

2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2023)

**Taipei, Taiwan
16-20 December 2023**

Pages 1-709



**IEEE Catalog Number: CFP23SRW-POD
ISBN: 979-8-3503-0690-3**

**Copyright © 2023 by the Institute of Electrical and Electronics Engineers, Inc.
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:	CFP23SRW-POD
ISBN (Print-On-Demand):	979-8-3503-0690-3
ISBN (Online):	979-8-3503-0689-7

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

CURRAN ASSOCIATES INC.
proceedings
.com

TABLE OF CONTENTS

Rescuespeech: A German Corpus for Speech Recognition in Search and Rescue Domain.....	1
<i>Sangeet Sagar, Mirco Ravanelli, Bernd Kiefer, Ivana Kruijff-Korbayová, Josef Van Genabith</i>	
Mask-Conformer: Augmenting Conformer with Mask-Predict Decoder	8
<i>Yosuke Higuchi, Andrew Rosenberg, Yuan Wang, Murali Karthick Baskar, Bhuvana Ramabhadran</i>	
ED-CEC: Improving Rare Word Recognition Using ASR Postprocessing Based on Error Detection and Context-Aware Error Correction.....	16
<i>Jiajun He, Zekun Yang, Tomoki Toda</i>	
Fast-Hubert: An Efficient Training Framework for Self-Supervised Speech Representation Learning	22
<i>Guanrou Yang, Ziyang Ma, Zhisheng Zheng, Yakun Song, Zhikang Niu, Xie Chen</i>	
Can Unpaired Textual Data Replace Synthetic Speech in ASR Model Adaptation?	29
<i>Pasquale D'Alterio, Christian Hensel, Bashar Awwad Shiekh Hasan</i>	
Leveraging Multilingual Self-Supervised Pretrained Models for Sequence-To-Sequence End-To- End Spoken Language Understanding.....	37
<i>Pavel Denisov, Ngoc Thang Vu</i>	
Preserving Phonemic Distinctions for Ordinal Regression: A Novel Loss Function for Automatic Pronunciation Assessment.....	45
<i>Bi-Cheng Yan, Hsin-Wei Wang, Yi-Cheng Wang, Jiun-Ting Li, Chi-Han Lin, Berlin Chen</i>	
Exploring Effective Distillation of Self-Supervised Speech Models for Automatic Speech Recognition	52
<i>Yujin Wang, Changli Tang, Ziyang Ma, Zhisheng Zheng, Xie Chen, Wei-Qiang Zhang</i>	
Efficient Cascaded Streaming ASR System Via Frame Rate Reduction	58
<i>Xingyu Cai, David Qiu, Shaojin Ding, Dongseong Hwang, Weiran Wang Antoine Bruguier, Rohit Prabhavalkar, Tara Sainath, Yanzhang He</i>	
Vsanet: Real-Time Speech Enhancement Based on Voice Activity Detection and Causal Spatial Attention.....	66
<i>Yuewei Zhang, Huanbin Zou, Jie Zhu</i>	
LC4SV: A Denoising Framework Learning to Compensate for Unseen Speaker Verification Models	74
<i>Chi-Chang Lee, Hong-Wei Chen, Chu-Song Chen, Hsin-Min Wang, Tsung-Te Liu, Yu Tsao</i>	
Maximizing Data Efficiency for Cross-Lingual TTS Adaptation by Self-Supervised Representation Mixing and Embedding Initialization.....	82
<i>Wei-Ping Huang, Sung-Feng Huang, Hung-Yi Lee</i>	
Meta-Learning Framework for End-To-End Imposter Identification in Unseen Speaker Recognition.....	90
<i>Ashutosh Chaubey, Sparsh Sinha, Susmita Ghose</i>	
Using Joint Training Speaker Encoder with Consistency Loss to Achieve Cross-Lingual Voice Conversion and Expressive Voice Conversion	98
<i>Houjian Guo, Chaoran Liu, Carlos Toshinori Ishi, Hiroshi Ishiguro</i>	

QUICKVC: A Lightweight VITS-Based Any-To-Many Voice Conversion Model Using ISTFT for Faster Conversion.....	106
<i>Houjian Guo, Chaoran Liu, Carlos Toshinori Ishi, Hiroshi Ishiguro</i>	
Multi Transcription-Style Speech Transcription Using Attention-Based Encoder-Decoder Model	113
<i>Yan Huang, Piyush Behre, Guoli Ye, Shawn Chang, Yifan Gong</i>	
Neuralkalman: A Learnable Kalman Filter for Acoustic Echo Cancellation	119
<i>Yixuan Zhang, Meng Yu, Hao Zhang, Dong Yu, Deliang Wang</i>	
Thai-Dialect: Low Resource Thai Dialectal Speech to Text Corpora.....	126
<i>Artit Suwanbandit, Jaturong Chitiyaphol, Sutthinan Chuenchom, Kanyarat Kwiecien, Husen Sawal, Ruslan Uthai, Orathai Sangpetch, Ekapol Chuangsuwanich</i>	
Deep Learning for Joint Acoustic Echo and Acoustic Howling Suppression in Hybrid Meetings.....	134
<i>Hao Zhang, Meng Yu, Dong Yu</i>	
On Time Domain Conformer Models for Monaural Speech Separation in Noisy Reverberant Acoustic Environments	141
<i>William Ravenscroft, Stefan Goetze, Thomas Hain</i>	
Neuralecho: Hybrid of Full-Band and Sub-Band Recurrent Neural Network for Acoustic Echo Cancellation and Speech Enhancement.....	148
<i>Meng Yu, Yong Xu, Chunlei Zhang, Shi-Xiong Zhang, Dong Yu</i>	
Combining Relative and Absolute Learning Formulations to Predict Emotional Attributes from Speech	156
<i>Abinay Reddy Naini, Shruthi Subramaniam, Seong-Gyun Leem, Carlos Busso</i>	
Espnet-Summ: Introducing a Novel Large Dataset, Toolkit, and a Cross-Corpora Evaluation of Speech Summarization Systems.....	164
<i>Roshan Sharma, William Chen, Takatomo Kano, Ruchira Sharma, Siddhant Arora, Shinji Watanabe, Atsunori Ogawa, Marc Delcroix, Rita Singh, Bhiksha Raj</i>	
Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data	172
<i>Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-Weon Jung, Soumi Maiti, Shinji Watanabe</i>	
Segment-Level Vectorized Beam Search Based on Partially Autoregressive Inference.....	180
<i>Masao Someki, Nicholas Eng, Yosuke Higuchi, Shinji Watanabe</i>	
Joint Prediction and Denoising for Large-Scale Multilingual Self-Supervised Learning.....	188
<i>William Chen, Jiatong Shi, Brian Yan, Dan Berrebbi, Wangyou Zhang, Yifan Peng, Xuankai Chang, Soumi Maiti, Shinji Watanabe</i>	
Findings of the 2023 ML-Superb Challenge: Pre-Training and Evaluation Over More Languages and Beyond.....	196
<i>Jiatong Shi, William Chen, Dan Berrebbi, Hsiu-Hsuan Wang, Wei-Ping Huang, En-Pei Hu, Ho-Lam Chuang, Xuankai Chang, Yuxun Tang, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, Shinji Watanabe</i>	
Diffusion-Based Mel-Spectrogram Enhancement for Personalized Speech Synthesis with Found Data	204
<i>Yusheng Tian, Wei Liu, Tan Lee</i>	

