24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)

Dublin, Ireland
20-24 August 2023

Volume 1 of 8

ISBN: 978-1-7138-8880-2

Printed from e-media with permission by:

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

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

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

KEYNOTE 1 ISCA MEDALLIST

Bridging Speech Science and Technology — Now and into the Future 1
Shrikanth Narayanan

SPEECH SYNTHESIS: PROSODY AND EMOTION

Emotional Talking Head Generation Based on Memory-Sharing and Attention-Augmented Networks 2
Jianrong Wang, Yaxin Zhao, Li Liu, Tianyi Xu, Qi Li, Sen Li

Speech Synthesis with Self-Supervisedly Learnt Prosodic Representations 7
Zhao-Ci Liu, Zhen-Hua Ling, Ya-Jun Hu, Jia Pan, Jin-Wei Wang, Yun-Di Wu

EmoMix: Emotion Mixing Via Diffusion Models for Emotional Speech Synthesis 12
Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, Jing Xiao

Laughter Synthesis Using Pseudo Phonetic Tokens with a Large-Scale In-The-Wild Laughter Corpus 17
Detai Xin, Shinnosuke Takamichi, Ai Morimatsu, Hiroshi Saruwatari

Explicit Intensity Control for Accented Text-To-Speech 22
Rui Liu, Haolin Zuo, De Hu, Guanglai Gao, Haizhou Li

Comparing Normalizing Flows and Diffusion Models for Prosody and Acoustic Modelling in Text-To-Speech 27
Guangyan Zhang, Thomas Merritt, Sam Ribeiro, Biel Tura-Vecino, Kayoko Yanagisawa, Kamil Pokora, Abdelhamid Ezzerg, Sebastian Cygert, Ammar Abbas, Piotr Bilinski, Roberto Barra-Chicote, Daniel Korzekwa, Jaime Lorenzo-Trueba

STATISTICAL MACHINE TRANSLATION

Modular Speech-To-Text Translation for Zero-Shot Cross-Modal Transfer 32
Paul-Ambroise Duquenne, Holger Schwenk, Benoît Sagot

Improving Isochronous Machine Translation with Target Factors and Auxiliary Counters 37
Proyag Pal, Brian Thompson, Yogesh Virkar, Prashant Mathur, Alexandra Chronopoulou, Marcello Federico

StyleS2ST: Zero-Shot Style Transfer for Direct Speech-To-Speech Translation 42
Kun Song, Yi Ren, Yi Lei, Chunfeng Wang, Kun Wei, Lei Xie, Xiang Yin, Zejun Ma

Joint Speech Translation and Named Entity Recognition 47
Marco Gaido, Sara Papi, Matteo Negri, Marco Turchi

Analysis of Acoustic Information in End-To-End Spoken Language Translation 52
Gerard Sant, Carlos Escolano

LAMASSU: A Streaming Language-Agnostic Multilingual Speech Recognition and Translation Model Using Neural Transducers 57
Peidong Wang, Eric Sun, Jian Xue, Yu Wu, Long Zhou, Yashesh Gaur, Shujie Liu, Jinyu Li

SELF-SUPERVISED LEARNING IN ASR

DPHuBERT: Joint Distillation and Pruning of Self-Supervised Speech Models 62
Yifan Peng, Yui Sudo, Shakeel Muhammad, Shinji Watanabe

Automatic Data Augmentation for Domain Adapted Fine-Tuning of Self-Supervised Speech Representations 67
Salah Zaiem, Titouan Parcollet, Slim Essid

Dual Acoustic Linguistic Self-Supervised Representation Learning for Cross-Domain Speech Recognition 72
Zhao Yang, Dianwen Ng, Chong Zhang, Xiao Fu, Rui Jiang, Wei Xi, Yukun Ma, Chongjia Ni, Eng Siong Chng, Bin Ma, Jizhong Zhao

O-1: Self-Training with Oracle and 1-Best Hypothesis 77
Murali Karthick Baskar, Andrew Rosenberg, Bhuvana Ramabhadran, Kartik Audhkhasi

MT4SSL: Boosting Self-Supervised Speech Representation Learning by Integrating Multiple Targets 82
Ziyang Ma, Zhisheng Zheng, Changli Tang, Yujin Wang, Xie Chen

Comparing Self-Supervised Pre-Training and Semi-Supervised Training for Speech Recognition in Languages with Weak Language Models 87
Léa-Marie Lam-Yee-Mui, Lucas Ondel Yang, Ondrej Klejch

PROSODY

Chinese EFL Learners' Perception of English Prosodic Focus 92
Xinya Zhang, Ying Chen

Pitch Accent Variation and the Interpretation of Rising and Falling Intonation in American English 97
Thomas Sostarics, Jennifer Cole

Tonal Coarticulation as a Cue for Upcoming Prosodic Boundary 102
Jianjing Kuang, May Pik Yu Chan, Nari Rhee

Alignment of Beat Gestures and Prosodic Prominence in German 107
Sophie Repp, Lara Muhtz, Johannes Heim

Creak Prevalence and Prosodic Context in Australian English 112
Hannah White, Joshua Penney, Andy Gibson, Anita Szakay, Felicity Cox

Speech Reduction: Position Within French Prosodic Structure 117
Kübra Bodur, Roxane Bertrand, James S. German, Stéphane Rauzy, Corinne Fredouille, Christine Meunier

SPEECH PRODUCTION

Transvelar Nasal Coupling Contributing to Speaker Characteristics in Non-Nasal Vowels 122
Ziyu Zhu, Yujie Chi, Zhao Zhang, Kiyoshi Honda, Jianguo Wei

Speech Synthesis from Articulatory Movements Recorded by Real-Time MRI 127
Yuto Otani, Shun Sawada, Hidefumi Ohmura, Kouichi Katsurada

The ART of Conversation: Measuring Phonetic Convergence and Deliberate Imitation in L2-Speech with a Siamese RNN 132
Zheng Yuan, Aldo Pastore, Dorina De Jong, Hao Xu, Luciano Fadiga, Alessandro D'Ausilio

Did You See That? Exploring the Role of Vision in the Development of Consonant Feature Contrasts in Children with Cochlear Implants 137
James Mahshie, Michael Larsen

DYSARTHRIC SPEECH ASSESSMENT

Automatic Assessments of Dysarthric Speech: The Usability of Acoustic-Phonetic Features 141
Loes Van Bemmel, Chiara Pesenti, Xue Wei, Helmer Strik

Classification of Multi-Class Vowels and Fricatives from Patients Having Amyotrophic Lateral Sclerosis with Varied Levels of Dysarthria Severity 146
Chowdam Venkata Thirumala Kumar, Tanuka Bhattacharjee, Yamini Belur, Atchayaram Nalini, Ravi Yadav, Prasanta Kumar Ghosh

Parameter-Efficient Dysarthric Speech Recognition Using Adapter Fusion and Householder Transformation 151
Jinzi Qi, Hugo Van Hamme

Few-Shot Dysarthric Speech Recognition with Text-To-Speech Data Augmentation 156
Enno Hermann, Mathew Magimai.-Doss

Latent Phrase Matching for Dysarthric Speech 161
Dianna Yee, Colin Lea, Jaya Narain, Zifang Huang, Lauren Tooley, Jeffrey P. Bigham, Leah Findlater

Speech Intelligibility Assessment of Dysarthric Speech by Using Goodness of Pronunciation with Uncertainty Quantification 166
Eun Jung Yeo, Kwanghee Choi, Sunhee Kim, Minhwa Chung

SPEECH CODING: TRANSMISSION AND ENHANCEMENT

CQNV: A Combination of Coarsely Quantized Bitstream and Neural Vocoder for Low Rate Speech Coding 171
Youqiang Zheng, Li Xiao, Weiping Tu, Yuhong Yang, Xinmeng Xu

Target Speech Extraction with Conditional Diffusion Model 176
Naoyuki Kamo, Marc Delcroix, Tomohiro Nakatani

Towards Fully Quantized Neural Networks for Speech Enhancement 181
Elad Cohen, Hai Victor Habi, Arnon Netzer

Complex Image Generation SwinTransformer Network for Audio Denoising 186
Youshan Zhang, Jialu Li

SPEECH RECOGNITION: SIGNAL PROCESSING, ACOUSTIC MODELING, ROBUSTNESS, ADAPTATION 1

Using Text Injection to Improve Recognition of Personal Identifiers in Speech 191
Yochai Blau, Rohan Agrawal, Lior Madmony, Gary Wang, Andrew Rosenberg, Zhehuai Chen, Zorik Gekhman, Genady Beryozkin, Parisa Haghani, Bhuvana Ramabhadran

Investigating Wav2vec2 Context Representations and the Effects of Fine-Tuning, a Case-Study of a Finnish Model 196
Tamas Grosz, Yaroslav Getman, Ragheb Al-Ghezi, Aku Rouhe, Mikko Kurimo

Transformer-Based Speech Recognition Models for Oral History Archives in English, German, and Czech 201
Jan Lehecka, Jan Švec, Josef V. Psutka, Pavel Ircing

Iteratively Improving Speech Recognition and Voice Conversion 206
Mayank Kumar Singh, Naoya Takahashi, Naoyuki Onoe

LABERT: A Combination of Local Aggregation and Self-Supervised Speech Representation Learning for Detecting Informative Hidden Units in Low-Resource ASR Systems 211
Kavan Fatehi, Ayse Kucukyilmaz

TranUSR: Phoneme-To-Word Transcoder Based Unified Speech Representation Learning for Cross-Lingual Speech Recognition 216
Hongfei Xue, Qijie Shao, Peikun Chen, Pengcheng Guo, Lei Xie, Jie Liu

Dual-Mode NAM: Effective Top-K Context Injection for End-To-End ASR 221
Zelin Wu, Tsendsuren Munkhdalai, Pat Rondon, Golan Pundak, Khe Chai Sim, Christopher Li

GhostRNN: Reducing State Redundancy in RNN with Cheap Operations 226
Hang Zhou, Xiaoxu Zheng, Yunhe Wang, Michael Bi Mi, Deyi Xiong, Kai Han

Task-Agnostic Structured Pruning of Speech Representation Models 231
Haoyu Wang, Siyuan Wang, Wei-Qiang Zhang, Suo Hongbin, Yulong Wan

Factual Consistency Oriented Speech Recognition 236
Naoyuki Kanda, Takuya Yoshioka, Yang Liu

Multi-Head State Space Model for Speech Recognition 241
Yassir Fathullah, Chunyang Wu, Yuan Shangguan, Junteng Jia, Wenhan Xiong, Jay Mahadeokar, Chunxi Liu, Yangyang Shi, Ozlem Kalinli, Mike Seltzer, Mark J. F. Gales

Cascaded Multi-Task Adaptive Learning Based on Neural Architecture Search 246
Yingying Gao, Shilei Zhang, Zihao Cui, Chao Deng, Junlan Feng

Probing Self-Supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration 251
Kinan Martin, Jon Gauthier, Canaan Breiss, Roger Levy

Selective Biasing with Trie-Based Contextual Adapters for Personalised Speech Recognition Using Neural Transducers 256
Philip Harding, Sibo Tong, Simon Wiesler

ANALYSIS OF SPEECH AND AUDIO SIGNALS 1

Robust Prototype Learning for Anomalous Sound Detection 261
Xiao-Min Zeng, Yan Song, Ian McLoughlin, Lin Liu, Li-Rong Dai

A Multimodal Prototypical Approach for Unsupervised Sound Classification 266
Saksham Singh Kushwaha, Magdalena Fuentes

Robust Audio Anti-Spoofing with Fusion-Reconstruction Learning on Multi-Order Spectrograms 271
Penghui Wen, Kun Hu, Wenxi Yue, Sen Zhang, Wanlei Zhou, Zhiyong Wang

Adapting Language-Audio Models as Few-Shot Audio Learners 276
Jinhua Liang, Xubo Liu, Haohe Liu, Huy Phan, Emmanouil Benetos, Mark D. Plumbley, Wenwu Wang

TFECN: Time-Frequency Enhanced ConvNet for Audio Classification 281
Mengwei Wang, Zhe Yang

Resolution Consistency Training on Time-Frequency Domain for Semi-Supervised Sound Event Detection 286
Won-Gook Choi, Joon-Hyuk Chang

Fine-Tuning Audio Spectrogram Transformer with Task-Aware Adapters for Sound Event Detection 291
Kang Li, Yan Song, Ian McLoughlin, Lin Liu, Jin Li, Li-Rong Dai

Small Footprint Multi-Channel Network for Keyword Spotting with Centroid Based Awareness 296
Dianwen Ng, Yang Xiao, Jia Qi Yip, Zhao Yang, Biao Tian, Qiang Fu, Eng Siong Chng, Bin Ma

Few-Shot Class-Incremental Audio Classification Using Adaptively-Refined Prototypes 301
Wei Xie, Yanxiong Li, Qianhua He, Wenchang Cao, Tuomas Virtanen

Interpretable Latent Space Using Space-Filling Curves for Phonetic Analysis in Voice Conversion 306
Mohammad Hassan Vali, Tom Bäckström

Topological Data Analysis for Speech Processing 311
Eduard Tulchinskii, Kristian Kuznetsov, Laida Kushnareva, Daniil Cherniavskii, Serguei Barannikov, Irina Piontkovskaya, Sergey Nikolenko, Evgeny Burnaev

Recycle-And-Distill: Universal Compression Strategy for Transformer-Based Speech SSL Models with Attention Map Reusing and Masking Distillation 316
Kangwook Jang, Sungnyun Kim, Se-Young Yun, Hoirin Kim

Personalized Acoustic Scene Classification in Ultra-Low Power Embedded Devices Using Privacy-Preserving Data Augmentation 321
Timm Koppelmann, Semih Agcaer, Rainer Martin

Background Domain Switch: A Novel Data Augmentation Technique for Robust Sound Event Detection 326
Wei-Cheng Lin, Luca Bondi, Shabnam Ghaffarzadegan

Joint Prediction of Audio Event and Annoyance Rating in an Urban Soundscape by Hierarchical Graph Representation Learning 331
Yuanbo Hou, Siyang Song, Cheng Luo, Andrew Mitchell, Qiaoqiao Ren, Weicheng Xie, Jian Kang, Wenwu Wang, Dick Botteldooren

Anomalous Sound Detection Using Self-Attention-Based Frequency Pattern Analysis of Machine Sounds 336
Hejing Zhang, Jian Guan, Qiaoxi Zhu, Feiyang Xiao, Youde Liu

Improving Audio-Text Retrieval Via Hierarchical Cross-Modal Interaction and Auxiliary Captions 341
Yifei Xin, Yuexian Zou

Differential Privacy Enabled Dementia Classification: An Exploration of the Privacy-Accuracy Trade-Off in Speech Signal Data 346
Suhas BN, Sarah Rajtmajer, Saeed Abdullah

Learning Emotional Representations from Imbalanced Speech Data for Speech Emotion Recognition and Emotional Text-To-Speech 351
Shijun Wang, Jón Guðnason, Damian Borth

Towards Multi-Lingual Audio Question Answering 356
Swarup Ranjan Behera, Pailla Balakrishna Reddy, Achyut Mani Tripathi, Megavath Bharadwaj Rathod, Tejesh Karavadi

SPEECH RECOGNITION: ARCHITECTURE, SEARCH, AND LINGUISTIC COMPONENTS 1

Diacritic Recognition Performance in Arabic ASR 361
Hanan Aldarmaki, Ahmad Ghannam

Personalization for BERT-Based Discriminative Speech Recognition Rescoring 366
Jari Kolehmainen, Yile Gu, Aditya Gourav, Prashanth Gurunath Shivakumar, Ankur Gandhe, Ariya Rastrow, Ivan Bulyko

On the N-Gram Approximation of Pre-Trained Language Models 371
Aravind Krishnan, Jesujoba O. Alabi, Dietrich Klakow

Record Deduplication for Entity Distribution Modeling in ASR Transcripts 376
Tianyu Huang, Chung Hoon Hong, Carl Wivagg, Kanna Shimizu

Learning When to Trust Which Teacher for Weakly Supervised ASR 381
Aakriti Agrawal, Milind Rao, Anit Kumar Sahu, Gopinath Chennupati, Andreas Stolcke

Text-Only Domain Adaptation Using Unified Speech-Text Representation in Transducer 386
Lu Huang, Boyu Li, Jun Zhang, Lu Lu, Zejun Ma

SPEECH RECOGNITION: TECHNOLOGIES AND SYSTEMS FOR NEW APPLICATIONS 1

Syllable Discovery and Cross-Lingual Generalization in a Visually Grounded, Self-Supervised Speech Model 391
Puyuan Peng, Shang-Wen Li, Okko Räsänen, Abdelrahman Mohamed, David Harwath

Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization 396
Puyuan Peng, Brian Yan, Shinji Watanabe, David Harwath

Progress and Prospects for Spoken Language Technology: Results from Five Sexennial Surveys 401
Roger K. Moore, Ricard Marxer

Acoustic Word Embeddings for Untranscribed Target Languages with Continued Pretraining and Learned Pooling 406
Ramon Sanabria, Ondrej Klejch, Hao Tang, Sharon Goldwater

CASA-ASR: Context-Aware Speaker-Attributed ASR 411
Mohan Shi, Zhihao Du, Qian Chen, Fan Yu, Yangze Li, Shiliang Zhang, Jie Zhang, Li-Rong Dai

Unsupervised Learning of Discrete Latent Representations with Data-Adaptive Dimensionality from Continuous Speech Streams 416
Shun Takahashi, Sakriani Sakti

AD-TUNING: An Adaptive CHILD-TUNING Approach to Efficient Hyperparameter Optimization of Child Networks for Speech Processing Tasks in the SUPERB Benchmark 421
Gaobin Yang, Jun Du, Maokui He, Shutong Niu, Baoxiang Li, Jiakui Li, Chin-Hui Lee

Distilling Knowledge from Gaussian Process Teacher to Neural Network Student 426
Jeremy H. M. Wong, Huayun Zhang, Nancy F. Chen

Segmental SpeechCLIP: Utilizing Pretrained Image-Text Models for Audio-Visual Learning 431
Saurabhchand Bhati, Jesús Villalba, Laureano Moro-Velazquez, Thomas Thebaud, Najim Dehak

Towards Hate Speech Detection in Low-Resource Languages: Comparing ASR to Acoustic Word Embeddings on Wolof and Swahili 436
Christiaan Jacobs, Nathanaël Carraz Rakotonirina, Everlyn Asiko Chimoto, Bruce A. Bassett, Herman Kamper

Mitigating Catastrophic Forgetting for Few-Shot Spoken Word Classification Through Meta-Learning 441
Ruan Van Der Merwe, Herman Kamper

Online Punctuation Restoration Using ELECTRA Model for Streaming ASR Systems 446
Martin Polácek, Petr Cerva, Jindrich Ždánský, Lenka Weingartová

Language Agnostic Data-Driven Inverse Text Normalization 451
Szu-Jui Chen, Debjyoti Paul, Yutong Pang, Peng Su, Xuedong Zhang

How to Estimate Model Transferability of Pre-Trained Speech Models? 456
Zih-Ching Chen, Chao-Han Huck Yang, Bo Li, Yu Zhang, Nanxin Chen, Shuo-Yiin Chang, Rohit Prabhavalkar, Hung-Yi Lee, Tara Sainath

Transcribing Speech as Spoken and Written Dual Text Using an Autoregressive Model 461
Mana Ihori, Hiroshi Sato, Tomohiro Tanaka, Ryo Masumura, Saki Mizuno, Nobukatsu Hojo

LEXICAL AND LANGUAGE MODELING FOR ASR

NoRefER: A Referenceless Quality Metric for Automatic Speech Recognition Via Semi-Supervised Language Model Fine-Tuning with Contrastive Learning 466
Kamer Ali Yuksel, Thiago Castro Ferreira, Golara Javadi, Mohamed Al-Badrashiny, Ahmet Gunduz

Scaling Laws for Discriminative Speech Recognition Rescoring Models 471
Yile Gu, Prashanth Gurunath Shivakumar, Jari Kolehmainen, Ankur Gandhe, Ariya Rastrow, Ivan Bulyko

Exploring Energy-Based Language Models with Different Architectures and Training Methods for Speech Recognition 476
Hong Liu, Zhaobiao Lv, Zhijian Ou, Wenbo Zhao, Qing Xiao

Memory Augmented Lookup Dictionary Based Language Modeling for Automatic Speech Recognition 481
Yukun Feng, Ming Tu, Rui Xia, Chuanzeng Huang, Yuxuan Wang

Memory Network-Based End-To-End Neural ES-KMeans for Improved Word Segmentation 486
Yu Iwamoto, Takahiro Shinozaki

Retraining-Free Customized ASR for Enharmonic Words Based on a Named-Entity-Aware Model and Phoneme Similarity Estimation 491
Yui Sudo, Kazuya Hata, Kazuhiro Nakadai

LANGUAGE IDENTIFICATION AND DIARIZATION

Lightweight and Efficient Spoken Language Identification of Long-Form Audio 496
Winstead Zhu, Md Iftekhar Tanveer, Yang Janet Liu, Seye Ojumu, Rosie Jones

End to End Spoken Language Diarization with Wav2vec Embeddings 501
Jagabandhu Mishra, Jayadev N Patil, Amartya Chowdhury, Mahadeva Prasanna

Efficient Spoken Language Recognition Via Multilabel Classification 506
Oriol Nieto, Zeyu Jin, Franck Dernoncourt, Justin Salamon

Description and Analysis of ABC Submission to NIST LRE 2022 511
Pavel Matejka, Anna Silnova, Josef Slavícek, Ladislav Mosner, Oldrich Plchot, Michal Klco, Junyi Peng, Themos Stafylakis, Lukáš Burget

Exploring the Impact of Pretrained Models and Web-Scraped Data for the 2022 NIST Language Recognition Evaluation 516
Tanel Alumäe, Kunnar Kukk, Viet-Bac Le, Claude Barras, Abdel Messaoudi, Waad Ben Kheder

Advances in Language Recognition in Low Resource African Languages: The JHU-MIT Submission for NIST LRE22 521
Jesús Villalba, Jonas Borgstrom, Maliha Jahan, Saurabh Kataria, Leibny Paola Garcia, Pedro Torres-Carrasquillo, Najim Dehak

SPEECH QUALITY ASSESSMENT

DeePMOS: Deep Posterior Mean-Opinion-Score of Speech 526
Xinyu Liang, Fredrik Cumlin, Christian Schüldt, Saikat Chatterjee

The Role of Formant and Excitation Source Features in Perceived Naturalness of Low Resource Tribal Language TTS: An Empirical Study 531
Ashwini Dasare, Pradyoth Hegde, Supritha Shetty, Deepak K T

A No-Reference Speech Quality Assessment Method Based on Neural Network with Densely Connected Convolutional Architecture 536
Wuxuan Gong, Jing Wang, Yitong Liu, Hongwen Yang

Probing Speech Quality Information in ASR Systems 541
Bao Thang Ta, Minh Tu Le, Nhat Minh Le, Van Hai Do

Preference-Based Training Framework for Automatic Speech Quality Assessment Using Deep Neural Network 546
Cheng-Hung Hu, Yusuke Yasuda, Tomoki Toda

Crowdsourced Data Validation for ASR Training 551
Wannaphong Phatthiyaphaibun, Chompakorn Chaksangchaichot, Thanawin Rakthammanon, Ekapol Chuangsuwanich, Sarana Nutanong

FEATURE MODELING FOR ASR

Re-Investigating the Efficient Transfer Learning of Speech Foundation Model Using Feature Fusion Methods 556
Zhouyuan Huo, Khe Chai Sim, Dongseong Hwang, Tsendsuren Munkhdalai, Tara Sainath, Pedro M. Mengibar

Robust Automatic Speech Recognition Via WavAugment Guided Phoneme Adversarial Training 561
Gege Qi, Yuefeng Chen, Xiaofeng Mao, Xiaojun Jia, Ranjie Duan, Rong Zhang, Hui Xue

InterFormer: Interactive Local and Global Features Fusion for Automatic Speech Recognition 566
Zhi-Hao Lai, Tian-Hao Zhang, Qi Liu, Xinyuan Qian, Li-Fang Wei, Feng Chen, Song-Lu Chen, Xu-Cheng Yin

Transductive Feature Space Regularization for Few-Shot Bioacoustic Event Detection 571
Yizhou Tan, Haojun Ai, Shengchen Li, Feng Zhang

Incorporating L2 Phonemes Using Articulatory Features for Robust Speech Recognition 576
Jisung Wang, Haram Lee, Myungwoo Oh

On the (In)Efficiency of Acoustic Feature Extractors for Self-Supervised Speech Representation Learning 581
Titouan Parcollet, Shucong Zhang, Rogier Van Dalen, Alberto Gil C. P. Ramos, Sourav Bhattacharya

INTERFACING SPEECH TECHNOLOGY AND PHONETICS

Phonemic Competition in End-To-End ASR Models 586
Louis Ten Bosch, Martijn Bentum, Lou Boves

Automatic Speaker Recognition with Variation Across Vocal Conditions: A Controlled Experiment with Implications for Forensics 591
Vincent Hughes, Jessica Wormald, Paul Foulkes, Philip Harrison, Finnian Kelly, David Van Der Vloed, Poppy Welch, Chenzi Xu

Exploring Graph Theory Methods for the Analysis of Pronunciation Variation in Spontaneous Speech 596
Bernhard C. Geiger, Barbara Schuppler

Automatic Speaker Recognition Performance with Matched and Mismatched Female Bilingual Speech Data 601
Bryony Nuttall, Philip Harrison, Vincent Hughes

SPEECH SYNTHESIS: MULTILINGUALITY

FACTSpeech: Speaking a Foreign Language Pronunciation Using Only Your Native Characters 606
Hong-Sun Yang, Ji-Hoon Kim, Yoon-Cheol Ju, Il-Hwan Kim, Byeong-Yeol Kim, Shuk-Jae Choi, Hyung-Yong Kim

Cross-Lingual Transfer Learning for Phrase Break Prediction with Multilingual Language Model.....611
Hoyeon Lee, Hyun-Wook Yoon, Jong-Hwan Kim, Jae-Min Kim

DSE-TTS: Dual Speaker Embedding for Cross-Lingual Text-To-Speech..... 616
Sen Liu, Yiwei Guo, Chenpeng Du, Xie Chen, Kai Yu

Generating Multilingual Gender-Ambiguous Text-To-Speech Voices 621
Konstantinos Markopoulos, Georgia Maniati, Georgios Vamvoukakis, Nikolaos Ellinas, Georgios Vardaxoglou, Panos Kakoulidis, Junkwang Oh, Gunu Jho, Inchul Hwang, Aimilios Chalamandaris, Pirros Tsiakoulis, Spyros Raptis

RAD-MMM: Multilingual Multiaccented Multispeaker Text to Speech..... 626
Rohan Badlani, Rafael Valle, Kevin J. Shih, João Felipe Santos, Siddharth Gururani, Bryan Catanzaro

Multilingual Context-Based Pronunciation Learning for Text-To-Speech 631
Giulia Comini, Sam Ribeiro, Fan Yang, Heereen Shim, Jaime Lorenzo-Trueba

SPEECH EMOTION RECOGNITION 1

Personalized Adaptation with Pre-Trained Speech Encoders for Continuous Emotion Recognition 636
Minh Tran, Yufeng Yin, Mohammad Soleymani

The Importance of Calibration: Rethinking Confidence and Performance of Speech Multi-Label Emotion Classifiers 641
Huang-Cheng Chou, Lucas Goncalves, Seong-Gyun Leem, Chi-Chun Lee, Carlos Busso

A Preliminary Study on Augmenting Speech Emotion Recognition Using a Diffusion Model 646
Mohammad Ibrahim Malik, Siddique Latif, Raja Jurdak, Björn W. Schuller

Privacy Risks in Speech Emotion Recognition: A Systematic Study on Gender Inference Attack 651
Basmah Alsenani, Tanaya Guha, Alessandro Vinciarelli

Episodic Memory for Domain-Adaptable, Robust Speech Emotion Recognition..... 656
James Tavernor, Matthew Perez, Emily Mower Provost

Stable Speech Emotion Recognition with Head-K-Pooling Loss..... 661
Chaoyue Ding, Jiakui Li, Daoming Zong, Baoxiang Li, Tian-Hao Zhang, Qunyan Zhou

SHOW AND TELL: HEALTH APPLICATIONS AND EMOTION RECOGNITION

A Personalised Speech Communication Application for Dysarthric Speakers 666
Matthew Gibson, Ievgen Karaulov, Oleksii Zhelo, Filip Jurcicek

Video Multimodal Emotion Recognition System for Real World Applications 668
Sun-Kyung Lee, Jong-Hwan Kim

Promoting Mental Self-Disclosure in a Spoken Dialogue System 670
Mahdin Rohmatillah, Bobbi Aditya, Li-Jen Yang, Bryan Gautama Ngo, Willianto Sulaiman, Jen-Tzung Chien

"Select Language, Modality Or Put on a Mask!" Experiments with Multimodal Emotion Recognition 672
Pawel Bujnowski, Bartlomiej Kuzma, Bartlomiej Paziewski, Jacek Rutkowski, Joanna Marhula, Zuzanna Bordzicka, Piotr Andruszkiewicz

My Vowels Matter: Formant Automation Tools for Diverse Child Speech 674
Hannah Valentine, Joel Macauslan, Maria Grigos, Marisha Speights

NEMA: An Ecologically Valid Tool for Assessing Hearing Devices, Advanced Algorithms, and Communication in Diverse Listening Environments 676
Nicky Chong-White, Arun Sebastian, Jorge Mejia

When Words Speak Just as Loudly as Actions: Virtual Agent Based Remote Health Assessment Integrating What Patients Say with What They Do 678
Vikram Ramanarayanan, David Pautler, Lakshmi Arbatti, Abhishek Hosamath, Michael Neumann, Hardik Kothare, Oliver Roesler, Jackson Liscombe, Andrew Cornish, Doug Habberstad, Vanessa Richter, David Fox, David Suendermann-Oeft, Ira Shoulson

Stuttering Detection Application 680
Kowshik Siva Sai Motepalli, Vamshiraghusimha Narasinga, Harsha Pathuri, Hina Khan, Sangeetha Mahesh, Ajish K. Abraham, Anil Kumar Vuppala

Providing Interpretable Insights for Neurological Speech and Cognitive Disorders from Interactive Serious Games 682
Mario Zusag, Laurin Wagner

Automated Neural Nursing Assistant (ANNA): An Over-The-Phone System for Cognitive Monitoring 684
Jacob Solinsky, Raymond Finzel, Martin Michalowski, Serguei Pakhomov

5G-IoT Cloud Based Demonstration of Real-Time Audio-Visual Speech Enhancement for Multimodal Hearing-Aids 686
Ankit Gupta, Abhijeet Bishnu, Mandar Gogate, Kia Dashtipour, Tughrul Arslan, Ahsan Adeel, Amir Hussain, Tharmalingam Ratnarajah, Mathini Sellathurai

Towards Two-Point Neuron-Inspired Energy-Efficient Multimodal Open Master Hearing Aid 688
Mohsin Raza, Adewale Adetomi, Khubaib Ahmed, Amir Hussain, Tughrul Arslan, Ahsan Adeel

SPOKEN DIALOG SYSTEMS AND CONVERSATIONAL ANALYSIS 1

FC-MTLF: A Fine- And Coarse-Grained Multi-Task Learning Framework for Cross-Lingual Spoken Language Understanding 690
Xuxin Cheng, Wanshi Xu, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, Yuexian Zou

VOLUME 2

C²A-SLU: Cross and Contrastive Attention for Improving ASR Robustness in Spoken Language Understanding 695
Xuxin Cheng, Ziyu Yao, Zhihong Zhu, Yaowei Li, Hongxiang Li, Yuexian Zou

Tri-Level Joint Natural Language Understanding for Multi-Turn Conversational Datasets 700
Henry Weld, Sijia Hu, Siqu Long, Josiah Poon, Soyeon Han

Semantic Enrichment Towards Efficient Speech Representations 705
Gaëlle Laperrière, Ha Nguyen, Sahar Ghannay, Bassam Jabaian, Yannick Estève

Tensor Decomposition for Minimization of E2E SLU Model Toward On-Device Processing 710
Yosuke Kashiwagi, Siddhant Arora, Hayato Futami, Jessica Huynh, Shih-Lun Wu, Yifan Peng, Brian Yan, Emiru Tsunoo, Shinji Watanabe

DiffSLU: Knowledge Distillation Based Diffusion Model for Cross-Lingual Spoken Language Understanding 715
Tianjun Mao, Chenghong Zhang

Integrating Pretrained ASR and LM to Perform Sequence Generation for Spoken Language Understanding 720
Siddhant Arora, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, Brian Yan, Shinji Watanabe

Contrastive Learning Based ASR Robust Knowledge Selection for Spoken Dialogue System 725
Zhiyuan Zhu, Yusheng Liao, Yu Wang, Yunfeng Guan

Unsupervised Dialogue Topic Segmentation in Hyperdimensional Space 730
Seongmin Park, Jinkyu Seo, Jihwa Lee

