

2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2025)

**Honolulu, Hawaii, USA
6-10 December 2025**

Pages 1–716

IEEE Catalog Number: CFP25SRW-POD
ISBN: 979-8-3315-4427-0



Copyright © 2025, IEEE

All Rights Reserved

Copyright and Reprint Permissions:

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

*** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.

IEEE Catalog Number:	CFP25SRW-POD
ISBN (Print-On-Demand):	979-8-3315-4427-0
ISBN (Online):	979-8-3315-4426-3
ISSN:	2997-6928

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
E-mail:	curran@proceedings.com
Web:	www.proceedings.com

TABLE OF CONTENTS

Sinba: Singing-To-Accompaniment Generation With Pitch Guidance Via Mamba-Based Language Model	1
<i>Jianwei Cui, Shihao Chen, Yu Gu, Jie Zhang, Liping Chen, Na Li, Chengxing Li, Shan Yang, Lirong Dai</i>	
SUTA-LM: Bridging Test-Time Adaptation and Language Model Rescoring for Robust ASR	9
<i>Wei-Ping Huang, Guan-Ting Lin, Hung-Yi Lee</i>	
Continual Pre-training for Codec-Based Speech LLMs: Balancing Understanding and Generation	16
<i>Jiatong Shi, Chunlei Zhang, Jinchuan Tian, Junrui Ni, Hao Zhang, Shinji Watanabe, Dong Yu</i>	
Conan: A Chunkwise Online Network for Zero-Shot Adaptive Voice Conversion	24
<i>Yu Zhang, Baotong Tian, Zhiyao Duan</i>	
Emphasis Sensitivity in Speech Representations	32
<i>Shaun Rafael Cassini, Thomas Hain, Anton Ragni</i>	
Layer-wise Analysis for Quality of Multilingual Synthesized Speech	40
<i>Erica Cooper, Takuma Okamoto, Yamato Ohtani, Tomoki Toda, Hisashi Kawai</i>	
Is Smaller Always Faster? Tradeoffs in Compressing Self-Supervised Speech Transformers	47
<i>Tzu-Quan Lin, Tsung-Huan Yang, Chun-Yao Chang, Kuang-Ming Chen, Tzu-Hsun Feng, Hung-Yi Lee Hao Tang</i>	
MBENet: Bone-conduction and Air-conduction Fusion Network for Target Speaker Extraction	54
<i>Chen Zhang, Linfeng Feng, Zhi Liu, Xiao-Lei Zhang, Xuelong Li</i>	
Revealing the Role of Audio Channels in ASR Performance Degradation	59
<i>Kuan-Tang Huang, Li-Wei Chen, Hung-Shin Lee, Berlin Chen, Hsin-Min Wang</i>	
EMO-Debias: Benchmarking Gender Debiasing Techniques in Multi-Label Speech Emotion Recognition	66
<i>Yi-Cheng Lin, Huang-Cheng Chou, Yu-Hsuan Li Liang, Hung-Yi Lee</i>	
Low-Resource Domain Adaptation for Speech LLMs via Text-Only Fine-Tuning	74
<i>Yangui Fang, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, Guohui Zhong, Kai Yu</i>	
Identifying and Calibrating Overconfidence in Noisy Speech Recognition	81
<i>Mingyue Huo, Yuheng Zhang, Yan Tang</i>	
Full-Duplex-Bench: A Benchmark to Evaluate Full-Duplex Spoken Dialogue Models on Turn-taking Capabilities	88
<i>Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu Hung-Yi Lee</i>	
AudioLens: A Closer Look at Auditory Attribute Perception of Large Audio-Language Models	96
<i>Chih-Kai Yang, Neo Ho, Yi-Jyun Lee, Hung-Yi Lee</i>	
WhisperNER: Unified Open Named Entity and Speech Recognition	104
<i>Gil Ayache, Menachem Pirchi, Aviv Navon, Aviv Shamsian, Gill Hetz, Joseph Keshet</i>	
Confidence-Based Self-Training for EMG-to-Speech: Leveraging Synthetic EMG for Robust Modeling	110
<i>Xiaodan Chen, Xiaoxue Gao, Mathias Quoy, Alexandre Pitti, Nancy F. Chen</i>	

Qieemo: Multimodal Emotion Recognition Based on the ASR Backbone	118
<i>Jinming Chen, Jingyi Fang, Yuanzhong Zheng, Yaoxuan Wang, Haojun Fei</i>	
Improving Noise Robust Audio-Visual Speech Recognition via Router-Gated Cross-Modal Feature Fusion	123
<i>DongHoon Lim, YoungChae Kim, Dong-Hyun Kim, Da-Hee Yang, Joon-Hyuk Chang</i>	
Bottleneck Transformer-Based Approach for Improved Automatic STOI Score Prediction	130
<i>Amartyaveer, Murali Kadambi, Chandra Mohan Sharma, Anupam Mandal, Prasanta Kumar Ghosh</i>	
TurboBias: Universal ASR Context-Biasing powered by GPU-accelerated Phrase-Boosting Tree	137
<i>Andrei Andrusenko, Vladimir Bataev, Lilit Grigoryan, Vitaly Lavrukhin, Boris Ginsburg</i>	
Benchmarking Rotary Position Embeddings for Automatic Speech Recognition	144
<i>Shucong Zhang, Titouan Parcollet, Rogier Van Dalen, Sourav Bhattacharya</i>	
EmoTale: An Enacted Speech-emotion Dataset in Danish	151
<i>Maja J. Hjulær, Harald V. Skat-Rørdam, Line H. Clemmensen, Sneha Das</i>	
DiffRhythm+: Controllable and Flexible Full-Length Song Generation with Preference Optimization	158
<i>Huakang Chen, Yuepeng Jiang, Guobin Ma, Chunbo Hao, Shuai Wang, Jixun Yao, Ziqian Ning, Meng Meng Jian Luan, Lei Xie</i>	
XEmoRAG: Cross-Lingual Emotion Transfer with Controllable Intensity Using Retrieval-Augmented Generation	166
<i>Tianlun Zuo, Jingbin Hu, Yuke Li, Xinfu Zhu, Hai Li, Ying Yan, Junhui Liu, Danming Xie, Lei Xie</i>	
Time-Frequency-Based Attention Cache Memory Model for Real-Time Speech Separation	173
<i>Guo Chen, Kai Li, Runxuan Yang, Xiaolin Hu</i>	
Llasa+: Free Lunch for Accelerated and Streaming Llama-Based Speech Synthesis	179
<i>Wenjie Tian, Xinfu Zhu, Hanke Xie, Zhen Ye, Wei Xue, Lei Xie</i>	
An Effective Strategy for Modeling Score Ordinality and Non-uniform Intervals in Automated Speaking Assessment	186
<i>Tien-Hong Lo, Szu-Yu Chen, Yao-Ting Sung, Berlin Chen</i>	
A Study of the Scale Invariant Signal to Distortion Ratio in Speech Separation with Noisy References*	193
<i>Simon Dahl Jepsen, Mads Græsbøll Christensen, Jesper Rindom Jensen</i>	
Improving Multimodal Speech-To-Slide Alignment for Academic Lectures with Vision LLMs	201
<i>Thomas Ranzenberger, Dominik Wagner, Steffen Freisinger, Tobias Bocklet, Korbinian Riedhammer</i>	
Selection of Layers from Self-supervised Learning Models for Predicting Mean-Opinion-Score of Speech	208
<i>Xinyu Liang, Fredrik Cumlin, Victor Ungureanu, Chandan K. A. Reddy, Christian Schüldt Saikat Chatterjee</i>	
Granite-speech: open-source speech-aware LLMs with strong English ASR capabilities	214
<i>George Saon, Avihu Dekel, Alexander Brooks, Tohru Nagano, Abraham Daniels, Aharon Satt, Ashish Mittal Brian Kingsbury, David Haws, Edmilson Moraes, Gakuto Kurata, Hagai Aronowitz, Ibrahim Ibrahim Jeff Kuo, Kate Soule, Luis Lastras, Masayuki Suzuki, Ron Hoory, Samuel Thomas, Sashi Novitasari Takashi Fukuda, Vishal Sunder, Xiaodong Cui, Zvi Kons</i>	