SQAT-LD: SPeech Quality Assessment Transformer Utilizing Listener Dependent Modeling for Zero-Shot Out-Of-Domain MOS Prediction	211
<i>Kailai Shen, Diqun Yan, Li Dong, Ying Ren, Xiaoxun Wu, Jing Hu</i>	
Scenario-Aware Audio-Visual TF-Gridnet for Target Speech Extraction.....	217
<i>Zexu Pan, Gordon Wichern, Yoshiki Masuyama, François G. Germain, Sameer Khurana, Chiori Hori, Jonathan Le Roux</i>	
Generative Speech Recognition Error Correction with Large Language Models and Task-Activating Prompting	225
<i>Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, Andreas Stolcke</i>	
Enhancing Task-Oriented Dialogues with Chitchat: A Comparative Study Based on Lexical Diversity and Divergence	233
<i>Armand Stricker, Patrick Paroubek</i>	
Token-Level Serialized Output Training for Joint Streaming ASR and ST Leveraging Textual Alignments	241
<i>Sara Papi, Peidong Wang, Junkun Chen, Jian Xue, Jinyu Li, Yashesh Gaur</i>	
LAE-ST-MOE: Boosted Language-Aware Encoder Using Speech Translation Auxiliary Task for E2E Code-Switching ASR.....	249
<i>Guodong Ma, Wenxuan Wang, Yuke Li, Yuting Yang, Binbin Du, Haoran Fu</i>	
A Token-Wise Beam Search Algorithm for RNN-T	257
<i>Gil Keren</i>	
Joint Federated Learning and Personalization for on-Device ASR	265
<i>Junteng Jia, Ke Li, Mani Malek, Kshitiz Malik, Jay Mahadeokar, Ozlem Kalinli, Frank Seide</i>	
MelHuBERT: A Simplified Hubert on Mel Spectrograms	273
<i>Tzu-Quan Lin, Hung-Yi Lee, Hao Tang</i>	
Exploring Data Augmentation in Bias Mitigation Against Non-Native-Accented Speech	281
<i>Yuan Yuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, Odette Scharenborg</i>	
AWMC: Online Test-Time Adaptation Without Mode Collapse for Continual Adaptation	289
<i>Jae-Hong Lee, Do-Hee Kim, Joon-Hyuk Chang</i>	
LE-SSL-MOS: Self-Supervised Learning MOS Prediction with Listener Enhancement.....	297
<i>Zili Qi, Xinhui Hu, Wangjin Zhou, Sheng Li, Hao Wu, Jian Lu, Xinkang Xu</i>	
Transcribing and Aligning Conversational Speech: A Hybrid Pipeline Applied to French Conversations	303
<i>Hiroyoshi Yamasaki, Jérôme Louradour, Julie Hunter, Laurent Prévot</i>	
FedCPC: An Effective Federated Contrastive Learning Method for Privacy Preserving Early-Stage Alzheimers Speech Detection.....	309
<i>Wenqing Wei, Zhengdong Yang, Yuan Gao, Jiyi Li, Chenhui Chu, Shogo Okada, Sheng Li</i>	
Towards General-Purpose Text-Instruction-Guided Voice Conversion.....	315
<i>Chun-Yi Kuan, Chen-An Li, Tsu-Yuan Hsu, Tse-Yang Lin, Ho-Lam Chung, Kai-Wei Chang, Shuo-Yiin Chang, Hung-Yi Lee</i>	
Av-Data2Vec: Self-Supervised Learning of Audio-Visual Speech Representations with Contextualized Target Representations.....	323
<i>Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, Michael Auli</i>	