An Investigation of the Combination of Rehearsal and Knowledge Distillation in Continual Learning for Spoken Language Understanding 735
Umberto Cappellazzo, Daniele Falavigna, Alessio Brutti

Enhancing New Intent Discovery Via Robust Neighbor-Based Contrastive Learning 740
Zhenhe Wu, Xiaoguang Yu, Meng Chen, Liangqing Wu, Jiahao Ji, Zhoujun Li

Personalized Predictive ASR for Latency Reduction in Voice Assistants 745
Andreas Schwarz, Di He, Maarten Van Segbroeck, Mohammed Hethnawi, Ariya Rastrow

Compositional Generalization in Spoken Language Understanding 750
Avik Ray, Yilin Shen, Hongxia Jin

Sampling Bias in NLU Models: Impact and Mitigation 755
Zefei Li, Anil Ramakrishna, Anna Rumshisky, Andy Rosenbaum, Saleh Solta, Rahul Gupta

5IDER: Unified Query Rewriting for Steering, Intent Carryover, Disfluencies, Entity Carryover and Repair 760
Jiarui Lu, Bo-Hsiang Tseng, Joel Ruben Antony Moniz, Site Li, Xueyun Zhu, Hong Yu, Murat Akbacak

Emotion Awareness in Multi-Utterance Turn for Improving Emotion Prediction in Multi-Speaker Conversation 765
Xiaohan Shi, Xingfeng Li, Tomoki Toda

WhiSLU: End-To-End Spoken Language Understanding with Whisper 770
Minghan Wang, Yinglu Li, Jiaxin Guo, Xiaosong Qiao, Zongyao Li, Hengchao Shang, Daimeng Wei, Shimin Tao, Min Zhang, Hao Yang

SPEECH CODING AND ENHANCEMENT 1

Biophysically-Inspired Single-Channel Speech Enhancement in the Time Domain 775
Chuan Wen, Sarah Verhulst

On-Device Speaker Anonymization of Acoustic Embeddings for ASR Based on Flexible Location Gradient Reversal Layer 780
Md Asif Jalal, Pablo Peso Parada, Jisi Zhang, Mete Ozay, Karthikeyan Saravanan, Myoungji Han, Jung In Lee, Seokyeong Jung

How to Construct Perfect and Worse-Than-Coin-Flip Spoofing Countermeasures: A Word of Warning on Shortcut Learning 785
Hye-Jin Shim, Rosa Gonzalez Hautamäki, Md Sahidullah, Tomi Kinnunen

CleanUNet 2: A Hybrid Speech Denoising Model on Waveform and Spectrogram 790
Zhifeng Kong, Wei Ping, Ambrish Dantrey, Bryan Catanzaro

A Two-Stage Progressive Neural Network for Acoustic Echo Cancellation 795
Zhuangqi Chen, Xianjun Xia, Cheng Chen, Xianke Wang, Yanhong Leng, Li Chen, Roberto Togneri, Yijian Xiao, Piao Ding, Shenyi Song, Pingjian Zhang

An Intra-BRNN and GB-RVQ Based END-TO-END Neural Audio Codec 800
Linping Xu, Jiawei Jiang, Dejun Zhang, Xianjun Xia, Li Chen, Yijian Xiao, Piao Ding, Shenyi Song, Sixing Yin, Ferdous Sohel

Real-Time Personalised Speech Enhancement Transformers with Dynamic Cross-Attended Speaker Representations 804
Shucong Zhang, Malcolm Chadwick, Alberto Gil C. P. Ramos, Titouan Parcollet, Rogier Van Dalen, Sourav Bhattacharya

CFTNet: Complex-Valued Frequency Transformation Network for Speech Enhancement 809
Nursadul Mamun, John H. L. Hansen

Feature Normalization for Fine-Tuning Self-Supervised Models in Speech Enhancement 814
Hejung Yang, Hong-Goo Kang

Multi-Mode Neural Speech Coding Based on Deep Generative Networks 819
Wei Xiao, Wenzhe Liu, Meng Wang, Shan Yang, Yupeng Shi, Yuyong Kang, Dan Su, Shidong Shang, Dong Yu

Streaming Dual-Path Transformer for Speech Enhancement 824
Soo Hyun Bae, Seok Wan Chae, Youngseok Kim, Keunsang Lee, Hyunjin Lim, Lae-Hoon Kim

Sequence-To-Sequence Multi-Modal Speech In-Painting 829
Mahsa Kadkhodaei Elyaderani, Shahram Shirani

Hybrid AHS: A Hybrid of Kalman Filter and Deep Learning for Acoustic Howling Suppression 834
Hao Zhang, Meng Yu, Yuzhong Wu, Tao Yu, Dong Yu

Differentially Private Adapters for Parameter Efficient Acoustic Modeling 839
Chun-Wei Ho, Chao-Han Huck Yang, Sabato Marco Siniscalchi

Incorporating Ultrasound Tongue Images for Audio-Visual Speech Enhancement Through Knowledge Distillation 844
Rui-Chen Zheng, Yang Ai, Zhen-Hua Ling

Consonant-Emphasis Method Incorporating Robust Consonant-Section Detection to Improve Intelligibility of Bone-Conducted Speech 849
Yasufumi Uezu, Sicheng Wang, Teruki Toya, Masashi Unoki

Downstream Task Agnostic Speech Enhancement with Self-Supervised Representation Loss 854
Hiroshi Sato, Ryo Masumura, Tsubasa Ochiai, Marc Delcroix, Takafumi Moriya, Takanori Ashihara, Kentaro Shinayama, Saki Mizuno, Mana Ihori, Tomohiro Tanaka, Nobukatsu Hojo

Perceptual Improvement of Deep Neural Network (DNN) Speech Coder Using Parametric and Non-Parametric Density Models 859
Joon Byun, Seungmin Shin, Jongmo Sung, Seungkwon Beack, Youngcheol Park

DeFT-AN RT: Real-Time Multichannel Speech Enhancement Using Dense Frequency-Time Attentive Network and Non-Overlapping Synthesis Window 864
Dongheon Lee, Dayun Choi, Jung-Woo Choi

SPEECH RECOGNITION: SIGNAL PROCESSING, ACOUSTIC MODELING, ROBUSTNESS, ADAPTATION 2

A More Accurate Internal Language Model Score Estimation for the Hybrid Autoregressive Transducer 869
Kyungmin Lee, Haeri Kim, Sichen Jin, Jinhwan Park, Youngho Han

Attention Gate Between Capsules in Fully Capsule-Network Speech Recognition 874
Kyungmin Lee, Hyeontaek Lim, Mun-Hwan Lee, Hong-Gee Kim

ML-SUPERB: Multilingual Speech Universal PERformance Benchmark 884
Jiatong Shi, Dan Berrebbi, William Chen, En-Pei Hu, Wei-Ping Huang, Ho-Lam Chung, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, Shinji Watanabe

General-Purpose Adversarial Training for Enhanced Automatic Speech Recognition Model Generalization 889
Dohee Kim, Daeyeol Shim, Joon-Hyuk Chang

Joint Instance Reconstruction and Feature Subspace Alignment for Cross-Domain Speech Emotion Recognition 894
Keke Zhao, Peng Song, Shaokai Li, Wenming Zheng

Knowledge Distillation for Neural Transducer-Based Target-Speaker ASR: Exploiting Parallel Mixture/Single-Talker Speech Data 899
Takafumi Moriya, Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, Takanori Ashihara, Kohei Matsuura, Tomohiro Tanaka, Ryo Masumura, Atsunori Ogawa, Taichi Asami

Random Utterance Concatenation Based Data Augmentation for Improving Short-Video Speech Recognition 904
Yist Y. Lin, Tao Han, Haihua Xu, Van Tung Pham, Yerbolat Khassanov, Tze Yuang Chong, Yi He, Lu Lu, Zejun Ma

Adapter Incremental Continual Learning of Efficient Audio Spectrogram Transformers 909
Nithish Muthuchamy Selvaraj, Xiaobao Guo, Adams Kong, Bingquan Shen, Alex Kot

Rethinking Speech Recognition with a Multimodal Perspective Via Acoustic and Semantic Cooperative Decoding 914
Tian-Hao Zhang, Hai-Bo Qin, Zhi-Hao Lai, Song-Lu Chen, Qi Liu, Feng Chen, Xinyuan Qian, Xu-Cheng Yin

Improving Code-Switching and Name Entity Recognition in ASR with Speech Editing Based Data Augmentation 919
Zheng Liang, Zheshu Song, Ziyang Ma, Chenpeng Du, Kai Yu, Xie Chen

Bypass Temporal Classification: Weakly Supervised Automatic Speech Recognition with Imperfect Transcripts 924
Dongji Gao, Matthew Wiesner, Hainan Xu, Leibny Paola Garcia, Daniel Povey, Sanjeev Khudanpur

DCCRN-KWS: An Audio Bias Based Model for Noise Robust Small-Footprint Keyword Spotting 929
Shubo Lv, Xiong Wang, Sining Sun, Long Ma, Lei Xie

OTF: Optimal Transport Based Fusion of Supervised and Self-Supervised Learning Models for Automatic Speech Recognition 934
Li Fu, Siqi Li, Qingtao Li, Fangzhu Li, Liping Deng, Lu Fan, Meng Chen, Youzheng Wu, Xiaodong He

Approximate Nearest Neighbour Phrase Mining for Contextual Speech Recognition 939
Maurits Bleeker, Pawel Swietojanski, Stefan Braun, Xiaodan Zhuang

Rehearsal-Free Online Continual Learning for Automatic Speech Recognition 944
Steven Vander Eeckt, Hugo Van Hamme

SPEECH RECOGNITION: TECHNOLOGIES AND SYSTEMS FOR NEW APPLICATIONS 2

Phonetic and Prosody-Aware Self-Supervised Learning Approach for Non-Native Fluency Scoring 949
Kaiqi Fu, Shaojun Gao, Shuju Shi, Xiaohai Tian, Wei Li, Zejun Ma

Disentangling the Contribution of Non-Native Speech in Automated Pronunciation Assessment 954
Shuju Shi, Kaiqi Fu, Yiwei Gu, Xiaohai Tian, Shaojun Gao, Wei Li, Zejun Ma

A Joint Model for Pronunciation Assessment and Mispronunciation Detection and Diagnosis with Multi-Task Learning 959
Hyungshin Ryu, Sunhee Kim, Minhwa Chung

Assessing Intelligibility in Non-Native Speech: Comparing Measures Obtained at Different Levels 964
Xing Wei, Roeland Van Hout, Catia Cucchiarini, Danielle Reuvekamp, Helmer Strik

End-To-End Word-Level Pronunciation Assessment with MASK Pre-Training 969
Yukang Liang, Kaitao Song, Shaoguang Mao, Huiqiang Jiang, Luna Qiu, Yuqing Yang, Dongsheng Li, Linli Xu, Lili Qiu

A Hierarchical Context-Aware Modeling Approach for Multi-Aspect and Multi-Granular Pronunciation Assessment 974
Fu-An Chao, Tien-Hong Lo, Tzu-I Wu, Yao-Ting Sung, Berlin Chen

Automatic Prediction of Language Learners' Listenability Using Speech and Text Features Extracted from Listening Drills 979
Yingxiang Gao, Jaehyun Choi, Nobuaki Minematsu, Noriko Nakanishi, Daisuke Saito

Assessment of Non-Native Speech Intelligibility Using Wav2vec2-Based Mispronunciation Detection and Multi-Level Goodness of Pronunciation Transformer 984
Ram C. M. C. Shekar, Mu Yang, Kevin Hirschi, Stephen Looney, Okim Kang, John H. L. Hansen

Adapting an Unadaptable ASR System 989
Rao Ma, Mengjie Qian, Mark J. F. Gales, Kate M. Knill

Addressing Cold Start Problem for End-To-End Automatic Speech Scoring 994
Jungbae Park, Seungtaek Choi

Improving Grapheme-To-Phoneme Conversion by Learning Pronunciations from Speech Recordings 999
Sam Ribeiro, Giulia Comini, Jaime Lorenzo-Trueba

Orthography-Based Pronunciation Scoring for Better CAPT Feedback 1004
Caitlin Richter, Ragnar Pálsson, Luke O'Brien, Kolbrún Friðriksdóttir, Branislav Bédi, Eydís Huld Magnúsdóttir, Jón Guðnason

Zero-Shot Automatic Pronunciation Assessment 1009
Hongfu Liu, Mingqian Shi, Ye Wang

Mispronunciation Detection and Diagnosis Model for Tonal Language, Applied to Vietnamese 1014
Tuong Tu Huu, Viet Thanh Pham, Thi Thu Trang Nguyen, Thai Lai Dao

KEYNOTE 2

Beyond the AI Hype: Balancing Innovation and Social Responsibility 1019
Virginia Dignum

PARALINGUISTICS 1

Detection of Emotional Hotspots in Meetings Using a Cross-Corpus Approach 1020
Georg Stemmer, Paulo Lopez Meyer, Juan Del Hoyo Ontiveros, Jose Lopez, Hector A. Cordourier, Tobias Bocklet

Detection of Laughter and Screaming Using the Attention and CTC Models 1025
Takuto Matsuda, Yoshiko Arimoto

Capturing Formality in Speech Across Domains and Languages 1030
Debasmita Bhattacharya, Jie Chi, Julia Hirschberg, Peter Bell

Towards Robust Family-Infant Audio Analysis Based on Unsupervised Pretraining of Wav2vec 2.0 on Large-Scale Unlabeled Family Audio 1035
Jialu Li, Mark Hasegawa-Johnson, Nancy L. McElwain

Cues to Next-Speaker Projection in Conversational Swedish: Evidence from Reaction Times 1040
Kathrin Feindt, Martina Rossi, Ghazaleh Esfandiari-Baiat, Axel G. Ekström, Margaret Zellers

Multiple Instance Learning for Inference of Child Attachment from Paralinguistic Aspects of Speech 1045
Areej Buker, Huda Alsofyani, Alessandro Vinciarelli

SPEECH ENHANCEMENT AND DENOISING

Real-Time Joint Personalized Speech Enhancement and Acoustic Echo Cancellation 1050
Sefik Emre Eskimez, Takuya Yoshioka, Alex Ju, Min Tang, Tanel Pärnamaa, Huaming Wang

TaylorBeamixer: Learning Taylor-Inspired All-Neural Multi-Channel Speech Enhancement from Beam-Space Dictionary Perspective 1055
Andong Li, Weixin Meng, Guochen Yu, Wenzhe Liu, Xiaodong Li, Chengshi Zheng

MFT-CRN:Multi-Scale Fourier Transform for Monaural Speech Enhancement 1060
Yulong Wang, Xueliang Zhang

Variance-Preserving-Based Interpolation Diffusion Models for Speech Enhancement 1065
Zilu Guo, Jun Du, Chin-Hui Lee, Yu Gao, Wenbin Zhang

Multi-Input Multi-Output Complex Spectral Mapping for Speaker Separation 1070
Hassan Taherian, Ashutosh Pandey, Daniel Wong, Buye Xu, Deliang Wang

Short-Term Extrapolation of Speech Signals Using Recursive Neural Networks in the STFT Domain 1075
Maurice Oberhag, Daniel Neudek, Rainer Martin, Tobias Rosenkranz, Henning Puder

SPEECH SYNTHESIS: EVALUATION

Listener Sensitivity to Deviating Obstruents in WaveNet 1080
Ayushi Pandey, Jens Edlund, Sébastien Le Maguer, Naomi Harte

How Generative Spoken Language Modeling Encodes Noisy Speech: Investigation from Phonetics to Syntactics 1085
Joonyong Park, Shinnosuke Takamichi, Tomohiko Nakamura, Kentaro Seki, Detai Xin, Hiroshi Saruwatari

MOS vs. AB: Evaluating Text-To-Speech Systems Reliably Using Clustered Standard Errors 1090
Joshua Camp, Tom Kenter, Lev Finkelstein, Rob Clark

RAMP: Retrieval-Augmented MOS Prediction Via Confidence-Based Dynamic Weighting 1095
Hui Wang, Shiwan Zhao, Xiguang Zheng, Yong Qin

Can Better Perception Become a Disadvantage? Synthetic Speech Perception in Congenitally Blind Users 1100
Gerda Ana Melnik-Leroy, Gediminas Navickas

Investigating Range-Equalizing Bias in Mean Opinion Score Ratings of Synthesized Speech 1104
Erica Cooper, Junichi Yamagishi

END-TO-END SPOKEN DIALOG SYSTEMS

Can ChatGPT Detect Intent? Evaluating Large Language Models for Spoken Language Understanding 1109
Mutian He, Philip N. Garner

