakes 1543
Viola Negroni, Davide Salvi, Paolo Bestagini, Stefano Tubaro

STOPA: A Dataset of Systematic VariaTion of DeePfake Audio for Open-Set Source Tracing and Attribution 1548
Anton Firc, Manasi Chhibber, Jagabandhu Mishra, Vishwanath Pratap Singh, Tomi Kinnunen, Kamil Malinka

Synthetic Speech Source Tracing Using Metric Learning 1553
Dimitrios Koutsianos, Stavros Zacharopoulos, Yannis Panagakis, Themos Stafylakis

Listen, Analyze, and Adapt to Learn New Attacks: An Exemplar-Free Class Incremental Learning Method for Audio Deepfake Source Tracing 1558
Yang Xiao, Rohan Kumar Das

VIB-Based Real Pre-Emphasis Audio Deepfake Source Tracing 1563
Thien-Phuc Doan, Kihun Hong, Souhwan Jung

Defending Unauthorized Voice Cloning with Watermark-Aware Codecs 1568
Jiankun Zhao, Lingwei Meng, Chengxi Deng, Helen Meng, Xixin Wu

Open-Set Source Tracing of Audio Deepfake Systems 1573
Nicholas Klein, Hemlata Tak, Elie Khoury

SPEAKER DIARIZATION 1

Fine-Tune Before Structured Pruning: Towards Compact and Accurate Self-Supervised Models for Speaker Diarization 1578
Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, Jan Cernocký, Lukáš Burget

Count Your Speakers! Multitask Learning for Multimodal Speaker Diarization 1583
Prabhav Singh, Jesus Villalba, Najim Dehak

End-To-End Diarization Utilizing Attractor Deep Clustering 1588
David Palzer, Matthew Maciejewski, Eric Fosler-Lussier

SDBench: A Comprehensive Benchmark Suite for Speaker Diarization 1593
Berkin Durmus, Blaise Munyampirwa, Eduardo Pacheco, Atila Orhon, Andrey Leonov

Enhancing Serialized Output Training for Multi-Talker ASR with Soft Monotonic Alignment and Utterance-Level Timestamp 1598
Fengyun Tan, Tao Wei, Kun Zou, Ning Cheng, Shaojun Wang, Jing Xiao

Pretraining Multi-Speaker Identification for Neural Speaker Diarization 1603
Shota Horiguchi, Atsushi Ando, Naohiro Tawara, Marc Delcroix

MULTILINGUAL SPEECH SYNTHESIS AND SPECIAL APPLICATIONS 1

Parameter-Efficient Fine-Tuning for Low-Resource Text-To-Speech Via Cross-Lingual Continual Learning 1608
Ki-Joong Kwon, Jun-Ho So, Sang-Hoon Lee

Accent Normalization Using Self-Supervised Discrete Tokens with Non-Parallel Data................................ 1613
Qibing Bai, Sho Inoue, Shuai Wang, Zhongjie Jiang, Yannan Wang, Haizhou Li

LIST: Language-Independent Speech Token for Multilingual Speech Synthesis with Language Models.. 1618
Chang Liu, Zhen-Hua Ling, Yu Gu

Developing High-Quality TTS for Punjabi and Urdu: Benchmarking Against MMS Models....................... 1623
Fatima Naseem, Maham Sajid, Farah Adeeba, Sahar Rauf, Asad Mustafa, Sarmad Hussain, Faisal Kamiran

Synthesizing Speech with Selected Perceptual Voice Qualities – a Case Study with Creaky Voice 1628
Frederik Rautenberg, Fritz Seebauer, Jana Wiechmann, Michael Kuhlmann, Petra Wagner, Reinhold Haeb-Umbach

Intrasentential English in Swedish TTS: Perceived English-Accentedness... 1633
Christina Tånnander, David House, Jonas Beskow, Jens Edlund

CHARACTERIZATION AND MULTIMODAL APPROACHES FOR SPEAKER RECOGNITION

Parameter-Efficient Fine-Tuning with Instance-Aware Prompt and Parallel Adapters for Speaker Verification 1638
Shengyu Peng, Wu Guo, Jie Zhang, Yu Guan, Lipeng Dai, Zuoliang Li

Unified Text and Speaker Verification Using SSL Model for Text-Dependent Speaker Verification............. 1643
Nathan Griot, Driss Matrouf, Raphael Blouet, Jean-François Bonastre, Ana Mantecon

Towards Robust Overlapping Speech Detection: A Speaker-Aware Progressive Approach Using WavLM.. 1648
Zhaokai Sun, Li Zhang, Qing Wang, Pan Zhou, Lei Xie

Towards Secure User Authentication for Headphones Via In-Ear Or In-Earcup Microphones...................... 1653
N Shashaank, Xiao Quan, Andrew Kaluzny, Leonard Varghese, Marko Stamenovic, Chuan-Che Huang

Mimic Blocker: Self-Supervised Adversarial Training for Voice Conversion Defense with Pretrained Feature Extractors 1658
Gwangyeol Yu, Junhyeok Lee, Seoryeong Kim, Jimin Lee, Jehyuk Lee

A Siamese Network-Based Framework for Voice Mimicry Proficiency Assessment Using X-Vector Embeddings 1663
Bhasi K. C., Rajeev Rajan

Towards Source Attribution of Singing Voice Deepfake with Multimodal Foundation Models 1668
Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Priyabrata Mallick, Pailla Balakrishna Reddy, Arun Balaji Buduru, Rajesh Sharma

Multimodal Zero-Shot Framework for Deepfake Hate Speech Detection in Low-Resource Languages... 1673
Rishabh Ranjan, Likhith Ayinala, Mayank Vatsa, Richa Singh

Joint Target-Speaker ASR and Activity Detection... 1678
Chikara Maeda, Muhammad Shakeel, Yui Sudo

DLF-EEND: Dynamic Layer Fusion for End-To-End Speaker Diarization 1683
Wooil Kim, Bongsu Jung

ACOUSTIC ANALYSIS AND BIOACOUSTICS

Analysis of Avian Biphonic Vocalization Using Computational Modelling 1688
Noumida A, Rajeev Rajan

Dog2vec: Self-Supervised Pre-Training for Canine Vocal Representation 1693
Xingyuan Li, Kenny Zhu, Mengyue Wu

Improving Bird Classification with Primary Color Additives 1698
Ezhini Rasendiran R, Chandresh Kumar Maurya

Exploring the Power of Empirical Mode Decomposition for Sensing the Sound of Silence: A Pilot Study on Mice Autism Detection Via Ultrasonic Vocalisation 1703
Chenhao Wu, Xiangjun Cai, Haojie Zhang, Tianrui Jia, Yilu Deng, Kun Qian, Björn W. Schuller, Yoshiharu Yamamoto, Jiang Liu

Exploring Pre-Trained Models on Ultrasound Modeling for Mice Autism Detection with Uniform Filter Bank and Attentive Scoring 1708
Yuchen Song, Yucong Zhang, Ming Li

MADUV: The 1st INTERSPEECH Mice Autism Detection Via Ultrasound Vocalization Challenge 1713
Zijiang Yang, Meishu Song, Xin Jing, Haojie Zhang, Kun Qian, Bin Hu, Kota Tamada, Toru Takumi, Björn W. Schuller, Yoshiharu Yamamoto

Significance of Time-Frequency Preprocessing for Automatic Ultrasonic Vocalization Classification in Autism Spectrum Disorder Model Detection 1718
Szymon Szmajdzinski, Juliusz Wójtowicz-Kruk, Ivan Ryzhankow, Lukasz Lazarski, Jakub Zak, Wladyslaw Sredniawa

Robust Vocal Intensity Prediction: Overcoming Dataset Bias with Pretrained Deep Models 1723
Quentin Le Tellier, Marc Evrard, Albert Rilliard, Jean-Sylvain Liénard

SLASH: Self-Supervised Speech Pitch Estimation Leveraging DSP-Derived Absolute Pitch 1728
Ryo Terashima, Yuma Shirahata, Masaya Kawamura

SPOKEN DIALOGUE SYSTEMS 1

PruneSLU: Efficient On-Device Spoken Language Understanding Through Vocabulary and Structural Pruning 1733
Truong Do, Minh-Phuong Nguyen, Le -Minh Nguyen

Leveraging LLMs for Written to Spoken Style Data Transformation to Enhance Spoken Dialog State Tracking 1738
Haris Gulzar, Monikka Roslianna Busto, Akiko Masaki, Takeharu Eda, Ryo Masumura

Approaching Dialogue State Tracking Via Aligning Speech Encoders and LLMs 1743
Šimon Sedlácek, Bolaji Yusuf, Ján Švec, Pradyoth Hegde, Santosh Kesiraju, Oldrich Plchot, Jan Cernocký

What Do Humans Hear When Interacting? Experiments on Selective Listening for Evaluating ASR of Spoken Dialogue Systems 1748
Kiyotada Mori, Seiya Kawano, Chaoran Liu, Carlos Toshinori Ishi, Angel García Contreras, Koichiro Yoshino

SpeechDialogueFactory: A Framework for Natural Speech Dialogue Generation 1753
Minghan Wang, Ye Bai, Yuxia Wang, Thuy-Trang Vu, Ehsan Shareghi, Gholamreza Haffari

Who, When, and What: Leveraging the ``Three Ws'' Concept for Emotion Recognition in Conversation 1758
Xiaohan Shi, Xingfeng Li, Tomoki Toda

``Alexa, Can You Forget Me?'' Machine Unlearning Benchmark in Spoken Language Understanding 1763
Alkis Koudounas, Claudio Savelli, Flavio Giobergia, Elena Baralis

Evaluating Large Language Models in Data Generation for Low-Resource Scenarios: A Case Study on Question Answering 1768
Ebru Arisoy, Merve Unlu Menevse, Yusufcan Manav, Arzucan Ozgur

I Want a Horror – Comedy – Movie: Slips-Of-The-Tongue Impact Conversational Recommender System Performance 1773
Maria Teleki, Lingfeng Shi, Chengkai Liu, James Caverlee

Towards a Japanese Full-Duplex Spoken Dialogue System 1778
Atsumoto Ohashi, Shinya IIzuka, Jingjing Jiang, Ryuichiro Higashinaka

SPEECH ASSESSMENT

SAKURA: On the Multi-Hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information 1783
Chih-Kai Yang, Neo Ho, Yen-Ting Piao, Hung-Yi Lee

Continual Speech Learning with Fused Speech Features 1788
Guitao Wang, Jinming Zhao, Hao Yang, Guilin Qi, Tongtong Wu, Gholamreza Haffari

Uni-VERSA: Versatile Speech Assessment with a Unified Network 1793
Jiatong Shi, Hye-Jin Shim, Shinji Watanabe

Evaluating ASR Robustness to Spontaneous Speech Errors: A Study of WhisperX Using a Speech Error Database 1798
John Alderete, Macarious Kin Fung Hui, Aanchan Mohan

Is Synthetic Data Truly Effective for Training Speech Language Models? 1803
Tomoya Mizumoto, Atsushi Kojima, Yusuke Fujita, Lianbo Liu, Yui Sudo

How to Connect Speech Foundation Models and Large Language Models? What Matters and What Does Not 1808
Francesco Verdini, Pierfrancesco Melucci, Stefano Perna, Francesco Cariaggi, Marco Gaido, Sara Papi, Szymon Mazurek, Marek Kasztelnik, Luisa Bentivogli, Sebastien Bratières, Paolo Merialdo, Simone Scardapane

AUDIO-VISUAL ASR AND MULTIMODAL SYSTEM

Text Entry for All: Towards Speech-Based Multimodal Interaction for Inclusion, Accessibility and the Preservation of the World's Linguistic Heritage 1813
Julian Zapata, Lara Hanna

Scaling and Enhancing LLM-Based AVSR: A Sparse Mixture of Projectors Approach 1818
Umberto Cappellazzo, Minsu Kim, Stavros Petridis, Daniele Falavigna, Alessio Brutti

Cocktail-Party Audio-Visual Speech Recognition 1823
Thai-Binh Nguyen, Ngoc-Quan Pham, Alexander Waibel

Efficient Noise-Robust Hybrid Audiovisual Encoder with Joint Distillation and Pruning for Audiovisual Speech Recognition 1828
Zhengyang Li, Pascal Reichert, Thomas Graave, Patrick Blumenberg, Tim Fingscheidt

Unified Audio-Visual Modeling for Recognizing Which Face Spoke When and What in Multi-Talker Overlapped Speech and Video 1833
Naoki Makishima, Naotaka Kawata, Taiga Yamane, Mana Ihori, Tomohiro Tanaka, Satoshi Suzuki, Shota Orihashi, Ryo Masumura

SPEECH AND VOICE DISORDERS 1

Leveraging LLM for Stuttering Speech: A Unified Architecture Bridging Recognition and Event Detection 1838
Shangkun Huang, Jing Deng, Jintao Kang, Rong Zheng

Seamless Dysfluent Speech Text Alignment for Disordered Speech Analysis 1843
Zongli Ye, Jiachen Lian, Xuanru Zhou, Jinming Zhang, Haodong Li, Shuhe Li, Chenxu Guo, Anaisha Das, Peter Park, Zoe Ezzes, Jet Vonk, Brittany Morin, Rian Bogley, Lisa Wauters, Zachary Miller, Maria Gorno-Tempini, Gopala Anumanchipalli

Analysis and Evaluation of Synthetic Data Generation in Speech Dysfluency Detection 1848
Jinming Zhang, Xuanru Zhou, Jiachen Lian, Shuhe Li, William Li, Zoe Ezzes, Rian Bogley, Lisa Wauters, Zachary Miller, Jet Vonk, Brittany Morin, Maria Gorno-Tempini, Gopala Anumanchipalli

Fine-Tuning Strategies for Automatic Speech Recognition of Low-Resource Speech with Autism Spectrum Disorder 1853
Yeseul Park, Bowon Lee

Identification of Pathological Pronunciation Profiles in ASR Transcription Errors 1858
Margot Masson, Isabelle Ferrané, Julie Mauclair

A Simple Method for Predicting Clinical Scores in Huntington's Disease by Leveraging ASR's Uncertainty on Spontaneous Speech 1863
Hadrien Titeux, Quang Tuan Rémy Nguyen, Andres Gil-Salcedo, Anne-Catherine Bachoud-Levi, Emmanuel Dupoux

Introducing EMOPARKNZ: The Emotional Speech Database from New Zealand English Speakers with Parkinson's Disease 1868
Itay Ben-Dom, Catherine I. Watson, Clare M. McCann

Revisiting WFST-Based Hybrid Japanese Speech Recognition System for Individuals with Organic Speech Disorders 1873
Naoki Hojo, Ryoichi Takashima, Chihiro Sugiyama, Nobukazu Tanaka, Kanji Nohara, Kazunori Nozaki, Tetsuya Takiguchi

MULTIMODAL INFORMATION BASED SPEECH PROCESSING (MISP) 2025 CHALLENGE

Pseudo Labels-Based Neural Speech Enhancement for the AVSR Task in the MISP-Meeting Challenge 1878
Longjie Luo, Shenghui Lu, Lin Li, Qingyang Hong

The Multimodal Information Based Speech Processing (MISP) 2025 Challenge: Audio-Visual Diarization and Recognition 1883
Ming Gao, Shilong Wu, Hang Chen, Jun Du, Chin-Hui Lee, Shinji Watanabe, Jingdong Chen, Sabato Marco Siniscalchi, Odette Scharenborg

Cross-Attention and Self-Attention for Audio-Visual Speaker Diarization in MISP-Meeting Challenge 1888
Zhaoyang Li, Haodong Zhou, Longjie Luo, Xiaoxiao Li, Yongxin Chen, Lin Li, Qingyang Hong

Multi-Channel Sequence-To-Sequence Neural Diarization: Experimental Results for the MISP 2025 Challenge 1893
Ming Cheng, Fei Su, Cancan Li, Juan Liu, Ming Li

Leveraging Self-Supervised Learning Based Speaker Diarization for MISP 2025 AVSD Challenge 1898
Zeyan Song, Tianchi Sun, Ronghui Hu, Kai Chen, Jing Lu

Overlap-Adaptive Hybrid Speaker Diarization and ASR-Aware Observation Addition for MISP 2025 Challenge 1903
Shangkun Huang, Yuxuan Du, Jingwen Yang, Dejun Zhang, Xupeng Jia, Jing Deng, Jintao Kang, Rong Zheng

SPEAKER EXTRACTION 1

Robust Target Speaker Diarization and Separation Via Augmented Speaker Embedding Sampling 1908
Md Asif Jalal, Luca Remaggi, Vasileios Moschopoulos, Thanasis Kotsiopoulos, Vandana Rajan, Karthikeyan Saravanan, Anastasis Drosou, Junho Heo, Hyuk Oh, Seokyeong Jeong

Inter-Speaker Relative Cues for Text-Guided Target Speech Extraction 1913
Wang Dai, Archontis Politis, Tuomas Virtanen

REAL-T: Real Conversational Mixtures for Target Speaker Extraction 1918
Shaole Li, Shuai Wang, Jiangyu Han, Ke Zhang, Wupeng Wang, Haizhou Li

Online Audio-Visual Autoregressive Speaker Extraction 1923
Zexu Pan, Wupeng Wang, Shengkui Zhao, Chong Zhang, Kun Zhou, Yukun Ma, Bin Ma

Plug-And-Play Co-Occurring Face Attention for Robust Audio-Visual Speaker Extraction 1928
Zexu Pan, Shengkui Zhao, Tingting Wang, Kun Zhou, Yukun Ma, Chong Zhang, Bin Ma

LOW RESOURCE SPEECH RECOGNITION

SardinianVoxes: A Speech Recognition Dataset for the Sardinian Languages 1933
Salvatore Carta, Alessandro Giuliani, Marco Manolo Manca, Mirko Marras, Leonardo Piano

Prompting Whisper for Improved Verbatim Transcription and End-To-End Miscue Detection.................... 1938
Griffin Smith, Dianna Yee, Jennifer King Chen, Leah Findlater

Automated Evaluation of Children's Speech Fluency for Low-Resource Languages 1943
Bowen Zhang, Nur Afiqah Abdul Latiff, Justin Kan, Rong Tong, Donny Soh, Xiaoxiao Miao, Ian McLoughlin

Cantonese Punctuation Restoration Using LLM Annotated Data 1948
King Yiu Suen, Rudolf Chow, Albert Y. S. Lam

Enhancing Speech Instruction Understanding and Disambiguation in Robotics Via Speech Prosody........... 1953
David Sasu, Benedict Quartey, Kweku Andoh Yamoah, Natalie Schluter

Beyond Traditional Speech Modifications : Utilizing Self Supervised Features for Enhanced Zero-Shot Children ASR 1958
Abhijit Sinha, Hemant Kumar Kathania, Mikko Kurimo

Spoken Language Modeling with Duration-Penalized Self-Supervised Units 1963
Nicol Visser, Herman Kamper

COMPUTATIONAL RESOURCE CONSTRAINED ASR

Towards One-Bit ASR: Extremely Low-Bit Conformer Quantization Using Co-Training and Stochastic Precision.. 1968
Zhaoqing Li, Haoning Xu, Zengrui Jin, Lingwei Meng, Tianzi Wang, Huimeng Wang, Youjun Chen, Mingyu Cui, Shujie Hu, Xunying Liu

Unfolding a Few Structures for the Many: Memory-Efficient Compression of Conformer and Speech Foundation Models 1973
Zhaoqing Li, Haoning Xu, Xurong Xie, Zengrui Jin, Tianzi Wang, Xunying Liu

Effective and Efficient One-Pass Compression of Speech Foundation Models Using Sparsity-Aware Self-Pinching Gates... 1978
Haoning Xu, Zhaoqing Li, Youjun Chen, Huimeng Wang, Guinan Li, Mengzhe Geng, Chengxi Deng, Xunying Liu

Ultra-Low Bit Post-Training Quantization of Large Speech Models Via K-Means Clustering and Mixed Precision Allocation 1983
Tianteng Gu, Bei Liu, Haoyu Wang, Yanmin Qian

Context-Driven Dynamic Pruning for Large Speech Foundation Models.. 1988
Masao Someki, Shikhar Bharadwaj, Atharva Anand Joshi, Chyi-Jiunn Lin, Jinchuan Tian, Jee-Weon Jung, Markus Müller, Nathan Susanj, Jing Liu, Shinji Watanabe

Analyzing the Importance of Blank for CTC-Based Knowledge Distillation 1993
Benedikt Hilmes, Nick Rossenbach, Ralf Schlüter

Speech LLMs in Low-Resource Scenarios: Data Volume Requirements and the Impact of Pretraining on High-Resource Languages 1998
Seraphina Fong, Marco Matassoni, Alessio Brutti

SPEECH AND LANGUAGE TECHNOLOGY FOR HEALTH APPLICATIONS

A Chinese Heart Failure Status Speech Database with Universal and Personalised Classification................ 2003
Yue Pan, Liwei Liu, Changxin Li, Xingyao Wang, Yili Xia, Hanyue Zhang, Ming Chu

Heart Rate as a Proxy Measure to Assess Human Confidence in Spoken Speech... 2008
Harish Battula, Gauri Deshpande, Yagna Gudipalli, Sachin Patel

Foundation Model Hidden Representations for Heart Rate Estimation from Auscultation............................ 2013
Jingping Nie, Tien Dung Tran, Karan Thakkar, Vasudha Kowtha, Jon Huang, Carlos Avendano, Erdrin Azemi, Vikramjit Mitra

Towards Fusion of Neural Audio Codec-Based Representations with Spectral for Heart Murmur Classification Via Bandit-Based Cross-Attention Mechanism 2018
Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Priyabrata Mallick, Santanu Roy, Arun Balaji Buduru, Rajesh Sharma

Perception of Emotional Speech by Individuals with High Borderline Personality Features......................... 2023
Yizhou Chen, Xiyu Wu

Visual Features of the Oral Region in Polish Sibilants Produced by Children with Various Sibilance Patterns 2028
Agata Sage, Zuzanna Miodonska, Michal Krecichwost, Ewa Kwasniok, Pawel Badura

Meta-Learning Approaches for Speaker-Dependent Voice Fatigue Models... 2033
Roseline Polle, Agnes Norbury, Alexandra Livia Georgescu, Nicholas Cummins, Stefano Goria

Decoding Alzheimer's: Interpretable Visual and Logical Attention in Picture Description Tasks 2038
Ning Wang, Bingyang Wen, Minghui Wu, Yang Sun, Zongru Shao, Haojie Zhou, K. P. Subbalakshmi

RESPONSIBLE SPEECH FOUNDATION MODELS + SUPERB CHALLENGE

Defending Speech-Enabled LLMs Against Adversarial Jailbreak Threats... 2043
Antonios Alexos, Raghuveer Peri, Sai Muralidhar Jayanthi, Metehan Cekic, Srikanth Vishnubhotla, Kyu J. Han, Srikanth Ronanki

Mitigating Subgroup Disparities in Multi-Label Speech Emotion Recognition: A Pseudo-Labeling and Unsupervised Learning Approach.. 2048
Yi-Cheng Lin, Huang-Cheng Chou, Hung-Yi Lee

Who Gets the Mic? Investigating Gender Bias in the Speaker Assignment of a Speech-LLM...................... 2053
Dariia Puhach, Amir H. Payberah, Éva Székely

Evaluating Speech Foundation Models for Automatic Speech Recognition in the Low-Resource Kanyen'kéha Language 2058
Mengzhe Geng, Patrick Littell, Aidan Pine, Robbie Jimerson, Gilles Boulianne, Vishwa Gupta, Rolando Coto-Solano, Anna Kazantseva, Marc Tessier, Delaney Lothian, Akwiratékha' Martin, Eric Joanis, Samuel Larkin, Roland Kuhn

Benchmarking and Confidence Evaluation of LALMs for Temporal Reasoning.. 2063
Debarpan Bhattacharya, Apoorva Kulkarni, Sriram Ganapathy

Teaching Audio-Aware Large Language Models What Does Not Hear: Mitigating Hallucinations Through Synthesized Negative Samples 2068
Chun-Yi Kuan, Hung-Yi Lee