Improving Stability in Simultaneous Speech Translation: A Revision-Controllable Decoding Approach.....	331
<i>Junkun Chen, Jian Xue, Peidong Wang, Jing Pan, Jinyu Li</i>	
Haha-POD: An Attempt for Laughter-Based Non-Verbal Speaker Verification.....	338
<i>Yuke Lin, Xiaoyi Qin, Ning Jiang, Guoqing Zhao, Ming Li</i>	
Variational Gaussian Process Data Uncertainty.....	345
<i>H. M. Jeremy Wong, Huayun Zhang, Nancy F. Chen</i>	
PP-MET: A Real-World Personalized Prompt Based Meeting Transcription System	353
<i>Xiang Lyu, Yuhang Cao, Qing Wang, Jingjing Yin, Yuguang Yang, Pengpeng Zou, Yanni Hu, Heng Lu</i>	
Brouhaha: Multi-Task Training for Voice Activity Detection, Speech-To-Noise Ratio, and C50 Room Acoustics Estimation	361
<i>Marvin Lavechin, Marianne Métais, Hadrien Titeux, Alodie Boissonnet, Jade Copet, Morgane Rivière, Erika Bergelson, Alejandrina Cristia, Emmanuel Dupoux, Hervé Bredin</i>	
Magnitude-And-Phase-Aware Speech Enhancement with Parallel Sequence Modeling	368
<i>Yuewei Zhang, Huanbin Zou, Jie Zhu</i>	
Speaker Adaptation for End-To-End Speech Recognition Systems in Noisy Environments.....	376
<i>Dominik Wagner, Ilja Baumann, Sebastian P. Bayerl, Korbinian Riedhammer, Tobias Bocklet</i>	
Improving Severity Preservation of Healthy-To-Pathological Voice Conversion with Global Style Tokens	382
<i>Bence Mark Halpern, Wen-Chin Huang, Lester Phillip Violeta, R. J. J. H. Van Son, Tomoki Toda</i>	
End-To-End Multichannel Speaker-Attributed ASR: Speaker Guided Decoder and Input Feature Analysis	389
<i>Can Cui, Imran Sheikh, Mostafa Sadeghi, Emmanuel Vincent</i>	
GPU-Accelerated Wfst Beam Search Decoder for CTC-Based Speech Recognition	397
<i>Daniel Galvez, Tim Kaldewey</i>	
Audio-Adapterfusion: A Task-Id-Free Approach for Efficient and Non-Destructive Multi-Task Speech Recognition	404
<i>Hillary Ngai, Rohan Agrawal, Neeraj Gaur, Ronny Huang, Parisa Haghani, Pedro Moreno Mengibar</i>	
CAMSAT: Augmentation Mix and Self-Augmented Training Clustering for Self-Supervised Speaker Recognition.....	412
<i>Abderrahim Fathan, Jahangir Alam</i>	
Toward Universal Speech Enhancement for Diverse Input Conditions.....	420
<i>Wangyou Zhang, Kohei Saijo, Zhong-Qiu Wang, Shinji Watanabe, Yanmin Qian</i>	
Adversarial Augmentation for Adapter Learning	426
<i>Jen-Tzung Chien, Wei-Yu Sun</i>	
Optimizing Two-Pass Cross-Lingual Transfer Learning: Phoneme Recognition and Phoneme to Grapheme Translation	433
<i>Wonjun Lee, Gary Geunbae Lee, Yunsu Kim</i>	

CTC Blank Triggered Dynamic Layer-Skipping for Efficient Ctc-Based Speech Recognition	441
<i>Junfeng Hou, Peiyao Wang, Jincheng Zhang, Meng Yang, Minwei Feng, Jingcheng Yin</i>	
Prompt Pool Based Class-Incremental Continual Learning for Dialog State Tracking	446
<i>Hong Liu, Yucheng Cai, Yuan Zhou, Zhijian Ou, Yi Huang, Junlan Feng</i>	
Model-Based Fairness Metric for Speaker Verification.....	454
<i>Maliha Jahan, Laureano Moro-Velazquez, Thomas Thebaud, Najim Dehak, Jesús Villalba</i>	
The Voicemos Challenge 2023: Zero-Shot Subjective Speech Quality Prediction for Multiple Domains	461
<i>Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, Junichi Yamagishi</i>	
Low-Rank Adaptation of Large Language Model Rescoring for Parameter-Efficient Speech Recognition	468
<i>Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G. Shivakumar, Yile Gu, Sungho Ryu, Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, Tuan Dinh, Ankur Gandhe, Denis Filimonov, Shalini Ghosh, Andreas Stolcke, Ariya Rastow, Ivan Bulyko</i>	
Cross-Modal Alignment with Optimal Transport for CTC-Based ASR	476
<i>Xugang Lu, Peng Shen, Yu Tsao, Hisashi Kawai</i>	
Study on the Correlation Between Objective Evaluations and Subjective Speech Quality and Intelligibility	483
<i>Hsin-Tien Chiang, Kuo-Hsuan Hung, Szu-Wei Fu, Heng-Cheng Kuo, Ming-Hsueh Tsai, Yu Tsao</i>	
Prompting and Adapter Tuning for Self-Supervised Encoder-Decoder Speech Model.....	490
<i>Kai-Wei Chang, Ming-Hsin Chen, Yun-Ping Lin, Jing Neng Hsu, Paul Kuo-Ming Huang, Chien-Yu Huang, Shang-Wen Li, Hung-Yi Lee</i>	
Voiceextender: Short-Utterance Text-Independent Speaker Verification with Guided Diffusion Model	498
<i>Yayun He, Zuheng Kang, Jianzong Wang, Junqing Peng, Jing Xiao</i>	
Zero-Shot Singing Voice Synthesis from Musical Score.....	506
<i>Jun-You Wang, Hung-Yi Lee, Jyh-Shing Roger Jang, Li Su</i>	
Crosssinger: A Cross-Lingual Multi-Singer High-Fidelity Singing Voice Synthesizer Trained on Monolingual Singers	514
<i>Xintong Wang, Chang Zeng, Jun Chen, Chunhui Wang</i>	
Permod: Perceptually Grounded Voice Modification with Latent Diffusion Models.....	520
<i>Robin Netzorg, Ajil Jalal, Luna McNulty, Gopala Krishna Anumanchipalli</i>	
Boosting Modality Representation with Pre-Trained Models and Multi-Task Training for Multimodal Sentiment Analysis	528
<i>Jiarui Hai, Yu-Jeh Liu, Mounya Elhilali</i>	
Efficient Text-Only Domain Adaptation for CTC-Based ASR.....	536
<i>Chang Chen, Xun Gong, Yanmin Qian</i>	
Adapting Pretrained Speech Model for Mandarin Lyrics Transcription and Alignment	543
<i>Jun-You Wang, Chon-In Leong, Yu-Chen Lin, Li Su, Jyh-Shing Roger Jang</i>	