Improving End-To-End SLU Performance with Prosodic Attention and Distillation 1114
Shangeth Rajaa

Modality Confidence Aware Training for Robust End-To-End Spoken Language Understanding 1119
Suyoun Kim, Akshat Shrivastava, Duc Le, Ju Lin, Ozlem Kalinli, Michael L. Seltzer

Cross-Modal Semantic Alignment Before Fusion for Two-Pass End-To-End Spoken Language Understanding 1124
Lingyan Huang, Tao Li, Haodong Zhou, Qingyang Hong, Lin Li

ConvKT: Conversation-Level Knowledge Transfer for Context Aware End-To-End Spoken Language Understanding 1129
Vishal Sunder, Eric Fosler-Lussier, Samuel Thomas, Hong-Kwang J Kuo, Brian Kingsbury

GhostT5: Generate More Features with Cheap Operations to Improve Textless Spoken Question Answering 1134
Xuxin Cheng, Zhihong Zhu, Ziyu Yao, Hongxiang Li, Yaowei Li, Yuexian Zou

BIOSIGNAL-ENABLED SPOKEN COMMUNICATION

Obstructive Sleep Apnea Detection Using Pre-Trained Speech Representations 1139
Kaibo Zhang, Lili Cao, Yiming Ding, Yanru Li, Chao Zhang, Ji Wu, Demin Han

EEG-Based Auditory Attention Detection with Spatiotemporal Graph and Graph Convolutional Network 1144
Ruicong Wang, Siqi Cai, Haizhou Li

Silent Speech Recognition with Articulator Positions Estimated from Tongue Ultrasound and Lip Video 1149
Rachel Beeson, Korin Richmond

Auditory Attention Detection in Real-Life Scenarios Using Common Spatial Patterns from EEG 1154
Kai Yang, Zhuang Xie, Di Zhou, Longbiao Wang, Gaoyan Zhang

Diff-E: Diffusion-Based Learning for Decoding Imagined Speech EEG 1159
Soowon Kim, Young-Eun Lee, Seo-Hyun Lee, Seong-Whan Lee

Towards Ultrasound Tongue Image Prediction from EEG During Speech Production 1164
Tamás Gábor Csapó, Frigyes Viktor Arthur, Péter Nagy, Ádám Boncz

Adaptation of Tongue Ultrasound-Based Silent Speech Interfaces Using Spatial Transformer Networks 1169
László Tóth, Amin Honarmandi Shandiz, Gábor Gosztolya, Tamás Gábor Csapó

STE-GAN: Speech-To-Electromyography Signal Conversion Using Generative Adversarial Networks 1174
Kevin Scheck, Tanja Schultz

Spanish Phone Confusion Analysis for EMG-Based Silent Speech Interfaces 1179
Inge Salomons, Eder Del Blanco, Eva Navas, Inma Hernáez

Hybrid Silent Speech Interface Through Fusion of Electroencephalography and Electromyography 1184
Huiyan Li, Mingyi Wang, Han Gao, Shuo Zhao, Guang Li, You Wang

NEURAL-BASED SPEECH AND ACOUSTIC ANALYSIS

Can Self-Supervised Neural Representations Pre-Trained on Human Speech Distinguish Animal Callers? 1189
Eklavya Sarkar, Mathew Magimai.-Doss

Discovering COVID-19 Coughing and Breathing Patterns from Unlabeled Data Using Contrastive Learning with Varying Pre-Training Domains 1194
Jinjin Cai, Sudip Vhaduri, Xiao Luo

Background-Aware Modeling for Weakly Supervised Sound Event Detection 1199
Yifei Xin, Dongchao Yang, Yuexian Zou

How to (Virtually) Train Your Speaker Localizer 1204
Prerak Srivastava, Antoine Deleforge, Archontis Politis, Emmanuel Vincent

MMER: Multimodal Multi-Task Learning for Speech Emotion Recognition 1209
Sreyan Ghosh, Utkarsh Tyagi, S Ramaneswaran, Harshvardhan Srivastava, Dinesh Manocha

A Multi-Task Learning Framework for Sound Event Detection Using High-Level Acoustic Characteristics of Sounds 1214
Tanmay Khandelwal, Rohan Kumar Das

DIGO - DIALOG FOR GOOD: SPEECH AND LANGUAGE TECHNOLOGY FOR SOCIAL GOOD

A Multimodal Investigation of Speech, Text, Cognitive and Facial Video Features for Characterizing Depression with and Without Medication 1219
Michael Neumann, Hardik Kothare, Doug Habberstad, Vikram Ramanarayanan

Understanding Disrupted Sentences Using Underspecified Abstract Meaning Representation 1224
Angus Addlesee, Marco Damonte

Developing Speech Processing Pipelines for Police Accountability 1229
Anjalie Field, Prateek Verma, Nay San, Jennifer L. Eberhardt, Dan Jurafsky

Prosody-Controllable Gender-Ambiguous Speech Synthesis: A Tool for Investigating Implicit Bias in Speech Perception 1234
Éva Székely, Joakim Gustafson, Ilaria Torre

Affective Attributes of French Caregivers' Professional Speech 1239
Jean-Luc Rouas, Yaru Wu, Takaaki Shochi

SPEECH RECOGNITION: SIGNAL PROCESSING, ACOUSTIC MODELING, ROBUSTNESS, ADAPTATION 3

ASR Data Augmentation in Low-Resource Settings Using Cross-Lingual Multi-Speaker TTS and Cross-Lingual Voice Conversion 1244
Edresson Casanova, Christopher Shulby, Alexander Korolev, Arnaldo Candido Junior, Anderson Da Silva Soares, Sandra Aluísio, Moacir Antonelli Ponti

Personality-Aware Training Based Speaker Adaptation for End-To-End Speech Recognition 1249
Yue Gu, Zhihao Du, Shiliang Zhang, Qian Chen, Jiqing Han

Target Vocabulary Recognition Based on Multi-Task Learning with Decomposed Teacher Sequences 1254
Aoi Ito, Tatsuya Komatsu, Yusuke Fujita, Yusuke Kida

Wave to Syntax: Probing Spoken Language Models for Syntax 1259
Gaofei Shen, Afra Alishahi, Arianna Bisazza, Grzegorz Chrupala

Effective Training of Attention-Based Contextual Biasing Adapters with Synthetic Audio for Personalised ASR 1264
Burin Naowarat, Philip Harding, Pasquale D'Alterio, Sibo Tong, Bashar Awwad Shiekh Hasan

Pushing the Limits of Unsupervised Unit Discovery for SSL Speech Representation 1269
Ziyang Ma, Zhisheng Zheng, Guanrou Yang, Yu Wang, Chao Zhang, Xie Chen

SlothSpeech: Denial-Of-Service Attack Against Speech Recognition Models 1274
Mirazul Haque, Rutvij Shah, Simin Chen, Berrak Sisman, Cong Liu, Wei Yang

CLRL-Tuning: A Novel Continual Learning Approach for Automatic Speech Recognition 1279
Zhihan Wang, Feng Hou, Ruili Wang

Exploring Sources of Racial Bias in Automatic Speech Recognition Through the Lens of Rhythmic Variation 1284
Li-Fang Lai, Nicole Holliday

Can Contextual Biasing Remain Effective with Whisper and GPT-2? 1289
Guangzhi Sun, Xianrui Zheng, Chao Zhang, Philip C. Woodland

Masked Modeling Duo for Speech: Specializing General-Purpose Audio Representation to Speech Using Denoising Distillation 1294
Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, Kunio Kashino

Improving RNN Transducer Acoustic Models for English Conversational Speech Recognition 1299
Xiaodong Cui, George Saon, Brian Kingsbury

MixRep: Hidden Representation Mixup for Low-Resource Speech Recognition 1304
Jiamin Xie, John H. L. Hansen

Adapting Multi-Lingual ASR Models for Handling Multiple Talkers 1314
Chenda Li, Yao Qian, Zhuo Chen, Naoyuki Kanda, Dongmei Wang, Takuya Yoshioka, Yanmin Qian, Michael Zeng

Adapter-Tuning with Effective Token-Dependent Representation Shift for Automatic Speech Recognition 1319
Dianwen Ng, Chong Zhang, Ruixi Zhang, Yukun Ma, Trung Hieu Nguyen, Chongjia Ni, Shengkui Zhao, Qian Chen, Wen Wang, Eng Siong Chng, Bin Ma

Model-Internal Slot-Triggered Biasing for Domain Expansion in Neural Transducer ASR Models 1324
Yiting Lu, Philip Harding, Kanthashree Mysore Sathyendra, Sibo Tong, Xuandi Fu, Jing Liu, Feng-Ju Chang, Simon Wiesler, Grant P. Strimel

Delay-Penalized CTC Implemented Based on Finite State Transducer 1329
Zengwei Yao, Wei Kang, Fangjun Kuang, Liyong Guo, Xiaoyu Yang, Yifan Yang, Long Lin, Daniel Povey

SPEECH RECOGNITION: ARCHITECTURE, SEARCH, AND LINGUISTIC COMPONENTS 2

Text-Only Domain Adaptation for End-To-End Speech Recognition Through Down-Sampling Acoustic Representation 1334
Jiaxu Zhu, Weinan Tong, Yaoxun Xu, Changhe Song, Zhiyong Wu, Zhao You, Dan Su, Dong Yu, Helen Meng

Knowledge Distillation Approach for Efficient Internal Language Model Estimation 1339
Zhipeng Chen, Haihua Xu, Yerbolat Khassanov, Yi He, Lu Lu, Zejun Ma, Ji Wu

Language Model Personalization for Improved Touchscreen Typing 1344
Jiban Adhikary, Keith Vertanen

Blank Collapse: Compressing CTC Emission for the Faster Decoding 1349
Minkyu Jung, Ohhyeok Kwon, Seunghyun Seo, Soonshin Seo

Improving Joint Speech-Text Representations Without Alignment..... 1354
Cal Peyser, Zhong Meng, Rohit Prabhavalkar, Andrew Rosenberg, Tara Sainath, Michael Picheny, Kyunghyun Cho, Ke Hu

Leveraging Cross-Utterance Context for ASR Decoding..... 1359
Robert Flynn, Anton Ragni

Knowledge Transfer from Pre-Trained Language Models to Cif-Based Speech Recognizers Via Hierarchical Distillation 1364
Minglun Han, Feilong Chen, Jing Shi, Shuang Xu, Bo Xu

Integration of Frame- And Label-Synchronous Beam Search for Streaming Encoder-Decoder Speech Recognition..... 1369
Emiru Tsunoo, Hayato Futami, Yosuke Kashiwagi, Siddhant Arora, Shinji Watanabe

A Neural Time Alignment Module for End-To-End Automatic Speech Recognition..... 1374
Dongcheng Jiang, Chao Zhang, Philip C. Woodland

Accelerating Transducers Through Adjacent Token Merging 1379
Yuang Li, Yu Wu, Jinyu Li, Shujie Liu

Language-Universal Phonetic Representation in Multilingual Speech Pretraining for Low-Resource Speech Recognition..... 1384
Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, Yuxuan Wang

Language-Routing Mixture of Experts for Multilingual and Code-Switching Speech Recognition 1389
Wenxuan Wang, Guodong Ma, Yuke Li, Binbin Du

Embedding Articulatory Constraints for Low-Resource Speech Recognition Based on Large Pre-Trained Model 1394
Jaeyoung Lee, Masato Mimura, Tatsuya Kawahara

VOLUME 3

Exploration of Efficient End-To-End ASR Using Discretized Input from Self-Supervised Learning..... 1399
Xuankai Chang, Brian Yan, Yuya Fujita, Takashi Maekaku, Shinji Watanabe

SpellMapper: A Non-Autoregressive Neural Spellchecker for ASR Customization with Candidate Retrieval Based on N-Gram Mappings..... 1404
Alexandra Antonova, Evelina Bakhturina, Boris Ginsburg

Text Injection for Capitalization and Turn-Taking Prediction in Speech Models..... 1409
Shaan Bijwadia, Shuo-Yiin Chang, Weiran Wang, Zhong Meng, Hao Zhang

Confidence-Based Ensembles of End-To-End Speech Recognition Models..... 1414
Igor Gitman, Vitaly Lavrukhin, Aleksandr Laptev, Boris Ginsburg

Unsupervised Code-Switched Text Generation from Parallel Text 1419
Jie Chi, Brian Lu, Jason Eisner, Peter Bell, Preethi Jyothi, Ahmed M. Ali

A Binary Keyword Spotting System with Error-Diffusion Based Feature Binarization..... 1424
Dingyi Wang, Mengjie Luo, Lin Li, Xiaoqin Wang, Shushan Qiao, Yumei Zhou

Language-Universal Phonetic Encoder for Low-Resource Speech Recognition 1429
Siyuan Feng, Ming Tu, Rui Xia, Chuanzeng Huang, Yuxuan Wang

A Lexical-Aware Non-Autoregressive Transformer-Based ASR Model 1434
Chong-En Lin, Kuan-Yu Chen

Improving Under-Resourced Code-Switched Speech Recognition: Large Pre-Trained Models Or Architectural Interventions 1439
Joshua Jansen Van Vüren, Thomas Niesler

SPOKEN LANGUAGE TRANSLATION, INFORMATION RETRIEVAL, SUMMARIZATION, RESOURCES, AND EVALUATION 1

Pragmatic Pertinence: A Learnable Confidence Metric to Assess the Subjective Quality of LM-Generated Text 1444
Jerome R. Bellegarda

ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition 1449
Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, Catherine Lai

BASS: Block-Wise Adaptation for Speech Summarization 1454
Roshan Sharma, Siddhant Arora, Kenneth Zheng, Shinji Watanabe, Rita Singh, Bhiksha Raj

Speaker Tracking Using Graph Attention Networks with Varying Duration Utterances Across Multi-Channel Naturalistic Data: Fearless Steps Apollo-11 Audio Corpus 1459
Meena M. C. Shekar, John H. L. Hansen

Combining Language Corpora in a Japanese Electromagnetic Articulography Database for Acoustic-To-Articulatory Inversion 1464
Tianfang Yan, Kikuo Maekawa, Yukiko Nota, Masayuki Hirata

A Dual Attention-Based Modality-Collaborative Fusion Network for Emotion Recognition 1468
Xiaoheng Zhang, Yang Li

Large Dataset Generation of Synchronized Music Audio and Lyrics at Scale Using Teacher-Student Paradigm 1473
Cristian Chivriga, Rinita Roy

Enc-Dec RNN Acoustic Word Embeddings Learned Via Pairwise Prediction 1478
Adhiraj Banerjee, Vipul Arora

Query Based Acoustic Summarization for Podcasts 1483
Samantha Kotey, Rozenn Dahyot, Naomi Harte

Spot Keywords from Very Noisy and Mixed Speech 1488
Ying Shi, Dong Wang, Lantian Li, Jiqing Han, Shi Yin

Knowledge Distillation on Joint Task End-To-End Speech Translation 1493
Khandokar Md. Nayem, Ran Xue, Ching-Yun Chang, Akshaya Vishnu Kudlu Shanbhogue

Investigating Pre-Trained Audio Encoders in the Low-Resource Condition 1498
Hao Yang, Jinming Zhao, Gholamreza Haffari, Ehsan Shareghi

Improving Textless Spoken Language Understanding with Discrete Units as Intermediate Target 1503
Guan-Wei Wu, Guan-Ting Lin, Shang-Wen Li, Hung-Yi Lee

SPEECH, VOICE, AND HEARING DISORDERS 1

Debiased Automatic Speech Recognition for Dysarthric Speech Via Sample Reweighting with Sample Affinity Test 1508
Eungbeom Kim, Yunkee Chae, Jaeheon Sim, Kyogu Lee

Multimodal Locally Enhanced Transformer for Continuous Sign Language Recognition 1513
Katerina Papadimitriou, Gerasimos Potamianos

Towards Supporting an Early Diagnosis of Multiple Sclerosis Using Vocal Features 1518
Monica Gonzalez-Machorro, Pascal Hecker, Uwe D. Reichel, Helly N. Hammer, Robert Hoepner, Lisa Pedrotti, Alisha Zmutt, Hesam Sagha, Johan Van Beek, Florian Eyben, Dagmar M. Schuller, Björn W. Schuller, Bert Arnrich

Whisper Features for Dysarthric Severity-Level Classification 1523
Siddharth Rathod, Monil Charola, Akshat Vora, Yash Jogi, Hemant A. Patil

A New Benchmark of Aphasia Speech Recognition and Detection Based on E-Branchformer and Multi-Task Learning 1528
Jiyang Tang, William Chen, Xuankai Chang, Shinji Watanabe, Brian Macwhinney