Speech-IFEval: Evaluating Instruction-Following and Quantifying Catastrophic Forgetting in Speech-Aware Language Models 2073
Ke-Han Lu, Chun-Yi Kuan, Hung-Yi Lee

Enhancing Low-Resource Language and Instruction Following Capabilities of Audio Language Models 2078
Potsawee Manakul, Guangzhi Sun, Warit Sirichotedumrong, Kasima Tharnpipitchai, Kunat Pipatanakul

Improving Multilingual Speech Models on ML-SUPERB 2.0: Fine-Tuning with Data Augmentation and LID-Aware CTC 2083
Qingzheng Wang, Jiancheng Sun, Yifan Peng, Shinji Watanabe

The ML-SUPERB 2.0 Challenge: Towards Inclusive ASR Benchmarking for All Language Varieties 2088
William Chen, Chutong Meng, Jiatong Shi, Martijn Bartelds, Shih-Heng Wang, Hsiu-Hsuan Wang, Rafael Mosquera, Sara Hincapie, Dan Jurafsky, Antonis Anastasopoulos, Hung-Yi Lee, Karen Livescu, Shinji Watanabe

TalTech Systems for the Interspeech 2025 ML-SUPERB 2.0 Challenge 2093
Tanel Alumäe, Artem Fedorchenko

DYSARTHRIC SPEECH ASSESSMENT 1

Regularized Federated Learning for Privacy-Preserving Dysarthric and Elderly Speech Recognition 2098
Tao Zhong, Mengzhe Geng, Shujie Hu, Guinan Li, Xunying Liu

Facilitating Personalized TTS for Dysarthric Speakers Using Knowledge Anchoring and Curriculum Learning 2103
Yejin Jeon, Solee Im, Youngjae Kim, Gary Geunbae Lee

DiffDSR: Dysarthric Speech Reconstruction Using Latent Diffusion Model 2108
Xueyuan Chen, Dongchao Yang, Wenxuan Wu, Minglin Wu, Jing Xu, Xixin Wu, Zhiyong Wu, Helen Meng

VOLUME 4

Improved Intelligibility of Dysarthric Speech Using Conditional Flow Matching 2113
Shoutrik Das, Nishant Singh, Arjun Gangwar, S Umesh

Bridging ASR and LLMs for Dysarthric Speech Recognition: Benchmarking Self-Supervised and Generative Approaches..........2118
Ahmed Aboeitta, Ahmed Sharshar, Youssef Nafea, Shady Shehata

Towards Inclusive ASR: Investigating Voice Conversion for Dysarthric Speech Recognition in Low-Resource Languages 2123
Chin-Jou Li, Eunjung Yeo, Kwanghee Choi, Paula Andrea Pérez-Toro, Masao Someki, Rohan Kumar Das, Zhengjun Yue, Juan Rafael Orozco-Arroyave, Elmar Nöth, David R. Mortensen

Mitigating Overfitting During Speech Foundation Model Fine-Tuning: Applications to Dysarthric Speech Detection 2128
Yan Xiong, Visar Berisha, Julie Liss, Chaitali Chakrabarti

Towards Temporally Explainable Dysarthric Speech Clarity Assessment 2133
Seohyun Park, Chitralekha Gupta, Michelle Kah Yian Kwan, Xinhui Fung, Alexander Wenjun Yip, Suranga Nanayakkara

SHOW AND TELL 2: SPEECH SYNTHESIS

Code Mix TTS: An Approach to Infer Human Like Speech for Multi-Lingual Input Texts 2138
Vishal Gourav, Phanindra Mankale

Turing's Echo: Investigating Linguistic Sensitivity of Deepfake Voice Detection Via Gamification 2140
Binh Nguyen, Thai Le

Unleashing the Inner Monster: Demonstrating High-Fidelity Human to Non-Human Voice Conversion 2142
Namhyun Cho, Sunmin Kim, Minsu Kang, Seolhee Lee, Choonghyeon Lee, Yangsun Lee

Tungnaá in Live Performance: An Implementation of Interactive Artistic Text-To-Voice 2144
Victor Shepardson, Jonathan Reus, Thor Magnusson

Hear Me Out: Interactive Evaluation and Bias Discovery Platform for Speech-To-Speech Conversational AI 2146
Shree Harsha Bokkahalli Satish, Gustav Eje Henter, Éva Székely

Vocal-Tract Model with Two Directions: Static Design for a Dummy Head and Dynamic Design for a Speaking Machine 2148
Takayuki Arai

DATABASES AND PROGRESS IN METHODOLOGY

Dhvani: A Weakly-Supervised Phonemic Error Detection and Personalized Feedback System for Hindi 2150
Arnav Rustagi, Satvik Bajpai, Nimrat Kaur, Siddharth Siddharth

Evaluating Wav2Vec2-Bert for Computer-Assisted Pronunciation Training for isiZulu 2155
Alexandra Fort, Francis Tyers

Towards Adaptable and Intelligible Speech Synthesis in Noisy Environments 2160
Lubos Marcinek, Jonas Beskow, Joakim Gustafsson

Harnessing Text-To-Speech Voice Cloning Models for Improved Audiological Speech Assessment 2165
Lidea Shahidi, Erdem Baha Topbas, Thu Ngan Dang, Tobias Goehring

75-Speaker Annot-16: A Benchmark Dataset for Speech Articulatory rt-MRI Annotation with Articulator Contours and Phonetic Alignment 2170
Xuan Shi, Yubin Zhang, Yijing Lu, Marcus Ma, Tiantian Feng, Asterios Toutios, Haley Hsu, Louis Goldstein, Shrikanth Narayanan

Representing Speech Through Autoregressive Prediction of Cochlear Tokens 2175
Greta Tuckute, Klemen Kotar, Evelina Fedorenko, Daniel Yamins

Reasoning-Based Approach with Chain-Of-Thought for Alzheimer's Detection Using Speech and Large Language Models 2180
Chanwoo Park, Anna Seo Gyeong Choi, Sunghye Cho, Chanwoo Kim

Finding the Human Voice in AI: Insights on the Perception of AI-Voice Clones from Naturalness and Similarity Ratings 2185
Linda Bakkouche, Charles McGhee, Emily Lau, Stephanie Cooper, Xinbing Luo, Madeleine Rees, Kai Alter, Brechtje Post, Julia Schwarz

Prosodically Enhanced Foreign Accent Simulation by Discrete Token-Based Resynthesis Only with Native Speech Corpora 2190
Kentaro Onda, Keisuke Imoto, Satoru Fukayama, Daisuke Saito, Nobuaki Minematsu

NOVEL ARCHITECTURES FOR ASR

Weight Factorization and Centralization for Continual Learning in Speech Recognition 2195
Enes Ugan, Ngoc-Quan Pham, Alexander Waibel

Dysfluent WFST: A Framework for Zero-Shot Speech Dysfluency Transcription and Detection 2200
Chenxu Guo, Jiachen Lian, Xuanru Zhou, Jinming Zhang, Shuhe Li, Zongli Ye, Peter Park, Anaisha Das, Zoe Ezzes, Jet Vonk, Brittany Morin, Rian Bogley, Lisa Wauters, Zachary Miller, Maria Gorno-Tempini, Gopala Anumanchipalli

Dysarthric Speech Recognition Using Curriculum Learning and Multi-Stream Architecture 2205
I-Ting Hsieh, Chung-Hsien Wu

DYNAC: Dynamic Vocabulary-Based Non-Autoregressive Contextualization for Speech Recognition 2210
Yui Sudo, Yosuke Fukumoto, Muhammad Shakeel, Yifan Peng, Chyi-Jiunn Lin, Shinji Watanabe

Beyond Hard Sharing: Efficient Multi-Task Speech-To-Text Modeling with Supervised Mixture of Experts 2215
Hojun Jin, Eunsoo Hong, Ziwon Hyung, Sungjun Lim, Seungjin Lee, Keunseok Cho

OWSM V4: Improving Open Whisper-Style Speech Models Via Data Scaling and Cleaning 2220
Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, Shinji Watanabe

DEEPFAKE DETECTION

Bona Fide Cross Testing Reveals Weak Spot in Audio Deepfake Detection Systems 2225
Chin Yuen Kwok, Jia Qi Yip, Zhen Qiu, Chi Hung Chi, Kwok Yan Lam

BiCrossMamba-ST: Speech Deepfake Detection with Bidirectional Mamba Spectro-Temporal Cross-Attention 2230
Yassine El Kheir, Tim Polzehl, Sebastian Möller

Few-Shot Speech Deepfake Detection Adaptation with Gaussian Processes 2235
Neta Glazer, David Chernin, Idan Achituve, Sharon Gannot, Ethan Fetaya

Replay Attacks Against Audio Deepfake Detection 2240
Nicolas Müller, Piotr Kawa, Wei-Herng Choong, Adriana Stan, Aditya Tirumala Bukkapatnam, Karla Pizzi, Alexander Wagner, Philip Sperl

Enhancing Audio Deepfake Detection by Improving Representation Similarity of Bonafide Speech........... 2245
Seung-Bin Kim, Hyun-Seo Shin, Jungwoo Heo, Chan-Yeong Lim, Kyo-Won Koo, Jisoo Son, Sanghyun Hong, Souhwan Jung, Ha-Jin Yu

Generalizable Audio Deepfake Detection Via Hierarchical Structure Learning and Feature Whitening in Poincaré Sphere 2250
Mingru Yang, Yanmei Gu, Qianhua He, Yanxiong Li, Peirong Zhang, Yongqiang Chen, Zhiming Wang, Huijia Zhu, Jian Liu, Weiqiang Wang

TOOLS FOR SPEECH ANALYSIS

VoiceNet: Multilingual On-Device Phoneme-To-Audio Alignment... 2255
Kun Jin, Siva Penke, Srinivasa Algubelli

Nosey: Open-Source Hardware for Acoustic Nasalance 2260
Maya Dewhurst, Jack Collins, Justin J. H. Lo, Roy Alderton, Sam Kirkham

Automatic Classification of Stop Realisation with Wav2vec2.0 2265
James Tanner, Morgan Sonderegger, Jane Stuart-Smith, Jeff Mielke, Tyler Kendall

TEXT PROCESSING AND EVALUATION FOR SPEECH SYNTHESIS 1

Acquiring Pronunciation from Speech Audio Via Multi-Task Learning 2270
Siqi Sun, Korin Richmond

Intelligibility of Text-To-Speech Systems for Mathematical Expressions.. 2275
Sujoy Roychowdhury, Ranjani H. G., Sumit Soman, Nishtha Paul, Subhadip Bandyopadhyay, Siddhanth Iyengar

The State of TTS: A Case Study with Human Fooling Rates 2280
Praveen Srinivasa Varadhan, Sherry Thomas, Sai Teja M S, Suvrat Bhooshan, Mitesh M. Khapra

Pairwise Evaluation of Accent Similarity in Speech Synthesis 2285
Jinzuomu Zhong, Suyuan Liu, Dan Wells, Korin Richmond

VoiceQualityVC: A Voice Conversion System for Studying the Perceptual Effects of Voice Quality in Speech 2290
Harm Lameris, Joakim Gustafsson, Éva Székely

Towards Frame-Level Quality Predictions of Synthetic Speech 2295
Michael Kuhlmann, Fritz Seebauer, Petra Wagner, Reinhold Haeb-Umbach

SEGMENTAL AND TONAL UNITS

Perception of Long and Short Vowel Contrast in Te Reo Māori in Clean and Everyday Listening Environments.. 2300
C. T. Justine Hui, Jenice Kuzhikombil, Isabella Shields, Hiraia Haami-Wells, Catherine I. Watson, Peter [j.] Keegan

The Function of Creaky Voice in South Korean: A Perception Study 2305
Patrik Hrabánek, Michaela Watkins, Silke Hamann

Talker Normalization in Chinese Bilinguals: A Comparative Study 2310
Mingxi Lu, Ran Tao, Yujia Tian

Coping with Segmental–prosodic Incongruity in Spoken Word Recognition in Japanese 2315
Terumichi Ariga

What the Filler? Both ASR Systems and Humans Struggle More with Other Kinds of Disfluencies than with Filler Particles 2320
Saskia Wepner, Lucas Eckert, Gernot Kubin, Barbara Schuppler

SPEECH QUALITY ASSESSMENT

Non-Intrusive Speech Quality Assessment with Diffusion Models Trained on Clean Speech 2325
Danilo De Oliveira, Julius Richter, Jean-Marie Lemercier, Simon Welker, Timo Gerkmann

SQ-AST: A Transformer-Based Model for Speech Quality Prediction 2330
Wafaa Wardah, Robert P. Spang, Vincent Barriac, Jan Reimes, Anna Llagostera, Jens Berger, Sebastian Möller

AttentiveMOS: A Lightweight Attention-Only Model forSpeech Quality Prediction 2335
Imran E Kibria, Donald S. Williamson

Universal Preference-Score-Based Pairwise Speech Quality Assessment.. 2340
Yu-Fei Shi, Yang Ai, Zhen-Hua Ling

FUSE-MOS: Fusion of Speech Embeddings for MOS Prediction with Uncertainty Quantification 2345
Enjamamul Hoq, Nikhil Gupta, Danielle Omondi, Ifeoma Nwogu

SHEET: A Multi-Purpose Open-Source Speech Human Evaluation Estimation Toolkit 2350
Wen-Chin Huang, Erica Cooper, Tomoki Toda

SPEECH ENHANCEMENT

Investigating Continuous Autoregressive Generative Speech Enhancement.. 2355
Haici Yang, Gordon Wichern, Ryo Aihara, Yoshiki Masuyama, Sameer Khurana, François G. Germain, Jonathan Le Roux

Dynamic Layer Gating for Speech Enhancement.. 2360
Venkatesh Parvathala, K. Sri Rama Murty

Model as Loss: A Self-Consistent Training Paradigm 2365
Saisamarth Rajesh Phaye, Milos Cernak, Andrew Harper

Test-Time Training for Speech Enhancement.. 2370
Avishkar Behera, Riya Ann Easow, Venkatesh Parvathala, K. Sri Rama Murty

Few-Step Adversarial Schrödinger Bridge for Generative Speech Enhancement 2375
Seungu Han, Sungho Lee, Juheon Lee, Kyogu Lee

Exploiting Bispectral Features for Single-Channel Speech Enhancement 2380
Venkatesh Parvathala, Ramesh Gundluru, Sreekanth Sankala, K. Sri Rama Murty

LANGUAGE LEARNING AND ASSESSMENT

Automatic Dialectal Transcription: An Evaluation on Finnish and Norwegian 2385
Olli Kuparinen

Can ASR Generate Valid Measures of Child Reading Fluency? 2390
Wieke Harmsen, Roeland Van Hout, Catia Cucchiarini, Helmer Strik

SGED-Probe: Probing E2E ASR Decoder and Aligner for Spoken Grammar Error Detection Under Three Speaking Practice Conditions.. 2395
Chowdam Venkata Thirumala Kumar, Chiranjeevi Yarra

Evaluating Logit-Based GOP Scores for Mispronunciation Detection 2400
Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik

Towards a Unified Benchmark for Arabic Pronunciation Assessment: Qur'anic Recitation as Case Study.. 2405
Yassine El Kheir, Omnia Ibrahim, Amit Meghanani, Nada Almarwani, Hawau Toyin, Sadeen Alharbi, Modar Alfadly, Lamya Alkanhal, Ibrahim Selim, Shehab Elbatal, Salima Mdhaffar, Thomas Hain, Yasser Hifny, Mostafa Shahin, Ahmed Ali

OMPAL: Bridging Speech and Learning with an Open-Source Mandarin Pronunciation Assessment Corpus for Global Learners 2410
Wen-Wei Hsieh, Hao-Wei Chi, Kuan-Chen Wang, Ping-Cheng Yeh, Te-Hsin Liu, Chen-Yu Chiang

A Perception-Based L2 Speech Intelligibility Indicator: Leveraging a Rater's Shadowing and Sequence-To-Sequence Voice Conversion 2415
Haopeng Geng, Daisuke Saito, Nobuaki Minematsu

Multimodal and Multitask Learning for Predicting Multiple Scores in L2 English Speech.......................... 2420
Sehyun Oh, Sunhee Kim, Minhwa Chung

Enhancing Generalization of Speech Large Language Models with Multi-Task Behavior Imitation and Speech-Text Interleaving 2425
Jingran Xie, Xiang Li, Hui Wang, Yue Yu, Yang Xiang, Xixin Wu, Zhiyong Wu

Mispronunciation Detection Without L2 Pronunciation Dataset in Low-Resource Setting: A Case Study in Finland Swedish 2430
Nhan Phan, Mikko Kuronen, Maria Kautonen, Riikka Ullakonoja, Anna Von Zansen, Yaroslav Getman, Ekaterina Voskoboinik, Tamás Grósz, Mikko Kurimo

SPEECH SYNTHESIS PARADIGMS AND METHODS 1

RapFlow-TTS: Rapid and High-Fidelity Text-To-Speech with Improved Consistency Flow Matching.. 2435
Hyun Joon Park, Jeongmin Liu, Jin Sob Kim, Jeong Yeol Yang, Sung Won Han, Eunwoo Song

Accelerating Flow-Matching-Based Text-To-Speech Via Empirically Pruned Step Sampling 2440
Qixi Zheng, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiaofei Wang, Kai Yu, Xie Chen

Differentiable Reward Optimization for LLM Based TTS System 2445
Changfeng Gao, Zhihao Du, Shiliang Zhang

Long-Context Speech Synthesis with Context-Aware Memory 2450
Zhipeng Li, Xiaofen Xing, Jingyuan Xing, Hangrui Hu, Heng Lu, Xiangmin Xu

Monotonic Attention for Robust Text-To-Speech Synthesis in Large Language Model Frameworks 2455
Yike Zhang, Yiming Li, Jie Chen, Qinghua Wu, Songjun Cao, Long Ma

Improving Noise Robustness of LLM-Based Zero-Shot TTS Via Discrete Acoustic Token Denoising 2460
Ye-Xin Lu, Hui-Peng Du, Fei Liu, Yang Ai, Zhen-Hua Ling

Bridging the Training–Inference Gap in TTS: Training Strategies for Robust Generative Postprocessing for Low-Resource Speakers 2465
Frank Zalkow, Paolo Sani, Kishor Kayyar Lakshminarayana, Emanuël A. P. Habets, Nicola Pia, Christian Dittmar

Robust Neural Codec Language Modeling with Phoneme Position Prediction for Zero-Shot TTS 2470
Chunhui Lu, Xue Wen, Liming Song, Junkwang Oh

SPATIAL AUDIO AND ACOUSTICS 2

SepVAC: Multitask Learning of Speaker Separation, Speaker Localization, Microphone Array Localization, and Room Acoustic Parameter Estimation in Various Acoustic Conditions 2475
Roland Hartanto, Sakriani Sakti, Koichi Shinoda

TA-RIR: Topology-Aware Neural Modeling of Acoustic Propagation for Room Impulse Response Synthesis........ 2480
Junhui Zhao, Hang Chen, Qing Wang, Jun Du, Yanhui Tu, Feng Ma

Spatially Weighted Contrastive Learning for Robust Sound Source Localization 2485
Hyun-Soo Kim, Da-Hee Yang, Joon-Hyuk Chang

Efficient and Microphone-Fault-Tolerant 3D Sound Source Localization 2490
Yiyuan Yang, Shitong Xu, Niki Trigoni, Andrew Markham

Joint Reference Microphone Selection and Filter Order Determination in Multi-Channel Active Noise Control 2495
De Hu, Shuyao Liu, Yanrong He

Direct-Path Relative Harmonic Coefficients Detection for Multi-Source Direction-Of-Arrival Estimation in Reverberant Environments 2500
Liang Tao, Maoshen Jia, Yonggang Hu

D-GAT: Dual Graph Attention Network for Global HRTF Interpolation 2505
Junsheng Hu, Shaojie Li, Qintuya Si, De Hu

Deep Learning Based Spatial Aliasing Reduction in Beamforming for Audio Capture 2510
Mateusz Guzik, Giulio Cengarle, Daniel Arteaga

SonarGuard2: Ultrasonic Face Liveness Detection Based on Adaptive Doppler Effect Feature Extraction 2515
Xiaoming Zhang, Ke-Yue Zhang, Taiping Yao, Songjun Cao, Shouhong Ding, Long Ma

TEXT PROCESSING AND EVALUATION FOR SPEECH SYNTHESIS 2

Grapheme-Coherent Phonemic and Prosodic Annotation of Speech by Implicit and Explicit Grapheme Conditioning 2520
Hien Ohnaka, Yuma Shirahata, Byeongseon Park, Ryuichi Yamamoto

Non-Standard Accent TTS Support Via Large Multi-Accent Frontend Pronunciation Knowledge Transfer 2525
Noe Berger, Siqi Sun, Korin Richmond

Speech-Guided Grapheme-To-Phoneme Conversion for Cantonese Text-To-Speech........ 2530
Timothy Shin Heng Mak, King Yiu Suen, Albert Y. S. Lam

Transcript-Prompted Whisper with Dictionary-Enhanced Decoding for Japanese Speech Annotation 2535
Rui Hu, Xiaolong Lin, Jiawang Liu, Shixi Huang, Zhenpeng Zhan

Enabling the Replicability of Speech Synthesis Perceptual Evaluations........ 2540
Sébastien Le Maguer, Gwénolé Lecorvé, Damien Lolive, Naomi Harte, Juraj Šimko

When the MOS Predictor Asks for Training Annotation in Cross Lingual/Domain Adaptation 2545
Natacha Miniconi, Meysam Shamsi, Anthony Larcher

Assessment of the Synthetic Quality and Controllability of Laughing Onset in Speech-Laugh Synthesis........ 2550
Ryo Setoguchi, Yoshiko Arimoto

GENERAL TOPICS IN ASR

Running Conventional Automatic Speech Recognition on Memristor Hardware: A Simulated Approach 2555
Nick Rossenbach, Benedikt Hilmes, Leon Brackmann, Moritz Gunz, Ralf Schlüter

Word Level Timestamp Generation for Automatic Speech Recognition and Translation 2560
Ke Hu, Krishna Puvvada, Elena Rastorgueva, Zhehuai Chen, He Huang, Shuoyang Ding, Kunal Dhawan, Hainan Xu, Jagadeesh Balam, Boris Ginsburg

Directional Speech Recognition with Full-Duplex Capability 2565
Ju Lin, Yiteng Huang, Ming Sun, Frank Seide, Florian Metze

CMT-LLM: Contextual Multi-Talker ASR Utilizing Large Language Models........ 2570
Jiajun He, Naoki Sawada, Koichi Miyazaki, Tomoki Toda

Selective Invocation for Multilingual ASR: A Cost-Effective Approach Adapting to Speech Recognition Difficulty........ 2575
Hongfei Xue, Yufeng Tang, Jun Zhang, Xuelong Geng, Lei Xie

ACOUSTIC EVENT DETECTION AND CLASSIFICATION

Improving Audio Classification by Transitioning from Zero- To Few-Shot 2580
James Taylor, Wolfgang Mack

Zero-Shot Learning for Acoustic Event Classification Using an Attribute Vector and Conditional GAN 2585
Kohei Uehara, Ryoichi Takashima, Tetsuya Takiguchi

Leveraging Multi-Level Features of ATST with Conformer-Based Dual-Branch Network for Sound Event Detection 2590
Lipeng Dai, Qing Wang, Jie Zhang, Shengyu Peng, Yu Guan, Wu Guo

Leveraging Unlabeled Audio for Audio-Text Contrastive Learning Via Audio-Composed Text Features 2595
Tatsuya Komatsu, Hokuto Munakata, Yuchi Ishikawa

Language-Guided Contrastive Audio-Visual Masked Autoencoder with Automatically Generated Audio-Visual-Text Triplets from Videos 2600
Yuchi Ishikawa, Shota Nakada, Hokuto Munakata, Kazuhiro Saito, Tatsuya Komatsu, Yoshimitsu Aoki

AC/DC: LLM-Based Audio Comprehension Via Dialogue Continuation 2605
Yusuke Fujita, Tomoya Mizumoto, Atsushi Kojima, Lianbo Liu, Yui Sudo

Anomalous Sound Detection Based Feature Fusion and Dual-Path Non-Linear Independent Components Estimation 2610
Yawei Wang, Qiaoling Zhang, Yi Zhang, Junyao Hu