Partial Rank Similarity Minimization Method for Quality MOS Prediction of Unseen Speech Synthesis Systems in Zero-Shot and Semi-Supervised Setting	551
<i>Hemant Yadav, Erica Cooper, Junichi Yamagishi, Sunayana Sitaram, Rajiv Ratn Shah</i>	
COCO-NUT: Corpus of Japanese Utterance and Voice Characteristics Description for Prompt- Based Control.....	558
<i>Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, Hiroshi Saruwatari</i>	
Generative Linguistic Representation for Spoken Language Identification	566
<i>Peng Shen, Xuguang Lu, Hisashi Kawai</i>	
Spike-Triggered Contextual Biasing for End-To-End Mandarin Speech Recognition	574
<i>Kaixun Huang, Ao Zhang, Binbin Zhang, Tianyi Xu, Xingchen Song, Lei Xie</i>	
Towards Robust Packet Loss Concealment System with ASR-Guided Representations	582
<i>Da-Hee Yang, Joon-Hyuk Chang</i>	
U2-KWS: Unified Two-Pass Open-Vocabulary Keyword Spotting with Keyword Bias	590
<i>Ao Zhang, Pan Zhou, Kaixun Huang, Yong Zou, Ming Liu, Lei Xie</i>	
Consistency Based Unsupervised Self-Training for ASR Personalisation	598
<i>Jisi Zhang, Vandana Rajan, Haaris Mehmood, David Tuckey, Pablo Peso Parada, Md Asif Jalal, Karthikeyan Saravanan, Gil Ho Lee, Jungin Lee, Seokyeong Jung</i>	
Towards a Unified End-To-End Language Understanding System for Speech and Text Inputs	606
<i>Mohan Li, Catalin Zorila, Cong-Thanh Do, Rama Doddipatla</i>	
After: Active Learning Based Fine-Tuning Framework for Speech Emotion Recognition.....	614
<i>Dongyuan Li, Yusong Wang, Kotaro Funakoshi, Manabu Okumura</i>	
On Decoder-Only Architecture for Speech-To-Text and Large Language Model Integration.....	622
<i>Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linqun Liu, Yu Wu</i>	
Paraconsistent Feature Analysis for the Competency Evaluation of Voice Impersonation.....	630
<i>Rajeev Rajan, Noumida A, Sreelakshmi S</i>	
Knowledge Distillation from Offline to Streaming Transducer: Towards Accurate and Fast Streaming Model by Matching Alignments.....	637
<i>Ji-Hwan Mo, Jae-Jin Jeon, Mun-Hak Lee, Joon-Hyuk Chang</i>	
Transformer Attractors for Robust and Efficient End-To-End Neural Diarization	644
<i>Lahiru Samarakoon, Samuel J Broughton, Marc Härkönen, Ivan Fung</i>	
Detection of Vowel Errors in Children’s Speech Using Synthetic Phonetic Transcripts	652
<i>Ilja Baumann, Dominik Wagner, Korbinian Riedhammer, Elmar Nöth, Tobias Bocklet</i>	
Invert-Classify: Recovering Discrete Prosody Inputs for Text-To-Speech.....	660
<i>Nicholas Sanders, Korin Richmond</i>	
KAQ: A Non-Intrusive Stacking Framework for Mean Opinion Score Prediction with Multi-Task Learning	667
<i>Chenglin Xu, Xiguang Zheng, Chen Zhang, Chao Zhou, Qi Huang, Bing Yu</i>	