Rapidly Adapting to New Voice Spoofing: Few-Shot Detection of Synthesized Speech Under Distribution Shifts	221
<i>Ashi Garg, Zexin Cai, Henry Li Xinyuan, Leibny Paola García-Perera, Kevin Duh Sanjeev Khudanpur Matthew Wiesner, Nicholas Andrews</i>	
Do Self-Supervised Speech Models Exhibit the Critical Period Effects in Language Acquisition?	228
<i>Yurie Koga, Shunsuke Kando, Yusuke Miyao</i>	
PARCO: Phoneme-Augmented Robust Contextual ASR via Contrastive Entity Disambiguation	236
<i>Jiajun He, Naoki Sawada, Koichi Miyazaki, Tomoki Toda</i>	
Diversity and complementarity of speech encoders across diverse tasks in a multi-modal large language model	243
<i>Jeremy H. M. Wong, Muhammad Huzaifah, Hardik B. Sailor, Shuo Sun, Kye Min Tan, Bin Wang Qiongqiong Wang, Wenyu Zhang, Xunlong Zou, Nancy F. Chen, Ai Ti Aw</i>	
On the Use of Self-Supervised Representation Learning for Speaker Diarization and Separation	251
<i>Séverin Baroudi, Hervé Bredin, Joseph Razik, Ricard Marxer</i>	
PRIME: Novel Prompting Strategies for Effective Biasing Word Recognition in Contextualized ASR	258
<i>Yu-Chun Liu, Li-Ting Pai, Yi-Cheng Wang, Bi-Cheng Yan, Hsin-Wei Wang, Chi-Han Lin, Juan-Wei Xu Berlin Chen</i>	
Deep Audio Zooming: Creating a Sound Barrier With Microphone Array Processing	265
<i>Meng Yu, Dong Yu</i>	
FlexCTC: GPU-powered CTC Beam Decoding With Advanced Contextual Abilities	273
<i>Lilit Grigoryan, Vladimir Bataev, Nikolay Karpov, Andrei Andrusenko, Vitaly Lavrukhin, Boris Ginsburg</i>	
JOOCI: a Novel Method for Learning Comprehensive Speech Representations	280
<i>Hemant Yadav, Sunayana Sitaram, Rajiv Ratn Shah</i>	
SENSE models: an open source solution for multilingual and multimodal semantic-based tasks	288
<i>Salima Mdhaffar, Haroun Elleuch, Chaimae Chellaf, Ha Nguyen, Yannick Estève</i>	
Streaming Endpointer for Spoken Dialogue using Neural Audio Codecs and Label-Delayed Training	296
<i>Sathvik Udupa, Shinji Watanabe, Petr Schwarz, Jan Cernocky</i>	
Codec2Vec: Self-Supervised Speech Representation Learning Using Neural Speech Codecs	304
<i>Wei-Cheng Tseng, David Harwath</i>	
Multi-Target Backdoor Attacks Against Speaker Recognition	311
<i>Alexandrine Fortier, Sonal Joshi, Thomas Thebaud, Jesus Villalba Lopez, Najim Dehak, Patrick Cardinal</i>	
EMO-Reasoning: Benchmarking Emotional Reasoning Capabilities in Spoken Dialogue Systems	318
<i>Jingwen Liu, Kan Jen Cheng, Jiachen Lian, Akshay Anand, Rishi Jain, Faith Qiao, Robin Netzorg Huang-Cheng Chou, Tingle Li, Guan-Ting, Gopala Anumanchipalli</i>	
Expressive Speech Retrieval using Natural Language Descriptions of Speaking Style	326
<i>Wonjune Kang, Deb Roy</i>	
L2 Vowel Acquisition Analysis at the Inventory Level	334
<i>Shuju Shi</i>	

Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio LLM?	341
<i>Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris James Glass</i>	
LCS-CTC: Leveraging Soft Alignments to Enhance Phonetic Transcription Robustness	348
<i>Zongli Ye, Jiachen Lian, Akshaj Gupta, Xuanru Zhou, Haodong Li, Krish Patel, Hwi Joo Park Dingkun Zhou, Chenxu Guo, Shuhe Li, Sam Wang, Iris Zhou, Cheol Jun Cho, Zoe Ezzes, Jet M.J. Vonk Brittany T. Morin, Rian Bogley, Lisa Wauters, Zachary A. Miller, Maria Luisa Gorno-Tempini Gopala Anumanchipalli</i>	
A Neural Model for Contextual Biasing Score Learning and Filtering	356
<i>Wanting Huang, Weiran Wang</i>	
MoLEx: Mixture of LoRA Experts in Speech Self-Supervised Models for Audio Deepfake Detection	363
<i>Zihan Pan, Sailor Hardik Bhupendra, Jinyang Wu</i>	
Towards Generalized Source Tracing for Codec-Based Deepfake Speech	371
<i>I-Ming Lin, Xuanjun Chen, Lin Zhang, Haibin Wu, Hung-yi Lee, Jyh-Shing Roger Jang</i>	
Efficient ASR Domain Adaptation with Long Noun Phrases: Harnessing the Linguistic Characteristics of Japanese	379
<i>Shusuke Komatsu, Kazuyo Onishi, Koki Tanaka, Dohyun Kim, Koichiro Yoshino</i>	
DarkStream: real-time speech anonymization with low latency	386
<i>Waris Quamer, Ricardo Gutierrez-Osuna</i>	
MMW: Side Talk Rejection Multi-Microphone Whisper On Smart Glasses	393
<i>Yang Liu, Li Wan, Yiteng Huang, Yong Xu, Yangyang Shi, Saurabh Adya, Ming Sun, Florian Metze</i>	
Advancing Controllable Music Generation with Latent Rectified Flow Guided by Rhythm and Harmony	401
<i>Haibin Yu, Jiayi Zhou, Wei Wang, Zhiming Wang, Huijia Zhu, Yanmin Qian</i>	
Predictive ASR and Turn-taking Prediction at Once: Towards More Responsive Spoken Dialog System	408
<i>Ryo Fukuda, Takatomo Kano, Naohiro Tawara, Marc Delcroix, Atsunori Ogawa, Yuya Chiba, Atsushi Ando</i>	
Efficient Scaling for LLM-based ASR	415
<i>Bingshen Mu, Yiwen Shao, Kun Wei, Dong Yu, Lei Xie</i>	
Post-training for Deepfake Speech Detection	422
<i>Wanying Ge, Xin Wang, Xuechen Liu, Junichi Yamagishi</i>	
Non-Autoregressive Multi-Speaker ASR with Decoupled Speaker Change Detection	430
<i>Yingke Zhu, Lahiru Samarakoon</i>	
KAN-AST: Kolmogorov-Arnold Network based Audio Spectrogram Transformer for Audio Classification	436
<i>Phuong Tuan Dat, Tran Huy Dat</i>	
Can We Really Repurpose Multi-Speaker ASR Corpus for Speaker Diarization?	442
<i>Shota Horiguchi, Naohiro Tawara, Takanori Ashihara, Atsushi Ando, Marc Delcroix</i>	
From Simulation to Strategy: Automating Personalized Interaction Planning for Conversational Agents	450
<i>Wen-Yu Chang, Tzu-Hung Huang, Chih-Ho Chen, Yun-Nung Chen</i>	

Learning Marmoset Vocal Patterns with a Masked Autoencoder for Robust Call Segmentation, Classification, and Caller Identification	457
<i>Bin Wu, Shinnosuke Takamichi, Sakriani Sakti, Satoshi Nakamura</i>	
Hybrid Decoding: Rapid Pass and Selective Detailed Correction for Sequence Models	464
<i>Yunkyu Lim, Jihwan Park, Hyung Yong Kim, Hanbin Lee, Byeong-Yeol Kim</i>	
ProtoCLAP – Prototypical Contrastive Language-Audio Pretraining	471
<i>Adria Mallol-Ragolta, Björn Schuller</i>	
SincQDR-VAD: A Noise-Robust Voice Activity Detection Framework Leveraging Learnable Filters and Ranking-Aware Optimization	478
<i>Chien-Chun Wang, En-Lun Yu, Jieh-Weih Hung, Shih-Chieh Huang, Berlin Chen</i>	
Evaluation of LLMs in Speech is Often Flawed: Test Set Contamination in Large Language Models for Speech Recognition	486
<i>Yuan Tseng, Titouan Parcollet, Rogier Van Dalen, Shucong Zhang, Sourav Bhattacharya</i>	
Beyond Modality Limitations: A Unified MLLM Approach to Automated Speaking Assessment with Effective Curriculum Learning	494
<i>Yu-Hsuan Fang, Tien-Hong Lo, Yao-Ting Sung, Berlin Chen</i>	
SEF-MK: Speaker-Embedding-Free Voice Anonymization through Multi-k-means Quantization	501
<i>Beilong Tang, Xiaoxiao Miao, Xin Wang, Ming Li</i>	
DeCRED: Decoder-Centric Regularization for Encoder-Decoder Based Speech Recognition	509
<i>Alexander Polok, Santosh Kesiraju, Karel Beneš, Bolaji Yusuf, Lukáš Burget, Jan Černocký</i>	
Whispering Context: Distilling Syntax and Semantics for Long Speech Transcripts	516
<i>Duygu Altinok</i>	
Interpreting the Role of Visemes in Audio-Visual Speech Recognition	523
<i>Aristeidis Papadopoulos, Naomi Harte</i>	
Benchmarking Prosody Encoding in Discrete Speech Tokens	531
<i>Kentaro Onda, Satoru Fukayama, Daisuke Saito, Nobuaki Minematsu</i>	
Masked Self-distilled Transducer-based Keyword Spotting with Semi-autoregressive Decoding	539
<i>Yu Xi, Xiaoyu Gu, Haoyu Li, Jun Song, Bo Zheng, Kai Yu</i>	
Personalized Federated Learning with Fuzzy Clustering for Dysarthric Speech Recognition	546
<i>Jie-Shiang Yang, Jing-Tong Tzeng, Chi-Chun Lee</i>	
MNSC: Advancing Singlish Speech Understanding with Carefully Curated Corpora	553
<i>Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F. Chen AiTi Aw</i>	
A Momentum-Based Framework with Contrastive Data Generation for Robust Sound Source Localization	561
<i>Hyun-Soo Kim, Da-Hee Yang, Joon-Hyuk Chang</i>	
All-in-One ASR: Unifying Encoder-Decoder Models of CTC, Attention, and Transducer in Dual-Mode ASR	568
<i>Takafumi Moriya, Masato Mimura, Tomohiro Tanaka, Hiroshi Sato, Ryo Masumura, Atsunori Ogawa</i>	