An Effective Anomalous Sound Detection Method Based on Global and Local Attribute Mining 2615
Nan Jiang, Yan Song, Qing Gu, Haoyu Song, Lirong Dai, Ian McLoughlin

Acoustic Scattering AI for Non-Invasive Object Classifications: A Case Study on Hair Assessment 2620
Long-Vu Hoang, Tuan Nguyen, Huy Dat Tran

KEYWORD SPOTTING AND RETRIEVAL

Language-Agnostic Speech Tokenizer for Spoken Term Detection with Efficient Retrieval 2625
Anup Singh, Kris Demuynck, Vipul Arora

H-QuEST: Accelerating Query-By-Example Spoken Term Detection with Hierarchical Indexing 2630
Akanksha Singh, Yi-Ping Phoebe Chen, Vipul Arora

Vela: Scalable Embeddings with Voice Large Language Models for Multimodal Retrieval 2635
Ruofan Hu, Yan Xia, Minjie Hong, Jieming Zhu, Bo Chen, Xiaoda Yang, Minghui Fang, Tao Jin

Adversarial Deep Metric Learning for Cross-Modal Audio-Text Alignment in Open-Vocabulary Keyword Spotting 2640
Youngmoon Jung, Yong-Hyeok Lee, Myunghun Jung, Jaeyoung Roh, Chang Woo Han, Hoon-Young Cho

GTA: Towards Generative Text-To-Audio Retrieval Via Multi-Scale Tokenizer 2645
Minghui Fang, Shengpeng Ji, Jialong Zuo, Xize Cheng, Wenrui Liu, Xiaoda Yang, Ruofan Hu, Jieming Zhu, Zhou Zhao

Enhancing Retrieval-Augmented Audio Captioning with Generation-Assisted Multimodal Querying and Progressive Learning 2650
Changin Choi, Sungjun Lim, Wonjong Rhee

On Retrieval of Long Audios with Complex Text Queries 2655
Ruochu Yang, Milind Rao, Harshavardhan Sundar, Anirudh Raju, Aparna Khare, Srinath Tankasala, Di He, Venkatesh Ravichandran

SIDC-KWS: Efficient Spiking Inception-Dilated Conformer with Self-Attention for Keyword Spotting 2660
Jin Gyo Lim, Seong Eun Kim

Multichannel Keyword Spotting for Noisy Conditions 2665
Dzmitry Saladukha, Ivan Koriabkin, Kanstantsin Artsiom, Aliaksei Rak, Nikita Ryzhikov

LLM-Synth4KWS: Scalable Automatic Generation and Synthesis of Confusable Data for Custom Keyword Spotting 2670
Pai Zhu, Quan Wang, Dhruuv Agarwal, Kurt Partridge

GraphemeAug: A Systematic Approach to Synthesized Hard Negative Keyword Spotting Examples 2675
Harry Zhang, Kurt Partridge, Pai Zhu, Neng Chen, Hyun Jin Park, Dhruuv Agarwal, Quan Wang

SpokenNativQA: Multilingual Everyday Spoken Queries for LLMs 2680
Firoj Alam, Md Arid Hasan, Shammur Absar Chowdhury

MULTIMODAL SYSTEMS

CAMER: Contribution-Aware Multimodal Emotion Recognition 2685
Sun-Kyung Lee, Jong-Hwan Kim

GIA-MIC: Multimodal Emotion Recognition with Gated Interactive Attention and Modality-Invariant Learning Constraints 2690
Jiajun He, Jinyi Mi, Tomoki Toda

SNIFR : Boosting Fine-Grained Child Harmful Content Detection Through Audio-Visual Alignment with Cascaded Cross-Transformer 2695
Orchid Chetia Phukan, Mohd Mujtaba Akhtar, Girish, Swarup Ranjan Behera, Abu Osama Siddiqui, Sarthak Jain, Priyabrata Mallick, Jaya Sai Kiran Patibandla, Pailla Balakrishna Reddy, Arun Balaji Buduru, Rajesh Sharma

CNVSRC 2024: The Second Chinese Continuous Visual Speech Recognition Challenge 2700
Zehua Liu, Xiaolou Li, Chen Chen, Lantian Li, Dong Wang

PAEFF: Precise Alignment and Enhanced Gated Feature Fusion for Face-Voice Association 2705
Abdul Hannan, Muhammad Arslan Manzoor, Shah Nawaz, Muhammad Irzam Liaqat, Markus Schedl, Mubashir Noman

Efficient and Direct Duplex Modeling for Speech-To-Speech Language Model 2710
Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Zelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, Boris Ginsburg

U-SAM: An Audio Language Model for Unified Speech, Audio, and Music Understanding 2715
Ziqian Wang, Xianjun Xia, Xinfa Zhu, Lei Xie

Enhanced Hybrid Transducer and Attention Encoder Decoder with Text Data 2720
Yun Tang, Eesung Kim, Vijendra Raj Apsingekar

The Role of Audio-Visual Integration in the Time Course of Phonetic Encoding in Self-Supervised Speech Models 2725
Yi Wang, Oli Danyi Liu, Peter Bell

DYSARTHRIC SPEECH ASSESSMENT 2

Voice Reconstruction Through Large-Scale TTS Models: Comparing Zero-Shot and Fine-Tuning Approaches to Personalise TTS in Assistive Communication 2730
Éva Székely, Péter Mihajlik, Máté Soma Kádár, László Tóth

Objective and Subjective Evaluation of Diffusion-Based Speech Enhancement for Dysarthric Speech 2735
Dimme De Groot, Tanvina Patel, Devendra Kayande, Odette Scharenborg, Zhengjun Yue

Data Augmentation Using Speech Synthesis for Speaker-Independent Dysarthria Severity Classification 2740
Minseop Kim, Minsu Han, Seokyoung Hong, Myoung-Wan Koo

Fairness in Dysarthric Speech Synthesis: Understanding Intrinsic Bias in Dysarthric Speech Cloning Using F5-TTS 2745
Anuprabha M, Krishna Gurugubelli, Anil Kumar Vuppala

Synthetic Dysarthric Speech: A Supplement, Not a Substitute for Authentic Data in Dysarthric Speech Recognition 2750
Jingting Li, Keyi Feng, Xinran Zhao, Yan Wang, Su-Jing Wang

Unsupervised Rhythm and Voice Conversion to Improve ASR on Dysarthric Speech 2755
Karl El Hajal, Enno Hermann, Sevada Hovsepyan, Mathew Magimai Doss

DIALECT IDENTIFICATION IN DIFFERENT LANGUAGES

Audio-Based Classification and Geographic Regression of Austrian Dialects 2760
Lorenz Gutscher, Michael Pucher

Jointly Improving Dialect Identification and ASR in Indian Languages Using Multimodal Feature Fusion 2765
Saurabh Kumar, Amartyaveer, Prasanta Kumar Ghosh

ADI-20: Arabic Dialect Identification Dataset and Models 2770
Haroun Elleuch, Salima Mdhaffar, Yannick Estève, Fethi Bougares

Improving Low-Resource Dialect Classification Using Retrieval-Based Voice Conversion 2775
Lea Fischbach, Akbar Karimi, Caroline Kleen, Alfred Lameli, Lucie Flek

Effects of Prosodic Information on Dialect Classification Using Whisper Features 2780
Phoebe Parsons, Heming Strømholt Bremnes, Knut Kvale, Torbjørn Svendsen, Giampiero Salvi

Voice Conversion Improves Cross-Domain Robustness for Spoken Arabic Dialect Identification 2785
Badr M. Abdullah, Matthew Baas, Bernd Möbius, Dietrich Klakow

CONNECTING SPEECH SCIENCE AND SPEECH TECHNOLOGY FOR CHILDREN'S SPEECH

Band-Split Self-Supervised Mamba for Infant-Centered Audio Analysis 2790
Xulin Fan, Jialu Li, Mark Hasegawa-Johnson, Nancy L. McElwain

Subtyping Speech Errors in Childhood Speech Sound Disorders with Acoustic-To-Articulatory Speech Inversion 2795
Nina R Benway, Saba Tabatabaee, Benjamin Munson, Jonathan Preston, Carol Espy-Wilson

PERCEPT-US: A Multimodal American English Child Speech Corpus Specialized for Articulatory Feedback........ 2800
Amanda Eads, Heather Kabakoff, Nina Benway, Elaine Hitchcock, Jonathan Preston, Tara McAllister

Children's Voice Privacy: First Steps and Emerging Challenges........ 2805
Ajinkya Kulkarni, Francisco Teixeira, Enno Hermann, Thomas Rolland, Isabel Trancoso, Mathew Magimai Doss

FT-Boosted SV: Towards Noise Robust Speaker Verification for English Speaking Classroom Environments........ 2810
Saba Tabatabaee, Jing Liu, Carol Espy-Wilson

Examining Test-Time Adaptation for Personalized Child Speech Recognition 2815
Zhonghao Shi, Xuan Shi, Anfeng Xu, Tiantian Feng, Harshvardhan Srivastava, Shrikanth Narayanan, Maja Mataric

Employing Self-Supervised Learning Models for Cross-Linguistic Child Speech Maturity Classification........ 2820
Theo Zhang, Madurya Suresh, Anne Warluamont, Kasia Hitczenko, Alejandrina Cristia, Margaret Cychosz

On Enhancing the Performance of Children's ASR Task in Limited Data Scenario........ 2825
Ankita Ankita, Shambhavi Shambhavi, Syed Shahnawazuddin

Egocentric Speaker Classification in Child-Adult Dyadic Interactions: From Sensing to Computational Modeling........ 2830
Tiantian Feng, Anfeng Xu, Xuan Shi, Somer Bishop, Shrikanth Narayanan

Large Language Models Based ASR Error Correction for Child Conversations 2835
Anfeng Xu, Tiantian Feng, So Hyun Kim, Somer Bishop, Catherine Lord, Shrikanth Narayanan

Challenges in Automated Processing of Speech from Child Wearables: The Case of Voice Type Classifier........ 2840
Tarek Kunze, Marianne Métais, Hadrien Titeux, Lucas Elbert, Joseph Coffey, Emmanuel Dupoux, Alejandrina Cristia, Marvin Lavechin

Improving Child Speech Recognition and Reading Mistake Detection by Using Prompts........ 2845
Lingyun Gao, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik

VOLUME 5

Improving Automatic Speech Recognition for Children's Reading Assessment with Disfluency-Aware Language Models........ 2850
Jazmín Vidal, Luciana Ferrer, Juan Esteban Kamienkowski, Pablo Riera

Oral Reading Errors by Grade 3 Children in Indian Schools: A Hindi-English Perspective 2855
Sneha Raman, Preeti Rao

Grammatical Error Detection on Spontaneous Children's Speech Using Iterative Pseudo Labeling 2860
Christopher Gebauer, Lars Rumberg, Lars Köhn, Hanna Ehlert, Edith Beaulac, Jörn Ostermann

Why is Children's ASR So Difficult? Analyzing Children's Phonological Error Patterns Using SSL-Based Phoneme Recognizers 2865
Koharu Horii, Naohiro Tawara, Atsunori Ogawa, Shoko Araki

Automatic Detection of Speech Sound Disorders in German-Speaking Children: Augmenting the Data with Typically Developed Speech 2870
Darline Monika Marx, Marco Matassoni, Alessio Brutti

Continuous Learning for Children's ASR: Overcoming Catastrophic Forgetting with Elastic Weight Consolidation and Synaptic Intelligence 2875
Edem Ahadzi, Vishwanath Pratap Singh, Tomi Kinnunen, Ville Hautamaki

Exploring Shared-Weight Mechanisms in Transformer and Conformer Architectures for Automatic Speech Recognition 2880
Thomas Rolland, Alberto Abad

Advancing Pediatric ASR: The Role of Voice Generation in Disordered Speech 2885
Karen Rosero, Ali N Salman, Shreeram Chandra, Berrak Sisman, Cortney Van'T Slot, Alex Kane, Rami R Hallac, Carlos Busso

CHSER: A Dataset and Case Study on Generative Speech Error Correction for Child ASR 2890
Natarajan Balaji Shankar, Zilai Wang, Kaiyuan Zhang, Mohan Shi, Abeer Alwan

Causal Structure Discovery for Error Diagnostics of Children's ASR.. 2895
Vishwanath Pratap Singh, Md Sahidullah, Tomi Kinnunen

BRAIN AND COGNITION

Brain-Tuned Speech Models Better Reflect Speech Processing Stages in the Brain 2900
Omer Moussa, Mariya Toneva

Enhancing Syllabic Recognition Via Speech-EEG Phase Analysis and Non-Activity State Modeling 2905
Rini Sharon, Hema A. Murthy

Functional Connectivity and Hilbert-Based Features for Covert Speech EEG Variability Analysis and Classification 2910
Saravanakumar Duraisamy, Maurice Rekrut, Luis A. Leiva

Neuro2Semantic: A Transfer Learning Framework for Semantic Reconstruction of Continuous Language from Human Intracranial EEG 2915
Siavash Shams, Richard Antonello, Gavin Mischler, Stephan Bickel, Ashesh Mehta, Nima Mesgarani

Selective Auditory Attention Decoding in Naturalistic Conversations Using EEG-Based Speech Envelope Tracking in Multi-Speaker Environments 2920
Gabriel Ivucic, Saurav Pahuja, Dashanka Da Silva, Tanja Schultz

MiSTR: Multi-Modal iEEG-To-Speech Synthesis with Transformer-Based Prosody Prediction and Neural Phase Reconstruction.. 2925
Mohammed Al-Radhi, Géza Németh, Branislav Gerazov

REGIONAL, SOCIAL AND DIACHRONIC VARIATION

Probing Prosodic Differences Between Two Regional Varieties of Brazilian Portuguese 2930
Gustavo Silveira, Aviad Albert, Martine Grice

Data-Driven Approaches to Pitch Modelling in Two Mexican Spanish Ethnolects: K-Means Clustering & GAMMs 2935
Gilly Marchini, Jeremy Steffman

Tracking /R/ Deletion: Forced Alignment of Pronunciation Variants and Sociophonetic Insights into Post-Obstruent Final /R/ in French 2940
Anisia Popescu, Lori Lamel, Marc Evrard, Ioana Vasilescu

Agent-Based Modelling, Sound Change, and Metaphony in Southern Italian Varieties of Italo-Romance. 2945
Lilian Von Bressensdorf, Pia Greca, Jonathan Harrington

Modeling Vowel System Typology Using Iterated Confusion Minimization 2950
John McGahay

Investigating Glottal Stop Coda Loss During Sound Change of Checked Syllables Based on Speech-EGG Voice Offset Alignment 2955
Bingliang Zhao, Xiyu Wu

SPEAKER EXTRACTION 2

FlowTSE: Target Speaker Extraction with Flow Matching 2960
Aviv Navon, Aviv Shamsian, Yael Segal-Feldman, Neta Glazer, Gil Hetz, Joseph Keshet

MTSE: Multi-Target Speaker Extraction for Conversation Scenarios.. 2965
Thomas Serre, Mathieu Fontaine, Eric Benhaim, Slim Essid

Location-Aware Target Speaker Extraction for Hearing Aids 2970
Daniel-José Alcala Padilla, Nils L. Westhausen, Swati Vivekananthan, Bernd T. Meyer

ClearerVoice-Studio: Bridging Advanced Speech Processing Research and Practical Deployment 2975
Shengkui Zhao, Zexu Pan, Bin Ma

Online AV-CrossNet: A Causal and Efficient Audiovisual System for Speech Enhancement and Target Speaker Extraction 2980
Cheng Yu, Vahid Ahmadi Kalkhorani, Buye Xu, Deliang Wang

Steering Deep Non-Linear Spatially Selective Filters for Weakly Guided Extraction of Moving Speakers in Dynamic Scenarios 2985
Jakob Kienegger, Timo Gerkmann

MULTIMODAL EMOTION RECOGNITION

RA-CLAP: Relation-Augmented Emotional Speaking Style Contrastive Language-Audio Pretraining for Speech Retrieval... 2990
Haoqin Sun, Jingguang Tian, Jiaming Zhou, Hui Wang, Jiabei He, Shiwan Zhao, Xiangyu Kong, Desheng Hu, Xinkang Xu, Xinhui Hu, Yong Qin

EmotionRankCLAP: Bridging Natural Language Speaking Styles and Ordinal Speech Emotion Via Rank-N-Contrast 2995
Shreeram Suresh Chandra, Lucas Goncalves, Junchen Lu, Carlos Busso, Berrak Sisman

Modality-Agnostic Multimodal Emotion Recognition Using a Contrastive Masked Autoencoder 3000
Georgios Chochlakis, Turab Iqbal, Woo Hyun Kang, Zhaocheng Huang

Multi-Modal Multi-Task Affective States Recognition Based on Label Encoder Fusion........ 3005
Maxim Markitantov, Elena Ryumina, Heysem Kaya, Alexey Karpov

CONVERSATION, COMMUNICATION AND INTERACTION 2

Triadic Multi-Party Voice Activity Projection for Turn-Taking in Spoken Dialogue Systems........ 3010
Mikey Elmers, Koji Inoue, Divesh Lala, Tatsuya Kawahara

Continuous Prediction of Backchannel Timing for Human-Robot Interaction 3015
Michael Paierl, Martin Hagmüller, Barbara Schuppler

Impact of Background Noise on Turn-Taking Dynamics in Triadic Conversations........ 3020
Valeska Slomianka, Tobias May, Torsten Dau

Multimodal Dynamics of Hand Gestures and Pauses in Multiparty Interactions........ 3025
Delphine Charuau, Naomi Harte

TinyClick: Single-Turn Agent for Empowering GUI Automation 3030
Pawel Pawlowski, Krystian Zawistowski, Wojciech Lapacz, Adam Wiacek, Marcin Skorupa, Sebastien Postansque, Jakub Hoscilowicz

Improving User Impression of Spoken Dialogue Systems by Controlling Para-Linguistic Expression Based on Intimacy 3035
Shoki Kawanishi, Akinori Ito, Yuya Chiba, Takashi Nose

Dialogue Response Prefetching Based on Semantic Similarity and Prediction Confidence of Language Model........ 3040
Kiyotada Mori, Seiya Kawano, Angel García Contreras, Koichiro Yoshino

MULTIMODAL SPEECH AND LANGUAGE PROCESSING IN HEALTHCARE SETTINGS

Can Multimodal Foundation Models Help Analyze Child-Inclusive Autism Diagnostic Videos?........ 3045
Aditya Kommineni, Digbalay Bose, Tiantian Feng, So Hyun Kim, Helen Tager-Flusberg, Somer Bishop, Catherine Lord, Sudarsana Kadiri, Shrikanth Narayanan

A Cascaded Multimodal Framework for Automatic Social Communication Severity Assessment in Children with Autism Spectrum Disorder 3050
Jihyun Mun, Sunhee Kim, Minhwa Chung

Accessible Real-Time Eye-Gaze Tracking for Neurocognitive Health Assessment: A Multimodal Web-Based Approach 3055
Daniel Tisdale, Jackson Liscombe, David Paulter, Michael Neumann, Vikram Ramanarayanan

Multimodal Biomarkers for Schizophrenia: Towards Individual Symptom Severity Estimation 3060
Gowtham Premananth, Philip Resnik, Sonia Bansal, Deanna L. Kelly, Carol Espy-Wilson

Fact-Controlled Diagnosis of Hallucinations in Medical Text Summarization 3065
Suhas Bn, Han-Chin Shing, Lei Xu, Mitch Strong, Jon Burnsky, Jessica Ofor, Jordan R. Mason, Susan Chen, Sundararajan Srinivasan, Chaitanya Shivade, Jack Moriarty, Joseph Paul Cohen

Evaluating Automatic Speech Recognition Pipelines for Mandarin-English Bilingual Child Language Assessment in Telehealth 3070
Hongchen Wu, Yao Du, Zirong Li, Yixin Gu, Disha Thotappala Jayaprakash, Li Sheng

MUSIC AND AUDIO ANALYSIS

Enhancing Lyrics Transcription on Music Mixtures with Consistency Loss.. 3075
Jiawen Huang, Felipe Sousa, Emir Demirel, Emmanouil Benetos, Igor Gadelha

Tonality-Based Accompaniment-Guided Automatic Singing Evaluation... 3080
Pei-Chin Hsieh, Yih-Liang Shen, Ngoc-Son Tran, Tai-Shih Chi

Investigating the Reasonable Effectiveness of Speaker Pre-Trained Models and Their Synergistic Power for SingMOS Prediction 3085
Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Swarup Ranjan Behera, Pailla Balakrishna Reddy, Arun Balaji Buduru, Rajesh Sharma

Focal Modulation Network: A Novel Solution for Polyphonic Music Instrument Recognition Without Attention and Aggregation Strategy.. 3090
Lekshmi C R, Rajeev Rajan

A Joint Network for Singing Melody Extraction from Polyphonic Music with Attention Aggregation and Self-Consistency Training... 3095
Jiabo Jing, Ying Hu, Hao Huang, Liang He, Zhijian Ou

Position Also Matters! Separating Same Instruments in String Quartet Using Timbral and Positional Cues 3100
Yuetonghui Xu, Yiwen Wang, Xihong Wu, Xiaobing Li, Feng Yu

WhisperMSS: A Two-Stage Framework for Mandarin Singing Transcription and Segmentation Using Pretrained Models 3105
Ruoxuan Liang, Xiangjian Zeng, Zhen Liu, Qingqiang Wu, Ruichen Zhang, Le Ren

Low Complex IIR Adaptive Hear-Through Ambient Filtering for Overcoming Practical Constraints in Earbuds..3110
Rishabh Gupta, Mlns Karthik, Yughendaran P

Sub-Band Based Adaptive IIR Algorithm with Biquad Filter Stability Constraints for Feedforward Hear-Through Equalization 3115
Rishabh Gupta, Mlns Karthik, Chelamkuri Omsrinath

AUDIO ANALYSIS, GENERATION AND ASSESSMENT

Discrete Audio Representations for Automated Audio Captioning 3120
Jingguang Tian, Haoqin Sun, Xinhui Hu, Xinkang Xu

CLAP-ART: Automated Audio Captioning with Semantic-Rich Audio Representation Tokenizer 3125
Daiki Takeuchi, Binh Thien Nguyen, Masahiro Yasuda, Yasunori Ohishi, Daisuke Niizumi, Noboru Harada

Temp4Cap: Temporally-Aligned Automated Audio Captioning 3130
Ho-Young Choi, Jae-Heung Cho, Pil Moo Byun, Won-Gook Choi, Joon-Hyuk Chang

Optimizing CLAP Reward with LLM Feedback for Semantically Aligned and Diverse Automated Audio Captioning 3135
Seyun Ahn, Pil Moo Byun, Won-Gook Choi, Joon-Hyuk Chang

Bridging Audio and Vision: Zero-Shot Audiovisual Segmentation by Connecting Pretrained Models 3140
Seung-Jae Lee, Paul Hongsuck Seo

DiffStereo: End-To-End Mono-To-Stereo Audio Generation with Diffusion Transformer 3145
Suqi Zhang, Zheqi Dai, Yongyi Zang, Yin Cao, Qiuqiang Kong

RELATE: Subjective Evaluation Dataset for Automatic Evaluation of Relevance Between Text and Audio 3150
Yusuke Kanamori, Yuki Okamoto, Taisei Takano, Shinnosuke Takamichi, Yuki Saito, Hiroshi Saruwatari

Crowdsourcing MUSHRA Tests in the Age of Generative Speech Technologies: A Comparative Analysis of Subjective and Objective Testing Methods 3155
Laura Lechler, Chamran Moradi, Ivana Balic

SMARTMOS: Modeling Subjective Audio Quality Evaluation for Real-Time Applications 3160
Sivakumar Balasubramanian, Jose Antonio Jimenez Amador, Kaustubh Kalgaonkar, King Wei Hor, Sriram Srinivasan

OTHER TOPICS IN SPEECH RECOGNITION

Effect of Loudspeaker Emitted Speech on ASR Performance... 3165
Vikram C M, Sanjoy Pal, Nidhi Mantri, Gopal Kumar Agrawal