Sa-Paraformer: Non-Autoregressive End-To-End Speaker-Attributed ASR	675
<i>Yangze Li, Fan Yu, Yuhao Liang, Pengcheng Guo, Mohan Shi, Zhihao Du, Shiliang Zhang, Lei Xie</i>	
Simulation of Teacher-Learner Interaction in English Language Pronunciation Learning	682
<i>Elaf Islam, Thomas Hain, Protima Nomo Sudro</i>	
Identifying People with Mild Cognitive Impairment at Risk of Developing Dementia Using Speech Analysis	688
<i>Bahman Mirheidari, Ronan O'Malley, Daniel Blackburn, Heidi Christensen</i>	
Ending the Blind Flight: Analyzing the Impact of Acoustic and Lexical Factors on WAV2VEC 2.0 in Air-Traffic Control	694
<i>Alexander Blatt, Badr M. Abdullah, Dietrich Klakow</i>	
Cross-Modal Learning for CTC-Based ASR: Leveraging CTC-Bertscore and Sequence-Level Training	702
<i>Mun-Hak Lee, Sang-Eon Lee, Ji-Eun Choi, Joon-Hyuk Chang</i>	
Clustering Unsupervised Representations as Defense Against Poisoning Attacks on Speech Commands Classification System	710
<i>Thomas Thebaud, Sonal Joshi, Henry Li, Martin Sustek, Jesús Villalba, Sanjeev Khudanpur, Najim Dehak</i>	
ECAPA2: A Hybrid Neural Network Architecture and Training Strategy for Robust Speaker Embeddings	718
<i>Jenthe Thienpondt, Kris Demuynck</i>	
Multitask Learning Model with Text and Speech Representation for Fine-Grained Speech Scoring	726
<i>Seongjin Park, Rutuja Ubale</i>	
LibriSpeech-PC: Benchmark for Evaluation of Punctuation and Capitalization Capabilities of End-To-End ASR Models	733
<i>Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg</i>	
Evaluating Self-Supervised Speech Models on a Taiwanese Hokkien Corpus	740
<i>Yi-Hui Chou, Calvin Chang, Meng-Ju Wu, Winston Ou, Alice Wen-Hsin Bi, Carol Yang, Bryan Y. Chen, Rong-Wei Pai, Po-Yen Yeh, Jo-Peng Chiang, Iu-Tshiann Phoann, Winnie Chang, Chenxuan Cui, Noel Chen, Jiatong Shi</i>	
Fast Conformer with Linearly Scalable Attention for Efficient Speech Recognition	747
<i>Dima Rekesh, Nithin Rao Koluguri, Samuel Krizan, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, Boris Ginsburg</i>	
Parameter-Efficient Cross-Language Transfer Learning for a Language-Modular Audiovisual Speech Recognition	755
<i>Zhengyang Li, Thomas Graave, Jing Liu, Timo Lohrenz, Siegfried Kunzmann, Tim Fingscheidt</i>	
Generalized Zero-Shot Audio-To-Intent Classification	763
<i>Veera Raghavendra Elluru, Devang Kulshreshtha, Rohit Paturi, Sravan Bodapati, Srikanth Ronanki</i>	

Investigating the Effect of Language Models in Sequence Discriminative Training for Neural Transducers.....	771
<i>Zijian Yang, Wei Zhou, Ralf Schlüter, Hermann Ney</i>	
TorchAudio 2.1: Advancing Speech Recognition, Self-Supervised Learning, and Audio Processing Components for Pytorch.....	779
<i>Jeff Hwang, Moto Hira, Caroline Chen, Xiaohui Zhang, Zhaoheng Ni, Guangzhi Sun, Pingchuan Ma, Ruizhe Huang, Vineel Pratap, Yuekai Zhang, Anurag Kumar, Chin-Yun Yu, Chuang Zhu, Chunxi Liu, Jacob Kahn, Mirco Ravanelli, Peng Sun, Shinji Watanabe, Yangyang Shi, Yumeng Tao</i>	
Deriving Translational Acoustic Sub-Word Embeddings	788
<i>Amit Meghanani, Thomas Hain</i>	
A Weakly-Supervised Streaming Multilingual Speech Model with Truly Zero-Shot Capability	796
<i>Jian Xue, Peidong Wang, Jinyu Li, Eric Sun</i>	
Transferring Speech-Generic and Depression-Specific Knowledge for Alzheimer’s Disease Detection	803
<i>Ziyun Cui, Wen Wu, Wei-Qiang Zhang, Ji Wu, Chao Zhang</i>	
Robust Logarithmic Champenowne Algorithm for Feedback Cancellation in Hearing Aids.....	811
<i>R Vanitha Devi, Vasundhara</i>	
Hierarchical Attention-Based Contextual Biasing for Personalized Speech Recognition Using Neural Transducers.....	816
<i>Sibo Tong, Philip Harding, Simon Wiesler</i>	
E3 TTS: Easy End-To-End Diffusion-Based Text to Speech	824
<i>Yuan Gao, Nobuyuki Morioka, Yu Zhang, Nanxin Chen</i>	
Building High-Accuracy Multilingual ASR with Gated Language Experts and Curriculum Training	832
<i>Eric Sun, Jinyu Li, Yuxuan Hu, Yimeng Zhu, Long Zhou, Jian Xue, Peidong Wang, Linqun Liu, Shujie Liu, Edward Lin, Yifan Gong</i>	
Bisinger: Bilingual Singing Voice Synthesis	839
<i>Huali Zhou, Yueqian Lin, Yao Shi, Peng Sun, Ming Li</i>	
Flap: Fast Language-Audio Pre-Training	847
<i>Ching-Feng Yeh, Po-Yao Huang, Vasu Sharma, Shang-Wen Li, Gargi Gosh</i>	
On the Relevance of Phoneme Duration Variability of Synthesized Training Data for Automatic Speech Recognition.....	855
<i>Nick Rossenbach, Benedikt Hilmes, Ralf Schlüter</i>	
Enabling Noisy Label Usage for Out-Of-Airspace Data in Read-Back Error Detection.....	863
<i>Lakshmi Rajendram Bashyam, Alexander Blatt, Dietrich Klakow</i>	
Enhancing Expressivity Transfer in Textless Speech-To-Speech Translation	871
<i>Jarod Duret, Benjamin O’Brien, Yannick Estève, Titouan Parcollet</i>	
Dialect Adaptation and Data Augmentation for Low-Resource ASR: Taltech Systems for the Madsr 2023 Challenge	879
<i>Tanel Alumäe, Jiaming Kong, Daniil Robnikov</i>	