Speech Synthesis From Continuous Features Using Per-Token Latent Diffusion	576
<i>Arnon Turetzky, Avihu Dekel, Nimrod Shabtay, Slava Shechtman, David Haws, Hagai Aronowitz Ron Hoory, Yossi Adi</i>	
SMILE: Speech Meta In-Context Learning for Low-Resource Language Automatic Speech Recognition	584
<i>Ming-Hao Hsu, Hung-Yi Lee</i>	
Joint Multimodal Contrastive Learning for Robust Spoken Term Detection and Keyword Spotting	591
<i>Ramesh Gundluru, Shubham Gupta, K Sri Rama Murty</i>	
TokenVerse++: Towards Flexible Multitask Learning with Dynamic Task Activation	598
<i>Shashi Kumar, Srikanth Madikeri, Esaú Villatoro-Tello, Sergio Burdisso, Pradeep Rangappa András Carofilis, Petr Motlicek, Karthik Pandia, Shankar Venkatesan, Kadri Hacıoğlu, Andreas Stolcke</i>	
Training and Inference Efficiency of Encoder-Decoder Speech Models	605
<i>Piotr Żelasko, Kunal Dhawan, Daniel Galvez, Krishna C. Puvvada, Ankita Pasad, Travis Bartley Nithin Rao Koluguri, Ke Hu, Vitaly Lavrukhin, Jagadeesh Balam, Boris Ginsburg</i>	
Mel-Refine: A Plug-and-Play Approach to Refine Mel-Spectrogram in Audio Generation	612
<i>Hongming Guo, Ruibo Fu, Yizhong Geng, Shuchen Shi, Tao Wang, Chunyu Qiang, Ya Li, Zhengqi Wen Yukun Liu, Xuefei Liu, Chenxing Li</i>	
AdaBit-TasNet: Speech Separation with Inference Adaptable Precision	618
<i>Mohamed Elminshawy, Srikanth Raj Chetupalli, Emanuël A. P. Habets</i>	
ASR for Affective Speech: Investigating Impact of Emotion and Speech Generative Strategy	624
<i>Ya-Tse Wu, Chi-Chun Lee</i>	
RE-LLM: Refining Empathetic Speech-LLM Responses by Integrating Emotion Nuance	631
<i>Jing-Han Chen, Bo-Hao Su, Ya-Tse Wu, Chi-Chun Lee</i>	
Reducing Object Hallucination in Large Audio-Language Models via Audio-Aware Decoding	638
<i>Tzu-Wen Hsu, Ke-Han Lu, Cheng-Han Chiang, Hung-Yi Lee</i>	
Graph Connectionist Temporal Classification for Phoneme Recognition	645
<i>Henry Grafé, Hugo Van Hamme</i>	
Speaker Style-Aware Phoneme Anchoring For Improved Cross-Lingual Speech Emotion Recognition	651
<i>Shreya G. Upadhyay, Carlos Busso, Chi-Chun Lee</i>	
Audio-CoT: Exploring Chain-of-Thought Reasoning in Large Audio Language Model	659
<i>Ziyang Ma, Zhuo Chen, Yuping Wang, Eng-Siong Chng, Xie Chen</i>	
Improving Speech Enhancement with Multi-Metric Supervision from Learned Quality Assessment	665
<i>Wei Wang, Wangyou Zhang, Chenda Li, Jaitong Shi, Shinji Watanabe, Yanmin Qian</i>	
MEAN-RIR: Multi-Modal Environment-Aware Network for Robust Room Impulse Response Estimation	673
<i>Jiajian Chen, Jiakang Chen, Hang Chen, Qing Wang, Yu Gao, Jun Du</i>	
A Self-Refining Framework for Enhancing ASR Using TTS-Synthesized Data	680
<i>Cheng-Kang Chou, Chan-Jan Hsu, Ho-Lam Chung, Liang-Hsuan Tseng, Hsi-Chun Cheng, Yu-Kuan Fu Kuan Po Huang, Hung-Yi Lee</i>	

Meta Audiobox Aesthetics: Unified Automatic Assessment for Speech, Music and Sound	688
<i>Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi</i>	
<i>Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, Wei-Ning Hsu</i>	
Iterative Feedback in the Online Active Learning Paradigm	696
<i>Mark Lindsey, Francis Kubala, Richard M. Stern</i>	
URGENT-PK: Perceptually-Aligned Ranking Model Designed for Speech Enhancement Competition	702
<i>Jiahe Wang, Chenda Li, Wei Wang, Wangyou Zhang, Samuele Cornell, Marvin Sach, Robin Scheibler</i>	
<i>Kohei Saijo, Yihui Fu, Zhaoheng Ni, Anurag Kumar, Tim Fingscheidt, Shinji Watanabe, Yanmin Qian</i>	
CO-VADA: A Confidence-Oriented Voice Augmentation Debiasing Approach for Fair Speech Emotion Recognition	709
<i>Yun-Shao Tsai, Yi-Cheng Lin, Huang-Cheng Chou, Hung-Yi Lee</i>	
Spiralformer: Low Latency Encoder for Streaming Speech Recognition with Circular Layer Skipping and Early Exiting	717
<i>Emiru Tsunoo, Hayato Futami, Yosuke Kashiwagi, Siddhant Arora, Shinji Watanabe</i>	
mSTEB: Massively Multilingual Evaluation of LLMs on Speech and Text Tasks	724
<i>Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O. Alabi, Fabian David Schmidt</i>	
<i>Joyce Nakatumba-Nabende, David Ifeoluwa Adelani</i>	
On the Difficulty of Token-Level Modeling of Dysfluency and Fluency Shaping Artifacts	732
<i>Kashaf Gulzar, Dominik Wagner, Sebastian P. Bayerl, Florian Hönig, Tobias Bocklet</i>	
<i>Korbinian Riedhammer</i>	
CAMÕES: A Comprehensive Automatic Speech Recognition Benchmark for European Portuguese	738
<i>Carlos Carvalho, Francisco Teixeira, Catarina Botelho, Anna Pompili, Rubén Solera-Ureña, Sérgio Paulo</i>	
<i>Mariana Julião, Thomas Rolland, John Mendonça, Diogo Pereira, Isabel Trancoso, Alberto Abad</i>	
GenVC: Self-Supervised Zero-Shot Voice Conversion	746
<i>Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur</i>	
<i>Matthew Wiesner, Nicholas Andrews</i>	
Less is More: Data Curation Matters in Scaling Speech Enhancement	754
<i>Chenda Li, Wangyou Zhang, Wei Wang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Yihui Fu</i>	
<i>Marvin Sach, Zhaoheng Ni, Anurag Kumar, Tim Fingscheidt, Shinji Watanabe, Yanmin Qian</i>	
Improving Streaming ASR via Differentially Private Fusion of Data from Multiple Sources	762
<i>Virat Shejwalkar, Om Thakkar, Steve Chien, Nicole Rafidi, Arun Narayanan</i>	
Multilingual Dataset Integration Strategies for Robust Audio Deepfake Detection: A SAFE Challenge System	768
<i>Hashim Ali, Surya Subramani, Nithin Sai Adupa, Lekha Bollinani, Sali El-Loh, Hafiz Malik</i>	
Enhancing Dialogue Annotation with Speaker Characteristics Leveraging a Frozen LLM	776
<i>Thomas Thebaud, Yen-Ju Lu, Matthew Wiesner, Peter Viechnicki, Najim Dehak</i>	
Whisper Has an Internal Word Aligner	783
<i>Sung-Lin Yeh, Yen Meng, Hao Tang</i>	
Utilizing Kolmogorov-Arnold Network in Self-Supervised Learning for Speaker Diarization	790
<i>Minh Vu, Phuong Tuan Dat, Kah Kuan Teh, Van Tuan Nguyen, Tran Huy Dat</i>	