Contextualized Automatic Speech Recognition with Dynamic Vocabulary Prediction and Activation 3169
Zhennan Lin, Kaixun Huang, Wei Ren, Linju Yang, Lei Xie

Character Error Rate Estimation for Semi-Supervised Training of Speech Recognition for Arabic Dialects.. 3174
Chanho Park, Oscar Saz

Unified Semi-Supervised Pipeline for Automatic Speech Recognition.. 3179
Nune Tadevosyan, Nikolay Karpov, Andrei Andrusenko, Vitaly Lavrukhin, Ante Jukic

Scaling Laws for Synthetic Speech for Model Training.. 3184
Christoph Minixhofer, Ondrej Klejch, Peter Bell

R2S: Real-To-Synthetic Representation Learning for Training Speech Recognition Models on Synthetic Data 3189
Minh Tran, Debjyoti Paul, Yutong Pang, Laxmi Pandey, Jinxi Guo, Ke Li, Shun Zhang, Xuedong Zhang, Xin Lei

Context is All You Need? Low-Resource Conversational ASR Profits from Context, Coming from the Same Or from the Other Speaker.. 3194
Julian Linke, Jana Winkler, Barbara Schuppler

Automatic Speech Recognition Biases in Newcastle English: An Error Analysis 3199
Dana Serditova, Kevin Tang, Jochen Steffens

PRIVACY AND ANONYMIZATION

Speech Unlearning 3204
Jiali Cheng, Hadi Amiri

Unlearning LLM-Based Speech Recognition Models 3209
Zhe Liu

EASY: Emotion-Aware Speaker Anonymization Via Factorized Distillation 3214
Jixun Yao, Hexin Liu, Eng Siong Chng, Lei Xie

Private kNN-VC: Interpretable Anonymization of Converted Speech 3219
Carlos Franzreb, Arnab Das, Tim Polzehl, Sebastian Möller

Legally Validated Evaluation Framework for Voice Anonymization 3224
Nathalie Vauquier, Brij Mohan Lal Srivastava, Seyed Ahmad Hosseini, Emmanuel Vincent

LANGUAGE MODELING FOR CONVERSATIONAL SYSTEMS

Analyzing Mitigation Strategies for Catastrophic Forgetting in End-To-End Training of Spoken Language Models 3229
Chi-Yuan Hsiao, Ke-Han Lu, Kai-Wei Chang, Chih-Kai Yang, Wei-Chih Chen, Hung-Yi Lee

Speechless: Speech Instruction Training Without Speech for Low Resource Languages 3234
Alan Dao, Dinh Bach Vu, Huy Hoang Ha, Tuan Le Duc Anh, Shreyas Gopal, Yue Heng Yeo, Warren Keng Hoong Low, Eng Siong Chng, Jia Qi Yip

LiSTEN: Learning Soft Token Embeddings for Neural Audio LLMs 3239
Pooneh Mousavi, Shubham Gupta, Cem Subakan, Mirco Ravanelli

Cryfish: On Deep Audio Analysis with Large Language Models 3244
Anton Mitrofanov, Sergey Novoselov, Tatiana Prisyach, Vladislav Marchevskiy, Arseniy Karelin, Nikita Khmelev, Dmitry Dutov, Stepan Malykh, Igor Agafonov, Aleksandr Nikitin, Oleg Petrov

Improving Linguistic Diversity of Large Language Models with Possibility Exploration Fine-Tuning 3249
Long Mai, Julie Carson-Berndsen

OpusLM: A Family of Open Unified Speech Language Models 3254
Jinchuan Tian, William Chen, Yifan Peng, Jiatong Shi, Siddhant Arora, Shikhar Bharadwaj, Takashi Maekaku, Yusuke Shinohara, Keita Goto, Xiang Yue, Huck Yang, Shinji Watanabe

CAPR: Confidence-Aware Prompt Refinement in Large Language Models 3259
Jen-Tzung Chien, Po-Chun Huang

SPEECH ACCESSIBILITY PROJECT CHALLENGE

The Interspeech 2025 Speech Accessibility Project Challenge 3264
Xiuwen Zheng, Bornali Phukon, Jonghwan Na, Ed Cutrell, Kyu J. Han, Mark Hasegawa-Johnson, Pan-Pan Jiang, Aadhrik Kuila, Colin Lea, Bob Macdonald, Gautam Mantena, Venkatesh Ravichandran, Leda Sari, Katrin Tomanek, Chang D. Yoo, Chris Zwilling

Towards Inclusive and Fair ASR: Insights from the SAPC Challenge for Optimizing Disordered Speech Recognition 3269
Nada Gohider, Otman Basir

Robust Fine-Tuning of Speech Recognition Models Via Model Merging: Application to Disordered Speech 3274
Alexandre Ducorroy, Rachid Riad

Exploring Generative Error Correction for Dysarthric Speech Recognition 3279
Moreno La Quatra, Alkis Koudounas, Valerio Mario Salerno, Sabato Marco Siniscalchi

Pathology-Aware Speech Encoding and Data Augmentation for Dysarthric Speech Recognition........ 3284
Ilja Baumann, Dominik Wagner, Korbinian Riedhammer, Tobias Bocklet

Personalized Fine-Tuning with Controllable Synthetic Speech from LLM-Generated Transcripts for Dysarthric Speech Recognition 3289
Dominik Wagner, Ilja Baumann, Natalie Engert, Seanie Lee, Elmar Nöth, Korbinian Riedhammer, Tobias Bocklet

A Self-Training Approach for Whisper to Enhance Long Dysarthric Speech Recognition 3294
Shiyao Wang, Jiaming Zhou, Shiwan Zhao, Yong Qin

Fine-Tuning Parakeet-TDT for Dysarthric Speech Recognition in the Speech Accessibility Project Challenge........ 3299
Kaito Takahashi, Keigo Hojo, Toshimitsu Sakai, Yukoh Wakabayashi, Norihide Kitaoka

CBA-Whisper: Curriculum Learning-Based AdaLoRA Fine-Tuning on Whisper for Low-Resource Dysarthric Speech Recognition 3304
Tianyi Tan, Xinan Chen, Xiaohuai Le, Wenzhi Fan, Xianjun Xia, Chuanzeng Huang, Jing Lu

NEURAL NETWORK TRAINING METHODS 1

Cross-Lingual Data Selection Using Clip-Level Acoustic Similarity for Enhancing Low-Resource Automatic Speech Recognition 3309
Shunsuke Mitsumori, Sara Kashiwagi, Keitaro Tanaka, Shigeo Morishima

Spot and Merge: A Hybrid Context Biasing Approach for Rare Word and Out of Vocabulary Recognition 3314
Jatin Agrawal, Bramhendra Koilakuntla, Srikanth Konjeti

Accurate, Fast, Cheap: Choose Three. Replacing Multi-Head-Attention with Bidirectional Recurrent Attention for Long-Form ASR 3319
Martin Ratajczak, Jean-Philippe Robichaud, Jennifer Drexler Fox

Improving Cross-Attention Based on Positional Alignment During Inference for Robust Long-Form Speech Recognition 3324
Changhan Oh, Kiyoung Park, Jeomja Kang, Woo Yong Choi, Hwa Jeon Song

Improving End-To-End Mixed-Case ASR with Knowledge Distillation and Integration of Voice Activity Cues........ 3329
Sashi Novitasari, Takashi Fukuda, Gakuto Kurata

Cross-Modal Knowledge Transfer Learning as Graph Matching Based on Optimal Transport for ASR 3334
Xugang Lu, Peng Shen, Yu Tsao, Hisashi Kawai

DIVERSITY: AGE, SEX, GENDER, ETHNICITY, AND MORE

Age-Related Changes in Multisensory Integration of Emotions in an Audiovisual Face-Prosody-Semantics Stroop Task 3339
Yi Lin, Shumeng Ni, Yangfan Lu

Investigating Effects of Sex Hormones, Cycle Phases and Age on Female Fundamental Frequency 3344
Melanie Weirich, Adrian Simpson

Pre-Aspiration in Iceland is Conditioned by Gender/Sex 3349
Meike Rommel, Míša Hejná, Nicole Dehé

Transcribing Diverse Voices: Using Whisper for ICE Corpora 3354
Andreas Weilinghoff

Is it All About Race?: A Cross-Examination of /S/ in a Multilingual (Nigerian) Context 3359
Oluwasegun Amoniyan

Investigating Gender Bias in Text-To-Audio Generation Models 3364
Aarish Shah Mohsin, Mohammad Nadeem, Shahab Saquib Sohail, Tughrul Arsalan, Mandar Gogate, Nasir Saleem, Amir Hussain

ANOMALOUS SOUND DETECTION

Dual Orthogonality Sub-Center Loss for Enhanced Anomalous Sound Detection 3369
Dong Wang, Jiqing Han, Tieran Zheng, Guibin Zheng, Yongjun He

Adaptive Across-Subcenter Representation Learning for Imbalanced Anomalous Sound Detection 3374
Dong Wang, Jiqing Han, Guibin Zheng, Tieran Zheng, Yongjun He

Towards Few-Shot Training-Free Anomaly Sound Detection 3379
Ho-Hsiang Wu, Wei-Cheng Lin, Abinaya Kumar, Luca Bondi, Shabnam Ghaffarzadegan, Juan Pablo Bello

Finetune Large Pre-Trained Model Based on Frequency-Wise Multi-Query Attention Pooling for Anomalous Sound Detection 3384
Nan Jiang, Yan Song, Qing Gu, Haoyu Song, Lirong Dai, Ian McLoughlin

Acoustic Detection of UAV Abnormality Using One Ground-Based Acoustic Vector Sensor 3389
Dengjian Zhou, Jianghan Hai, Sijia Liao, Yue Ivan Wu, Kainam Thomas Wong, Xiujuan Zheng

StarGAN-Aug: A Cross-Domain Fault Audio Generation Method for High-Performance Fault Diagnosis of Power Transformers 3394
Ben Niu, Yangjie Wei, Gang Yang, Yuqiao Wang, Shengling Yu

FAR-FIELD AND ROBUST SPEECH RECOGNITION

SuPseudo: A Pseudo-Supervised Learning Method for Neural Speech Enhancement in Far-Field Speech Recognition 3399
Longjie Luo, Lin Li, Qingyang Hong

Lightweight Front-End Enhancement for Robust ASR Via Frame Resampling and Sub-Band Pruning 3404
Siyi Zhao, Wei Wang, Yanmin Qian

Calm-Whisper: Reduce Whisper Hallucination on Non-Speech by Calming Crazy Heads Down 3409
Yingzhi Wang, Anas Alhmoud, Saad Alsahly, Muhammad Alqurishi, Mirco Ravanelli

HuBERT-VIC: Improving Noise-Robust Automatic Speech Recognition of Speech Foundation Model Via Variance-Invariance-Covariance Regularization 3414
Hyebin Ahn, Kangwook Jang, Hoirin Kim

MOVER: Combining Multiple Meeting Recognition Systems 3419
Naoyuki Kamo, Tsubasa Ochiai, Marc Delcroix, Tomohiro Nakatani

EmbedAug: An Augmentation Scheme for End-To-End Automatic Speech Recognition 3424
Ashish Panda, Sunil Kumar Kopparapu

Attention Models and Auditory Transduction Features for Noise Robustness 3429
Cathal Ó Faoláin, Andrew Hines

Lightweight and Robust Multi-Channel End-To-End Speech Recognition with Spherical Harmonic Transform 3434
Xiangzhu Kong, Hao Huang, Zhijian Ou

On the Design of a Robust Superdirective Beamformer and Topology Parameter Optimization with Frustum-Shaped Microphone Arrays Featuring Multiple Rings 3439
Kunlong Zhao, Gongping Huang, Xudong Zhao, Jingdong Chen, Jacob Benesty, Zoran Cvetkovic

SPEECH SYNTHESIS PARADIGMS AND METHODS 2

Efficient Neural and Numerical Methods for High-QualityOnline Speech Spectrogram Inversion Via Gradient Theorem 3444
Andres Fernandez, Juan Azcarreta Ortiz, Çagdas Bilen, Jesus Monge Alvarez

Fine-Tuning Text-To-Speech Diffusion Models Using Reinforcement Learning with Human Feedback 3449
Jingyi Chen, Ju Seung Byun, Micha Elsner, Pichao Wang, Andrew Perrault

Accelerating Diffusion-Based Text-To-Speech Model Trainingwith Dual Modality Alignment 3454
Jeongsoo Choi, Zhikang Niu, Ji-Hoon Kim, Chunhui Wang, Joon Son Chung, Xie Chen

SpeechSEC: A Unified Multi-Task Framework for Speech Synthesis, Editing, and Continuation 3459
Liming Liang, Dongchao Yang, Xianwei Zhuang, Yuxin Xie, Luo Chen, Yuehan Jin, Yuexian Zou

VoiceNoNG: Robust High-Quality Speech Editing Model Without Hallucinations 3464
Sung-Feng Huang, Heng-Cheng Kuo, Zhehuai Chen, Xuesong Yang, Pin-Jui Ku, Ante Jukic, Huck Yang, Yu Tsao, Yu-Chiang Frank Wang, Hung-Yi Lee, Szu-Wei Fu

A Watermark for Auto-Regressive Speech Generation Models 3469
Yihan Wu, Ruibo Chen, Georgios Milis, Junfeng Guo, Heng Huang

ARTICULATORY ANALYSES

Influence of Wall Coverings of 3D-Printed Vocal Tract Models on Measured Transfer Functions 3474
Peter Birkholz, Dominik Schäfer, Patrick Häsner, Jihyeon Yun, Iris Kruppke, Rémi Blandin

Supralaryngeal Kinematics of Implosives in Central Vietnamese: An EMA Study 3479
Paul McGuire, Kye Shibata, Thanh Viet Cao, Feng-Fan Hsieh, Yueh-Chin Chang

Lateral Channel Formation in Australian English /L/: Insights from Magnetic Resonance Imaging 3484
Tünde Szalay, Michael Proctor, Amelia Gully, Tharinda Piyadasa, Craig Jin, David Waddington, Naeim Sanaei, Sheryl Foster, Kirrie Ballard

Articulatory Variations in Apical Vowels in Southwestern Mandarin 3489
Jing Huang, Feng-Fan Hsieh, Yueh-Chin Chang

Rhotic Articulation in Australian English: Insights from MRI 3494
Michael Proctor, Tünde Szalay, Tharinda Piyadasa, Craig Jin, Naeim Sanaei, Amelia Gully, David Waddington, Sheryl Foster, Kirrie Ballard

Articulatory Strategy in Vowel Production as a Basis for Speaker Discrimination 3499
Justin J. H. Lo, Patrycja Strycharczuk, Sam Kirkham

SPEECH AND AUDIO ANALYSIS AND REPRESENTATION

PAST: Phonetic-Acoustic Speech Tokenizer 3504
Nadav Har-Tuv, Or Tal, Yossi Adi

Factorized RVQ-GAN for Disentangled Speech Tokenization... 3509
Sameer Khurana, Dominik Klement, Antoine Laurent, Dominik Boboš, Juraj Novosad, Peter Gazdik, Ellen Zhang, Zili Huang, Amir Hussein, Ricard Marxer, Yoshiki Masuyama, Ryo Aihara, Chiori Hori, François G. Germain, Gordon Wichern, Jonathan Le Roux

EnCodecMAE: Leveraging Neural Codecs for Universal Audio Representation Learning 3514
Leonardo Pepino, Pablo Riera, Luciana Ferrer

AxLSTMs: Learning Self-Supervised Audio Representations with xLSTMs 3519
Sarthak Yadav, Sergios Theodoridis, Zheng-Hua Tan

SHOW AND TELL 3: SIGNAL PROCESSING / MULTIMODAL PROCESSING

Real-Time TSE Demonstration Via SoundBeam with KD 3524
Keigo Wakayama, Tomoko Kawase, Takafumi Moriya, Marc Delcroix, Hiroshi Sato, Tsubasa Ochiai, Masahiro Yasuda, Shoko Araki

Real-Time Diffusion Buffer for Speech Enhancement on a Laptop 3526
Bunlong Lay, Rostilav Makarov, Timo Gerkmann

Co-Speech Motion for Virtual Agents in Dialogue Using LLM-Driven Primitive Action Selection 3528
Muhammad Yeza Baihaqi, Angel García Contreras, Seiya Kawano, Koichiro Yoshino

TargetVoice: Single Channel Low-Latency Target Speaker Extraction 3530
Arun Kumar Pallala, Nivedita Chennupati, Balaji Padmanaban, Rakesh Pogula, Uma Subhashini Ravuri, Naveen Ellanki, Harish Rajamani, Naveen Ambati

Rollback Speech: Smart Feedback Prompts for Lost Utterances in Unstable Online Calls 3532
Yuni Amaloa Quintero Villalobos, Wafaa Wardah, Sebastian Möller, Robert P. Spang

Simultaneous Speech Translation Integrated Compact Multiple Sound Spot Synthesis System on a Laptop Carried Out with a Backpack 3534
Takuma Okamoto, Michiyo Kono

GenECA: A General-Purpose Framework for Real-Time Adaptive Multimodal Embodied Conversational Agents 3536
Santosh Patapati, Aashrith Tatineni, Trisanth Srinivasan

SPEECH AND VOICE DISORDERS 2

Hybrid Expert Knowledge and Self-Supervised Learning for Diagnostic Modeling of Adductor Spasmodic and Primary Myotonic Dysphonia 3538
Zhou Du, Hang Chen, Huijun Ding, Jun Du, Zhen Chen

MVP: Multi-Source Voice Pathology Detection 3543
Alkis Koudounas, Moreno La Quatra, Gabriele Ciravegna, Marco Fantini, Erika Crosetti, Giovanni Succo, Tania Cerquitelli, Sabato Marco Siniscalchi, Elena Baralis

Phonetic Posteriorgram-Based Phoneme Selection for Vocal Cord Disorder Classification in Continuous Mandarin Speech 3548
Chih-Ning Chen, Yu-Lan Chuang, Ming-Jhang Yang, Wei-Cheng Hsu, Yung-An Tsou, Yi-Wen Liu

Articulatory Clarity and Variability Before and After Surgery for Tongue Cancer 3553
Thomas Tienkamp, Fleur Van Ast, Roos Van Der Veen, Teja Rebernik, Raoul Buurke, Nikki Hoekzema, Katharina Polsterer, Hedwig Sekeres, Rob Van Son, Martijn Wieling, Max Witjes, Sebastiaan De Visscher, Defne Abur

Hybrid HMM-SVM Classifier Using Frication-Based Features for Detection of Non-Normative Sibilant Articulation Patterns in Polish Children's Speech 3558
Zuzanna Miodonska

Fine-Tuning ASR for Stuttered Speech: Personalized Vs. Generalized Approaches 3563
Dena Mujtaba, Nihar R. Mahapatra

NEURAL NETWORK TRAINING METHODS 2

SiamCTC: Learning Speech Representations Through Monotonic Temporal Alignment 3568
Soohwan Eom, Mark Hasegawa-Johnson, Chang D. Yoo

Improving Generalization of End-To-End ASR Through Diversity and Independence Regularization 3573
Ye-Eun Ko, Mun-Hak Lee, Dong-Hyun Kim, Joon-Hyuk Chang

Exploring Linear Variant Transformers and k-NN Memory Inference for Long-Form ASR 3578
Carlos Carvalho, Jinchuan Tian, William Chen, Yifan Peng, Alberto Abad, Shinji Watanabe

Attention-Free Dual-Mode ASR with Latency-Controlled Selective State Spaces 3583
Takafumi Moriya, Masato Mimura, Kiyoaki Matsui, Hiroshi Sato, Kohei Matsuura

VOLUME 6

Thinking Fast and Slow: Robust Speech Recognition Via Deep Filter-Tuning 3588
Dianwen Ng, Kun Zhou, Bin Ma, Eng Siong Chng

Towards Efficiently Whisper Fine-Tuning with Monotonic Alignments.. 3593
Ziyang Zhuang, Tao Wei, Ming Fang, Ning Cheng, Shaojun Wang, Jing Xiao

Dynamic Acoustic Model Architecture Optimization in Training for ASR 3598
Jingjing Xu, Zijian Yang, Albert Zeyer, Eugen Beck, Ralf Schlüter, Hermann Ney

Knowledge Distillation Method for Pruned RNN-T Models Via Pruning Bounds Sharing and Losses Confusion 3603
Xiaocan Zhang, Weiwei Jiang, Guibin Zheng, Chenhao Jing, Jiqing Han, Tieran Zheng

An Effective Training Framework for Light-Weight Automatic Speech Recognition Models....................... 3608
Abdul Hannan, Alessio Brutti, Shah Nawaz, Mubashir Noman

Better Semi-Supervised Learning for Multi-Domain ASR Through Incremental Retraining and Data Filtering.. 3613
Andrés Carofilis, Pradeep Rangappa, Srikanth Madikeri, Shashi Kumar, Sergio Burdisso, Jeena Prakash, Esaú Villatoro-Tello, Petr Motlicek, Bidisha Sharma, Kadri Hacioglu, Shankar Venkatesan, Saurabh Vyas, Andreas Stolcke

DISENTANGLEMENT OF INFORMATION FOR SPEAKER RECOGNITION

LASPA: Language Agnostic Speaker Disentanglement with Prefix-Tuned Cross-Attention......................... 3618
Aditya Srinivas Menon, Raj Prakash Gohil, Kumud Tripathi, Pankaj Wasnik

SCD-Conformer: Semantic Content Disentanglement for Text-Independent Speaker Verification 3623
Shanshan Yao, Dianlong Liu, Tian Li

Universal Semantic Disentangled Privacy-Preserving Speech Representation Learning.............................. 3628
Biel Tura-Vecino, Subhadeep Maji, Aravind Varier, Antonio Bonafonte, Ivan Valles, Michael Owen, Constantinos Papayiannis, Leif Radel, Grant Strimel, Oluwaseyi Feyisetan, Roberto Barra-Chicote, Ariya Rastrow, Volker Leutnant, Trevor Wood

Using Gender, Phonation and Age to Interpret Automatically Discovered Speech Attributes for Explainable Speaker Recognition.. 3633
Carole Millot, Clara Ponchard, Cédric Gendrot, Jean-François Bonastre, Orane Dufour

Do You Read Me? - Flow of Speech Effect on Speaker Recognition Systems 3638
Alicja Martinek, Joanna Gajewska, Ewelina Bartuzi-Trokielewicz

VoxAging: Continuously Tracking Speaker Aging with a Large-Scale Longitudinal Dataset in English and Mandarin.. 3643
Zhiqi Ai, Meixuan Bao, Zhiyong Chen, Zhi Yang, Xinnuo Li, Shugong Xu

ERROR CORRECTION AND CONFIDENCE ESTIMATION

LLM-Based Generative Error Correction for Rare Words with Synthetic Data and Phonetic Context.......... 3648
Natsuo Yamashita, Masaaki Yamamoto, Hiroaki Kokubo, Yohei Kawaguchi

ASR Confidence Estimation Using True Class Lexical Similarity Score 3653
Nagarathna Ravi, Thishyan Raj T, Ravi Teja Chaganti, Vipul Arora

Semi-Supervised Learning for Automatic Speech Recognition with Word Error Rate Estimation and Targeted Domain Data Selection 3658
Chanho Park, Thomas Hain

Voice Activity-Based Text Segmentation for ASR Text Denormalization 3663
Sashi Novitasari, Takashi Fukuda, Gakuto Kurata

Phonetically-Augmented Discriminative Rescoring for Voice Search Error Correction 3668
Christophe Van Gysel, Maggie Wu, Lyan Verwimp, Caglar Tirkaz, Marco Bertola, Zhihong Lei, Youssef Oualil

From Weak Labels to Strong Results: Utilizing 5,000 Hours of Noisy Classroom Transcripts with Minimal Accurate Data 3673
Ahmed Attia, Dorottya Demszky, Jing Liu, Carol Espy-Wilson

TRAINING AND SCORING METHODS FOR SPEAKER RECOGNITION

Boundary-Conscious Pruning: Hard Set-Aware Model Compression for Efficient Speaker Recognition 3678
Seongkyu Mun, Jubum Han