Reducing the Cost of Spoof Detection Labeling Using Mixed-Strategy Active Learning and Pretrained Models.....	886
<i>Mark Lindsey, Nathaniel R. Robinson, Francis Kubala, Richard M. Stern</i>	
A Single Speech Enhancement Model Unifying Dereverberation, Denoising, Speaker Counting, Separation, and Extraction.....	893
<i>Kohei Saijo, Wangyou Zhang, Zhong-Qiu Wang, Shinji Watanabe, Tetsunori Kobayashi, Tetsuji Ogawa</i>	
Two-Pass Endpoint Detection for Speech Recognition	899
<i>Anirudh Raju, Aparna Khare, Di He, Ilya Sklyar, Long Chen, Sam Alptekin, Viet Anh Trinh, Zhe Zhang, Colin Vaz, Venkatesh Ravichandran, Roland Maas, Ariya Rastrow</i>	
Improved Long-Form Speech Recognition by Jointly Modeling the Primary and Non-Primary Speakers	907
<i>Guru Prakash Arumugam, Shuo-Yiin Chang, Tara N. Sainath Rohit Prabhavalkar, Quan Wang, Shaan Bijwadia</i>	
Exploring Time-Frequency Domain Target Speaker Extraction for Causal and Non-Causal Processing.....	915
<i>Wangyou Zhang, Lei Yang, Yanmin Qian</i>	
Joint Energy-Based Model for Robust Speech Classification System Against Dirty-Label Backdoor Poisoning Attacks	921
<i>Martin Sustek, Sonal Joshi, Henry Li, Thomas Thebaud, Jesús Villalba, Sanjeev Khudanpur, Najim Dehak</i>	
Importance of Smoothness Induced by Optimizers in F14Asr: Towards Understanding Federated Learning for End-To-End ASR.....	929
<i>Sheikh Shams Azam, Tatiana Likhomanenko, Martin Pelikan, Jan “honza” Silovsky</i>	
Joint Audio and Speech Understanding	937
<i>Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, James Glass</i>	
No Pitch Left Behind: Addressing Gender Unbalance in Automatic Speech Recognition Through Pitch Manipulation	945
<i>Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, Luisa Bentivogli</i>	
Improving Audiovisual Active Speaker Detection in Egocentric Recordings with the Data-Efficient Image Transformer	953
<i>Jason Clarke, Yoshihiko Gotoh, Stefan Goetze</i>	
Yodas: Youtube-Oriented Dataset for Audio and Speech	961
<i>Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, Shinji Watanabe</i>	
Discriminative Speech Recognition Rescoring with Pre-Trained Language Models.....	969
<i>Prashanth Gurunath Shivakumar, Jari Kolehmainen, Yile Gu, Ankur Gandhe, Ariya Rastrow, Ivan Bulyko</i>	
Unconstrained Dysfluency Modeling for Dysfluent Speech Transcription and Detection	976
<i>Jiachen Lian, Carly Feng, Naasir Farooqi, Steve Li, Anshul Kashyap, Cheol Jun Cho, Peter Wu, Robbie Netzorg, Tingle Li, Gopala Krishna Anumanchipalli</i>	
Minisuperb: Lightweight Benchmark for Self-Supervised Speech Models	984
<i>Yu-Hsiang Wang, Huang-Yu Chen, Kai-Wei Chang, Winston Hsu, Hung-Yi Lee</i>	

MASR: Multi-Label Aware Speech Representation	992
<i>Anjali Raj, Shikhar Bharadwaj, Sriram Ganapathy, Min Ma, Shikhar Vashishth</i>	
A Comparative Study of Voice Conversion Models with Large-Scale Speech and Singing Data: The T13 Systems for the Singing Voice Conversion Challenge 2023	1000
<i>Ryuichi Yamamoto, Reo Yoneyama, Lester Phillip Violeta, Wen-Chin Huang, Tomoki Toda</i>	
Extending Self-Distilled Self-Supervised Learning for Semi-Supervised Speaker Verification	1006
<i>Jeong-Hwan Choi, Jehyun Kyung, Ju-Seok Seong, Ye-Rin Jeoung, Joon-Hyuk Chang</i>	
Pseudo-Label Based Supervised Contrastive Loss for Robust Speech Representations	1014
<i>Varun Krishna, Sriram Ganapathy</i>	
Audio-Visual Neural Syntax Acquisition	1022
<i>Cheng-I Jeff Lai, Freda Shi, Puyuan Peng, Yoon Kim, Kevin Gimpel, Shiyu Chang, Yung-Sung Chuang, Saurabhchand Bhati, David Cox, David Harwath, Yang Zhang, Karen Livescu, James Glass</i>	
Improving Speech Enhancement Using Audio Tagging Knowledge from Pre-Trained Representations and Multi-Task Learning.....	1030
<i>Shaoxiong Lin, Chao Zhang, Yanmin Qian</i>	
BA-MoE: Boundary-Aware Mixture-Of-Experts Adapter for Code-Switching Speech Recognition	1037
<i>Peikun Chen, Fan Yu, Yuhao Liang, Hongfei Xue, Xucheng Wan, Naijun Zheng, Huan Zhou, Lei Xie</i>	
Zero-Shot Domain-Sensitive Speech Recognition with Prompt-Conditioning Fine-Tuning.....	1044
<i>Feng-Ting Liao, Yung-Chieh Chan, Yi-Chang Chen, Chan-Jan Hsu, Da-Shan Shiu</i>	
Few-Shot Spoken Language Understanding Via Joint Speech-Text Models.....	1052
<i>Chung-Ming Chien, Mingjiamai Zhang, Ju-Chieh Chou, Karen Livescu</i>	
Summarize While Translating: Universal Model with Parallel Decoding for Summarization and Translation.....	1060
<i>Takatomo Kano, Atsunori Ogawa, Marc Delcroix, Kohei Matsuura, Takanori Ashihara, William Chen, Shinji Watanabe</i>	
Acoustic Model Fusion for End-To-End Speech Recognition.....	1068
<i>Zhihong Lei, Mingbin Xu, Shiyi Han, Leo Liu, Zhen Huang, Tim Ng, Yuanyuan Zhang, Ernest Pusateri, Mirko Hannemann, Yaqiao Deng, Man-Hung Siu</i>	
Domain Adaptation by Data Distribution Matching Via Submodularity for Speech Recognition	1075
<i>Yusuke Shinohara, Shinji Watanabe</i>	
The Second Multi-Channel Multi-Party Meeting Transcription Challenge (M2MeT 2.0): A Benchmark for Speaker-Attributed ASR	1082
<i>Yuhao Liang, Mohan Shi, Fan Yu, Yangze Li, Shiliang Zhang, Zhihao Du, Qian Chen, Lei Xie, Yanmin Qian, Jian Wu, Zhuo Chen, Kong Aik Lee, Zhijie Yan, Hui Bu</i>	
SLM: Bridge the Thin Gap Between Speech and Text Foundation Models	1090
<i>Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K. Rubenstein, Lukas Zilka, Dian Yu, Golan Pundak, Nikhil Siddhartha, Johan Schalkwyk, Yonghui Wu</i>	
An Exploration of Task-Decoupling on Two-Stage Neural Post Filter for Real-Time Personalized Acoustic Echo Cancellation.....	1098
<i>Zihan Zhang, Jiayao Sun, Xianjun Xia, Ziqian Wang, Xiaopeng Yan, Yijian Xiao, Lei Xie</i>	