Dysarthric Speech Recognition, Detection and Classification Using Raw Phase and Magnitude Spectra 1533
Zhengjun Yue, Erfan Loweimi, Zoran Cvetkovic

A Stutter Seldom Comes Alone – Cross-Corpus Stuttering Detection as a Multi-Label Problem 1538
Sebastian P. Bayerl, Dominik Wagner, Ilja Baumann, Florian Hönig, Tobias Bocklet, Elmar Nöth, Korbinian Riedhammer

Transfer Learning to Aid Dysarthria Severity Classification for Patients with Amyotrophic Lateral Sclerosis 1543
Tanuka Bhattacharjee, Anjali Jayakumar, Yamini Belur, Atchayaram Nalini, Ravi Yadav, Prasanta Kumar Ghosh

DuTa-VC: A Duration-Aware Typical-To-Atypical Voice Conversion Approach with Diffusion Probabilistic Model 1548
Helin Wang, Thomas Thebaud, Jesús Villalba, Myra Sydnor, Becky Lammers, Najim Dehak, Laureano Moro-Velazquez

CNVVE: Dataset and Benchmark for Classifying Non-Verbal Voice 1553
Ramin Hedeshy, Raphael Menges, Steffen Staab

Arabic Dysarthric Speech Recognition Using Adversarial and Signal-Based Augmentation 1558
Massa Baali, Ibrahim Almakky, Shady Shehata, Fakhri Karray

Weakly-Supervised Forced Alignment of Disfluent Speech Using Phoneme-Level Modeling 1563
Theodoros Kouzelis, Georgios Paraskevopoulos, Athanasios Katsamanis, Vassilis Katsouros

Glottal Source Analysis of Voice Deficits in Basal Ganglia Dysfunction: Evidence from De Novo Parkinson's Disease and Huntington's Disease 1568
Michal Novotný, Tereza Tykalová, Michal Šimek, Tomáš Kouba, Jan Rusz

An Analysis of Glottal Features of Chronic Kidney Disease Speech and Its Application to CKD Detection 1573
Jihyun Mun, Sunhee Kim, Myeong Ju Kim, Jiwon Ryu, Sejoong Kim, Minhwa Chung

Weakly Supervised Glottis Segmentation in High-Speed Videoendoscopy Using Bounding Box Labels 1578
Varun Belagali, Achuth Rao, Prasanta Kumar Ghosh

SPEECH RECOGNITION: TECHNOLOGIES AND SYSTEMS FOR NEW APPLICATIONS 3

An Efficient and Noise-Robust Audiovisual Encoder for Audiovisual Speech Recognition 1583
Zhengyang Li, Chenwei Liang, Timo Lohrenz, Marvin Sach, Björn Möller, Tim Fingscheidt

A Novel Self-Training Approach for Low-Resource Speech Recognition 1588
Satwinder Singh, Feng Hou, Ruili Wang

FunASR: A Fundamental End-To-End Speech Recognition Toolkit 1593
Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Shiliang Zhang

Streaming Audio-Visual Speech Recognition with Alignment Regularization 1598
Pingchuan Ma, Niko Moritz, Stavros Petridis, Christian Fuegen, Maja Pantic

SparseVSR: Lightweight and Noise Robust Visual Speech Recognition 1603
Adriana Fernandez-Lopez, Honglie Chen, Pingchuan Ma, Alexandros Haliassos, Stavros Petridis, Maja Pantic

Multimodal Speech Recognition for Language-Guided Embodied Agents 1608
Allen Chang, Xiaoyuan Zhu, Aarav Monga, Seoho Ahn, Tejas Srinivasan, Jesse Thomason

SPOKEN TERM DETECTION AND VOICE SEARCH

Matching Latent Encoding for Audio-Text Based Keyword Spotting 1613
Kumari Nishu, Minsik Cho, Devang Naik

Self-Paced Pattern Augmentation for Spoken Term Detection in Zero-Resource 1618
Sudhakar P, Sreenivasa K. Rao, Pabitra Mitra

On-Device Constrained Self-Supervised Speech Representation Learning for Keyword Spotting Via Knowledge Distillation 1623
Gene-Ping Yang, Yue Gu, Qingming Tang, Dongsu Du, Yuzong Liu

Online Continual Learning in Keyword Spotting for Low-Resource Devices Via Pooling High-Order Temporal Statistics 1628
Umberto Michieli, Pablo Peso Parada, Mete Ozay

Improving Small Footprint Few-Shot Keyword Spotting with Supervision on Auxiliary Data 1633
Seunghan Yang, Byeonggeun Kim, Kyuhong Shim, Simyoung Chang

Robust Keyword Spotting for Noisy Environments by Leveraging Speech Enhancement and Speech Presence Probability 1638
Chouchang Yang, Yashas Malur Saidutta, Rakshith Sharma Srinivasa, Ching-Hua Lee, Yilin Shen, Hongxia Jin

MODELS FOR STREAMING ASR

Enhancing the Unified Streaming and Non-Streaming Model with Contrastive Learning 1643
Yuting Yang, Yuke Li, Binbin Du

ZeroPrompt: Streaming Acoustic Encoders Are Zero-Shot Masked LMs 1648
Xingchen Song, Di Wu, Binbin Zhang, Zhendong Peng, Bo Dang, Fuping Pan, Zhiyong Wu

Improved Training for End-To-End Streaming Automatic Speech Recognition Model with Punctuation 1653
Hanbyul Kim, Seunghyun Seo, Lukas Lee, Seolki Baek

DCTX-Conformer: Dynamic Context Carry-Over for Low Latency Unified Streaming and Non-Streaming Conformer 1658
Goeric Huybrechts, Srikanth Ronanki, Xilai Li, Hadis Nosrati, Sravan Bodapati, Katrin Kirchhoff

Knowledge Distillation from Non-Streaming to Streaming ASR Encoder Using Auxiliary Non-Streaming Layer 1663
Kyuhong Shim, Jinkyu Lee, Simyoung Chang, Kyuwoong Hwang

Adaptive Contextual Biasing for Transducer Based Streaming Speech Recognition 1668
Tianyi Xu, Zhanheng Yang, Kaixun Huang, Pengcheng Guo, Ao Zhang, Biao Li, Changru Chen, Chao Li, Lei Xie

SOURCE SEPARATION

Audio-Visual Speech Separation in Noisy Environments with a Lightweight Iterative Model 1673
Héctor Martel, Julius Richter, Kai Li, Xiaolin Hu, Timo Gerkmann

Remixing-Based Unsupervised Source Separation from Scratch 1678
Kohei Saijo, Tetsuji Ogawa

CAPTDURE: Captioned Sound Dataset of Single Sources 1683
Yuki Okamoto, Kanta Shimonishi, Keisuke Imoto, Kota Dohi, Shota Horiguchi, Yohei Kawaguchi

Recursive Sound Source Separation with Deep Learning-Based Beamforming for Unknown Number of Sources 1688
Hokuto Munakata, Ryu Takeda, Kazunori Komatani

Multi-Channel Speech Separation with Cross-Attention and Beamforming 1693
Ladislav Mosner, Oldrich Plchot, Junyi Peng, Lukáš Burget, Jan "Honza" Cernocký

Background-Sound Controllable Voice Source Separation 1698
Deokjun Eom, Woo Hyun Nam, Kyung-Rae Kim

SPEECH AND LANGUAGE IN HEALTH: FROM REMOTE MONITORING TO MEDICAL CONVERSATIONS 1

An Automatic Multimodal Approach to Analyze Linguistic and Acoustic Cues on Parkinson's Disease Patients 1703
Daniel Escobar-Grisales, Tomás Arias-Vergara, Cristian David Ríos-Urrego, Elmar Nöth, Adolfo M. García, Juan Rafael Orozco-Arroyave

Personalization for Robust Voice Pathology Detection in Sound Waves 1708
Khanh-Tung Tran, Truong Hoang, Duy Khuong Nguyen, Hoang D. Nguyen, Xuan-Son Vu

Integrated and Enhanced Pipeline System to Support Spoken Language Analytics for Screening Neurocognitive Disorders 1713
Helen Meng, Brian Mak, Man-Wai Mak, Helene Fung, Xianmin Gong, Timothy Kwok, Xunying Liu, Vincent Mok, Patrick Wong, Jean Woo, Xixin Wu, Ka Ho Wong, Shensheng Xu, Naijun Zheng, Ranzo Huang, Jiawen Kang, Xiaoquan Ke, Junan Li, Jinchao Li, Yi Wang

Capturing Mismatch Between Textual and Acoustic Emotion Expressions for Mood Identification in Bipolar Disorder 1718
Minxue Niu, Amrit Romana, Mimansa Jaiswal, Melvin McInnis, Emily Mower Provost

FTA-Net: A Frequency and Time Attention Network for Speech Depression Detection 1723
Qifei Li, Dong Wang, Yiming Ren, Yingming Gao, Ya Li

Bayesian Networks for the Robust and Unbiased Prediction of Depression and Its Symptoms Utilizing Speech and Multimodal Data 1728
Salvatore Fara, Orlaith Hickey, Alexandra Georgescu, Stefano Goria, Emilia Molimpakis, Nicholas Cummins

Hyper-Parameter Adaptation of Conformer ASR Systems for Elderly and Dysarthric Speech Recognition 1733
Tianzi Wang, Shoukang Hu, Jiajun Deng, Zengrui Jin, Mengzhe Geng, Yi Wang, Helen Meng, Xunying Liu

Classifying Depression Symptom Severity: Assessment of Speech Representations in Personalized and Generalized Machine Learning Models. 1738
Edward L. Campbell, Judith Dineley, Pauline Conde, Faith Matcham, Katie M. White, Carolin Oetzmann, Sara Simblett, Stuart Bruce, Amos A. Folarin, Til Wykes, Srinivasan Vairavan, Richard J. B. Dobson, Laura Docio-Fernandez, Carmen Garcia-Mateo, Vaibhav A. Narayan, Matthew Hotopf, Nicholas Cummins

Active Learning for Abnormal Lung Sound Data Curation and Detection in Asthma 1743
Shabnam Ghaffarzadegan, Luca Bondi, Ho-Hsiang Wu, Sirajum Munir, Kelly J. Shields, Samarjit Das, Joseph Aracri

Automatic Assessment of Alzheimer's Across Three Languages Using Speech and Language Features 1748
Paula A. Pérez-Toro, Tomás Arias-Vergara, Franziska Braun, Florian Hönig, Carlos A. Tobón-Quintero, David Aguillón, Francisco Lopera, Liliana Hincapié-Henao, Maria Schuster, Korbinian Riedhammer, Andreas Maier, Elmar Nöth, Juan Rafael Orozco-Arroyave

On-The-Fly Feature Based Rapid Speaker Adaptation for Dysarthric and Elderly Speech Recognition 1753
Mengzhe Geng, Xurong Xie, Rongfeng Su, Jianwei Yu, Zengrui Jin, Tianzi Wang, Shujie Hu, Zi Ye, Helen Meng, Xunying Liu

Relationship Between LTAS-Based Spectral Moments and Acoustic Parameters of Hypokinetic Dysarthria in Parkinson’s Disease 1758
Jan Svihlik, Vojtech Illner, Petr Kryze, Mário Sousa, Paul Krack, Elina Tripoliti, Robert Jech, Jan Rusz

Respiratory Distress Estimation in Human-Robot Interaction Scenario 1763
Eduardo Alvarado, Nicolás Grágeda, Alejandro Luzanto, Rodrigo Mahu, Jorge Wuth, Laura Mendoza, Richard Stern, Néstor Becerra Yoma

Prediction of the Gender-Based Violence Victim Condition Using Speech: What Do Machine Learning Models Rely On? 1768
Emma Reyner-Fuentes, Esther Rituerto-González, Isabel Trancoso, Carmen Peláez-Moreno

Whisper Encoder Features for Infant Cry Classification 1773
Monil Charola, Aastha Kachhi, Hemant A. Patil

SPEECH PERCEPTION

A Neural Architecture for Selective Attention to Speech Features 1778
Nika Jurov, William Idsardi, Naomi H. Feldman

Quantifying Informational Masking Due to Masker Intelligibility in Same-Talker Speech-In-Speech Perception 1783
Mingyue Huo, Yinglun Sun, Dan Fogerty, Yan Tang

On the Benefits of Self-Supervised Learned Speech Representations for Predicting Human Phonetic Misperceptions 1788
Santiago Cuervo, Ricard Marxer

Predicting Perceptual Centers Located at Vowel Onset in German Speech Using Long Short-Term Memory Networks 1793
Felicia Schulz, Mirella De Sisto, M. Paula Roncaglia-Denissen, Peter Hendrix

Exploring the Mutual Intelligibility Breakdown Caused by Sculpting Speech from a Competing Speech Signal 1798
Martin Cooke, María Luisa García Lecumberri

Perception of Incomplete Voicing Neutralization of Obstruents in Tohoku Japanese 1803
Mafuyu Kitahara, Naoya Watabe, Hiroto Noguchi, Chuyu Huang, Ayako Hashimoto, Ai Mizoguchi

PHONETICS AND PHONOLOGY: LANGUAGES AND VARIETIES

The Emergence of Obstruent-Intrinsic F0 and VOT as Cues to the Fortis/Lenis Contrast in West Central Bavarian 1808
Jasmin Pöhnlein, Felicitas Kleber

⟨'⟩ in Tsimane': A Preliminary Investigation 1813
William N. Havard, Yaya Sy, Camila Scaff, Loann Peurey, Alejandrina Cristia

Segmental Features of Brazilian (Santa Catarina) Hunsrik 1818
Dennis Hoffmann, Maria O'Reilly

Opening or Closing? An Electroglottographic Analysis of Voiceless Coda Consonants in Australian English 1823
Louise Ratko, Joshua Penney, Felicity Cox

Increasing Aspiration of Word-Medial Fortis Plosives in Swiss Standard German 1828
Franka Zebe

Lexical Stress and Velar Palatalization in Italian: A Spatio-Temporal Interaction 1833
Bowei Shao, Philipp Buech, Anne Hermes, Maria Giavazzi

PARALINGUISTICS 2

Speaker Embeddings as Individuality Proxy for Voice Stress Detection 1838
Zihan Wu, Neil Scheidwasser-Clow, Karl El Hajal, Milos Cernak

From Interval to Ordinal: A HMM Based Approach for Emotion Label Conversion 1843
Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, Eliathamby Ambikairajah

Turbo Your Multi-Modal Classification with Contrastive Learning 1848
Zhiyu Zhang, Da Liu, Shengqiang Liu, Anna Wang, Jie Gao, Yali Li

Towards Paralinguistic-Only Speech Representations for End-To-End Speech Emotion Recognition 1853
Georgios Ioannides, Michael Owen, Andrew Fletcher, Viktor Rozgic, Chao Wang

SOT: Self-Supervised Learning-Assisted Optimal Transport for Unsupervised Adaptive Speech Emotion Recognition 1858
Ruiteng Zhang, Jianguo Wei, Xugang Lu, Yongwei Li, Junhai Xu, Di Jin, Jianhua Tao

On the Efficacy and Noise-Robustness of Jointly Learned Speech Emotion and Automatic Speech Recognition 1863
Lokesh Bansal, S. Pavankumar Dubagunta, Malolan Chetlur, Pushpak Jagtap, Aravind Ganapathiraju

Speaking State Decoder with Transition Detection for Next Speaker Prediction 1868
Shao-Hao Lu, Yun-Shao Lin, Chi-Chun Lee

What Are Differences? Comparing DNN and Human by Their Performance and Characteristics in Speaker Age Estimation 1873
Yuki Kitagishi, Naohiro Tawara, Atsunori Ogawa, Ryo Masumura, Taichi Asami

Effects of Perceived Gender on the Perceived Social Function of Laughter 1878
Joop Arts, Khiet P. Truong

Implicit Phonetic Information Modeling for Speech Emotion Recognition 1883
Tilak Purohit, Bogdan Vlasenko, Mathew Magimai.-Doss

Computation and Memory Efficient Noise Adaptation of Wav2Vec2.0 for Noisy Speech Emotion Recognition with Skip Connection Adapters 1888
Seong-Gyun Leem, Daniel Fulford, Jukka-Pekka Onnela, David Gard, Carlos Busso

Multi-Level Knowledge Distillation for Speech Emotion Recognition in Noisy Conditions 1893
Yang Liu, Haoqin Sun, Geng Chen, Qingyue Wang, Zhen Zhao, Xugang Lu, Longbiao Wang

Preference Learning Labels by Anchoring on Consecutive Annotations 1898
Abinay Reddy Naini, Ali N. Salman, Carlos Busso