Pushing the Frontiers of Self-Distillation Prototypes Network with Dimension Regularization and Score Normalization 3683
Yafeng Chen, Chong Deng, Hui Wang, Yiheng Jiang, Han Yin, Qian Chen, Wen Wang

A Domain Robust Pre-Training Method with Local Prototypes for Speaker Verification 3688
Qing Gu, Yan Song, Haoyu Song, Nan Jiang, Lirong Dai, Ian McLoughlin

Clustering-Based Hard Negative Sampling for Supervised Contrastive Speaker Verification 3693
Piotr Masztalski, Michal Romaniuk, Jakub Zak, Mateusz Matuszewski, Konrad Kowalczyk

MASV: Speaker Verification with Global and Local Context Mamba 3698
Yang Liu, Li Wan, Yiteng Huang, Ming Sun, Xinhao Mei, Xubo Liu, Yangyang Shi, Florian Metze

ATMM-SAGA: Alternating Training for Multi-Module with Score-Aware Gated Attention SASV System 3703
Amro Asali, Yehuda Ben-Shimol, Itshak Lapidot

Rethinking Leveraging Pre-Trained Multi-Layer Representations for Speaker Verification 3708
Jin Sob Kim, Hyun Joon Park, Wooseok Shin, Sung Won Han

SEED: Speaker Embedding Enhancement Diffusion Model 3713
Kihyun Nam, Jungwoo Heo, Jee-Weon Jung, Gangin Park, Chaeyoung Jung, Ha-Jin Yu, Joon Son Chung

A Copula-Based Generative Score-Level Fusion Model for Speaker Verification 3718
Sandro Cumani

PATHOLOGICAL SPEECH ANALYSIS 2

Acoustic Similarities, Articulatory Uniqueness: Speech Production Mechanisms in Individuals with Congenital Lip Paralysis 3723
Anne Hermes, Ivana Didirková, Philipp Buech, Gilles Vannuscorps

Relationship Between Objective and Subjective Perceptual Measures of Speech in Individuals with Head and Neck Cancer 3728
Bence Mark Halpern, Thomas Tienkamp, Teja Rebernik, Rob J. J. H. Van Son, Martijn Wieling, Defne Abur, Tomoki Toda

Evaluating the Usefulness of Non-Diagnostic Speech Data for Developing Parkinson's Disease Classifiers 3733
Terry Yi Zhong, Esther Janse, Cristian Tejedor-Garcia, Louis Ten Bosch, Martha Larson

Multimodal Assessment of Speech Impairment in Amyotrophic Lateral Sclerosis Using Audio-Visual and Machine Learning Approaches 3738
Francesco Pierotti, Andrea Bandini

Development and Validation of a Wav2Vec 2.0-Based Cross-Language Methodology for Measurement of Articulatory Precision 3743
Tanya Talkar, Kan Kawabata, Connor Higgins, Sean Tobyne

J-J-J-Just Stutter: Benchmarking Whisper's Performance Disparities on Different Stuttering Patterns 3748
Charan Sridhar, Shaomei Wu

MULTIMODAL AND VISUAL SPEECH SYNTHESIS

MM-MovieDubber: Towards Multi-Modal Learning for Multi-Modal Movie Dubbing 3753
Junjie Zheng, Zihao Chen, Chaofan Ding, Yunming Liang, Yihan Fan, Huan Yang, Lei Xie, Xinhan Di

Learning Phonetic Context-Dependent Viseme for Enhancing Speech-Driven 3D Facial Animation 3758
Hyung Kyu Kim, Hak Gu Kim

Face2VoiceSync: Lightweight Face-Voice Consistency for Text-Driven Talking Face Generation 3763
Fang Kang, Yin Cao, Haoyu Chen

Revival with Voice: Multi-Modal Controllable Text-To-Speech Synthesis 3768
Minsu Kim, Pingchuan Ma, Honglie Chen, Stavros Petridis, Maja Pantic

VisualSpeech: Enhancing Prosody Modeling in TTS Using Video 3773
Shumin Que, Anton Ragni

LightL2S: Ultra-Low Complexity Lip-To-Speech Synthesis for Multi-Speaker Scenarios 3778
Yifan Liang, Kang Yang, Fangkun Liu, Andong Li, Xiaodong Li, Chengshi Zheng

LEXICON AND GRAMMAR

Processing of Grammatical Information in Cochlear Implant Simulated Speech by German Adult Listeners 3783
Atty Schouwenaars, Esther Ruigendijk

A Gradient Effect of Hand Beat Timing on Spoken Word Recognition 3788
Chengjia Ye, James M. McQueen, Hans Rutger Bosker

The Effect of Word Predictability on Spoken Cross-Language Intelligibility 3793
Wei Xue, Iuliia Zaitova, Bernd Möbius

Sentence-Final Particles in Mandarin Child-Directed Speech: Frequency and Impact on Speech Rate 3798
Yizhi Liu, Luyuan Geng, Yan Gu, Mengru Han

Bilingual Speakers Exhibit Cognitive Fatigue: A Speech Disfluencies Case Study on Research Talks 3803
Ashwin Ram, Marisol Muñoz, Zoi Gkalitsiou, Alexandros G. Dimakis

NOISE REDUCTION AND DEREVERBERATION

Boosting StoRM Convergence with Metric Guidance and Non-Uniform State-Sampling for Optimal Dereverberation 3808
Chandra Mohan Sharma, Arnab Kumar Roy, Anupam Mandal, Prasanta Kumar Ghosh, Prasanna Kumar Kr

Unified Variational and Physics-Aware Model for Room Impulse Response Estimation 3813
Louis Lalay, Mathieu Fontaine, Roland Badeau

MelRe: Vision-Based Mel-Spectrogram Restoration 3818
Kaixuan Luan, Xiaoda Yang, Shile Cai, Ruofan Hu, Minghui Fang, Wenrui Liu, Jialong Zuo, Jiaqi Duan, Yuhang Ma, Junyu Lu

SpeechRefiner: Towards Perceptual Quality Refinement for Front-End Algorithms 3823
Sirui Li, Shuai Wang, Zhijun Liu, Zhongjie Jiang, Yannan Wang, Haizhou Li

Modality-Specific Speech Enhancement and Noise-Adaptive Fusion for Acoustic and Body-Conduction Microphone Framework 3828
Yunsik Kim, Yoonyoung Chung

Joint Rate Allocation and Sensor Selection for Speech Enhancement in Wireless Acoustic Sensor Networks 3833
De Hu, Qilong Li

Individualized Speech Enhancement for Hearing-Impaired Listeners 3838
Chuan Wen, Sarah Verhulst

First Analyze Then Enhance: A Task-Aware System for Speech Separation, Denoising, and Dereverberation 3843
Shaoxiang Dang, Li Li, Shogo Seki, Hiroaki Kudo

A Robust Hybrid ACC-PM Approach for Personal Sound Zones 3848
Yaqi Zhu, Lei Zhou, Hongqing Liu, Liming Shi, Lu Gan

NEURAL NETWORK TRAINING METHODS AND ARCHITECTURES

Distilling a Speech and Music Encoder with Task Arithmetic 3853
Fabian Ritter-Gutierrez, Yi-Cheng Lin, Jui-Chiang Wei, Jeremy H. M Wong, Eng Siong Chng, Nancy F. Chen, Hung-Yi Lee

MSDA: Combining Pseudo-Labeling and Self-Supervision for Unsupervised Domain Adaptation in ASR 3858
Dimitrios Damianos, Georgios Paraskevopoulos, Alexandros Potamianos

REB-Former: RWKV-Enhanced E-Branchformer for Speech Recognition 3863
Jie Song, Wang Xiang, Jian Zhou, Cunhang Fan, Zhao Lv

PredTrAD – Prediction-Based Transformer for Anomaly Detection in Multivariate Time Series Data 3868
Jan Schuster, Alexander Wölfel, Fabian Brunner, Christian Bergler

FairASR: Fair Audio Contrastive Learning for Automatic Speech Recognition 3873
Jongsuk Kim, Jaemyung Yu, Minchan Kwon, Junmo Kim

Automatic Speech Recognition of African American English: Lexical and Contextual Effects 3878
Hamid Mojarad, Kevin Tang

Improving Synthetic Data Training for Contextual Biasing Models with a Keyword-Aware Cost Function 3883
Chin Yuen Kwok, Jia Qi Yip, Eng Siong Chng

SOMSRED-SVC: Sequential Output Modeling with Speaker Vector Constraints for Joint Multi-Talker Overlapped ASR and Speaker Diarization 3888
Naoki Makishima, Naotaka Kawata, Taiga Yamane, Mana Ihori, Tomohiro Tanaka, Satoshi Suzuki, Shota Orihashi, Ryo Masumura

Thinking in Directivity: Speech Large Language Model for Multi-Talker Directional Speech Recognition 3893
Jiamin Xie, Ju Lin, Yiteng Huang, Tyler Vuong, Zhaojiang Lin, Zhaojun Yang, Peng Su, Prashant Rawat, Sangeeta Srivastava, Ming Sun, Florian Metze

CHALLENGES IN SPEECH DATA COLLECTION, CURATION AND ANNOTATION - PART 1

A Study of Real-World Audio-Visual Corpus Design and Production: A Perspective from MISP Challenges 3898
Hang Chen, Jun Du, Qing Wang, Juan Xie, Shi-Fu Xiong

VCapAV: A Video-Caption Based Audio-Visual Deepfake Detection Dataset........ 3903
Yuxi Wang, Yikang Wang, Qishan Zhang, Hiromitsu Nishizaki, Ming Li

J-SPAW: Japanese Speaker Verification and Spoofing Attacks Recorded In-The-Wild Dataset........ 3908
Sayaka Shiota, Suzuka Horie, Kouta Kanno, Shinnosuke Takamichi

CommissionsQC: A Québec French Speech Corpus for Automatic Speech Recognition 3913
Coralie Serrand, Amira Morsli, Gilles Boulianne

Granary: Speech Recognition and Translation Dataset in 25 European Languages 3918
Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov#, Jagadeesh Balam, Vitaly Lavrukhin, Yifan Peng, Sara Papi, Marco Gaido, Alessio Brutti, Boris Ginsburg

Collecting, Curating, and Annotating Good Quality Speech Deepfake Dataset for Famous Figures: Process and Challenges 3923
Hashim Ali, Surya Subramani, Raksha Varahamurthy, Nithin Adupa, Lekha Bollinani, Hafiz Malik

Quantifying and Reducing Speaker Heterogeneity Within the Common Voice Corpus for Phonetic Analysis 3928
Miao Zhang, Aref Farhadipour, Annie Baker, Jiachen Ma, Bogdan Pricop, Eleanor Chodroff

The Speech Accessibility Project: Best Practices for Collection and Curation of Disordered Speech 3933
Chris Zwilling, Mark Hasegawa-Johnson, Heather Hodges, Lorraine Ramig, Adina Bradshaw, Clarion Mendes, Heejin Kim, Alexandria Barkhimer, Laura Mattie, Meg Dickinson, Shawnise Carter, Marie Moore Channell

Challenges and Practical Guidelines for Atypical Speech Data Collection, Annotation, Usage and Sharing: A Multi-Project Perspective 3938
Zhengjun Yue, Mara Barberis, Tanvina Patel, Judith Dineley, Willemijn Doedens, Lottie Stipdonk, Yuanyuan Zhang, Elke De Witte, Erfan Loweimi, Hugo Van Hamme, Djaina Satoer, Marina Ruiter, Laureano Moro Velazquez, Nicholas Cummins, Odette Scharenborg

Fifteen Years of Child-Centered Long-Form Recordings: Promises, Resources, and Remaining Challenges to Validity 3943
Loann Peurey, Marvin Lavechin, Tarek Kunze, Manel Khentout, Lucas Gautheron, Emmanuel Dupoux, Alejandrina Cristia

Contextual Paralinguistic Data Creation for Multi-Modal Speech-LLM: Data Condensation and Spoken QA Generation 3948
Qiongqiong Wang, Hardik B. Sailor, Tianchi Liu, Ai Ti Aw

Investigating Affect Mining Techniques for Annotation Sample Selection in the Creation of Finnish Affective Speech Corpus 3953
Kalle Lahtinen, Einari Vaaras, Liisa Mustanoja, Okko Räsänen

Scalable Spontaneous Speech Dataset (SSSD): Crowdsourcing Data Collection to Promote Dialogue Research 3958
Zaid Sheikh, Shuichiro Shimizu, Siddhant Arora, Jiatong Shi, Samuele Cornell, Xinjian Li, Shinji Watanabe

A Multimodal Chinese Dataset for Cross-Lingual Sarcasm Detection 3963
Xiyuan Gao, Bruce Xiao Wang, Meiling Zhang, Shuming Huang, Zhu Li, Shekhar Nayak, Matt Coler

Leveraging Large Language Models for Sarcastic Speech Annotation in Sarcasm Detection 3968
Zhu Li, Yuqing Zhang, Xiyuan Gao, Shekhar Nayak, Matt Coler

EVALUATION AND FORENSIC APPLICATIONS OF SPEAKER RECOGNITION

Analysis of the ABC Classification Backends for NIST SRE24 3973
Sandro Cumani, Anna Silnova, Sara Barahona, Ladislav Mošner, Oldrich Plchot, Johan Rohdin

STCON NIST SRE24 System: Composite Speaker Recognition Solution for Challenging Scenarios 3978
Stepan Malykh, Alexander Anikin, Nikita Khmelev, Anastasia Korenevskaya, Anastasia Zorkina, Sergey Novoselov, Vladislav Marchevskiy, Vladimir Volokhov, Andrey Shulipa, Alexander Kozlov, Alexander Melnikov, Vasiliy Galyuk, Timur Pekhovskiy

Vo-Ve: An Explainable Voice-Vector for Speaker Identity Evaluation 3983
Jaejun Lee, Kyogu Lee

Variability in Performance Across Four Generations of Automatic Speaker Recognition Systems 3988
Lauren Harrington, Vincent Hughes, Philip Harrison, Paul Foulkes, Jessica Wormald, Finnian Kelly, David Van Der Vloed

On the Influence of Language Similarity in Non-Target Speaker Verification Trials.................................... 3993
Paul M. Reuter, Michael Jessen

The Sub-3Sec Problem: From Text-Independent to Text-Dependent Corpus 3998
Ruichen Zuo, Kong Aik Lee, Zilong Huang, Man-Wai Mak

LANGUAGE RESOURCES

ToxicTone: A Mandarin Audio Dataset Annotated for Toxicity and Toxic Utterance Tonality 4003
Yu-Xiang Luo, Yi-Cheng Lin, Ming-To Chuang, Jia-Hung Chen, I-Ning Tsai, Pei Xing Kiew, Yueh-Hsuan Huang, Chien-Feng Liu, Yu-Chen Chen, Bo-Han Feng, Wenze Ren, Hung-Yi Lee

ViToSA: Audio-Based Toxic Spans Detection on Vietnamese Speech Utterances 4008
Huy Ba Do, Vy Le-Phuong Huynh, Luan Thanh Nguyen

Self-Supervised Models of Speech Processing for Haitian Creole... 4013
William N. Havard, Renauld Govain, Benjamin Lecouteux, Emmanuel Schang

AfriHuBERT: A Self-Supervised Speech Representation Model for African Languages 4018
Jesujoba O. Alabi, Xuechen Liu, Dietrich Klakow, Junichi Yamagishi

The Faetar Speech Recognition Benchmark.. 4023
Michael Ong, Sean Robertson, Leo Peckham, Alba Jorquera Jimenez De Aberasturi, Paula Arkhangorodsky, Robin Huo, Aman Sakhardande, Mark Hallap, Naomi Nagy, Ewan Dunbar

LHCP-ASR: An English Speech Corpus of High-Energy Particle Physics Talks for Narrow-Domain ASR Benchmarking... 4028
Jaume Santamaría-Jordà, Pablo Segovia-Martínez, Gonçal V. Garcés Díaz-Munío, Joan Albert Silvestre-Cerdà, Adrià Giménez, Rubén Gaspar Aparicio, René Fernández Sánchez, Jorge Civera, Albert Sanchis, Alfons Juan

Towards High-Quality LLM-Based Data for French Spontaneous Speech Simplification: An Exo-Refinement Approach.. 4033
Lucía Ormaechea, Nikos Tsourakis, Pierrette Bouillon, Benjamin Lecouteux, Didier Schwab

BR-ASR: Efficient and Scalable Bias Retrieval Framework for Contextual Biasing ASR in Speech LLM 4038
Xun Gong, Anqi Lv, Wangyou Zhang, Zhiming Wang, Huijia Zhu, Yanmin Qian

SPGISpeech 2.0: Transcribed Multi-Speaker Financial Audio for Speaker-Tagged Transcription 4043
Raymond Grossman, Taejin Park, Kunal Dhawan, Andrew Titus, Sophia Zhi, Yulia Shchadilova, Weiqing Wang, Jagadeesh Balam, Boris Ginsburg

Loquacious Set: 25,000 Hours of Transcribed and Diverse English Speech Recognition Data for Research and Commercial Use.. 4048
Titouan Parcollet, Yuan Tseng, Shucong Zhang, Rogier C. Van Dalen

CEREALES : A New Dataset of Quebec French Accented Speech with Applications to Speech Recognition 4053
Lucas Maison, Thomas Soulas, Marie-Jean Meurs

BANDWIDTH EXPANSION AND DIFFUSION-BASED SPEECH ENHANCEMENT

A Semantic Information-Based Hierarchical Speech Enhancement Method Using Factorized Codec and Diffusion Model 4058
Yang Xiang, Canan Huang, Desheng Hu, Jingguang Tian, Xinhui Hu, Chao Zhang

Voice-ENHANCE: Speech Restoration Using a Diffusion-Based Voice Conversion Framework 4063
Kyungguen Byun, Jason Filos, Erik Visser, Sunkuk Moon

SNR-Aligned Consistent Diffusion for Adaptive Speech Enhancement 4068
Yonghyeon Jun, Beomjun Woo, Myeonghun Jeong, Namsoo Kim

MDDM: A Multi-View Discriminative Enhanced Diffusion-Based Model for Speech Enhancement 4073
Nan Xu, Zhaolong Huang, Xiaonan Zhi

A Neural Codec Approach for Noise-Robust Bandwidth Expansion 4078
Xi Liu, Mu Yang, Szu-Jui Chen, John H. L. Hansen

HWB-Net: A Novel High-Performance and Efficient Hybrid Waveform Bandwidth Extension Method 4083
Xin Liu, Shulin He, Xueliang Zhang

Frequency-Domain Enhanced Extreme Bandwidth Extension Network with ICCRN for Superior Speech Quality 4088
Hongtao Bao, Xueliang Zhang

SPOKEN LANGUAGE UNDERSTANDING

QUADS: Quantized Distillation Framework for Efficient Speech Language Understanding 4093
Subrata Biswas, Mohammad Nur Hossain Khan, Bashima Islam

Spoken Language Understanding on Unseen Tasks with In-Context Learning 4098
Neeraj Agrawal, Sriram Ganapathy

Leveraging Information Retrieval to Enhance Spoken Language Understanding Prompts in Few-Shot Learning 4103
Pierre Lepagnol, Sahar Ghannay, Thomas Gerald, Christophe Servan, Sophie Rosset

Modeling Multi-Turn Spoken Language Understanding with Dynamic Graph Convolutional Networks 4108
Yi Huang, Si Chen, Jingyu Yao, Junlan Feng

DRI-GAN: A Novel Dual Real Input GAN with Triplet Loss for Cross-Lingual and Noisy SLU 4113
Ankit Kumar, Munir Georges

“KAN You Hear Me?” Exploring Kolmogorov-Arnold Networks for Spoken Language Understanding 4118
Alkis Koudounas, Moreno La Quatra, Eliana Pastor, Sabato Marco Siniscalchi, Elena Baralis

MULTILINGUAL SPEECH SYNTHESIS AND SPECIAL APPLICATIONS 2

Rasmalai : Resources for Adaptive Speech Modeling in IndiAn Languages with Accents and Intonations 4123
Ashwin Sankar, Yoach Lacombe, Sherry Thomas, Praveen Srinivasa Varadhan, Sanchit Gandhi, Mitesh M. Khapra

Kinship in Speech: Leveraging Linguistic Relatedness for Zero-Shot TTS in Indian Languages 4128
Utkarsh Pathak, Chandra Sai Krishna Gunda, Anusha Prakash, Keshav Agarwal, Hema A. Murthy

Can We Reconstruct a Dysarthric Voice with the Large Speech Model Parler TTS? 4133
Ariadna Sanchez, Simon King

Voice Adaptation for Swiss German 4138
Samuel Stucki, Jan Deriu, Mark Cieliebak

Gradual Modeling of the Lombard Effect by Modifying Speaker Embeddings from a Text-To-Speech Model 4143
Thiago Henrique Gomes Lobato, Magnus Schäfer

When Humans Growl and Birds Speak: High-Fidelity Voice Conversion from Human to Animal and Designed Sounds 4148
Minsu Kang, Seolhee Lee, Choonghyeon Lee, Namhyun Cho

EEG-Based Voice Conversion : Hearing the Voice of Your Brain 4153
Yizhong Geng, Wenxin Fu, Qihang Lu, Bingsong Bai, Cong Wang, Yingming Gao, Ya Li

Streaming Non-Autoregressive Model for Accent Conversion and Pronunciation Improvement 4158
Tuan-Nam Nguyen, Ngoc-Quan Pham, Seymanur Akti, Alexander Waibel

Zero-Shot Mono-To-Binaural Speech Synthesis 4163
Alon Levkovitch, Julian Salazar, Soroosh Mariooryad, Rj Skerry-Ryan, Nadav Bar, Bastiaan Kleijn, Eliya Nachmani

PROSODY AND VOICE QUALITY

Tone Recognition in Low-Resource Languages of North-East India: Peeling the Layers of SSL-Based Speech Models 4168
Parismita Gogoi, Sishir Kalita, Wendy Lalhminghlui, Viyazonuo Terhiija, Moakala Tzudir, Priyankoo Sarmah, S. R. M. Prasanna

Corpus-Based Insights into Mandarin Neutral Tone: Effects of Tonal Context and Structural Patterns in Spontaneous Speech 4173
Jingyi Sun, Nicolas Audibert, Yaru Wu, Martine Adda-Decker

Tonal Variation and Word Meaning in Taiwanese 4178
Yu-Ying Chuang, Sheng-Fu Wang

Sounding Like a Winner? Prosodic Differences in Post-Match Interviews 4183
Sofoklis Kakouros, Haoyu Chen

Exploratory Study of Filled Pauses in Ukrainian Language: Phonetic Properties of Filled Pauses 4188
Anna Havras, Carlos Mendes, Helena Moniz, Gueorgui Hristovsky, João Miranda

Evaluating the Suitability of Acoustic Parameters for Capturing Breathy Voice in Non-Pathological Female Speakers 4193
Chloe Patman, Paul Foulkes, Kirsty McDougall

Robustness of F0 Ratio as a Diagnostic: Comparing Creaky Voice in Danish and Seoul Korean 4198
Michaela Watkins, Rasmus Puggaard-Rode, Paul Boersma, Silke Hamann

GENERATIVE MODELS FOR AUDIO

Discovering Directions of Uncertainty in Speech Inpainting 4203
Kfir Cohen, Lior Wolf, Bracha Laufer-Goldshtein

InfiniteAudio: Infinite-Length Audio Generation with Consistency 4208
Chaeyoung Jung, Hojoon Ki, Ji-Hoon Kim, Junmo Kim, Joon Son Chung

FoleyMaster: High-Quality Video-To-Audio Synthesis Via MLLM-Augmented Prompt Tuning and Joint Semantic-Temporal Adaptation 4213
Liming Liang, Luo Chen, Yuehan Jin, Xianwei Zhuang, Yuxin Xie, Yongkang Yin, Yuexian Zou

Video-To-Audio Generation with Fine-Grained Temporal Semantics 4218
Yuchen Hu, Yu Gu, Chenxing Li, Rilin Chen, Dong Yu

TTMBA: Towards Text to Multiple Sources Binaural Audio Generation 4223
Yuxuan He, Xiaoran Yang, Ningning Pan, Gongping Huang

EzAudio: Enhancing Text-To-Audio Generation with Efficient Diffusion Transformer 4228
Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, Dong Yu