Improving Whispered Speech Recognition Performance Using Pseudo-Whispered Based Data Augmentation	1105
<i>Zhaofeng Lin, Tanvina Patel, Odette Scharenborg</i>	
Leveraging the Multilingual Indonesian Ethnic Languages Dataset in Self-Supervised Models for Low-Resource ASR Task.....	1113
<i>Sakriani Sakti, Benita Angela Titalim</i>	
Promptspeaker: Speaker Generation Based on Text Descriptions	1121
<i>Yongmao Zhang, Guanghou Liu, Yi Lei, Yunlin Chen, Hao Yin, Lei Xie, Zhifei Li</i>	
HIGNN-TTS: Hierarchical Prosody Modeling with Graph Neural Networks for Expressive Long-Form TTS	1128
<i>Dake Guo, Xinfu Zhu, Liumeng Xue, Tao Li, Yuanjun Lv, Yuepeng Jiang, Lei Xie</i>	
Zero-Shot Emotion Transfer for Cross-Lingual Speech Synthesis	1135
<i>Yuke Li, Xinfu Zhu, Yi Lei, Hai Li, Junhui Liu, Danming Xie, Lei Xie</i>	
VITS-Based Singing Voice Conversion Leveraging Whisper and Multi-Scale F0 Modeling	1143
<i>Ziqian Ning, Yuepeng Jiang, Zhichao Wang, Bin Zhang, Lei Xie</i>	
Salt: Distinguishable Speaker Anonymization Through Latent Space Transformation.....	1151
<i>Yuanjun Lv, Jixun Yao, Peikun Chen, Hongbin Zhou, Heng Lu, Lei Xie</i>	
Robust Recognition of Speaker Emotion with Difference Feature Extraction Using a Few Enrollment Utterances	1159
<i>Daichi Hayakawa, Takehiko Kagoshima, Kenji Iwata, Norbert Braunschweiler, Rama Doddipatla</i>	
MUST: A Multilingual Student-Teacher Learning Approach for Low-Resource Speech Recognition	1166
<i>Muhammad Umar Farooq, Rehan Ahmad, Thomas Hain</i>	
WaveNeXt: ConvNeXt-Based Fast Neural Vocoder Without ISTFT Layer	1172
<i>Takuma Okamoto, Haruki Yamashita, Yamato Ohtani, Tomoki Toda, Hisashi Kawai</i>	
Spectral Tilt May Have a Smaller Impact on the Intelligibility of Speech in Noise.....	1180
<i>Yoshiki Sato, Julián Villegas</i>	
H _{eval} : A New Hybrid Evaluation Metric for Automatic Speech Recognition Tasks.....	1185
<i>Zitha Sasindran, Harsha Yelchuri, T. V. Prabhakar, Supreeth Rao</i>	
Towards Developing State-Of-The-Art TTS Synthesisers for 13 Indian Languages with Signal Processing Aided Alignments.....	1192
<i>Anusha Prakash, S Umesh, Hema A Murthy</i>	
Parameter-Efficient Tuning with Adaptive Bottlenecks for Automatic Speech Recognition	1200
<i>Geoffroy Vanderreydt, Amrutha Prasad, Driss Khalil, Srikanth Madikeri, Kris Demuyneck, Petr Motlicek</i>	
Semi-Supervised Multi-Channel Speaker Diarization with Cross-Channel Attention	1207
<i>Shilong Wu, Jun Du, Mao-Kui He, Shutong Niu, Hang Chen, Haitao Tang, Chin-Hui Lee</i>	
Exploring the Viability of Synthetic Audio Data for Audio-Based Dialogue State Tracking	1215
<i>Jihyun Lee, Yejin Jeon, Wonjun Lee, Yunsu Kim, Gary Geunbae Lee</i>	