Transforming the Embeddings: A Lightweight Technique for Speech Emotion Recognition Tasks 1903
Orchid Chetia Phukan, Arun Balaji Buduru, Rajesh Sharma

Learning Local to Global Feature Aggregation for Speech Emotion Recognition 1908
Cheng Lu, Hailun Lian, Wenming Zheng, Yuan Zong, Yan Zhao, Sunan Li

Supervised Contrastive Learning with Nearest Neighbor Search for Speech Emotion Recognition 1913
Xuechen Wang, Shiwan Zhao, Yong Qin

SPEAKER AND LANGUAGE IDENTIFICATION 1

Vietnam-Celeb: A Large-Scale Dataset for Vietnamese Speaker Recognition 1918
Viet Thanh Pham, Xuan Thai Hoa Nguyen, Vu Hoang, Thi Thu Trang Nguyen

What Can an Accent Identifier Learn? Probing Phonetic and Prosodic Information in a Wav2vec2-Based Accent Identification Model 1923
Mu Yang, Ram C. M. C. Shekar, Okim Kang, John H. L. Hansen

The 2022 NIST Language Recognition Evaluation 1928
Yooyoung Lee, Craig Greenberg, Eliot Godard, Asad A. Butt, Elliot Singer, Trang Nguyen, Lisa Mason, Douglas Reynolds

Description and Analysis of the KPT System for NIST Language Recognition Evaluation 2022 1933
Salvatore Sarni, Sandro Cumani, Sabato Marco Siniscalchi, Andrea Bottino

ACA-Net: Towards Lightweight Speaker Verification Using Asymmetric Cross Attention 1938
Jia Qi Yip, Duc-Tuan Truong, Dianwen Ng, Chong Zhang, Yukun Ma, Trung Hieu Nguyen, Chongjia Ni, Shengkui Zhao, Eng Siong Chng, Bin Ma

Branch-ECAPA-TDNN: A Parallel Branch Architecture to Capture Local and Global Features for Speaker Verification 1943
Jiadi Yao, Chengdong Liang, Zhendong Peng, Binbin Zhang, Xiao-Lei Zhang

Speaker Verification Across Ages: Investigating Deep Speaker Embedding Sensitivity to Age Mismatch in Enrollment and Test Speech 1948
Vishwanath Pratap Singh, Md Sahidullah, Tomi Kinnunen

Wavelet Scattering Transform for Improving Generalization in Low-Resourced Spoken Language Identification 1953
Spandan Dey, Premjeet Singh, Goutam Saha

A Parameter-Efficient Learning Approach to Arabic Dialect Identification with Pre-Trained General-Purpose Speech Model 1958
Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Narsis A. Kiani, David Gomez-Cabrero, Jesper N. Tegner

HABLA: A Dataset of Latin American Spanish Accents for Voice Anti-Spoofing 1963
Pablo Andrés Tamayo Flórez, Rubén Manrique, Bernardo Pereira Nunes

Self-Supervised Learning Representation Based Accent Recognition with Persistent Accent Memory 1968
Rui Li, Zhiwei Xie, Haihua Xu, Yizhou Peng, Hexin Liu, Hao Huang, Eng Siong Chng

Extremely Low Bit Quantization for Mobile Speaker Verification Systems Under 1MB Memory 1973
Bei Liu, Haoyu Wang, Yanmin Qian

Unsupervised Out-Of-Distribution Dialect Detection with Mahalanobis Distance 1978
Sourya Dipta Das, Yash Vadi, Abhishek Unnam, Kuldeep Yadav

Pyannote.audio 2.1 Speaker Diarization Pipeline: Principle, Benchmark, and Recipe 1983
Hervé Bredin

Model Compression for DNN-Based Speaker Verification Using Weight Quantization 1988
Jingyu Li, Wei Liu, Zhaoyang Zhang, Jiong Wang, Tan Lee

Multi-Resolution Approach to Identification of Spoken Languages and to Improve Overall Language Diarization System Using Whisper Model 1993
Bhavik Vachhani, Dipesh Singh, Rustom Lawyer

Improving Generalization Ability of Countermeasures for New Mismatch Scenario by Combining Multiple Advanced Regularization Terms 1998
Chang Zeng, Xin Wang, Xiaoxiao Miao, Erica Cooper, Junichi Yamagishi

Dynamic Fully-Connected Layer for Large-Scale Speaker Verification 2003
Zhida Song, Liang He, Baowei Zhao, Minqiang Xu, Yu Zheng

SHOW AND TELL: SPEECH TOOLS, SPEECH ENHANCEMENT, SPEECH SYNTHESIS

DeepFilterNet: Perceptually Motivated Real-Time Speech Enhancement 2008
Hendrik Schröter, Alberto N. Escalante-B., Tobias Rosenkranz, Andreas Maier

Nkululeko: Machine Learning Experiments on Speaker Characteristics Without Programming 2010
Felix Burkhardt, Florian Eyben, Björn W. Schuller

Sp1NY: A Quick and Flexible Speech Visualisation Tool in Python 2012
Sébastien Le Maguer, Mark Anderson, Naomi Harte

Intonation Control for Neural Text-To-Speech Synthesis with Polynomial Models of F0 2014
Niamh Corkey, Johannah O'Mahony, Simon King

So-To-Speak: An Exploratory Platform for Investigating the Interplay Between Style and Prosody in TTS 2016
Éva Székely, Siyang Wang, Joakim Gustafson

Comparing /b/ and /d/ with a Single Physical Model of the Human Vocal Tract to Visualize Droplets Produced While Speaking 2018
Takayuki Arai, Tsukasa Yoshinaga, Akiyoshi Iida

Show & Tell: Voice Activity Projection and Turn-Taking 2020
Erik Ekstedt, Gabriel Skantze

Real Time Detection of Soft Voice for Speech Enhancement 2022
Hector A. Cordourier, Georg Stemmer, Sinem Aslan, Tobias Bocklet, Himanshu Bhalla

Data Augmentation for Diverse Voice Conversion in Noisy Environments 2024
Avani Tanna, Michael Saxon, Amr El Abbadi, William Yang Wang

Application for Real-Time Audio-Visual Speech Enhancement 2026
Mandar Gogate, Kia Dashtipour, Amir Hussain

SPEECH SYNTHESIS AND VOICE CONVERSION

Mitigating the Exposure Bias in Sentence-Level Grapheme-To-Phoneme (G2P) Transduction 2028
Eunseop Yoon, Hee Suk Yoon, Dhananjaya Gowda, Soohwan Eom, Daehyeok Kim, John Harvill, Heting Gao, Mark Hasegawa-Johnson, Chanwoo Kim, Chang D. Yoo

Streaming Parrotron for On-Device Speech-To-Speech Conversion 2033
Oleg Rybakov, Fadi Biadsy, Xia Zhang, Liyang Jiang, Phoenix Meadowlark, Shivani Agrawal

Exploiting Emotion Information in Speaker Embeddings for Expressive Text-To-Speech 2038
Zein Shaheen, Tasnima Sadekova, Yulia Matveeva, Alexandra Shirshova, Mikhail Kudinov

E2E-S2S-VC: End-To-End Sequence-To-Sequence Voice Conversion 2043
Takuma Okamoto, Tomoki Toda, Hisashi Kawai

DC CoMix TTS: An End-To-End Expressive TTS with Discrete Code Collaborated with Mixer 2048
Yerin Choi, Myoung-Wan Koo

Voice Conversion with Just Nearest Neighbors 2053
Matthew Baas, Benjamin Van Niekerk, Herman Kamper

CFVC: Conditional Filtering for Controllable Voice Conversion 2058
Kou Tanaka, Takuhiro Kaneko, Hirokazu Kameoka, Shogo Seki

DualVC: Dual-Mode Voice Conversion Using Intra-Model Knowledge Distillation and Hybrid Predictive Coding 2063
Ziqian Ning, Yuepeng Jiang, Pengcheng Zhu, Jixun Yao, Shuai Wang, Lei Xie, Mengxiao Bi

Attention-Based Interactive Disentangling Network for Instance-Level Emotional Voice Conversion 2068
Yun Chen, Lingxiao Yang, Qi Chen, Jian-Huang Lai, Xiaohua Xie

ALO-VC: Any-To-Any Low-Latency One-Shot Voice Conversion 2073
Bohan Wang, Damien Ronssin, Milos Cernak

Evaluating and Reducing the Distance Between Synthetic and Real Speech Distributions 2078
Christoph Minixhofer, Ondrej Klejch, Peter Bell

Decoupling Segmental and Prosodic Cues of Non-Native Speech Through Vector Quantization 2083
Waris Quamer, Anurag Das, Ricardo Gutierrez-Osuna

VC-T: Streaming Voice Conversion Based on Neural Transducer 2088
Hiroki Kanagawa, Takafumi Moriya, Yusuke Ijima

VOLUME 4

Emo-StarGAN: A Semi-Supervised Any-To-Many Non-Parallel Emotion-Preserving Voice Conversion 2093
Suhita Ghosh, Arnab Das, Yamini Sinha, Ingo Siegert, Tim Polzehl, Sebastian Stober

ControlVC: Zero-Shot Voice Conversion with Time-Varying Controls on Pitch and Speed 2098
Meiying Chen, Zhiyao Duan

Reverberation-Controllable Voice Conversion Using Reverberation Time Estimator 2103
Yeonjong Choi, Chao Xie, Tomoki Toda

Cross-Utterance Conditioned Coherent Speech Editing 2108
Cheng Yu, Yang Li, Weiqin Zu, Fanglei Sun, Zheng Tian, Jun Wang

SPOKEN LANGUAGE TRANSLATION, INFORMATION RETRIEVAL, SUMMARIZATION, RESOURCES, AND EVALUATION 2

MAVD: The First Open Large-Scale Mandarin Audio-Visual Dataset with Depth Information 2113
Jianrong Wang, Yuchen Huo, Li Liu, Tianyi Xu, Qi Li, Sen Li

CN-Celeb-AV: A Multi-Genre Audio-Visual Dataset for Person Recognition 2118
Lantian Li, Xiaolou Li, Haoyu Jiang, Chen Chen, Ruihai Hou, Dong Wang

Improving Zero-Shot Cross-Domain Slot Filling Via Transformer-Based Slot Semantics Fusion 2123
Yuhang Li, Xiao Wei, Yuke Si, Longbiao Wang, Xiaobao Wang, Jianwu Dang

Rethinking Transfer and Auxiliary Learning for Improving Audio Captioning Transformer..... 2128
Wooseok Shin, Hyun Joon Park, Jin Sob Kim, Dongwon Kim, Seungjin Lee, Sung Won Han

Boosting Punctuation Restoration with Data Generation and Reinforcement Learning 2133
Viet Dac Lai, Abel Salinas, Hao Tan, Trung Bui, Quan Tran, Seunghyun Yoon, Hanieh Deilamsalehy, Franck Dernoncourt, Thien Huu Nguyen

J-ToneNet: A Transformer-Based Encoding Network for Improving Tone Classification in Continuous Speech Via F0 Sequences..... 2138
Yi-Fen Liu, Xiang-Li Lu

Towards Cross-Language Prosody Transfer for Dialog..... 2143
Jonathan E. Avila, Nigel G. Ward

Strategies for Improving Low Resource Speech to Text Translation Relying on Pre-Trained ASR Models..... 2148
Santosh Kesiraju, Marek Sarvaš, Tomáš Pavlícek, Cécile Macaire, Alejandro Ciuba

ITALIC: An Italian Intent Classification Dataset 2153
Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, Elena Baralis

Perceptual and Task-Oriented Assessment of a Semantic Metric for ASR Evaluation..... 2158
Janine Rugayan, Giampiero Salvi, Torbjørn Svendsen

How ChatGPT is Robust for Spoken Language Understanding? 2163
Guangpeng Li, Lu Chen, Kai Yu

GigaST: A 10,000-Hour Pseudo Speech Translation Corpus..... 2168
Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, Jun Cao

Boosting Chinese ASR Error Correction with Dynamic Error Scaling Mechanism..... 2173
Jiaxin Fan, Yong Zhang, Hanzhang Li, Jianzong Wang, Zhitao Li, Sheng Ouyang, Ning Cheng, Jing Xiao

Crowdsource-Based Validation of the Audio Cocktail as a Sound Browsing Tool 2178
Per Fallgren, Jens Edlund

PunCantonese: A Benchmark Corpus for Low-Resource Cantonese Punctuation Restoration from Speech Transcripts..... 2183
Yunxiang Li, Pengfei Liu, Xixin Wu, Helen Meng

Speech-To-Face Conversion Using Denoising Diffusion Probabilistic Models..... 2188
Shuhei Kato, Taiichi Hashimoto

Inter-Connection: Effective Connection Between Pre-Trained Encoder and Decoder for Speech Translation..... 2193
Yuta Nishikawa, Satoshi Nakamura

NOVEL TRANSFORMER MODELS FOR ASR

Conmer: Streaming Conformer Without Self-Attention for Interactive Voice Assistants 2198
Martin Radfar, Paulina Lyskawa, Brandon Trujillo, Yi Xie, Kai Zhen, Jahn Heymann, Denis Filimonov, Grant P. Strimel, Nathan Susanj, Athanasios Mouchtaris

Intra-Ensemble: A New Method for Combining Intermediate Outputs in Transformer-Based Automatic Speech Recognition 2203
Dohee Kim, Jieun Choi, Joon-Hyuk Chang

A Comparative Study on E-Branchformer Vs Conformer in Speech Recognition, Translation, and Understanding Tasks..... 2208
Yifan Peng, Kwangyoun Kim, Felix Wu, Brian Yan, Siddhant Arora, William Chen, Jiyang Tang, Suwon Shon, Prashant Sridhar, Shinji Watanabe

HyperConformer: Multi-Head HyperMixer for Efficient Speech Recognition 2213
Florian Mai, Juan Zuluaga-Gomez, Titouan Parcollet, Petr Motlicek

Memory-Augmented Conformer for Improved End-To-End Long-Form ASR..... 2218
Carlos Carvalho, Alberto Abad

Towards Effective and Compact Contextual Representation for Conformer Transducer Speech Recognition Systems 2223
Mingyu Cui, Jiawen Kang, Jiajun Deng, Xi Yin, Yutao Xie, Xie Chen, Xunying Liu

SPEAKER RECOGNITION 1

An Enhanced Res2Net with Local and Global Feature Fusion for Speaker Verification 2228
Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Jiajun Qi

A Study on Visualization of Voiceprint Feature..... 2233
Jian Zhang, Liang He, Xiaochen Guo, Jing Ma

VoxTube: A Multilingual Speaker Recognition Dataset 2238
Ivan Yakovlev, Anton Okhotnikov, Nikita Torgashov, Rostislav Makarov, Yuri Voevodin, Konstantin Simonchik

Visualizing Data Augmentation in Deep Speaker Recognition 2243
Pengqi Li, Lantian Li, Askar Hamdulla, Dong Wang

CROSS-LINGUAL AND MULTILINGUAL ASR

Fast and Efficient Multilingual Self-Supervised Pre-Training for Low-Resource Speech Recognition 2248
Zhilong Zhang, Wei Wang, Yanmin Qian

UniSplice: Universal Cross-Lingual Data Splicing for Low-Resource ASR..... 2253
Wei Wang, Yanmin Qian

Allophant: Cross-Lingual Phoneme Recognition with Articulatory Attributes 2258
Kevin Glocker, Aaricia Herygers, Munir Georges

Phonetic-Assisted Multi-Target Units Modeling for Improving Conformer-Transducer ASR System 2263
Li Li, Dongxing Xu, Haoran Wei, Yanhua Long

Comparison of Multilingual Self-Supervised and Weakly-Supervised Speech Pre-Training for Adaptation to Unseen Languages 2268
Andrew Rouditchenko, Sameer Khurana, Samuel Thomas, Rogerio Feris, Leonid Karlinsky, Hilde Kuehne, David Harwath, Brian Kingsbury, James Glass

DistilXLSR: A Light Weight Cross-Lingual Speech Representation Model 2273
Haoyu Wang, Siyuan Wang, Wei-Qiang Zhang, Jinfeng Bai

VOICE CONVERSION

Emotional Voice Conversion with Semi-Supervised Generative Modeling 2278
Hai Zhu, Huayi Zhan, Hong Cheng, Ying Wu

Diff-HierVC: Diffusion-Based Hierarchical Voice Conversion with Robust Pitch Generation and Masked Prior for Zero-Shot Speaker Adaptation 2283
Ha-Yeong Choi, Sang-Hoon Lee, Seong-Whan Lee

S2CD: Self-Heuristic Speaker Content Disentanglement for Any-To-Any Voice Conversion..... 2288
Pengfei Wei, Xiang Yin, Chunfeng Wang, Zhonghao Li, Xinghua Qu, Zhiqiang Xu, Zejun Ma