CHALLENGES IN SPEECH DATA COLLECTION, CURATION AND ANNOTATION - PART 2

You Are What You Say: Exploiting Linguistic Content for VoicePrivacy Attacks 4233
Ünal Ege Gaznepoglu, Anna Leschanowsky, Ahmad Aloradi, Prachi Singh, Daniel Tenbrinck, Emanuël A. P. Habets, Nils Peters

Recognizing Every Voice: Towards Inclusive ASR for Rural Bhojpuri Women 4238
Sakshi Joshi, Eldho Ittan George, Tahir Javed, Kaushal Bhogale, Nikhil Narasimhan, Mitesh M. Khapra

Augment Mandarin to Cantonese Speech Databases Via Retrieval-Augmented Generation and Speech Synthesis 4243
Fan Liu, Cheng Gong, Boyu Zhu, Ruihao Jing, Chunyu Qiang, Tianrui Wang, Xiao-Lei Zhang, Xuelong Li

An Exploratory Framework for LLM-Assisted Human Annotation of Speech Datasets 4248
Alexander Johnson, Harsh Deshpande, Emmy Phung, Ahmad Emami

Automatic Labeling and Correction of Noisy Labels for Robust Self-Supervised Speaker Verification 4253
Abderrahim Fathan, Jahangir Alam

Auto-Landmark: Acoustic Landmark Dataset and Open-Source Toolkit for Landmark Extraction 4258
Xiangyu Zhang, Daijiao Liu, Tianyi Xiao, Cihan Xiao, Tünde Szalay, Mostafa Shahin, Beena Ahmed, Julien Epps

AusKidTalk: Using Strategic Data Collection and Out-Of-Domain Tools to Semi-Automate Novel Corpora Annotation 4263
Tünde Szalay, Mostafa Shahin, Tharmakulasingam Sirojan, Zheng Nan, Renata Huang, Kirrie Ballard, Beena Ahmed

ASR-Based Segmentation for the Analysis of Larger Child-Speech Datasets: Performance Evaluation on Vowels from Australian-English Speaking Children Aged 4 to 11 Years 4268
Rui Cai, Titia Benders

A Semi-Automatic Pipeline for Transcribing and Segmenting Child Speech 4273
Polychronia Christodoulidou, James Tanner, Jane Stuart-Smith, Michael McAuliffe, Mridhula Murali, Amy Smith, Lauren Taylor, Joanne Cleland, Anja Kuschmann

Hybrid Data Sampling for ASR: Integrating Acoustic Diversity and Transcription Uncertainty 4278
Komei Hiruta, Yosuke Yamano, Hideaki Tamori

Whilter: A Whisper-Based Data Filter for "In-The-Wild" Speech Corpora Using Utterance-Level Multi-Task Classification 4283
William Ravenscroft, George Close, Kit Bower-Morris, Jamie Stacey, Dmitry Sityaev, Kris Y. Hong

Adapting Whisper for Low-Resource Hindi-English Code-Mix Speech with On-The-Fly Augmentation & LLM-Synthesised Data 4288
Astik Biswas, Oleg Shevelev, Amine Abdaoui, Vivek Tyagi, Abdelmoumene Boumadane

Optimizing ASR for Catalan-Spanish Code-Switching: A Comparative Analysis of Methodologies 4293
Carlos Mena, Pol Serra, Jacobo Romero, Abir Messaoudi, Jose Giraldo, Carme Armentano-Oller, Rodolfo Zevallos, Ivan Meza, Javier Hernando

From Scarcity to Sufficiency: Speech Recognition Pipeline for Zero-Resource Language 4298
Nikolay Karpov, Sofia Kostandian, Nune Tadevosyan, Alexan Ayrapetyan, Andrei Andrusenko, Ara Yeroyan, Mher Yerznkanyan, Vitaly Lavrukhin

MIKU-PAL: An Automated and Standardized Multimodal Method for Speech Paralinguistic and Affect Labeling 4303
Yifan Cheng, Ruoyi Zhang, Jiatong Shi

Multimodal Fusion with Semi-Supervised Learning Minimizes Annotation Quantity for Modeling Videoconference Conversation Experience 4308
Andrew Chang, Chenkai Hu, Ji Qi, Zhuojian Wei, Kexin Zhang, Viswadruth Akkaraju, David Poeppel, Dustin Freeman

Clinical Annotations for Automatic Stuttering Severity Assessment 4313
Ana Valente, Rufael Marew, Hawau Toyin, Hamdan Al-Ali, Anelise Bohnen, Inma Becerra, Elsa Soares, Gonçalo Leal, Hanan Aldarmaki

SPEECH EMOTION RECOGNITION 3

Frozen Large Language Models Can Perceive Paralinguistic Aspects of Speech 4318
Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, Ozlem Kalinli

VOLUME 7

AA-SLLM: An Acoustically Augmented Speech Large Language Model for Speech Emotion Recognition 4323
Jialong Mai, Xiaofen Xing, Weidong Chen, Yuanbo Fang, Xiangmin Xu

Speaker-Aware Multi-Task Learning for Speech Emotion Recognition........ 4328
Xiaohan Shi, Xingfeng Li, Tomoki Toda

Multimodal Emotion Diarization: Frame-Wise Integration of Text and Audio Representations........ 4333
Ziv Tamir, Thomas Thebaud, Jesus Villalba, Najim Dehak, Oren Kurland

Analysis of Phonetic Level Similarities Across Languages in Emotional Speech 4338
Pravin Mote, Abinay Reddy Naini, Donita Robinson, Elizabeth Richerson, Carlos Busso

Label Semantic-Driven Contrastive Learning for Speech Emotion Recognition........ 4343
Jiaxi Hu, Leyuan Qu, Haoxun Li, Taihao Li

Pitch Contour Model (PCM) with Transformer Cross-Attention for Speech Emotion Recognition 4348
Minji Ryu, Ji-Hyeon Hur, Sung Heuk Kim, Gahgene Gweon

EMOTION AND EXPRESSIVITY IN SPEECH SYNTHESIS AND VOICE CONVERSION

EATS-Speech: Emotion-Adaptive Transformation and Priority Synthesis for Zero-Shot Text-To-Speech 4353
Jingyuan Xing, Zhipeng Li, Shuaiqi Chen, Xiaofen Xing, Xiangmin Xu

Voice Impression Control in Zero-Shot TTS 4358
Kenichi Fujita, Shota Horiguchi, Yusuke Ijima

EME-TTS: Unlocking the Emphasis and Emotion Link in Speech Synthesis 4363
Haoxun Li, Leyuan Qu, Jiaxi Hu, Taihao Li

DiEmo-TTS: Disentangled Emotion Representations Via Self-Supervised Distillation for Cross-Speaker Emotion Transfer in Text-To-Speech........ 4368
Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Seong-Whan Lee

Spotlight-TTS: Spotlighting the Style Via Voiced-Aware Style Extraction and Style Direction Adjustment for Expressive Text-To-Speech 4373
Nam-Gyu Kim, Deok-Hyeon Cho, Seung-Bin Kim, Seong-Whan Lee

Speaker-Agnostic Emotion Vector for Cross-Speaker Emotion Intensity Control 4378
Masato Murata, Koichi Miyazaki, Tomoki Koriyama

SA-RAS: Speaker-Aware Style Retrieval Augmented Generation for Expressive Zero-Shot Text-To-Speech Synthesis........ 4383
Xueru Li, Jingyuan Xing, Xiaofen Xing, Zhipeng Li, Xiangmin Xu

DiffEmotionVC: A Dual-Granularity Disentangled Diffusion Framework for Any-To-Any Emotional Voice Conversion 4388
Xiaosu Su, Bowen Yang, Xiaowei Yi, Yun Cao

ZSDEVC: Zero-Shot Diffusion-Based Emotional Voice Conversion with Disentangled Mechanism........ 4393
Hsing-Hang Chou, Yun-Shao Lin, Ching-Chin Sung, Yu Tsao, Chi-Chun Lee

MPE-TTS: Customized Emotion Zero-Shot Text-To-Speech Using Multi-Modal Prompt 4398
Zhichao Wu, Yueteng Kang, Songjun Cao, Long Ma, Qiulin Li, Qun Yang

STREAMING ASR

MFLA: Monotonic Finite Look-Ahead Attention for Streaming Speech Recognition 4403
Yinfeng Xia, Huiyan Li, Chenyang Le, Manhong Wang, Yutao Sun, Xingyang Ma, Yanmin Qian

Delayed-KD: Delayed Knowledge Distillation Based CTC for Low-Latency Streaming ASR 4408
Longhao Li, Yangze Li, Hongfei Xue, Jie Liu, Shuai Fang, Kai Wang, Lei Xie

Parameter-Efficient Fine-Tuning of Conformer-Based Streaming Speech Recognition into Non-Streaming Models 4413
Yunjae Nam, Jeong U Han, Kiyeon Kim, Jaemin Lim

On-Device Streaming Discrete Speech Units 4418
Kwanghee Choi, Masao Someki, Emma Strubell, Shinji Watanabe

Adapting Whisper for Streaming Speech Recognition Via Two-Pass Decoding 4423
Haoran Zhou, Xingchen Song, Brendan Fahy, Qiaochu Song, Binbin Zhang, Zhendong Peng, Anshul Wadhawan, Denglin Jiang, Apurv Verma, Vinay Ramesh, Srivas Prasad, Michele M. Franceschini

Dynamic Context-Aware Streaming Pretrained Language Model for Inverse Text Normalization 4428
Luong Ho, Khanh Le, Vinh Pham, Bao Nguyen, Tan Tran, Duc Chau

L1 AND L2 ACQUISITION, PERCEPTION AND PROCESSING

Evaluating Progress of CALL System Users on Accentedness and Comprehensibility: An Acoustic and ASR-Based Approach 4433
Wenwei Dong, Catia Cucchiarini, Roeland Van Hout, Helmer Strik

Does English Fish Sound Like French Fiche? Perceptual Similarity Judgments Versus Acoustic Similarity 4438
Rory Turnbull, Elisa Kiefer, Sharon Peperkamp

Acoustic Features of Mandarin Tone Production in Noise: A Comparison Between Chinese Native Speakers and Korean L2 Learners 4443
Jinxin Ji, Yiying Hu, Xiaohu Yang, Gang Peng

The Role of Contextual Variation in Learning Cantonese Tones from Naturalistic Speech 4448
Fengyue Lisa Zhao, Jennifer Kuo

Pitch Target Realization in Putonghua Tone Production of Children from Dialect-Speaking Regions 4453
Mengxue Cao, Tianxin Zheng, Jiewen Zheng

The Development of Speech Rhythm in Putonghua-Learning Preschool Children in South Xinjiang Uyghur Autonomous Region of China 4458
Aijun Li, Zhiwei Wang, Jun Gao, Xin Zhou

SPEECH EMOTION RECOGNITION 2

PARROT: Synergizing Mamba and Attention-Based SSL Pre-Trained Models Via Parallel Branch Hadamard Optimal Transport for Speech Emotion Recognition 4463
Orchid Chetia Phukan, Mohd Mujtaba Akhtar, Girish, Swarup Ranjan Behera, Jaya Sai Kiran Patibandla, Arun Balaji Buduru, Rajesh Sharma

Towards Machine Unlearning for Paralinguistic Speech Processing 4468
Orchid Chetia Phukan, Girish, Mohd Mujtaba Akhtar, Shubham Singh, Swarup Ranjan Behera, Vandana Rajan, Muskaan Singh, Arun Balaji Buduru, Rajesh Sharma

Infant Cry Emotion Recognition Using Improved ECAPA-TDNN with Multi-Scale Feature Fusion and Attention Enhancement 4473
Junyu Zhou, Yanxiong Li, Haolin Yu

Speech Mutil-Label Emotion Recognition Using Asymmetric Class Loss Function Based on Effective Samples 4478
Shanshan Xiang, Hankiz Yilahun, Askar Hamdulla

EmoDB 2.0: A Database of Emotional Speech in a World that is Not Black Or White but Grey 4483
Felix Burkhardt, Oliver Schrüfer, Uwe Reichel, Hagen Wierstorf, Anna Derington, Florian Eyben, Björn W. Schuller

Cross-Corpus Open-Set Speech Emotion Recognition Method Based on Spatiotemporal Features with Inverse-Entropy Regularization 4488
Zhaohui Zhou, Hui Luo

CLEP-DG: Contrastive Learning for Speech Emotion Domain Generalization Via Soft Prompt Tuning 4493
Jiacheng Shi, Yanfu Zhang, Ye Gao

Leveraging Unlabeled Audio-Visual Data in Speech Emotion Recognition Using Knowledge Distillation 4498
Varsha Pendyala, Pedro Morgado, William Sethares

SPEAKER TRAITS RECOGNITION

Who Knows Best? Effects of Speech Disfluencies on Incentivized Decision-Making 4503
Ambika Kirkland, Jens Edlund

Enhancing Transcripts of Open-Source Automatic Speech Recognition Models Through Fine-Tuning with Laughter and Speech-Laugh 4508
Phuoc Hoang Ho, Drago? Alexandru Balan, Dirk K. J. Heylen, Khiet P. Truong

Investigating the Reasoning Abilities of Large Language Models for Understanding Spoken Language in Interpersonal Interactions 4513
Pranjal Aggarwal, Ghritachi Mahajani, Pavan Kumar Malasani, Vaibhav Jamadagni, Caroline J. Wendt, Ehsanul Haque Nirjhar, Theodora Chaspari

A Naturally Elicited Multimodal Stress Database and Speech Breathing Based Stress Detection 4518
Karumannil Mohamed Ismail Yasar Arafath, Mohammed Abeer K. C., Aurobinda Routray

From Context to Code-Switching: Examining the Interplay of Language Proficiency and Multilingualism in Speech 4523
Debasmita Bhattacharya, Aanya Tolat, Julia Hirschberg

Extending the Fongbe to French Speech Translation Corpus: Resources, Models and Benchmark 4528
D. Fortuné Kponou, Salima Mdhaffar, Fréjus A. A. Laleye, Eugène C. Ezin, Yannick Estève

On the Relationship Between Accent Strength and Articulatory Features 4533
Kevin Huang, Sean Foley, Jihwan Lee, Yoonjeong Lee, Dani Byrd, Shrikanth Narayanan

A Multi-Stream Framework Utilizing 3D Human Reconstruction for Cued Speech Recognition 4538
Katerina Papadimitriou, Gerasimos Potamianos

On the Cross-Modal Makeup of Charisma: Insights from a Field-Data Analysis 4543
Oliver Niebuhr

SPOOFING AND ADVERSARIAL ATTACKS

Generalizable Audio Spoofing Detection Using Non-Semantic Representations 4548
Arnab Das, Yassine El Kheir, Carlos Franzreb, Tim Herzig, Tim Polzehl, Sebastian Möller

Adversarial Attacks on Text-Dependent Speaker Verification System 4553
Sreekanth Sankala, Venkatesh Parvathala, Ramesh Gundluru, K. Sri Rama Murty

Beyond Attacks: Advancing Fake Speech Detection with Attack-Agnostic Methods 4558
Shilpa Chandra, Akansha Tyagi, Shiven Patel, Padmanabhan Rajan

ASVspoof2019 Vs. ASVspoof5: Assessment and Comparison... 4563
Avishai Weizman, Yehuda Ben-Shimol, Itshak Lapidot

Evaluating Parameter Sharing for Spoofing-Aware Speaker Verification: A Case Study on the ASVspoof 5 Dataset 4568
Aykut Büker, Oguzhan Kurnaz, Sule Bekiryazici, Selim Can Demirtas, Cemal Hanilçi

Can Quantized Audio Language Models Perform Zero-Shot Spoofing Detection? 4573
Bikash Dutta, Rishabh Ranjan, Shyam Sathvik, Mayank Vatsa, Richa Singh

VOICE CONVERSION 2

ClapFM-EVC: High-Fidelity and Flexible Emotional Voice Conversion with Dual Control from Natural Language and Speech 4578
Yu Pan, Yanni Hu, Yuguang Yang, Jixun Yao, Jianhao Ye, Hongbin Zhou, Lei Ma, Jianjun Zhao

PromptEVC: Controllable Emotional Voice Conversion with Natural Language Prompts............................ 4583
Tianhua Qi, Shiyan Wang, Cheng Lu, Tengfei Song, Hao Yang, Zhanglin Wu, Wenming Zheng

StarVC: A Unified Auto-Regressive Framework for Joint Text and Speech Generation in Voice Conversion.. 4588
Fengjin Li, Jie Wang, Yadong Niu, Yongqing Wang, Meng Meng, Jian Luan, Zhiyong Wu

FasterVoiceGrad: Faster One-Step Diffusion-Based Voice Conversion with Adversarial Diffusion Conversion Distillation.. 4593
Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Yuto Kondo

PATHOLOGICAL SPEECH ANALYSIS 3

Evaluating the Effectiveness of Pre-Trained Audio Embeddings for Classification of Parkinson's Disease Speech Data........ 4598
Emmy Postma, Cristian Tejedor-Garcia

On-The-Fly Routing for Zero-Shot MoE Speaker Adaptation of Speech Foundation Models for Dysarthric Speech Recognition 4603
Shujie Hu, Xurong Xie, Mengzhe Geng, Jiajun Deng, Huimeng Wang, Guinan Li, Chengxi Deng, Tianzi Wang, Mingyu Cui, Helen Meng, Xunying Liu

Lightweight Speech Enhancement for Mandarin Esophageal Speech........ 4608
Jia-Jyu Su, Yen-Ting Lin, Wu-Hao Li, Chao-Kai Chang, Yan-Zhi Chen, Chen-Yu Chiang

VocalAgent: Large Language Models for Vocal Health Diagnostics with Safety-Aware Evaluation........ 4613
Yubin Kim, Taehan Kim, Wonjune Kang, Eugene Park, Joonsik Yoon, Dongjae Lee, Xin Liu, Daniel McDuff, Hyeonhoon Lee, Cynthia Breazeal, Hae Won Park

A Cookbook for Community-Driven Data Collection of Impaired Speech in Low-Resource Languages........ 4618
Sumaya Ahmed Salihs, Isaac Wiafe, Jamal-Deen Abdulai, Elikem Doe Atsakpo, Gifty Ayoka, Richard Cave, Akon Obu Ekpezu, Catherine Holloway, Katrin Tomanek, Fiifi Baffoe Payin Winful

Voice Quality Dimensions as Interpretable Primitives for Speaking Style for Atypical Speech and Affect........ 4623
Jaya Narain, Vasudha Kowtha, Colin Lea, Lauren Tooley, Dianna Yee, Vikramjit Mitra, Zifang Huang, Miquel Espi Marques, Jon Huang, Carlos Avendano, Shirley Ren

SPEECH EMOTION RECOGNITION IN NATURALISTIC CONDITIONS CHALLENGE

Towards LLM-Empowered Fine-Grained Speech Descriptors for Explainable Emotion Recognition........ 4628
Youjun Chen, Xurong Xie, Haoning Xu, Mengzhe Geng, Guinan Li, Chengxi Deng, Huimeng Wang, Shujie Hu, Xunying Liu

From Pretraining to Performance: Benchmarking Self-Supervised Speech Models for Interspeech-25 SER Challenge 4633
Drishya Uniyal, Vinayak Abrol

Developing a Top-Tier Framework in Naturalistic Conditions Challenge for Categorized Emotion Prediction: From Speech Foundation Models and Learning Objective to Data Augmentation and Engineering Choices........ 4638
Tiantian Feng, Thanathai Lertpetchpun, Dani Byrd, Shrikanth Narayanan

Developing a High-Performance Framework for Speech Emotion Recognition in Naturalistic Conditions Challenge for Emotional Attribute Prediction........ 4643
Thanathai Lertpetchpun, Tiantian Feng, Dani Byrd, Shrikanth Narayanan

EmoSphere-SER: Enhancing Speech Emotion Recognition Through Spherical Representation with Auxiliary Classification........ 4648
Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Seong-Whan Lee

Explainable Speech Emotion Recognition Through Attentive Pooling: Insights from Attention-Based Temporal Localization 4653
Tahitoa Leygue, Astrid Sabourin, Christian Bolzmacher, Sylvain Bouchigny, Margarita Anastassova, Quoc-Cuong Pham

ABHINAYA - A System for Speech Emotion Recognition in Naturalistic Conditions Challenge 4658
Soumya Dutta, Smruthi Balaji, Varada R, Viveka Salinamakki, Sriram Ganapathy

The Interspeech 2025 Challenge on Speech Emotion Recognition in Naturalistic Conditions 4663
Abinay Reddy Naini, Lucas Goncalves, Ali N. Salman, Pravin Mote, Ismail R. Ulgen, Thomas Thebaud, Laureano Moro Velazquez, Leibny Paola Garcia, Najim Dehak, Berrak Sisman, Carlos Busso

MATER: Multi-Level Acoustic and Textual Emotion Representation for Interpretable Speech Emotion Recognition 4668
Hyo Jin Jon, Longbin Jin, Hyuntaek Jung, Hyunseo Kim, Donghun Min, Eun Yi Kim

Multi-Task Learning for Speech Emotion Recognition in Naturalistic Conditions 4673
Bartlomiej Zgórzynski, Juliusz Wójtowicz-Kruk, Piotr Masztalski, Wladyslaw Sredniawa

Medusa: A Multimodal Deep Fusion Multi-Stage Training Framework for Speech Emotion Recognition in Naturalistic Conditions 4678
Georgios Chatzichristodoulou, Despoina Kosmopoulou, Antonios Kritikos, Anastasia Poulopoulou, Efthymios Georgiou, Athanasios Katsamanis, Vassilis Katsouros, Alexandros Potamianos

Interactive Fusion of Multi-View Speech Embeddings Via Pretrained Large-Scale Speech Models for Speech Emotional Attribute Prediction in Naturalistic Conditions 4683
Yuyun Liu, Yujia Gu, Jiahao Luo, Wenming Zheng, Cheng Lu, Yuan Zong

Advancing Emotion Recognition Via Ensemble Learning: Integrating Speech, Context, and Text Representations 4688
Xiaohan Shi, Jinyi Mi, Xingfeng Li, Tomoki Toda

Improving Speech Emotion Recognition Through Cross Modal Attention Alignment and Balanced Stacking Model 4693
Lucas Ueda, João Lima, Leonardo Marques, Paula Costa

EmoJudge: LLM Based Post-Hoc Refinement for Multimodal Speech Emotion Recognition 4698
Prabhav Singh, Jesus Villalba

Lessons Learnt: Revisit Key Training Strategies for Effective Speech Emotion Recognition in the Wild 4703
Jing-Tong Tzeng, Bo-Hao Su, Ya-Tse Wu, Hsing-Hang Chou, Chi-Chun Lee

Enhancing Speech Emotion Recognition with Multi-Task Learning and Dynamic Feature Fusion 4708
Honghong Wang, Jing Deng, Fanqin Meng, Rong Zheng

PROSODY, PHONEME AND STRESS MODELING IN ASR

WhiStress: Enriching Transcriptions with Sentence Stress Detection 4713
Iddo Yosha, Dorin Shteyman, Yossi Adi

Prosodic Structure Beyond Lexical Content: A Study of Self-Supervised Learning 4718
Sarenne Wallbridge, Christoph Minixhofer, Catherine Lai, Peter Bell

Learning Optimal Prosody Embedding Codebook Based on F0 and Energy 4723
David Porteš, Aleš Horák

Pitch Accent Detection Improves Pretrained Automatic Speech Recognition... 4728
David Sasu, Natalie Schluter

Towards Accurate Phonetic Error Detection Through Phoneme Similarity Modeling.................................. 4733
Xuanru Zhou, Jiachen Lian, Cheol Jun Cho, Tejas Prabhune, Shuhe Li, William Li, Rodrigo Ortiz, Zoe Ezzes, Jet Vonk, Brittany Morin, Rian Bogley, Lisa Wauters, Zachary Miller, Maria Gorno-Tempini, Gopala Anumanchipalli

Exploring Auditory Feedback Mechanisms in Speech Recognition... 4738
Louise Coppieters De Gibson, Philip N. Garner