Joint ASR and Speech Attribute Prediction for Conversational Dysarthric Speech Analysis with Multimodal Language Models	796
<i>Dominik Wagner, Ilja Baumann, Natalie Engert, Elmar Nöth, Korbinian Riedhammer, Tobias Bocklet</i>	
Enhancing In-the-Wild Speech Emotion Conversion with Resynthesis-based Duration Modeling	804
<i>Navin Raj Prabhu, Danilo De Oliveira, Nale Lehmann-Willenbrock, Timo Gerkmann</i>	
Unifying Model and Layer Fusion for Speech Foundation Models	811
<i>Yi-Jen Shih, David Harwath</i>	
Reliability of Lexical Richness Measures for ASR-Based Children’s Speech Assessment	818
<i>Imen Talbi, Christopher Gebauer, Lars Rumberg, Edith Beaulac, Hanna Ehlert, Jörn Ostermann</i>	
SSVD: Structured SVD for Parameter-Efficient Fine-Tuning and Benchmarking under Domain Shift in ASR	825
<i>Pu Wang, Shinji Watanabe, Hugo Van Hamme</i>	
CASPER: A Large Scale Spontaneous Speech Dataset	832
<i>Cihan Xiao, Ruixing Liang, Xiangyu Zhang, Mehmet Emre Tiryaki, Veronica Bae, Lavanya Shankar Rong Yang, Ethan Poon, Emmanuel Dupoux, Sanjeev Khudanpur, Leibny Paola Garcia Perera</i>	
PURE Codec: Progressive Unfolding of Residual Entropy for Speech Codec Learning	839
<i>Jiatong Shi, Haoran Wang, William Chen, Chenda Li, Wangyou Zhang, Jinchuan Tian, Shinji Watanabe</i>	
Analysis of Domain Shift across ASR Architectures via TTS-Enabled Separation of Target Domain and Acoustic Conditions	847
<i>Tina Raissi, Nick Rossenbach, Ralf Schlüter</i>	
LLM-Based Dictation Detection from Doctor-Patient Conversations	854
<i>Siyuan Chen, Mojtaba Kadkhodaie Elyaderani, Jing Su, Susanne Burger, Thomas Schaaf</i>	
USAD: Universal Speech and Audio Representation via Distillation	861
<i>Heng-Jui Chang, Saurabhchand Bhati, James Glass, Alexander H. Liu</i>	
State-of-the-art Embeddings with Video-free Segmentation of the Source VoxCeleb Data	869
<i>Sara Barahona, Ladislav Mošnery, Themos Stafylakis, Oldřich Plchotý, Junyi Pengy, Lukáš Burgety Jan Černocký</i>	
More Similar than Dissimilar: Modeling Annotators for Cross-Corpus Speech Emotion Recognition	876
<i>James Tavernor, Emily Mower Provost</i>	
The JHU-MIT System for NIST SRE24: Post-Evaluation Analysis	883
<i>Jesús Villalba, Jonas Borgstrom, Prabhav Singh, Leibny Paola García, Pedro A. Torres-Carrasquillo Najim Dehak</i>	
Emotional Styles Hide in Deep Speaker Embeddings: Disentangle Deep Speaker Embeddings for Speaker Clustering	890
<i>Chaohao Lin, Xu Zheng, Kaida Wu, Peihao Xiang, Ou Bai</i>	
Improving Resource-Efficient Speech Enhancement via Neural Differentiable DSP Vocoder Refinement	896
<i>Heitor R. Guimarães, Ke Tan, Juan Azcarreta, Jesus M. Alvarez, Prabhav Agrawal, Ashutosh Pandey Buye Xu</i>	