Flow-VAE VC: End-To-End Flow Framework with Contrastive Loss for Zero-Shot Voice Conversion..... 2293
Le Xu, Rongxiu Zhong, Ying Liu, Huibao Yang, Shilei Zhang

Automatic Speech Disentanglement for Voice Conversion Using Rank Module and Speech Augmentation 2298
Zhonghua Liu, Shijun Wang, Ning Chen

End-To-End Zero-Shot Voice Conversion with Location-Variable Convolutions 2303
Wonjune Kang, Mark Hasegawa-Johnson, Deb Roy

SPEECH AND LANGUAGE IN HEALTH: FROM REMOTE MONITORING TO MEDICAL CONVERSATIONS 2

Classifying Dementia in the Presence of Depression: A Cross-Corpus Study..... 2308
Franziska Braun, Sebastian P. Bayerl, Paula A. Pérez-Toro, Florian Hönig, Hartmut Lehfeld, Thomas Hillemacher, Elmar Nöth, Tobias Bocklet, Korbinian Riedhammer

Exploiting Cross-Domain and Cross-Lingual Ultrasound Tongue Imaging Features for Elderly and Dysarthric Speech Recognition 2313
Shujie Hu, Xurong Xie, Mengzhe Geng, Mingyu Cui, Jiajun Deng, Guinan Li, Tianzi Wang, Helen Meng, Xunying Liu

Multi-Class Detection of Pathological Speech with Latent Features: How Does it Perform on Unseen Data?..... 2318
Dominik Wagner, Ilja Baumann, Franziska Braun, Sebastian P. Bayerl, Elmar Nöth, Korbinian Riedhammer, Tobias Bocklet

Responsiveness, Sensitivity and Clinical Utility of Timing-Related Speech Biomarkers for Remote Monitoring of ALS Disease Progression 2323
Hardik Kothare, Michael Neumann, Jackson Liscombe, Jordan Green, Vikram Ramanarayanan

Use of Speech Impairment Severity for Dysarthric Speech Recognition 2328
Mengzhe Geng, Zengrui Jin, Tianzi Wang, Shujie Hu, Jiajun Deng, Mingyu Cui, Guinan Li, Jianwei Yu, Xurong Xie, Xunying Liu

MMLung: Moving Closer to Practical Lung Health Estimation Using Smartphones 2333
Mohammed Mosuily, Lindsay Welch, Jagmohan Chauhan

Investigating the Utility of Synthetic Data for Doctor-Patient Conversation Summarization 2338
Siyuan Chen, Colin A. Grambow, Mojtaba Kadkhodaie Elyaderani, Alireza Sadeghi, Federico Fancellu, Thomas Schaaf

Non-Uniform Speaker Disentanglement for Depression Detection from Raw Speech Signals 2343
Jinhan Wang, Vijay Ravi, Abeer Alwan

PoCaPNet: A Novel Approach for Surgical Phase Recognition Using Speech and X-Ray Images..... 2348
Kubilay Can Demir, Tobias Weise, Matthias May, Axel Schmid, Andreas Maier, Seung Hee Yang

Combining Multiple Multimodal Speech Features into an Interpretable Index Score for Capturing Disease Progression in Amyotrophic Lateral Sclerosis 2353
Michael Neumann, Hardik Kothare, Vikram Ramanarayanan

The MASCFLICHT Corpus: Face Mask Type and Coverage Area Recognition from Speech 2358
Adria Mallol-Ragolta, Nils Urbach, Shuo Liu, Anton Batliner, Björn W. Schuller

Towards Reference Speech Characterization for Health Applications 2363
Catarina Botelho, Alberto Abad, Tanja Schultz, Isabel Trancoso

Automatic Classification of Hypokinetic and Hyperkinetic Dysarthria Based on GMM-Supervectors 2368
Cristian David Ríos-Urrego, Jan Rusz, Elmar Nöth, Juan Rafael Orozco-Arroyave

Towards Robust Paralinguistic Assessment for Real-World Mobile Health (mHealth) Monitoring: An Initial Study of Reverberation Effects on Speech 2373
Judith Dineley, Ewan Carr, Faith Matcham, Johnny Downs, Richard J. B. Dobson, Thomas F. Quatieri, Nicholas Cummins

PATHOLOGICAL SPEECH ANALYSIS 1

Multimodal Assessment of Bulbar Amyotrophic Lateral Sclerosis (ALS) Using a Novel Remote Speech Assessment App 2378
Leif Simmatis, Timothy Pommeé, Yana Yunusova

On the Use of High Frequency Information for Voice Pathology Classification..... 2383
David Martínez, Dayana Ribas, Eduardo Lleida

Do Phonatory Features Display Robustness to Characterize Parkinsonian Speech Across Corpora?..... 2388
Anna Favaro, Tianyu Cao, Thomas Thebaud, Jesus Villalba, Ankur Butala, Najim Dehak, Laureano Moro-Velazquez

Severity Classification of Parkinson's Disease from Speech Using Single Frequency Filtering-Based Features..... 2393
Sudarsana Reddy Kadiri, Manila Kodali, Paavo Alku

Comparison of Acoustic Measures of Dysphonia in Parkinson's Disease and Huntington's Disease: Effect of Sex and Speaking Task 2398
Michal Šimek, Tomáš Kouba, Michal Novotný, Tereza Tykalová, Jan Rusz

Alzheimer Disease Classification Through ASR-Based Transcriptions: Exploring the Impact of Punctuation and Pauses 2403
Lucía Gómez-Zaragozá, Simone Wills, Cristian Tejedor-Garcia, Javier Marín-Morales, Mariano Alcañiz, Helmer Strik

MULTIMODAL SPEECH EMOTION RECOGNITION

LanSER: Language-Model Supported Speech Emotion Recognition 2408
Taesik Gong, Josh Belanich, Krishna Somandepalli, Arsha Nagrani, Brian Eoff, Brendan Jou

Fine-Tuned RoBERTa Model with a CNN-LSTM Network for Conversational Emotion Recognition 2413
Jiachen Luo, Huy Phan, Joshua Reiss

Emotion Label Encoding Using Word Embeddings for Speech Emotion Recognition 2418
Eimear Stanley, Eric Demattos, Anita Klementiev, Piotr Ozimek, Georgia Clarke, Michael Berger, Dimitri Palaz

Discrimination of the Different Intents Carried by the Same Text Through Integrating Multimodal Information 2423
Zhongjie Li, Gaoyan Zhang, Longbiao Wang, Jianwu Dang

Meta-Domain Adversarial Contrastive Learning for Alleviating Individual Bias in Self-Sentiment Predictions 2428
Zhi Li, Ryu Takeda, Takahiro Hara

SWRR: Feature Map Classifier Based on Sliding Window Attention and High-Response Feature Reuse for Multimodal Emotion Recognition 2433
Ziping Zhao, Tian Gao, Haishuai Wang, Björn W. Schuller

SPEECH CODING AND ENHANCEMENT 2

PCNN: A Lightweight Parallel Conformer Neural Network for Efficient Monaural Speech Enhancement 2438
Xinmeng Xu, Weiping Tu, Yuhong Yang

Exploring the Interactions Between Target Positive and Negative Information for Acoustic Echo Cancellation 2443
Chang Han, Xinmeng Xu, Weiping Tu, Yuhong Yang, Yajie Liu

Iterative Autoregression: A Novel Trick to Improve Your Low-Latency Speech Enhancement Model 2448
Pavel Andreev, Nicholas Babaev, Azat Saginbaev, Ivan Shchekotov, Aibek Alanov

A Multi-Dimensional Deep Structured State Space Approach to Speech Enhancement Using Small-Footprint Models 2453
Pin-Jui Ku, Chao-Han Huck Yang, Sabato Siniscalchi, Chin-Hui Lee

Domain Adaptation for Speech Enhancement in a Large Domain Gap 2458
Lior Frenkel, Jacob Goldberger, Shlomo E. Chazan

SCP-GAN: Self-Correcting Discriminator Optimization for Training Consistency Preserving Metric GAN on Speech Enhancement Tasks 2463
Vasily Zadorozhnyy, Qiang Ye, Kazuhito Koishida

A Mask Free Neural Network for Monaural Speech Enhancement 2468
Liang Liu, Haixin Guan, Jinlong Ma, Wei Dai, Guangyong Wang, Shaowei Ding

A Training and Inference Strategy Using Noisy and Enhanced Speech as Target for Speech Enhancement Without Clean Speech 2473
Li-Wei Chen, Yao-Fei Cheng, Hung-Shin Lee, Yu Tsao, Hsin-Min Wang

A Simple RNN Model for Lightweight, Low-Compute and Low-Latency Multichannel Speech Enhancement in the Time Domain 2478
Ashutosh Pandey, Ke Tan, Buye Xu

High Fidelity Speech Enhancement with Band-Split RNN 2483
Jianwei Yu, Hangting Chen, Yi Luo, Rongzhi Gu, Chao Weng

Focus on the Sound Around You: Monaural Target Speaker Extraction Via Distance and Speaker Information 2488
Jiuxin Lin, Peng Wang, Heinrich Dinkel, Jun Chen, Zhiyong Wu, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang

DFSNet: A Steerable Neural Beamformer Invariant to Microphone Array Configuration for Real-Time, Low-Latency Speech Enhancement 2493
Anton Kovalyov, Kashyap Patel, Issa Panahi

Speaker-Aware Anti-Spoofing 2498
Xuechen Liu, Md Sahidullah, Kong Aik Lee, Tomi Kinnunen

Impact of Residual Noise and Artifacts in Speech Enhancement Errors on Intelligibility of Human and Machine 2503
Shoko Araki, Ayako Yamamoto, Tsubasa Ochiai, Kenichi Arai, Atsunori Ogawa, Tomohiro Nakatani, Toshio Irino

EffCRN: An Efficient Convolutional Recurrent Network for High-Performance Speech Enhancement 2508
Marvin Sach, Jan Franzen, Bruno Defraene, Kristoff Fluyt, Maximilian Strake, Wouter Tirry, Tim Fingscheidt

HAD-ANC: A Hybrid System Comprising an Adaptive Filter and Deep Neural Networks for Active Noise Control 2513
Jungphil Park, Jeong-Hwan Choi, Yungyeo Kim, Joon-Hyuk Chang

MSAF: A Multiple Self-Attention Field Method for Speech Enhancement 2518
Minghang Chu, Jing Wang, Yaoyao Ma, Zhiwei Fan, Mengtao Yang, Chao Xu, Zhi Tao, Di Wu

Ultra Dual-Path Compression for Joint Echo Cancellation and Noise Suppression 2523
Hangting Chen, Jianwei Yu, Yi Luo, Rongzhi Gu, Weihua Li, Zhuocheng Lu, Chao Weng

ABC-KD: Attention-Based-Compression Knowledge Distillation for Deep Learning-Based Noise Suppression 2528
Yixin Wan, Yuan Zhou, Xiulian Peng, Kai-Wei Chang, Yan Lu

PLCMOS – a Data-Driven Non-Intrusive Metric for the Evaluation of Packet Loss Concealment Algorithms 2533
Lorenz Diener, Marju Purin, Sten Sootla, Ando Saabas, Robert Aichner, Ross Cutler

PHONETICS, PHONOLOGY, AND PROSODY 1

Effects of Meter, Genre and Experience on Pausing, Lengthening and Prosodic Phrasing in German Poetry Reading 2538
Petra Wagner, Simon Betz

Comparing First Spectral Moment of Australian English /s/ Between Straight and Gay Voices Using Three Analysis Window Sizes 2543
Tünde Szalay, John Holik, Duy Duong Nguyen, James Morandini, Catherine J. Madill

Universal Automatic Phonetic Transcription into the International Phonetic Alphabet 2548
Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, David Chiang

Voice Twins: Discovering Extremely Similar-Sounding, Unrelated Speakers 2553
Linda Gerlach, Kirsty McDougall, Finnian Kelly, Anil Alexander

Filling the Population Statistics Gap: Swiss German Reference Data on F0 and Speech Tempo for Forensic Contexts 2558
Hannah Hedegard, Andrea Fröhlich, Fabian Tomaschek, Carina Steiner, Adrian Leemann

Investigating the Syntax-Discourse Interface in the Phonetic Implementation of Discourse Markers 2563
Mathilde Hutin, Liesbeth Degand, Marc Allassonnière-Tang

Evaluation of a Forensic Automatic Speaker Recognition System with Emotional Speech Recordings 2568
Robert Essery, Philip Harrison, Vincent Hughes

An Outlier Analysis of Vowel Formants from a Corpus Phonetics Pipeline 2573
Emily P. Ahn, Gina-Anne Levow, Richard A. Wright, Eleanor Chodroff

The Hidden Dance of Phonemes and Visage: Unveiling the Enigmatic Link Between Phonemes and Facial Features 2578
Liao Qu, Xianwei Zou, Xiang Li, Yandong Wen, Rita Singh, Bhiksha Raj

Beatboxing Kick Drum Kinematics 2583
Reed Blaylock, Shrikanth Narayanan

Effects of Hearing Loss and Amplification on Mandarin Consonant Perception 2588
Huali Zhou, Xianming Bei, Nengheng Zheng, Qinglin Meng

An Acoustic Analysis of Fricative Variation in Three Accents of English 2593
Roland Adams, Calbert Graham

Acoustic Cues to Stress Perception in Spanish – a Mismatch Negativity Study 2598
Karolina Bros

Bulgarian Unstressed Vowel Reduction: Received Views Vs Corpus Findings 2603
Mitko Sabev, Bistra Andreeva, Christoph Gabriel, Jonas Gruenke

An Investigation of Indian Native Language Phonemic Influences on L2 English Pronunciations 2608
Shelly Jain, Priyanshi Pal, Anil Kumar Vuppala, Prasanta Kumar Ghosh, Chiranjeevi Yarra

Identifying Stable Sections for Formant Frequency Extraction of French Nasal Vowels Based on Difference Thresholds 2613
Hye-Sook Park, Sunhee Kim

Evaluation of Delexicalization Methods for Research on Emotional Speech 2618
Nicolas Audibert, Francesca Carbone, Maud Champagne-Lavau, Aurélien Said Housseini, Caterina Petrone

SPOKEN DIALOG SYSTEMS AND CONVERSATIONAL ANALYSIS 2

Relationship Between Auditory and Semantic Entrainment Using Deep Neural Networks (DNN) 2623
Jay Kejriwal, Štefan Benuš

Unsupervised Auditory and Semantic Entrainment Models with Deep Neural Networks 2628
Jay Kejriwal, Štefan Benuš, Lina M. Rojas-Barahona

Prosodic Features Improve Sentence Segmentation and Parsing in English 2633
Elizabeth Nielsen, Mark Steedman, Sharon Goldwater

Estimation of Listening Response Timing by Generative Model and Parameter Control of Response Substantialness Using Dynamic-Prompt-Tune 2638
Toshiki Muromachi, Yoshinobu Kano

Parameter Selection for Analyzing Conversations with Autism Spectrum Disorder 2643
Tahiya Chowdhury, Veronica Romero, Amanda Stent

Efficient Multimodal Neural Networks for Trigger-Less Voice Assistants 2648
Sai Srujana Buddi, Utkarsh Oggy Sarawgi, Tashweena Heeramun, Karan Sawnhey, Ed Yanosik, Saravana Rathinam, Saurabh Adya

Rapid Lexical Alignment to a Conversational Agent 2653
Rachel Ostrand, Victor S. Ferreira, David Piorkowski

Multimodal Turn-Taking Model Using Visual Cues for End-Of-Utterance Prediction in Spoken Dialogue Systems 2658
Fuma Kurata, Mao Saeki, Shinya Fujie, Yoichi Matsuyama

Audio-Visual Praise Estimation for Conversational Video Based on Synchronization-Guided Multimodal Transformer 2663
Nobukatsu Hojo, Saki Mizuno, Satoshi Kobashikawa, Ryo Masumura, Mana Ihori, Hiroshi Sato, Tomohiro Tanaka

Improving the Response Timing Estimation for Spoken Dialogue Systems by Reducing the Effect of Speech Recognition Delay 2668
Jin Sakuma, Shinya Fujie, Huaibo Zhao, Tetsunori Kobayashi

Focus-Attention-Enhanced Crossmodal Transformer with Metric Learning for Multimodal Speech Emotion Recognition 2673
Keulbit Kim, Namhyun Cho

A Multiple-Teacher Pruning Based Self-Distillation (MT-PSD) Approach to Model Compression for Audio-Visual Wake Word Spotting 2678
Haotian Wang, Jun Du, Hengshun Zhou, Chin-Hui Lee, Yuling Ren, Jiangjiang Zhao