SEGMENTS

French Schwa is Not Acoustically Distinct from Its Two Lexical Neighbors /ø/ and /œ/ 4743
Mathilde Hutin, Mélanie Lancien, Noam Faust

Apical Vs. Regular Vowel Duration: A Corpus-Based Analysis of Contextual Influences in Standard Mandarin 4748
Jingyi Sun, Bowei Shao, Martine Adda-Decker

On Apical Vowels in Eastern Zhenjiang Mandarin.. 4753
Xuying Wang, Fang Hu

Equivalence and Differences: Formant Patterns of Labialization and Pharyngealization in Tashlhiyt 4758
Philipp Buech, Anne Hermes, Rachid Ridouane

Temporal Organization of Prenuclear Glides in Hefei Mandarin 4763
Yifan Yang, Zhiheng Qian

Speaker-Specific Patterns of Phonetic Covariation in Korean Word-Medial Stops and the Role of Phonological and Morphological Contexts 4768
Chloe D. Kwon

DATASETS AND TOOLS FOR SPEECH SYNTHESIS

HiFiTTS-2: A Large-Scale High Bandwidth Speech Dataset... 4773
Ryan Langman, Xuesong Yang, Paarth Neekhara, Shehzeen Hussain, Edresson Casanova, Evelina Bakhturina, Jason Li

JIS: A Speech Corpus of Japanese Idol Speakers with Various Speaking Styles.. 4778
Yuto Kondo, Hirokazu Kameoka, Kou Tanaka, Takuhiro Kaneko

FaVC: A Validated, Transcribed, Parallel Farsi Speech Dataset for Voice Conversion 4783
Mina Serajian, Saeed Najafzadeh Rahaghi, Hadi Veisi, Saman Haratizadeh

SawtArabi: A Benchmark Corpus for Arabic TTS. Standard, Dialectal and Code-Switching........................ 4788
Vasista Sai Lodagala, Lamya Alkanhal, Daniel Izham, Shivam Mehta, Shammur Absar Chowdhury, Aqeelah Makki, Hamdy S. Hussein, Gustav Eje Henter, Ahmed Ali

The Text-To-Speech in the Wild (TITW) Database 4793
Jee-Weon Jung, Wangyou Zhang, Soumi Maiti, Yihan Wu, Xin Wang, Ji-Hoon Kim, Yuta Matsunaga, Seyun Um, Jinchuan Tian, Hye-Jin Shim, Nicholas Evans, Joon Son Chung, Shinnosuke Takamichi, Shinji Watanabe

Towards Emotionally Consistent Text-Based Speech Editing: Introducing EmoCorrector and the ECD-TSE Dataset 4798
Rui Liu, Pu Gao, Jiatian Xi, Berrak Sisman, Carlos Busso, Haizhou Li

ArVoice: A Multi-Speaker Dataset for Arabic Speech Synthesis 4803
Hawau Toyin, Rufael Marew, Humaid Alblooshi, Samar M. Magdy, Hanan Aldarmaki

A Dataset for Automatic Assessment of TTS Quality in Spanish 4808
Alejandro Sosa Welford, Leonardo Pepino

SPOKEN DIALOGUE SYSTEMS 2

Factors Affecting the In-Context Learning Abilities of LLMs for Dialogue State Tracking 4813
Pradyoth Hegde, Santosh Kesiraju, Ján Švec, Šimon Sedlácek, Bolaji Yusuf, Oldrich Plchot, Deepak K T, Jan Cernocký

Spoken Question Answering for Visual Queries 4818
Nimrod Shabtay, Zvi Kons, Avihu Dekel, Hagai Aronowitz, Ron Hoory, Assaf Arbelle

Towards Human-Like Multimodal Conversational Agent by Generating Engaging Speech 4823
Taesoo Kim, Yongsik Jo, Hyunmin Song, Taehwan Kim

Chain-Of-Thought Training for Open E2E Spoken Dialogue Systems 4828
Siddhant Arora, Jinchuan Tian, Hayato Futami, Jee-Weon Jung, Jiatong Shi, Yosuke Kashiwagi, Emiru Tsunoo, Shinji Watanabe

SPEECH ENHANCEMENT AND REPRESENTATION LEARNING

SaD: A Scenario-Aware Discriminator for Speech Enhancement 4833
Xihao Yuan, Siqi Liu, Yan Chen, Hang Zhou, Chang Liu, Hanting Chen, Jie Hu

Listen Through the Sound: Generative Speech Restoration Leveraging Acoustic Context Representation 4838
Soo-Whan Chung, Min-Seok Choi

Efficient Speech Enhancement Via Embeddings from Pre-Trained Generative Audioencoders 4843
Xingwei Sun, Heinrich Dinkel, Yadong Niu, Linzhang Wang, Junbo Zhang, Jian Luan

Towards Personalised Audio Visual Speech Enhancement 4848
Mandar Gogate, Kia Dashtipour, Amir Hussain

FlowSE: Efficient and High-Quality Speech Enhancement Via Flow Matching 4853
Ziqian Wang, Zikai Liu, Xinfa Zhu, Yike Zhu, Mingshuai Liu, Jun Chen, Longshuai Xiao, Chao Weng, Lei Xie

Speech Enhancement Based on Cascaded Two Flows 4858
Seonggyu Lee, Sein Cheong, Sangwook Han, Kihyuk Kim, Jong Won Shin

X-ARES: A Comprehensive Framework for Assessing Audio Encoder Performance 4863
Junbo Zhang, Heinrich Dinkel, Yadong Niu, Chenyu Liu, Si Cheng, Anbei Zhao, Jian Luan

WavShape: Information-Theoretic Speech Representation Learning for Fair and Privacy-Aware Audio Processing 4868
Oguzhan Baser, Ahmet Ege Tanriverdi, Kaan Kale, Sandeep Chinchali, Sriram Vishwanath

NEURAL CODECS AND VOCODERS

FreeCodec: A Disentangled Neural Speech Codec with Fewer Tokens 4873
Youqiang Zheng, Weiping Tu, Yueteng Kang, Jie Chen, Yike Zhang, Li Xiao, Yuhong Yang, Long Ma

DualCodec: A Low-Frame-Rate, Semantically-Enhanced Neural Audio Codec for Speech Generation 4878
Jiaqi Li, Xiaolong Lin, Zhekai Li, Shixi Huang, Yuancheng Wang, Chaoren Wang, Zhenpeng Zhan, Zhizheng Wu

Comparative Analysis of Fast and High-Fidelity Neural Vocoders for Low-Latency Streaming Synthesis in Resource-Constrained Environments 4883
Reo Yoneyama, Masaya Kawamura, Ryo Terashima, Ryuichi Yamamoto, Tomoki Toda

Prosody-Adaptable Audio Codecs for Zero-Shot Voice Conversion Via In-Context Learning 4888
Junchuan Zhao, Xintong Wang, Ye Wang

Vocoder-Projected Feature Discriminator 4893
Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Yuto Kondo

AF-Vocoder: Artifact-Free Neural Vocoder with Global Artifact Filter 4898
Zhuangqi Chen, Xianjun Xia, Xiaohuai Le, Siyu Sun, Chuanzeng Huang

DS-Codec: Dual-Stage Training with Mirror-To-NonMirror Architecture Switching for Speech Codec 4903
Peijie Chen, Wenhao Guan, Kaidi Wang, Weijie Wu, Hukai Huang, Qingyang Hong, Lin Li

PeriodCodec: A Pitch-Controllable Neural Audio Codec Using Periodic Signals for Singing Voice Synthesis 4908
Masato Takagi, Miku Nishihara, Yukiya Hono, Kei Hashimoto, Yoshihiko Nankaku, Keiichi Tokuda

ADAPTATION AND TARGET-SPEAKER ASR

Enhancing Target-Speaker Automatic Speech Recognition Using Multiple Speaker Embedding Extractors with Virtual Speaker Embedding 4913
Ju-Seok Seong, Jeong-Hwan Choi, Ye-Rin Jeoung, Ilseok Kim, Joon-Hyuk Chang

SC-SOT: Conditioning the Decoder on Diarized Speaker Information for End-To-End Overlapped Speech Recognition 4918
Yuta Hirano, Sakriani Sakti

Efficient Data Selection for Domain Adaptation of ASR Using Pseudo-Labels and Multi-Stage Filtering 4923
Pradeep Rangappa, Andrés Carofilis, Jeena Prakash, Shashi Kumar, Sergio Burdisso, Srikanth Madikeri, Esaú Villatoro-Tello, Bidisha Sharma, Petr Motlicek, Kadri Hacioglu, Shankar Venkatesan, Saurabh Vyas, Andreas Stolcke

MOPSA: Mixture of Prompt-Experts Based Speaker Adaptation for Elderly Speech Recognition 4928
Chengxi Deng, Xurong Xie, Shujie Hu, Mengzhe Geng, Yicong Jiang, Jiankun Zhao, Jiajun Deng, Guinan Li, Youjun Chen, Huimeng Wang, Haoning Xu, Mingyu Cui, Xunying Liu

Visually-Adaptive Guided Robust Speech Recognition with Parameter-Efficient Adaptation 4933
Zhao Yang, Rui Jiang, Yue Heng Yeo, Xiao Fu, Wei Xi, Jizhong Zhao

Regularizing Learnable Feature Extraction for Automatic Speech Recognition 4938
Peter Vieting, Maximilian Kannen, Benedikt Hilmes, Ralf Schlüter, Hermann Ney

MMLoRA: Multitask Memory Parameter-Efficient Fine-Tuning for Multimodal SER 4943
Yuanbo Fang, Xiaofen Xing, Xueru Li, Weibin Zhang, Xiangmin Xu

Robust Unsupervised Adaptation of a Speech Recogniser Using Entropy Minimisation and Speaker Codes 4948
Rogier C. Van Dalen, Shucong Zhang, Titouan Parcollet, Sourav Bhattacharya

SHOW AND TELL 4: EDUCATION / ASSISTIVE TECHNOLOGY

SCRIBAL: A Digital Transcription Tool in Higher Education 4953
Javier Román, Pol Pastells, Mauro Vázquez, Clara Puigventós, Montserrat Nofre, Mariona Taulé, Mireia Farrús

From Static to Dynamic: Enhancing AAC with Generative Imagery and Zero-Shot TTS 4955
Juliana Francis, Joakim Gustafsson, Éva Székely

Concurrent Speech and Auditory Tag Clouds for Non-Visual Web Interaction 4958
Dhia Eddine Merzougui, Nilesh Tete, Fabrice Maurel, Gaël Dias, Mohammed Hasanuzzaman, Aurélien Bournonville, Edgar Madelaine, Thomas Berthelin Le Tellier, François Ledoyen, Laure Poutrain-Lejeune, François Rioult, Jérémie Pantin

Towards Domain-Specific Spoken Language Understanding for a Catalan Voice-Controlled Video Game 4960
Alex Peiró-Lilja, Rodolfo Zevallos, Carme Armentano-Oller, Jose Giraldo, Cristina España-Bonet, Mireia Farrús

Accessible Delivery of Visual-Acoustic Biofeedback for Speech Sound Disorder 4962
Tara McAllister, Peter Traver, Amanda Eads, William Haack, Helen Carey, Yi Shan, Wendy Liang, Tae Hong Park

End-To-End Indian Language Dubbing with Zero-Shot Speaker Preservation 4964
Giri Raju, Sandeep Konam

SOURCE SEPARATION 2

Band-SCNet: A Causal, Lightweight Model for High-Performance Real-Time Music Source Separation 4966
Junqi Yang, Yuhong Yang, Weiping Tu, Xin Zhao, Cedar Lin

CabinSep: IR-Augmented Mask-Based MVDR for Real-Time In-Car Speech Separation with Distributed Heterogeneous Arrays 4971
Runduo Han, Yanxin Hu, Yihui Fu, Zihan Zhang, Yukai Jv, Li Chen, Lei Xie

DGMO: Training-Free Audio Source Separation Through Diffusion-Guided Mask Optimization 4976
Geonyoung Lee, Geonhee Han, Paul Hongsuck Seo

Cross-Attention-Based Target Sound Extraction by Fully Leveraging Enrollment in a Shared Latent Space 4981
Xue Yang, Guiru Shen, Yu Yang

DnR-Nonverbal: Cinematic Audio Source Separation DatasetContaining Non-Verbal Sounds 4986
Takuya Hasumi, Yusuke Fujita

Neural Speech Extraction with Human Feedback 4991
Malek Itani, Ashton Graves, Sefik Emre Eskimez, Shyamnath Gollakota

SPEECH CODING

Unlocking Temporal Flexibility: Neural Speech Codec with Variable Frame Rate 4996
Hanglei Zhang, Yiwei Guo, Zhihan Li, Xiang Hao, Xie Chen, Kai Yu

SPCODEC: Split and Prediction for Neural Speech Codec 5001
Liang Wen, Lizhong Wang, Yuxing Zheng, Weijing Shi, Kwang Pyo Choi

Probing the Robustness Properties of Neural Speech Codecs 5006
Wei-Cheng Tseng, David Harwath

LSCodec: Low-Bitrate and Speaker-Decoupled Discrete Speech Codec..........5011
Yiwei Guo, Zhihan Li, Chenpeng Du, Hankun Wang, Xie Chen, Kai Yu

Bringing Interpretability to Neural Audio Codecs.......... 5016
Samir Sadok, Julien Hauret, Éric Bavu

NanoCodec: Towards High-Quality Ultra Fast Speech LLM Inference.......... 5021
Edresson Casanova, Paarth Neekhara, Ryan Langman, Shehzeen Hussain, Subhankar Ghosh, Xuesong Yang, Ante Jukic, Jason Li, Boris Ginsburg

MULTIMODALITY

Beat Gestures Made by Human-Like Avatars Affect Speech Perception 5026
Matteo Maran, Renske Rötjes, Anna R. E. Schreurs, Hans Rutger Bosker

The Mutual Exclusivity Bias of Bilingual Visually Grounded Speech Models.......... 5031
Dan Oneata, Leanne Nortje, Yevgen Matusevych, Herman Kamper

MultiActor-Audiobook: Zero-Shot Audiobook Generation with Faces and Voices of Multiple Speakers 5036
Kyeongman Park, Seongho Joo, Kyomin Jung

Incorporating Linguistic Constraints from External Knowledge Source for Audio-Visual Target Speech Extraction.......... 5041
Wenxuan Wu, Shuai Wang, Xixin Wu, Helen Meng, Haizhou Li

Multimodal Silent Recognition of Phonemes Using Radar and Optopalatographic Silent Speech Interfaces 5046
João Menezes, Aubin Mouras, Arne-Lukas Fietkau, Dani Kazzy, Peter Birkholz

SPEECH ASSESSMENT AND LANGUAGE LEARNING

GoP2Vec: A Few Shot Learning for Pronunciation Assessment with Goodness of Pronunciation (GoP) Based Representations from an I-Vector Framework and Augmentation 5051
Meenakshi Sirigiaju, Chiranjeevi Yarra

Enhancing GOP in CTC-Based Mispronunciation Detection with Phonological Knowledge 5056
Aditya Kamlesh Parikh, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik

VOLUME 8

Multilingual Speech Assessment Using Cross-Attention and Multitask Learning 5061
Sehyun Oh, Minhwa Chung, Sunhee Kim

Assessment of L2 Oral Proficiency Using Speech Large Language Models 5066
Rao Ma, Mengjie Qian, Siyuan Tang, Stefano Bannò, Kate M. Knill, Mark J. F. Gales

Scaling and Prompting for Improved End-To-End Spoken Grammatical Error Correction 5071
Mengjie Qian, Rao Ma, Stefano Bannò, Kate M. Knill, Mark J. F. Gales

Bidirectional Spoken-Written Text Conversion with Large Language Models 5076
Muyeol Choi, Hyunjung Choi, Yohan Lim, Jeonguk Bang, Minkyu Lee, Seonhui Kim, Seung Yun, Donghyun Kim, Minsoo Kim, Sanghun Kim

WATERMARKING AND ANONYMIZATION

WAKE: Watermarking Audio with Key Enrichment 5081
Yaoxun Xu, Jianwei Yu, Hangting Chen, Zhiyong Wu, Xixin Wu, Dong Yu, Rongzhi Gu, Yi Luo

Defend for Self-Vocoding: A Novel Enhanced Decoder Network for Watermark Recovery 5086
Yu-Sheng Lin, Ching-Yu Yang, Hsing-Hang Chou, Ya-Tse Wu, Bo-Hao Su, Chi-Chun Lee

Cross-Modal Watermarking for Authentic Audio Recovery and Tamper Localization in Synthesized Audiovisual Forgeries 5091
Minyoung Kim, Sehwan Park, Sungmin Cha, Paul Hongsuck Seo

VoiceMark: Zero-Shot Voice Cloning-Resistant Watermarking Approach Leveraging Speaker-Specific Latents 5096
Haiyun Li, Zhiyong Wu, Xiaofeng Xie, Jingran Xie, Yaoxun Xu, Hanyang Peng

A Comprehensive Real-World Assessment of Audio Watermarking Algorithms: Will They Survive Neural Codecs?.. 5101
Yigitcan Özer, Woosung Choi, Joan Serrà, Mayank Kumar Singh, Wei-Hsiang Liao, Yuki Mitsufuji

How to Recover Long Audio Sequences Through Gradient Inversion Attack with Dynamic Segment-Based Reconstruction... 5106
Xijie Zeng, Frank Rudzicz

First Steps Towards Voice Anonymization for Code-Switching Speech 5111
Sarina Meyer, Ekaterina Kolos, Ngoc Thang Vu

Exploiting Context-Dependent Duration Features for Voice Anonymization Attack Systems 5116
Natalia Tomashenko, Emmanuel Vincent, Marc Tommasi

Mitigating Language Mismatch in SSL-Based Speaker Anonymization 5121
Zhe Zhang, Wen-Chin Huang, Xin Wang, Xiaoxiao Miao, Junichi Yamagishi

SINGLE-CHANNEL SPEECH ENHANCEMENT

MSFNet: A Nested Model for Multi-Sampling-Frequency Speech Enhancement 5126
Venkatesh Parvathala, K. Sri Rama Murty

TF-SkiMNet: Speech Enhancement Based on Inplace Modeling and Skipping Memory in Time-Frequency Domain 5131
Zixuan Li, Shulin He, Jinglin Bai, Xueliang Zhang

XLSTM-SENet: xLSTM for Single-Channel Speech Enhancement.. 5136
Nikolai Lund Kühne, Jan Østergaard, Jesper Jensen, Zheng-Hua Tan

From KAN to GR-KAN: Advancing Speech Enhancement with KAN-Based Methodology........................ 5141
Haoyang Li, Yuchen Hu, Chen Chen, Sabato Marco Siniscalchi, Songting Liu, Eng Siong Chng

Stack Less, Repeat More: A Block Reusing Approach for Progressive Speech Enhancement....................... 5146
Jangyeon Kim, Ui-Hyeop Shin, Jaehyun Ko, Hyung-Min Park

Mamba-Based Hybrid Model for Speech Enhancement 5151
Se-Ha Kim, Tae-Gyeong Kim, Chang-Jae Chun

Restoring Harmonics: Enhancing Speech Quality with Deep Mask and Harmonic Restoration Network.. 5156
Yu Zhao, Zengqiang Shang, Mou Wang, Xin Liu, Pengyuan Zhang

CONTEXTUAL BIASING AND ADAPTATION

GLCLAP: A Novel Contrastive Learning Pre-Trained Model for Contextual Biasing in ASR...................... 5161
Yuxiang Kong, Fan Cui, Liyong Guo, Heinrich Dinkel, Lichun Fan, Junbo Zhang, Jian Luan

WCTC-Biasing: Retraining-Free Contextual Biasing ASR with Wildcard CTC-Based Keyword Spotting and Inter-Layer Biasing.. 5166
Yu Nakagome, Michael Hentschel

Ranking and Selection of Bias Words for Contextual Bias Speech Recognition 5171
Haoxiang Hou, Xun Gong, Wangyou Zhang, Wei Wang, Yanmin Qian

OWSM-Biasing: Contextualizing Open Whisper-Style Speech Models for Automatic Speech Recognition with Dynamic Vocabulary 5176
Yui Sudo, Yusuke Fujita, Atsushi Kojima, Tomoya Mizumoto, Lianbo Liu

Label-Context-Dependent Internal Language Model Estimation for CTC.. 5181
Zijian Yang, Minh-Nghia Phan, Ralf Schlüter, Hermann Ney

Assessing the Performance and Efficiency of Mamba ASR in Low-Resource Scenarios............................. 5186
Rodolfo Zevallos, Martí Cortada Garcia, Sarah Solito, Carlos Mena, Alex Peiró-Lilja, Javier Hernando

Adapting Whisper for Parameter-Efficient Code-Switching Speech Recognition Via Soft Prompt Tuning 5191
Hongli Yang, Yizhou Peng, Hao Huang, Sheng Li

SPEAKER DIARIZATION 2

Mitigating Non-Target Speaker Bias in Guided Speaker Embedding 5196
Shota Horiguchi, Takanori Ashihara, Marc Delcroix, Atsushi Ando, Naohiro Tawara

Speaker Diarization with Overlapping Community Detection Using Graph Attention Networks and Label Propagation Algorithm 5201
Zhaoyang Li, Jie Wang, Xiaoxiao Li, Wangjie Li, Longjie Luo, Lin Li, Qingyang Hong

Pushing the Limits of End-To-End Diarization 5206
Samuel J. Broughton, Lahiru Samarakoon

Spatio-Spectral Diarization of Meetings by Combining TDOA-Based Segmentation and Speaker Embedding-Based Clustering 5211
Tobias Cord-Landwehr, Tobias Gburrek, Marc Deegen, Reinhold Haeb-Umbach

Selective Channel Attention Based Target Speaker Voice Activity Detection for Speaker Diarization Under AD-HOC Microphone Array Settings 5216
Hongyu Zhang, Ming Cheng, Jing Feng, Ming Li

Diarization-Guided Multi-Speaker Embeddings 5221
Joonas Kalda, Clément Pagés, Tanel Alumäe, Hervé Bredin

Streaming Sortformer: Speaker Cache-Based Online Speaker Diarization with Arrival-Time Ordering 5226
Ivan Medennikov, Taejin Park, Weiqing Wang, He Huang, Kunal Dhawan, Jinhan Wang, Jagadeesh Balam, Boris Ginsburg

A Hybrid Approach to Combining Role Diarization with ASR for Professional Conversations........ 5231
Bongjun Kim, Arindam Ghosh, Mark C. Fuhs, Anurag Chowdhury, Deblin Bagchi, Monika Woszczyna

DEPRESSION DETECTION AND ASSESSMENT 2

An Interpretable Speech Foundation Model for Depression Detection by Revealing Prediction-Relevant Acoustic Features from Long Speech........ 5236
Qingkun Deng, Saturnino Luz, Sofia De La Fuente Garcia

Speech and Text Foundation Models for Depression Detection: Cross-Task and Cross-Language Evaluation........ 5241
Lucía Gómez-Zaragozá, Javier Marín-Morales, Mariano Alcañiz, Mohammad Soleymani

A Study on the Impact of Foundation Models on Automatic Depression Detection from Speech Signals 5246
Bubai Maji, Monorama Swain, Shazia Nasreen, Debabrata Majumdar, Rajlakshmi Guha, Aurobinda Routray, Anders Søgaard

Identifying Vocal and Facial Biomarkers of Depression in Large-Scale Remote Recordings: A Multimodal Study Using Mixed-Effects Modeling........ 5251
Nelson Hidalgo Julia, Robert Lewis, Craig Ferguson, Simon Goldberg, Wendy Lau, Caroline Swords, Gabriela Valdivia, Christine Wilson-Mendenhall, Raquel Tartar, Rosalind Picard, Richard Davidson

M3L: A Multi-Modal and Multi-Lingual Depression Detection Framework 5256
Jiajun You, Shuai Wang, Xun Gong, Xiang Wan

PATHOLOGICAL SPEECH ANALYSIS 4

On the Relevance of Clinical Assessment Tasks for the Automatic Detection of Parkinson’s Disease Medication State from Speech 5261
David Gimeno-Gómez, Rubén Solera-Ureña, Anna Pompili, Carlos-D. Martínez-Hinarejos, Rita Cardoso, Isabel Guimarães, Joaquim J. Ferreira, Alberto Abad