Can self-supervised speech models predict the perceived acceptability of prosodic variation?	903
<i>Sarenne Wallbridge, Adaeze Adigwe, Peter Bell</i>	
Maestro-EVC: Controllable Emotional Voice Conversion Guided by References and Explicit Prosody	911
<i>Jinsung Yoon, Wooyeol Jeong, Jio Gim, Young-Joo Suh</i>	
Few-shot Personalization via In-Context Learning for Speech Emotion Recognition based on Speech-Language Model	918
<i>Mana Ihori, Taiga Yamane, Naotaka Kawata, Naoki Makishima, Tomohiro Tanaka, Satoshi Suzuki Shota Orihashi, Ryo Masumura</i>	
Acoustic to Articulatory Speech Inversion for Children with Velopharyngeal Insufficiency	926
<i>Saba Tabatabaee, Suzanne Boyce, Liran Oren, Mark Tiede, Carol Espy-Wilson</i>	
Phoneme Overlapping-Aware Pre-Training with External Text Resources for Multi-Talker ASR	933
<i>Ryo Masumura, Tomohiro Tanaka, Naoki Makishima, Mana Ihori, Shota Orihashi, Naotaka Kawata Taiga Yamane, Satoshi Suzuki, Takafumi Moriya</i>	
Transcribe, Translate, or Transliterate: An Investigation of Intermediate Representations in Spoken Language Models	941
<i>Tolúlope Ógúnremí, Christopher D. Manning, Dan Jurafsky, Karen Livescu</i>	
Unifying Diarization, Separation, and ASR with Multi-Speaker Encoder	948
<i>Muhammad Shakeel, Yui Sudo, Yifan Peng, Chyi-Jiunn Lin, Shinji Watanabe</i>	
Scalable Controllable Accented TTS	955
<i>Henry Li Xinyuan, Zexin Cai, Ashi Garg, Kevin Duh, Leibny Paola García-Perera, Sanjeev Khudanpur Nicholas Andrews, Matthew Wiesner</i>	
Analysing the Language of Neural Audio Codecs	963
<i>Joonyong Park, Shinnosuke Takamichi, David M. Chan, Shunsuke Kando, Yuki Saito, Hiroshi Saruwatari</i>	
Robust Training of Singing Voice Synthesis Using Prior and Posterior Uncertainty	970
<i>Yiwen Zhao, Jiatong Shi, Yuxun Tang, William Chen, Shinji Watanabe</i>	
SLM-S2ST: A multimodal language model for direct speech-to-speech translation	978
<i>Yuxuan Hu, Haibin Wu, Ruchao Fan, Xiaofei Wang, Heng Lu, Yao Qian, Jinyu Li</i>	
Geolocation-Aware Robust Spoken Language Identification	986
<i>Qingzheng Wang, Hye-Jin Shim, Jiancheng Sun, Shinji Watanabe</i>	
Robot Confirmation Generation and Action Planning Using Long-context Q-Former Integrated with Multimodal LLM	993
<i>Chiori Hori, Yoshiki Masuyama, Siddarth Jain, Radu Corcodel, Devesh Jha, Diego Romeres Jonathan Le Roux</i>	
Towards Efficient Speech-Text Jointly Decoding within One Speech Language Model	1000
<i>Haibin Wu, Yuxuan Hu, Ruchao Fan, Xiaofei Wang, Kenichi Kumatani, Bo Ren, Jianwei Yu, Heng Lu Lijuan Wang, Yao Qian, Jinyu Li</i>	
Customizing Speech Recognition Model with Large Language Model Feedback	1007
<i>Shaoshi Ling, Guoli Ye</i>	

CoLMbo: Speaker Language Model for Descriptive Profiling	1013
<i>Massa Baali, Shuo Han, Syed Abdul Hannan, Purusottam Samal, Karanveer Singh, Soham Deshmukh Rita Singh, Bhiksha Raj</i>	
Recognizing Dementia from Neuropsychological Tests with State Space Models	1020
<i>Liming Wang, Saurabhchand Bhati, Cody Karjadi, Rhoda Au, James Glass</i>	
WST: Weakly Supervised Transducer for Automatic Speech Recognition	1027
<i>Dongji Gao, Chenda Liao, Changliang Liu, Matthew Wiesner, Leibny Paola Garcia, Daniel Povey Sanjeev Khudanpur, Jian Wu</i>	
Group Relative Policy Optimization for Speech Recognition	1034
<i>Prashanth Gurunath Shivakumar, Yile Gu, Ankur Gandhe, Ivan Bulyko</i>	
Fake-Mamba: Real-Time Speech Deepfake Detection Using Bidirectional Mamba as Self-Attention's Alternative	1041
<i>Xi Xuan, Zimo Zhu, Wenxin Zhang, Yi-Cheng Lin, Tomi Kinnunen</i>	
CLAIRA: Leveraging Large Language Models to Judge Audio Captions	1049
<i>Tsung-Han Wu, Joseph E. Gonzalez, Trevor Darrell, David M. Chan</i>	
Evaluating Japanese Dialect Robustness Across Speech and Text-based Large Language Models	1056
<i>Tomoya Mizumoto, Yusuke Fujita, Hao Shi, Lianbo Liu, Atsushi Kojima, Yui Sudo</i>	
ULTRAS - Unified Learning of Transformer Representations for Audio and Speech Signals	1063
<i>P E Ameenudeen, Charumathi Narayanan, Sriram Ganapathy</i>	
Robust Speech Emotion Recognition via Classifier Retraining on Mixup-Augmented Representations	1070
<i>Shi-Wook Lee</i>	
ZO-ASR: Zeroth-Order Fine-Tuning of Speech Foundation Models without Back-Propagation	1077
<i>Yuezhang Peng, Yuxin Liu, Yao Li, Sheng Wang, Fei Wen, Xie Chen</i>	
Flow-SLM: Joint Learning of Linguistic and Acoustic Information for Spoken Language Modeling	1084
<i>Ju-Chieh Chou, Jiawei Zhou, Karen Livescu</i>	
ZipVoice: Fast and High-Quality Zero-Shot Text-to-Speech with Flow Matching	1091
<i>Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin Daniel Povey</i>	
Pitch-Assistant Harmonic Recovery for Efficient Speech Enhancement	1099
<i>Biao Liu, Zengqiang Shang, Haoyuan Xie, Mou Wang, Xin Liu, Pengyuan Zhang</i>	
Bridging the Modality Gap: Softly Discretizing Audio Representation for LLM-based Automatic Speech Recognition	1104
<i>Mu Yang, Szu-Jui Chen, Jiamin Xie, H. L. John Hansen</i>	
Incorporating Contextual Paralinguistic Understanding in Large Speech-Language Models	1111
<i>Qiongqiong Wang, Hardik B. Sailor, Jeremy H. M. Wong, Tianchi Liu, Shuo Sun, Wenyu Zhang Muhammad Huzaifah, Nancy Chen, Ai Ti Aw</i>	
Lightweight Prompt Biasing for Contextualized End-to-End ASR Systems	1119
<i>Bo Ren, Yu Shi, Jinyu Li</i>	