Gated Multi Encoders and Multitask Objectives for Dialectal Speech Recognition in Indian Languages.....	1223
<i>Sathvik Udupa, Jesuraja Bandekar, G Deekshitha, Saurabh Kumar, Prasanta Kumar Ghosh, Sandhya Badiger, Abhayjeet Singh, Savitha Murthy, Priyanka Pai, Srinivasa Raghavan, Raoul Nanavati</i>	
VITS-Based Singing Voice Conversion System with DSPGAN Post-Processing for SVCC2023.....	1231
<i>Yiquan Zhou, Meng Chen, Yi Lei, Jihua Zhu, Weifeng Zhao</i>	
Prompting Large Language Models for Zero-Shot Domain Adaptation in Speech Recognition.....	1239
<i>Yuang Li, Yu Wu, Jinyu Li, Shujie Liu</i>	
MBTFNET: Multi-Band Temporal-Frequency Neural Network for Singing Voice Enhancement.....	1247
<i>Weiming Xu, Zhouxuan Chen, Zhili Tan, Shubo Lv, Runduo Han, Wenjiang Zhou, Weifeng Zhao, Lei Xie</i>	
Can We Use Speaker Embeddings on Spontaneous Speech Obtained from Medical Conversations to Predict Intelligibility?	1255
<i>Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, Julien Pinquier</i>	
End-To-End Training of a Neural HMM with Label and Transition Probabilities	1262
<i>Daniel Mann, Tina Raissi, Wilfried Michel, Ralf Schlüter, Hermann Ney</i>	
Wiki-En-ASR-Adapt: Large-Scale Synthetic Dataset for English ASR Customization.....	1270
<i>Alexandra Antonova</i>	
The Role of Feature Correlation on Quantized Neural Networks	1278
<i>David Qiu, Shaojin Ding, Yanzhang He</i>	
LV-CTC: Non-Autoregressive ASR with CTC and Latent Variable Models.....	1285
<i>Yuya Fujita, Shinji Watanabe, Xuankai Chang, Takashi Maekaku</i>	
The Singing Voice Conversion Challenge 2023	1291
<i>Wen-Chin Huang, Lester Phillip Violeta, Songxiang Liu, Jiatong Shi, Tomoki Toda</i>	
Improving Multilingual and Code-Switching ASR Using Large Language Model Generated Text	1299
<i>Ke Hu, Tara N. Sainath, Bo Li, Yu Zhang, Yong Cheng, Tao Wang, Yujing Zhang, Frederick Liu</i>	
Pareto Efficiency of Learning-Forgetting Trade-Off in Neural Language Model Adaptation.....	1306
<i>Jerome R. Bellegarda</i>	
Improved Multi-Modal Emotion Recognition Using Squeeze-And-Excitation Block in Cross-Modal Attention.....	1314
<i>Junchen Liu, Jesin James, Karan Nathwani</i>	
Improving Large-Scale Deep Biasing with Phoneme Features and Text-Only Data in Streaming Transducer.....	1322
<i>Jin Qiu, Lu Huang, Boyu Li, Jun Zhang, Lu Lu, Zejun Ma</i>	
Locality Enhanced Dynamic Biasing and Sampling Strategies for Contextual ASR	1330
<i>Md Asif Jalal, Pablo Peso Parada, George Pavlidis, Vasileios Moschopoulos, Karthikeyan Saravanan, Chrysovalantis-Giorgos Kontoulis, Jisi Zhang, Anastasios Drosou, Gil Ho Lee, Jungin Lee, Seokyeong Jung</i>	
Robust End-To-End Diarization with Domain Adaptive Training and Multi-Task Learning	1338
<i>Ivan Fung, Lahiru Samarakoon, Samuel J. Broughton</i>	

Whisper-Slu: Extending a Pretrained Speech-To-Text Transformer for Low Resource Spoken Language Understanding.....	1345
<i>Quentin Meeus, Marie-Francine Moens, Hugo Van Hamme</i>	
Detecting Speech Abnormalities with a Perceiver-Based Sequence Classifier that Leverages a Universal Speech Model.....	1351
<i>Hagen Soltau, Izhak Shafran, Alex Ottenwess, Joseph R. JR Duffy, Rene L. Utianski, Leland R. Barnard, John L. Stricker, Daniela Wiepert, David T. Jones, Hugo Botha</i>	
Contextual Spelling Correction with Large Language Models	1358
<i>Gan Song, Zelin Wu, Golan Pundak, Angad Chandorkar, Kandarp Joshi, Xavier Velez, Diamantino Caseiro, Ben Haynor, Weiran Wang, Nikhil Siddhartha, Pat Rondon, Khe Chai Sim</i>	
Not All Errors Are Created Equal: Evaluating the Impact of Model and Speaker Factors on ASR Outcomes in Clinical Populations	1366
<i>Daniela A. Wiepert, Rene L. Utianski, Joseph R. Duffy, John L. Stricker, Leland Barnard, Keith A. Josephs, Jennifer L. Whitwell, David T. Jones, Hugo Botha</i>	
The Gift of Feedback: Improving ASR Model Quality by Learning from User Corrections Through Federated Learning.....	1372
<i>Lillian Zhou, Yuxin Ding, Mingqing Chen, Harry Zhang, Rohit Prabhavalkar, Dhruv Guliani, Giovanni Motta, Rajiv Mathews</i>	
Transduce and Speak: Neural Transducer for Text-To-Speech with Semantic Token Prediction.....	1379
<i>Minchan Kim, Myeonghun Jeong, Byoung Jin Choi, Dongjune Lee, Nam Soo Kim</i>	
FAT-HuBERT: Front-End Adaptive Training of Hidden-Unit BERT for Distortion-Invariant Robust Speech Recognition.....	1386
<i>Dongning Yang, Wei Wang, Yanmin Qian</i>	
Speech Emotion Diarization: Which Emotion Appears When?	1394
<i>Yingzhi Wang, Mirco Ravanelli, Alya Yacoubi</i>	
Learning from Flawed Data: Weakly Supervised Automatic Speech Recognition.....	1401
<i>Dongji Gao, Hainan Xu, Desh Raj, Leibny Paola Garcia Perera, Daniel Povey, Sanjeev Khudanpur</i>	
Acoustics-Text Dual-Modal Joint Representation Learning for Cover Song Identification.....	1409
<i>Yanmei Gu, Jingli, Jiayizhou, Zhiming Wang, Huijia Zhu</i>	
Towards Matching Phones and Speech Representations	1417
<i>Gene-Ping Yang, Hao Tang</i>	

Author Index