Speech Power Spectra: A Window into Neural Oscillations in Parkinson's Disease 5266
Sevada Hovsepyan, Mathew Magimai Doss

Synchronous Analysis of Abnormal Acoustic and Linguistic Production in Parkinson's Speech.. 5271
Daniel Escobar-Grisales, Cristian David Ríos-Urrego, Sabato Marco Siniscalchi, Adolfo M. Garcia, Yamile Bocanegra, Leonardo Moreno, Elmar Nöth, Juan Rafael Orozco-Arroyave

Automatic Detection and Sub-Typing of Primary Progressive Aphasia from Speech: Integrating Task-Specific Features and Spatio-Semantic Graphs 5276
Fritz Peters, W Richard Bevan-Jones, Grace Threlfall, Jenny M Harris, Julie S Snowden, Matthew Jones, Jennifer C Thompson, Daniel J Blackburn, Heidi Christensen

Towards Classification of Typical and Atypical Disfluencies: A Self Supervised Representation Approach 5281
Priyanka Kommagouni, Pragya Khanna, Vamshiraghusimha Narasinga, Anirudh Bocha, Anil Kumar Vuppala

Stuttering Detection Based on Self-Attention Weights of Temporal Acoustic Vector Sequence.. 5286
Genzo Miyahara, Tsuneo Kato, Akihiro Tamura

Speech-Based Automatic Chronic Kidney Disease Diagnosis Via Transformer Fusion of Glottal and Spectrogram Features 5291
Jihyun Mun, Minhwa Chung, Sunhee Kim

Influence of Room Acoustics on Objective Voice Assessment Methods in the Context of Speech and Language Therapy 5296
Sven Franz, Tanja Grewe, Bernd T. Meyer, Jörg Bitzer

Multimodal Speech-Based Biomarkers Outperform the ALS Functional Rating Scale in Predicting Individual Disease Progression in ALS 5301
Hardik Kothare, Michael Neumann, Vikram Ramanarayanan

SPEECH DEEPFAKES

Naturalness-Aware Curriculum Learning with Dynamic Temperature for Speech Deepfake Detection 5306
Taewoo Kim, Guisik Kim, Choongsang Cho, Young Han Lee

Leveraging SSL Speech Features and Mamba for Enhanced DeepFake Detection..5311
Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, Pierre-François Marteau

A Comparative Study on Proactive and Passive Detection of Deepfake Speech 5316
Chia-Hua Wu, Wanying Ge, Xin Wang, Junichi Yamagishi, Yu Tsao, Hsin-Min Wang

PhonemeFake: Redefining Deepfake Realism with Language-Driven Segmental Manipulation and Adaptive Bilevel Detection 5321
Oguzhan Baser, Ahmet Ege Tanriverdi, Sriram Vishwanath, Sandeep Chinchali

From Sharpness to Better Generalization for Speech Deepfake Detection 5326
Wen Huang, Xuechen Liu, Xin Wang, Junichi Yamagishi, Yanmin Qian

Unmasking Real-World Audio Deepfakes: A Data-Centric Approach 5331
David Combei, Adriana Stan, Dan Oneata, Nicolas Müller, Horia Cucu

A Data-Driven Diffusion-Based Approach for Audio Deepfake Explanations... 5336
Petr Grinberg, Ankur Kumar, Surya Koppisetti, Gaurav Bharaj

PartialEdit: Identifying Partial Deepfakes in the Era of Neural Speech Editing 5341
You Zhang, Baotong Tian, Lin Zhang, Zhiyao Duan

Rehearsal with Auxiliary-Informed Sampling for Audio Deepfake Detection 5346
Falih Gozi Febrinanto, Kristen Moore, Chandra Thapa, Jiangang Ma, Vidya Saikrishna, Feng Xia

PROSODY

The Prosodic Characteristics of Standard Chinese Rhetorical Questions in Naturalistic Settings 5351
Shuwen Chen, Qingke Sun, Yue Huang, Yingyi Luo

ProBiEM: Acoustic and Lexical Correlates of Prosodic Prominence in English-Malayalam Bilingual Speech.. 5356
Anindita Mondal, Rahul Biju, Anil Kumar Vuppala, Reni K Cherian, Chiranjeevi Yarra

Are You Being Sarcastic? Prosodic Cues to Irony Perception in German... 5361
Sophia Fünfgeld, Angelika Braun, Katharina Zahner-Ritter

Can AI Understand Mandarin Speech Prosody? a Framework and Benchmark Showcase........................... 5366
Zilong Wang, Xiaoxue Zhang, Xinyang Jiang, Kaitao Song, Jue Yu

Generating Consistent Prosodic Patterns from Open-Source TTS Systems 5371
Ha Eun Shim, Olivia Yung, Paige Tuttösí, Boey Kwan, Angelica Lim, Yue Wang, H. Henny Yeung

Multimodal Prosody Modeling: A Use Case for Multilingual Sentence Mode Prediction............................ 5376
Bogdan Vlasenko, Mathew Magimai Doss

SPEECH ANALYSIS AND QUALITY ASSESSMENT

HASRD: Hierarchical Acoustic and Semantic Representation Disentanglement.. 5381
Amir Hussein, Sameer Khurana, Gordon Wichern, François G. Germain, Jonathan Le Roux

Performance of Montreal Forced Aligner on Cantonese Spontaneous Speech... 5386
Ka Ki So, Chenzi Xu, Grace Wenling Cao, Peggy Mok

Segmentation-Variant Codebooks for Preservation of Paralinguistic and Prosodic Information 5391
Nicholas Sanders, Yuanchao Li, Korin Richmond, Simon King

AdaKWS: Towards Robust Keyword Spotting with Test-Time Adaptation... 5396
Yang Xiao, Tianyi Peng, Yanghao Zhou, Rohan Kumar Das

Multivariate Probabilistic Assessment of Speech Quality 5401
Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan K. A. Reddy, Christian Schüldt, Saikat Chatterjee

A Study on Speech Assessment with Visual Cues 5406
Shafique Ahmed, Ryandhimas E. Zezario, Nasir Saleem, Amir Hussain, Hsin-Min Wang, Yu Tsao

Efficient Streaming Speech Quality Prediction with Spiking Neural Networks 5411
Mattias Nilsson, Riccardo Miccini, Julian Rossbroich, Clément Laroche, Tobias Piechowiak, Friedemann Zenke

Unifying Listener Scoring Scales: Comparison Learning Framework for Speech Quality Assessment and Continuous Speech Emotion Recognition... 5416
Cheng Hung Hu, Yusuke Yasuda, Akifumi Yoshimoto, Tomoki Toda

EMOTIONS AND FOUNDATIONAL MODELS

EAA: Emotion-Aware Audio Large Language Models with Dual Cross-Attention and Context-Aware Instruction Tuning 5421
Hongfei Du, Sidi Lu, Gang Zhou, Ye Gao

Chain-Of-Thought Distillation with Fine-Grained Acoustic Cues for Speech Emotion Recognition 5426
Jialong Mai, Xiaofen Xing, Yangbiao Li, Xiangmin Xu

Exploring the Limits of Conformer CTC-Encoder for Speech Emotion Recognition Using Large Language Models 5431
Edmilson Morais, Hagai Aronowitz, Aharon Satt, Ron Hoory, Avihu Dekel, Brian Kingsbury, George Saon

Token-Level Logits Matter: A Closer Look at Speech Foundation Models for Ambiguous Emotion Recognition 5436
Jule Valendo Halim, Siyi Wang, Hong Jia, Ting Dang

Assessing the Feasibility of Large Language Models for Detecting Micro-Behaviors in Team Interactions During Space Missions 5441
Ankush Raut, Projna Paromita, Sydney Begerowski, Suzanne Bell, Theodora Chaspari

A-SMiLE: Affective Sparse Mixture-Of-Experts Adapter with Multi-Task Learning for Spoken Dialogue Models 5446
Yi-Wen Chao, Yizhou Peng, Dianwen Ng, Yukun Ma, Chongjia Ni, Eng Siong Chng, Eng Siong Chng

PREDICTION AND EVALUATION OF SPEECH QUALITY AND INTELLIGIBILITY

Non-Intrusive Binaural Speech Intelligibility Prediction Using Mamba for Hearing-Impaired Listeners 5451
Katsuhiko Yamamoto, Koichi Miyazaki

No Audiogram: Leveraging Existing Scores for Personalized Speech Intelligibility Prediction.................... 5456
Haoshuai Zhou, Changgeng Mo, Boxuan Cao, Linkai Li, Shan Xiang Wang

Feature Importance Across Domains for Improving Non-Intrusive Speech Intelligibility Prediction in Hearing Aids.. 5461
Ryandhimas E. Zezario, Sabato M. Siniscalchi, Fei Chen, Hsin-Min Wang, Yu Tsao

Intelligibility Prediction for Time-Modified Speech Signals Using Spectro-Temporal Modulation Features 5466
Aymen Bashir, Haolan Wang, Amin Edraki, Wai-Yip Chan, Jesper Jensen

French Listening Tests for the Assessment of Intelligibility, Quality, and Identity of Body-Conducted Speech Enhancement.......... 5471
Thomas Joubaud, Julien Hauret, Véronique Zimpfer, Éric Bavu

Benchmarking Neural Speech Codec Intelligibility with SITool 5476
Anna Leschanowsky, Kishor Kayyar Lakshminarayana, Anjana Rajasekhar, Lyonel Behringer, Ibrahim Kilinc, Guillaume Fuchs, Emanuël A. P. Habets

MULTI-TALKER ASR

AISHELL-5: The First Open-Source In-Car Multi-Channel Multi-Speaker Speech Dataset for Automatic Speech Diarization and Recognition.......... 5481
Yuhang Dai, He Wang, Xingchen Li, Zihan Zhang, Shuiyuan Wang, Lei Xie, Xin Xu, Hongxiao Guo, Shaoji Zhang, Hui Bu, Wei Chen

Speaker Targeting Via Self-Speaker Adaptation for Multi-Talker ASR 5486
Weiqing Wang, Taejin Park, Ivan Medennikov, Jinhan Wang, Kunal Dhawan, He Huang, Nithin Rao Koluguri, Jagadeesh Balam, Boris Ginsburg

Speaker-Distinguishable CTC: Learning Speaker Distinction Using CTC for Multi-Talker Speech Recognition 5491
Asahi Sakuma, Hiroaki Sato, Ryuga Sugano, Tadashi Kumano, Yoshihiko Kawai, Tetsuji Ogawa

Improving Practical Aspects of End-To-End Multi-Talker Speech Recognition for Online and Offline Scenarios 5496
Aswin Shanmugam Subramanian, Amit Das, Naoyuki Kanda, Jinyu Li, Xiaofei Wang, Yifan Gong

SPEECH SYNTHESIS PARADIGMS AND METHODS 3

Efficient Streaming TTS Acoustic Model with Depthwise RVQ Decoding Strategies in a Mamba Framework.......... 5501
Joun Yeop Lee, Sangjun Park, Byoung Jin Choi, Ji-Hyun Lee, Min-Kyung Kim, Hoon-Young Cho

APTTS: Adversarial Post-Training in Latent Flow Matching for Fast and High-Fidelity Text-To-Speech 5506
Hyungchan Yoon, Chanwoo Lee, Hoodong Lee, Stanley Jungkyu Choi

Eigenvoice Synthesis Based on Model Editing for Speaker Generation 5511
Masato Murata, Koichi Miyazaki, Tomoki Koriyama, Tomoki Toda

Score-Based Training for Energy-Based TTS Models 5516
Wanli Sun, Anton Ragni

Accelerating Autoregressive Speech Synthesis Inference with Speech Speculative Decoding.......... 5521
Zijian Lin, Yang Zhang, Yougen Yuan, Yuming Yan, Jinjiang Liu, Zhiyong Wu, Pengfei Hu, Qun Yu

BitTTS: Highly Compact Text-To-Speech Using 1.58-Bit Quantization and Weight Indexing 5526
Masaya Kawamura, Takuya Hasumi, Yuma Shirahata, Ryuichi Yamamoto

BIOSIGNAL-ENABLED SPOKEN COMMUNICATION

GTAnet: Geometry-Guided Temporal Attention for EEG-Based Sound Source Tracking in Cocktail Party Scenarios 5531
Saurav Pahuja, Gabriel Ivucic, Siqi Cai, Dashanka Da Silva, Haizhou Li, Tanja Schultz

Decoding Listener's Identity: Person Identification from EEG Signals Using a Lightweight Spiking Transformer 5536
Zheyuan Lin, Siqi Cai, Haizhou Li

Recreating Neural Activity During Speech Production with Language and Speech Model Embeddings 5541
Owais Mujtaba Khanday, Pablo Rodríguez San Esteban, Zubair Ahmad Lone, Marc Ouellet, Jose A. Gonzalez-Lopez

Towards Sentence Level Imagined Speech Generation from EEG Signals 5546
Sparsh Rastogi, Harsh Dadwal, Khushboo Modi, Jatin Bedi, Jasmeet Singh

Word-Level Error Analysis in Decoding Systems: From Speech Recognition to Brain-Computer Interfaces 5551
Jingya Huang, Aashish N. Patel, Sowmya Manojna Narasimha, Gal Mishne, Vikash Gilja

NeuroSpex+: Dual-Task Training of Neuro-Guided Speaker Extraction with Speech Envelope and Waveform 5556
Dashanka Da Silva, Siqi Cai, Saurav Pahuja, Tanja Schultz, Haizhou Li

DiffMV-ETS: Diffusion-Based Multi-Voice Electromyography-To-Speech Conversion Using Speaker-Independent Speech Training Targets 5561
Kevin Scheck, Tom Dombeck, Zhao Ren, Peter Wu, Michael Wand, Tanja Schultz

Conformer-Based Ultrasound-To-Speech Conversion 5566
Ibrahim Ibrahimov, Csaba Zainkó, Gábor Gosztolya

Training Articulatory Inversion Models for Interspeaker Consistency 5571
Charles McGhee, Mark J. F. Gales, Kate M. Knill

Enhancing Acoustic-To-Articulatory Inversion with Multi-Target Pretraining for Low-Resource Settings 5576
Jesuraj Bandekar, Prasanta Kumar Ghosh

Articulatory Vowel Distinctiveness in Spanish 5581
Kristin Teplansky, Emily Rangel, Mimi Lavalley, Jinuk Kwon, Beiming Cao, Jun Wang

EEG-Based Speech Decoding Based on Multi-Mode Joint Modeling 5586
Peiran Li, Fei Chen, Xixin Wu

A Silent Speech Decoding System from EEG and EMG with Heterogenous Electrode Configurations 5591
Masakazu Inoue, Motoshige Sato, Kenichi Tomeoka, Nathania Nah, Eri Hatakeyama, Kai Arulkumaran, Ilya Horiguchi, Shuntaro Sasai

NAM-To-Speech Conversion with Multitask-Enhanced Autoregressive Models 5596
Neil Shah, Shirish Karande, Vineet Gandhi

RESOUND: Speech Reconstruction from Silent Videos Via Acoustic-Semantic Decomposed Modeling 5601
Long-Khanh Pham, Thanh V. T. Tran, Minh-Tan Pham, Van Nguyen

SPEECH DEEPFAKES, ANTISPOOFING AND BACKDOOR ATTACKS

Layer-Wise Decision Fusion for Fake Audio Detection Using XLS-R 5606
Yixuan Xiao, Ngoc Thang Vu

SynHate: Detecting Hate Speech in Synthetic Deepfake Audio 5611
Rishabh Ranjan, Kishan Pipariya, Mayank Vatsa, Richa Singh

Can Emotion Fool Anti-Spoofing? 5616
Aurosweta Mahapatra, Ismail R. Ulgen, Abinay Reddy Naini, Carlos Busso, Berrak Sisman

Pushing the Performance of Synthetic Speech Detection with Kolmogorov-Arnold Networks and Self-Supervised Learning Models 5621
Tuan Dat Phuong, Long-Vu Hoang, Huy Dat Tran

Amplifying Artifacts with Speech Enhancement in Voice Anti-Spoofing 5626
Thanapat Trachu, Thanathai Lertpetchpun, Ekapol Chuangsuwanich

Thai Speech Spoofing Detection Dataset with Variations in Speaking Styles 5631
Ticho Urai, Pachara Boonsarngsuk, Ekapol Chuangsuwanich

CBA: Backdoor Attack on Deep Speech Classification Via Audio Compression 5636
Yuheng Huang, Ying Ren, Wenjie Zhang, Diqun Yan

LRBA: Stealthy Backdoor Attacks on Speech Classification Via Latent Rearrangement in VITS 5641
Zexin Li, Wenhan Yao, Ye Xiao, Jinsu Yang, Fen Xiao, Weiping Wen

LitMAS: A Lightweight and Generalized Multi-Modal Anti-Spoofing Framework for Biometric Security 5646
Nidheesh Gorthi, Kartik Thakral, Rishabh Ranjan, Richa Singh, Mayank Vatsa

PATHOLOGICAL SPEECH ANALYSIS 5

Pitfalls and Limits in Automatic Dementia Assessment 5651
Franziska Braun, Christopher Witzl, Andreas Erzigkeit, Hartmut Lehfeld, Thomas Hillemacher, Tobias Bocklet, Korbinian Riedhammer

On the Within-Class Variation Issue in Alzheimer's Disease Detection 5656
Jiawen Kang, Dongrui Han, Lingwei Meng, Jingyan Zhou, Jinchao Li, Xixin Wu, Helen Meng

Alzheimer’s Disease Detection Using Co-Attention Mechanism for Acoustic and ASR-Transcribed Text Features 5661
Yongqi Shao, Tao Fang

Beyond Manual Transcripts: The Potential of Automated Speech Recognition Errors in Improving Alzheimer’s Disease Detection 5666
Yin-Long Liu, Rui Feng, Jia-Xin Chen, Yi-Ming Wang, Jia-Hong Yuan, Zhen-Hua Ling

Voice-Based Dysphagia Detection: Leveraging Self-Supervised Speech Representation 5671
Injune Hwang, Jung-Min Kim, Ju Seok Ryu, Kyogu Lee

ADCeleb: A Longitudinal Speech Dataset from Public Figures for Early Detection of Alzheimer's Disease 5676
Kunxiao Gao, Anna Favaro, Najim Dehak, Laureano Moro Velazquez

Anne Rowling Neurological Speech Corpus: Clinically Annotated Longitudinal Dataset for Developing Speech Biomarkers in Neurodegenerative Disorders 5681
Johnny Tam, Christine Weaver, Oliver Watts, Siddharthan Chandran, Suvankar Pal, Rowling Speech Consortium

Multitask Learning with Fused Attention for Improved ASR and Mispronunciation Detection in Children's Speech Sound Disorders 5686
Selina S. Sung, Seunghee Ha, Tae-Jin Yoon, Jungmin So

Multimodal Speech, Language and Orofacial Analysis for Remote Assessment of Positive, Negative and Cognitive Symptoms in Schizophrenia 5691
Michael Neumann, Hardik Kothare, Beverly Insel, Anzalee Khan, Danyah Nadim, Jean-Pierre Lindenmayer, Vikram Ramanarayanan

ASR ASSESSMENT AND FOUNDATIONAL MODELS

Aligning ASR Evaluation with Human and LLM Judgments: Intelligibility Metrics Using Phonetic, Semantic, and NLI Approaches 5696
Bornali Phukon, Xiuwen Zheng, Mark Hasegawa-Johnson

SOVA-Bench: Benchmarking the Speech Conversation Ability for LLM-Based Voice Assistant 5701
Yixuan Hou, Heyang Liu, Yuhao Wang, Ziyang Cheng, Ronghua Wu, Qunshan Gu, Yanfeng Wang, Yu Wang

An Approach to Measuring the Performance of Automatic Speech Recognition(ASR) Models in the Context of Large Language Model(LLM) Powered Applications 5706
Sujith Pulikodan, Sahapthan K, Prasanta Kumar Ghosh, Visruth Sanka, Nihar Desai

DC-Spin: A Speaker-Invariant Speech Tokenizer for Spoken Language Models 5711
Heng-Jui Chang, Hongyu Gong, Changhan Wang, James Glass, Yu-An Chung

Exploring the Effect of Segmentation and Vocabulary Size on Speech Tokenization for Speech Language Models 5716
Shunsuke Kando, Yusuke Miyao, Shinnosuke Takamichi

Hearing Deficits of Transformer-Based ASR for Anechoic and Spatial Signals 5721
Dirk Hoffner, Simon Weihe, Thomas Brand, Bernd T. Meyer

SPEAKER RECOGNITION

TELVID: A Multilingual Multi-Modal Corpus for Speaker Recognition 5726
Karen Jones, Kevin Walker, Christopher Caruso, Elliot Singer, Trang Nguyen, Robert Dunn, Stephanie Strassel

EmoSpeechAuth: Emotion-Aware Speaker Verification 5731
Magdalena Golebiowska, Piotr Syga

The 2024 NIST Speaker Recognition Evaluation 5736
Craig Greenberg, Lukas Diduch, Audrey Tong, Elliot Singer, Trang Nguyen, Robert Dunn, Lisa Mason, Beth Matys

A Simple-Yet-Effective Data Augmentation Method for Speaker Identification in Novels 5741
Wenjie Zhong, Jason Naradowsky, Yusuke Miyao

IDIR: Identifying and Distilling Informative Relations for Speaker Verification 5746
Chong-Xin Gan, Zhe Li, Zezhong Jin, Zilong Huang, Man-Wai Mak, Kong Aik Lee

Analysis of ABC Frontend Audio Systems for the NIST-SRE24 5751
Sara Barahona, Anna Silnova, Ladislav Mošner, Junyi Peng, Oldrich Plchot, Johan Rohdin, Lin Zhang, Jiangyu Han, Petr Palka, Federico Landini, Lukáš Burget, Themos Stafylakis, Sandro Cumani, Dominik Boboš, Miroslav Hlavacek, Martin Kodovsky, Tomaš Pavlicek

SPEECH ANALYSIS, DETECTION AND CLASSIFICATION 2

Identifying Primary Stress Across Related Languages and Dialects with Transformer-Based Speech Encoder Models 5756
Nikola Ljubešic, Ivan Porupski, Peter Rupnik

SupraDoRAL: Automatic Word Prominence Detection Using Suprasegmental Dependencies of Representations with Acoustic and Linguistic Context 5761
Jhansi Mallela, Upendra Vishwanath Y. S., Sankara Bharadwaj Rangavajjala, Bhaskar Bhatt, Chiranjeevi Yarra

LombardTokenizer: Disentanglement and Control of Vocal Effort in a Neural Speech Codec 5766
Maxime Jacquelin, Maëva Garnier, Laurent Girin, Rémy Vincent, Olivier Perrotin

Robust Personal Voice Activity Detection for Mitigating Domain Mismatch and False Acceptance Scenarios 5771
Yuke Lin, Jun Chen, Wenjie Li, Longshuai Xiao, Chao Weng

Adaptive Knowledge Distillation for Device-Directed Speech Detection 5776
Hyung-Gun Chi, Florian Pesce, Wonil Chang, Oggi Rudovic, Arturo Argueta, Stefan Braun, Vineet Garg, Ahmed Hussen Abdelaziz

Flexible VAD-PVAD Transition: A Detachable PVAD Module for Dynamic Encoder RNN VAD 5781
En-Lun Yu, Chien-Chun Wang, Jeih-Weih Hung, Shih-Chieh Huang, Berlin Chen

Speaker Conditioning of Voice Activity Detection Via Implicit Separation 5786
Matthew Maciejewski

ASDA: Audio Spectrogram Differential Attention Mechanism for Self-Supervised Representation Learning 5791
Junyu Wang, Tianrui Wang, Meng Ge, Longbiao Wang, Jianwu Dang

DuRep: Dual-Mode Speech Representation Learning Via ASR-Aware Distillation 5796
Prabash Reddy Male, Swayambhu Nath Ray, Harish Arsikere, Akshat Jaiswal, Prakhar Swarup, Prantik Sen, Debmalya Chakrabarty, K V Vijay Girish, Nikhil Bhave, Frederick Weber, Sambuddha Bhattacharya, Sri Garimella

Author Index