Text-Guided Speech Representations for Language Acquisition Assessment	1126
<i>Ilja Baumann, Dominik Wagner, Philipp Seeberger, Korbinian Riedhammer, Tobias Bocklet</i>	
Towards General Discrete Speech Codec for Complex Acoustic Environments: A Study of Reconstruction and Downstream Task Consistency	1134
<i>Haoran Wang, Guanyu Chen, Bohan Li, Hankun Wang, Yiwei Guo, Zhihan Li, Xie Chen, Kai Yu</i>	
Speech in-context learning of paralinguistic tasks	1140
<i>Jeremy H. M. Wong, Muhammad Huzaifah, Nancy F. Chen, Ai Ti Aw</i>	
EmoBiMamba-TTS: Bidirectional State Space Model for Emotion-Intensity Controllable Text-to-Speech	1147
<i>Insung Ham, Bonwha Ku, Hanseok Ko</i>	
Evaluating Self-Supervised Speech Models Via Text-Based LLMs	1155
<i>Takashi Maekaku, Keita Goto, Jinchuan Tian, Yusuke Shinohara, Shinji Watanabe</i>	
Controllable Singing Voice Synthesis using Phoneme-Level Energy Sequence	1163
<i>Yerin Ryu, Inseop Shin, Chanwoo Kim</i>	
Obtaining objective labels and analysing annotator subjectivity by using a Rasch model for ordinal speech processing	1170
<i>Jeremy H. M. Wong, Nancy F. Chen</i>	
Audio Aesthetics Prediction System QAM16k Based on Pre-trained Audio Encoder	1177
<i>Linping Xu, Ziqian Wu, Dejun Zhang</i>	
QAMRO: Quality-aware Adaptive Margin Ranking Optimization for Human-aligned Assessment of Audio Generation Systems	1181
<i>Chien-Chun Wang, Kuan-Tang Huang, Cheng-Yeh Yang, Hung-Shin Lee, Hsin-Min Wang, Berlin Chen</i>	
Efficient Deployment of Large Speech Recognition Models on GPU	1185
<i>Yuekai Zhang, Shuang Yu, Junjie Lai</i>	
Multi-Sampling-Frequency Naturalness MOS Prediction Using Self-Supervised Learning Model with Sampling-Frequency-Independent Layer	1189
<i>Go Nishikawa, Wataru Nakata, Yuki Saito, Kanami Imamura, Hiroshi Saruwatari, Tomohiko Nakamura</i>	
Multi-Distillation from Speech and Music Representation Models	1193
<i>Jui-Chiang Wei, Yi-Cheng Lin, Fabian Ritter-Gutierrez, Hung-Yi Lee</i>	
Voice Factor Control Using FIR-Based Fast Neural Vocoder for Speech Generation Applications	1201
<i>Yamato Ohtani, Takuma Okamoto, Tomoki Toda, Hisashi Kawai</i>	
Towards Scalable and Robust Multilingual ASR for Indian Languages with MixLoRA-Whisper	1205
<i>Yeseul Park, Bowon Lee</i>	
The T12 System for AudioMOS Challenge 2025: Audio Aesthetics Score Prediction System Using KAN- and VERSA-based Models	1209
<i>Katsuhiko Yamamoto, Koichi Miyazaki, Shogo Seki</i>	
MADASR 2.0: Multi-Lingual Multi-Dialect ASR Challenge in 8 Indian Languages	1213
<i>Saurabh Kumar, Sumit Sharma, Deekshitha G, Abhayjeet Singh, Amartyaveer, Sathvik Udupa Sandhya Badiger, Sanjeev Khudanpur, Sunayana Sitaram, S. Umesh, Bhuvana Ramabhadran Brian Kingsbury, Hema A. Murthy, Srikanth S. Narayanan, Howard Lakounga, Prasanta Kumar Ghosh</i>	

DyMEvalNet: Dynamic Text-Audio-Personalization Fusion for Multimodal Music Quality Assessment	1220
<i>Xiaoxun Wu, Kailai Shen, Yuheng Huang, Naiyuan Li, Diqun Yan</i>	
ASTAR-NTU solution to AudioMOS Challenge 2025 Track1	1224
<i>Fabian Ritter-Gutierrez, Yi-Cheng Lin, Jui-Chiang Wei, Jeremy H.M. Wong, Nancy F. Chen, Hung-Yi Lee</i>	
Improving Perceptual Audio Aesthetic Assessment via Triplet Loss and Self-Supervised Embeddings	1228
<i>Dyah A. M. G. Wisnu, Ryandhimas E. Zezario, Stefano Rini, Hsin-Min Wang, Yu Tsao</i>	
VERSA-v2: A Modular and Scalable Toolkit for Speech and Audio Evaluation with Expanded Metrics, Visualization, and LLM Integration	1232
<i>Jiatong Shi, Bo-Hao Su, Shikhar Bharadwaj, Yiwen Zhao, Shih-Heng Wang, Jionghao Hang, Haoran Wang Wei Wang, Wenhao Feng, Yuxun Tang, Nezih Topaloğlu, Siddhant Arora, Jinchuan Tian, William Chen Hye-jin Shim, Wangyou Zhang, Wen-Chin Huang, Shinji Watanabe</i>	
WhisQ: Cross-Modal Representation Learning for Text-to-Music MOS Prediction	1236
<i>Jakaria Islam Emon, Md. Abu Salek, Kazi Tamanna Alam</i>	
KyotoMOS2: MOS Prediction for Speech Across Multiple Sampling Rates	1240
<i>Wangjin Zhou, Yizhou Zhang, Keisuke Imoto, Tatsuya Kawahara</i>	
The AudioMOS Challenge 2025	1244
<i>Wen-Chin Huang, Hui Wang, Cheng Liu, Yi-Chiao Wu, Andros Tjandra, Wei-Ning Hsu, Erica Cooper Yong Qin, Tomoki Toda</i>	
HighRateMOS: Sampling-Rate Aware Modeling for Speech Quality Assessment	1252
<i>Wenze Ren, Yi-Cheng Lin, Wen-Chin Huang, Ryandhimas E. Zezario, Szu-Wei Fu, Sung-Feng Huang Erica Cooper, Haibin Wu, Hung-Yu Wei, Hsin-Min Wang, Hung-Yi Lee, Yu Tsao</i>	
AURA: Agent for Understanding, Reasoning, and Automated Tool Use in Voice-Driven Tasks	1256
<i>Leander Melroy Maben, Gayathri Ganesh Lakshmy, Srijith Radhakrishnan, Siddhant Arora Shinji Watanabe</i>	
AsyncVoice Agent: Real-Time Explanation for LLM Planning and Reasoning	1260
<i>Yueqian Lin, Zhengmian Hu, Jayakumar Subramanian, Qinsi Wang, Nikos Vlassis, Hai “Helen” Li Yiran Chen</i>	
ChipChat: Low-Latency Cascaded Conversational Agent in MLX	1264
<i>Tatiana Likhomanenko, Luke Carlson, Richard He Bai, Zijin Gu, Han Tran, Zakaria Aldeneh, Yizhe Zhang Ruixiang Zhang, Huangjie Zheng, Navdeep Jaitly</i>	
Open Full-duplex Voice Agent with Speech-to-Speech Language Model	1268
<i>Edresson Casanova, Chen Chen, Kevin Hu, Ankita Pasad, Elena Rastorgueva, Seelan Lakshmi Narasimhan Slyne Deng, Ehsan Hosseini-Asl, Piotr Zelasko, Valentin Mendelev, Subhankar Ghosh, Yifan Peng Zhehuai Chen, Jason Li, Jagadeesh Balam, Vitaly Lavrukhin, Boris Ginsburg</i>	
CAVIARES: Corpus for Audio-Visual Expressive Voice Agent	1272
<i>Jinsheng Chen, Yuki Saito, Dong Yang, Naoko Tanji, Hironori Doi, Byeongseon Park, Yuma Shirahata Kentaro Tachibana, Hiroshi Saruwatari</i>	
Speech Masking System Based on Spatially Separated Multiple TTS Maskers With A Compact Circular Loudspeaker Array	1276
<i>Takuma Okamoto</i>	

MMMOS: Multi-domain Multi-axis Audio Quality Assessment	1280
<i>Yi-Cheng Lin, Jia-Hung Chen, Hung-Yi Lee</i>	
Long-Form Fuzzy Speech-to-Text Alignment for 1000+ Languages	1284
<i>Ruizhe Huang, Xiaohui Zhang, Zhaoheng Ni, Moto Hira, Jeff Hwang, Vineel Pratap, Ju Lin, Ming Sun Florian Metze</i>	
Competitive Audio-Language Models with Data-Efficient Single-Stage Training on Public Data	1287
<i>Gokul Karthik Kumar, Rishabh Saraf, Ludovick Lepauloux, Abdul Muneer, Billel Mokeddem, Hakim Hacid</i>	
Efficient Speech Watermarking for Speech Synthesis via Progressive Knowledge Distillation	1295
<i>Yang Cui, Peter Pan, Lei He, Sheng Zhao</i>	
Acoustic Phonetic Temporal Speech Representation	1302
<i>Yunbin Deng</i>	
Adaptive Audio-Visual Speech Recognition via Matryoshka-Based Multimodal LLMs	1309
<i>Umberto Cappellazzo, Minsu Kim, Stavros Petridis</i>	
PhysMVNet: Physics-Informed End-to-End MVDR Beamformer with Residual Spectral Mapping for Multichannel Speech Enhancement	1317
<i>Xingyu Shen, Wei-Ping Zhu, Benoit Champagne</i>	
A correlation-permutation approach for speech-music encoders model merging	1324
<i>Fabian Ritter-Gutierrez, Yi-Cheng Lin, Jeremy H.M Wong, Hung-Yi Lee, Eng Siong Chng, Nancy F. Chen</i>	
AsyncSwitch: Asynchronous Text-Speech Adaptation for Code-Switched ASR	1331
<i>Tuan Nguyen, Huy-Dat Tran</i>	
LauraTSE: Target Speaker Extraction using Auto-Regressive Decoder-Only Language Models	1338
<i>Beilong Tang, Bang Zeng, Ming Li</i>	
Enhancing Code-switched Text-to-Speech Synthesis Capability in Large Language Models with only Monolingual Corpora	1346
<i>Jing Xu, Daxin Tan, Jiaqi Wang, Xiao Chen</i>	
Serialized Output Prompting for Large Language Model-based Multi-Talker Speech Recognition	1354
<i>Hao Shi, Yusuke Fujita, Tomoya Mizumoto, Lianbo Liu, Atsushi Kojima, Yui Sudo</i>	
SV-Mixer: Replacing the Transformer Encoder with Lightweight MLPs for Self-Supervised Model Compresison in Speaker Verification	1362
<i>Jungwoo Heo, Hyun-Seo Shin, Chan-Yeong Lim, Kyo-Won Koo, Seung-Bin Kim, Jisoo Son, Ha-Jin Yu</i>	
Token-based Attractors and Cross-attention in Spoof Diarization	1370
<i>Kyo-Won Koo, Chan-Yeong Lim, Jee-Weon Jun, Hye-Jin Shim, Ha-Jin Yu</i>	
REF-VC: Robust, Expressive and Fast Zero-Shot Voice Conversion with Diffusion Transformers	1377
<i>Yuepeng Jiang, Ziqian Ning, Shuai Wang, Chengjia Wang, Mengxiao Bi, Pengcheng Zhu, Zhonghua Fu Lei Xie</i>	
Lightweight Wasserstein Audio-Visual Model for Unified Speech Enhancement and Separation	1383
<i>Jisoo Park, Seonghak Lee, Guisik Kim, Taewoo Kim, Junseok Kwon</i>	

Enhancing Fully Formatted End-to-End Speech Recognition with Knowledge Distillation via Multi-Codebook Vector Quantization	1391
<i>Jian You, Xiangfeng Li, Erwan Zerhouni</i>	
Fewer Hallucinations, More Verification: A Three-Stage LLM-Based Framework for ASR Error Correction	1397
<i>Yangui Fang, Baixu Chen, Jing Peng, Xu Li, Yu Xi, Chengwei Zhang, Guohui Zhong</i>	
Intermediate-Selective Feature Enhancement for Speech Emotion Recognition	1404
<i>Yangbiao Li, Xiaofen Xing, Jialong Mai, Jingyuan Xing, Xiangmin Xu</i>	
OOQ: Outlier-Oriented Quantization for Efficient Large Language Models	1411
<i>Haoyu Wang, Bei Liu, Hang Shao, Bo Xiao, Ke Zeng, Guanglu Wan, Yanmin Qian</i>	
Omni-Router: Sharing Routing Decisions in Sparse Mixture-of-Experts for Speech Recognition	1418
<i>Zijin Gu, Tatiana Likhomanenko, Navdeep Jaitly</i>	