Abusive Speech Detection in Indic Languages Using Acoustic Features 2683
Anika A. Spiesberger, Andreas Triantafyllopoulos, Iosif Tsangko, Björn W. Schuller

Listening to Silences in Contact Center Conversations Using Textual Cues 2688
Digvijay Anil Ingle, Ayush Kumar, Jithendra Vepa

I Learned Error, I Can Fix It! : A Detector-Corrector Structure for ASR Error Calibration 2693
Heui-Yeen Yeen, Min-Ju Kim, Myoung-Wan Koo

Verbal and Nonverbal Feedback Signals in Response to Increasing Levels of Miscommunication 2698
Maeva Garnier, Eric Le Ferrand, Fabien Ringeval

Speech-Based Classification of Defensive Communication: A Novel Dataset and Results 2703
Shahin Amiriparian, Lukas Christ, Regina Kushtanova, Maurice Gerczuk, Alexandra Teynor, Björn W. Schuller

Quantifying the Perceptual Value of Lexical and Non-Lexical Channels in Speech 2708
Sarenne Wallbridge, Peter Bell, Catherine Lai

Relationships Between Gender, Personality Traits and Features of Multi-Modal Data to Responses to Spoken Dialog Systems Breakdown 2713
Kazuya Tsubokura, Yurie Iribe, Norihide Kitaoka

Speaker-Aware Cross-Modal Fusion Architecture for Conversational Emotion Recognition 2718
Huan Zhao, Bo Li, Zixing Zhang

ANALYSIS OF SPEECH AND AUDIO SIGNALS 2

Blind Estimation of Room Impulse Response from Monaural Reverberant Speech with Segmental Generative Neural Network 2723
Zhiheng Liao, Feifei Xiong, Juan Luo, Minjie Cai, Eng Siong Chng, Jinwei Feng, Xionghu Zhong

Emotion-Aware Audio-Driven Face Animation Via Contrastive Feature Disentanglement 2728
Xin Ren, Juan Luo, Xionghu Zhong, Minjie Cai

Anomalous Sound Detection Based on Sound Separation 2733
Kanta Shimonishi, Kota Dohi, Yohei Kawaguchi

Random Forest Classification of Breathing Phases from Audio Signals Recorded Using Mobile Devices 2738
Vitória S. Fahed, Emer P Doheny, Madeleine M Lowery

GRAVO: Learning to Generate Relevant Audio from Visual Features with Noisy Online Videos 2743
Youngdo Ahn, Chengyi Wang, Yu Wu, Jong Won Shin, Shujie Liu

Wav2ToBI: A New Approach to Automatic ToBI Transcription 2748
Wanyue Zhai, Mark Hasegawa-Johnson

Joint-Former: Jointly Regularized and Locally Down-Sampled Conformer for Semi-Supervised Sound Event Detection 2753
Lijian Gao, Qirong Mao, Ming Dong

Towards Attention-Based Contrastive Learning for Audio Spoof Detection 2758
Chirag Goel, Surya Koppisetti, Ben Colman, Ali Shahriyari, Gaurav Bharaj

Masked Audio Modeling with CLAP and Multi-Objective Learning 2763
Yifei Xin, Xiulian Peng, Yan Lu

Few-Shot Open-Set Learning for On-Device Customization of KeyWord Spotting Systems 2768
Manuele Rusci, Tinne Tuytelaars

Self-Supervised Dataset Pruning for Efficient Training in Audio Anti-Spoofing 2773
Abdul Hameed Azeemi, Ihsan Ayyub Qazi, Agha Ali Raza

Semantic Segmentation with Bidirectional Language Models Improves Long-Form ASR 2778
W. Ronny Huang, Hao Zhang, Shankar Kumar, Shuo-Yiin Chang, Tara Sainath

Multi-Microphone Automatic Speech Segmentation in Meetings Based on Circular Harmonics Features 2783
Théo Mariotte, Anthony Larcher, Silvio Montrésor, Jean-Hugh Thomas

VOLUME 5

Advanced RawNet2 with Attention-Based Channel Masking for Synthetic Speech Detection 2788
Jing Li, Yanhua Long, Yijie Li, Dongxing Xu

Insights into End-To-End Audio-To-Score Transcription with Real Recordings: A Case Study with Saxophone Works 2793
Juan Carlos Martínez-Sevilla, María Alfaro-Contreras, Jose J. Valero-Mas, Jorge Calvo-Zaragoza

Whisper-AT: Noise-Robust Automatic Speech Recognizers Are Also Strong General Audio Event Taggers 2798
Yuan Gong, Sameer Khurana, Leonid Karlinsky, James Glass

Synthetic Voice Spoofing Detection Based on Feature Pyramid Conformer 2803
Jingran Gong, Ning Chen

Learning a Self-Supervised Domain-Invariant Feature Representation for Generalized Audio Deepfake Detection 2808
Yuankun Xie, Haonan Cheng, Yutian Wang, Long Ye

Application of Knowledge Distillation to Multi-Task Speech Representation Learning 2813
Mine Kerpicci, Van Nguyen, Shuhua Zhang, Erik Visser

DeCoR: Defy Knowledge Forgetting by Predicting Earlier Audio Codes 2818
Xilin Jiang, Yinghao Aaron Li, Nima Mesgarani

Variational Classifier for Unsupervised Anomalous Sound Detection Under Domain Generalization 2823
Antonio Almudévar, Alfonso Ortega, Luis Vicente, Antonio Miguel, Eduardo Lleida

FlexiAST: Flexibility is What AST Needs 2828
Jiu Feng, Mehmet Hamza Erol, Joon Son Chung, Arda Senocak

MCR-Data2vec 2.0: Improving Self-Supervised Speech Pre-Training Via Model-Level Consistency Regularization 2833
Ji Won Yoon, Seok Min Kim, Nam Soo Kim

Visually-Aware Audio Captioning with Adaptive Audio-Visual Attention 2838
Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H. Tang, Mark D. Plumbley, Volkan Kiliç, Wenwu Wang

SPEECH CODING: PRIVACY

Masking Kernel for Learning Energy-Efficient Representations for Speaker Recognition and Mobile Health 2843
Apiwat Ditthapron, Emmanuel O. Agu, Adam C. Lammert

ESTImate: A Real-Time Speech Transmission Index Estimator with Speech Enhancement Auxiliary Task Using Self-Attention Feature Pyramid Network 2848
Bajian Xiang, Hongkun Liu, Zedong Wu, Su Shen, Xiangdong Zhang

Efficient Encoder-Decoder and Dual-Path Conformer for Comprehensive Feature Learning in Speech Enhancement 2853
Junyu Wang

Privacy-Preserving Representation Learning for Speech Understanding 2858
Minh Tran, Mohammad Soleymani

Vocoder Drift in X-Vector–based Speaker Anonymization 2863
Michele Panariello, Massimiliano Todisco, Nicholas Evans

Malafide: A Novel Adversarial Convolutive Noise Attack Against Deepfake and Spoofing Detection Systems 2868
Michele Panariello, Wanying Ge, Hemlata Tak, Massimiliano Todisco, Nicholas Evans

ANALYSIS OF NEURAL SPEECH REPRESENTATIONS

Speech Self-Supervised Representation Benchmarking: Are We Doing it Right? 2873
Salah Zaiem, Youcef Kemiche, Titouan Parcollet, Slim Essid, Mirco Ravanelli

An Extension of Disentanglement Metrics and Its Application to Voice 2878
Olivier Zhang, Olivier Le Blouch, Nicolas Gengembre, Damien Lolive

An Information-Theoretic Analysis of Self-Supervised Discrete Representations of Speech 2883
Badr M. Abdullah, Mohammed Maqsood Shaik, Bernd Möbius, Dietrich Klakow

SpeechGLUE: How Well Can Self-Supervised Speech Models Capture Linguistic Knowledge? 2888
Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, Tomohiro Tanaka, Yusuke Ijima, Taichi Asami, Marc Delcroix, Yukinori Honma

Comparison of GIF- And SSL-Based Features in Pathological-Voice Detection 2893
Akira Sasou, Yang Chen

What is Learnt by the LEArnable Front-End (LEAF)? Adapting Per-Channel Energy Normalisation (PCEN) to Noisy Conditions 2898
Hanyu Meng, Vidhyasaharan Sethu, Eliathamby Ambikairajah

END-TO-END ASR

End-To-End Joint Target and Non-Target Speakers ASR 2903
Ryo Masumura, Naoki Makishima, Taiga Yamane, Yoshihiko Yamazaki, Saki Mizuno, Mana Ihori, Mihiro Uchida, Keita Suzuki, Hiroshi Sato, Tomohiro Tanaka, Akihiko Takashima, Satoshi Suzuki, Takafumi Moriya, Nobukatsu Hojo, Atsushi Ando

Improving Frame-Level Classifier for Word Timings with Non-Peaky CTC in End-To-End Automatic Speech Recognition 2908
Xianzhao Chen, Yist Y. Lin, Kang Wang, Yi He, Zejun Ma

Joint Autoregressive Modeling of End-To-End Multi-Talker Overlapped Speech Recognition and Utterance-Level Timestamp Prediction 2913
Naoki Makishima, Keita Suzuki, Satoshi Suzuki, Atsushi Ando, Ryo Masumura

Dual-Path Style Learning for End-To-End Noise-Robust Speech Recognition 2918
Yuchen Hu, Nana Hou, Chen Chen, Eng Siong Chng

Multi-Pass Training and Cross-Information Fusion for Low-Resource End-To-End Accented Speech Recognition 2923
Xuefei Wang, Yanhua Long, Yijie Li, Haoran Wei

Text-Only Domain Adaptation for End-To-End ASR Using Integrated Text-To-Mel-Spectrogram Generator 2928
Vladimir Bataev, Roman Korostik, Evgeny Shabalin, Vitaly Lavrukhin, Boris Ginsburg

SPOKEN LANGUAGE UNDERSTANDING, SUMMARIZATION, AND INFORMATION RETRIEVAL

Leveraging Pretrained ASR Encoders for Effective and Efficient End-To-End Speech Intent Classification and Slot Filling 2933
He Huang, Jagadeesh Balam, Boris Ginsburg

Relation-Based Counterfactual Data Augmentation and Contrastive Learning for Robustifying Natural Language Inference Models 2938
Heerin Yang, Seung-Won Hwang, Jungmin So

Transfer Learning from Pre-Trained Language Models Improves End-To-End Speech Summarization 2943
Kohei Matsuura, Takanori Ashihara, Takafumi Moriya, Tomohiro Tanaka, Takatomo Kano, Atsunori Ogawa, Marc Delcroix

Audio Retrieval with WavText5K and CLAP Training 2948
Soham Deshmukh, Benjamin Elizalde, Huaming Wang

Sequence-Level Knowledge Distillation for Class-Incremental End-To-End Spoken Language Understanding 2953
Umberto Cappellazzo, Muqiao Yang, Daniele Falavigna, Alessio Brutti

Contrastive Disentangled Learning for Memory-Augmented Transformer 2958
Jen-Tzung Chien, Shang-En Li

INVARIANT AND ROBUST PRE-TRAINED ACOUSTIC MODELS

ProsAudit, a Prosodic Benchmark for Self-Supervised Speech Models 2963
Maureen De Seyssel, Marvin Lavechin, Hadrien Titeux, Arthur Thomas, Gwendal Virlet, Andrea Santos Revilla, Guillaume Wisniewski, Bogdan Ludusan, Emmanuel Dupoux

Self-Supervised Predictive Coding Models Encode Speaker and Phonetic Information in Orthogonal Subspaces 2968
Oli Danyi Liu, Hao Tang, Sharon Goldwater

Evaluating Context-Invariance in Unsupervised Speech Representations 2973
Mark Hallap, Emmanuel Dupoux, Ewan Dunbar

CoBERT: Self-Supervised Speech Representation Learning Through Code Representation Learning 2978
Chutong Meng, Junyi Ao, Tom Ko, Mingxuan Wang, Haizhou Li

Self-Supervised Fine-Tuning for Improved Content Representations by Speaker-Invariant Clustering 2983
Heng-Jui Chang, Alexander H. Liu, James Glass

Self-Supervised Acoustic Word Embedding Learning Via Correspondence Transformer Encoder 2988
Jingru Lin, Xianghu Yue, Junyi Ao, Haizhou Li

PATHOLOGICAL SPEECH ANALYSIS 2

A Pipeline to Evaluate the Effects of Noise on Machine Learning Detection of Laryngeal Cancer 2993
Mary Paterson, James Moor, Luisa Cutillo

ReCLR: Reference-Enhanced Contrastive Learning of Audio Representation for Depression Detection 2998
Pingyue Zhang, Mengyue Wu, Kai Yu

Automated Multiple Sclerosis Screening Based on Encoded Speech Representations 3003
José Egas-López, Veronika Svindt, Judit Bóna, Ildikó Hoffmann, Gábor Gosztolya

Cross-Lingual Features for Alzheimer's Dementia Detection from Speech 3008
Thomas Melistas, Lefteris Kapelonis, Nikos Antoniou, Petros Mitseas, Dimitris Sgouropoulos, Theodoros Giannakopoulos, Athanasios Katsamanis, Shrikanth Narayanan

Careful Whisper - Leveraging Advances in Automatic Speech Recognition for Robust and Interpretable Aphasia Subtype Classification 3013
Mario Zusag, Laurin Wagner, Theresa Bloder

Behavioral Analysis of Pathological Speaker Embeddings of Patients During Oncological Treatment of Oral Cancer 3018
Jenthe Thienpondt, Caroline M. Speksnijder, Kris Demuynck

SPEECH SYNTHESIS: REPRESENTATION LEARNING

Adversarial Learning of Intermediate Acoustic Feature for End-To-End Lightweight Text-To-Speech 3023
Hyungchan Yoon, Seyun Um, Changhwan Kim, Hong-Goo Kang

Adapter-Based Extension of Multi-Speaker Text-To-Speech Model for New Speakers 3028
Cheng-Ping Hsieh, Subhankar Ghosh, Boris Ginsburg

SALTTS: Leveraging Self-Supervised Speech Representations for Improved Text-To-Speech Synthesis 3033
Ramanan Sivaguru, Vasista Sai Lodagala, S Umesh

UnitSpeech: Speaker-Adaptive Speech Synthesis with Untranscribed Data 3038
Heeseung Kim, Sungwon Kim, Jiheum Yeom, Sungroh Yoon

LightVoc: An Upsampling-Free GAN Vocoder Based on Conformer and Inverse Short-Time Fourier Transform 3043
Dinh Son Dang, Tung Lam Nguyen, Bao Thang Ta, Tien Thanh Nguyen, Thi Ngoc Anh Nguyen, Dang Linh Le, Nhat Minh Le, Van Hai Do

ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-Derived Context Word Embeddings 3048
Yuki Saito, Shinnosuke Takamichi, Eiji Iimori, Kentaro Tachibana, Hiroshi Saruwatari

SPEECH PERCEPTION, PRODUCTION, AND ACQUISITION 1

Human Transcription Quality Improvement 3053
Jian Gao, Hanbo Sun, Cheng Cao, Zheng Du

The Effect of Masking Noise on Listeners' Spectral Tilt Preferences 3058
Olympia Simantiraki, Yannis Pantazis, Martin Cooke

The Effect of Whistled Vowels on Whistled Word Categorization for Naive Listeners 3063
Anais Tran Ngoc, Fanny Meunier, Julien Meyer

Automatic Deep Neural Network-Based Segmental Pronunciation Error Detection of L2 English Speech (L1 Bengali) 3068
Puja Bharati, Sabyasachi Chandra, Shayamal Kumar Das Mandal

The Effect of Stress on Mandarin Tonal Perception in Continuous Speech for Spanish-Speaking Learners 3073
Lixia Hao, Qi Gong, Jinsong Zhang

Combining Acoustic and Aerodynamic Data Collection: A Perceptual Evaluation of Acoustic Distortions 3078
Amélie Elmerich, Jiayin Gao, Angelique Amelot, Lise Crevier-Buchman, Shinji Maeda

Estimating Virtual Targets for Lingual Stop Consonants Using General Tau Theory 3083
Benjamin Elie, Alice Turk

Using Random Forests to Classify Language as a Function of Syllable Timing in Two Groups: Children with Cochlear Implants and with Normal Hearing 3088
Mark Gibson

An Improved End-To-End Audio-Visual Speech Recognition Model 3093
Sheng Yang, Zheng Gong, Jia Kang

What Influences the Foreign Accent Strength? Phonological and Grammatical Errors in the Perception of Accentedness 3098
Sarah Wesolek, Piotr Gulgowski, Joanna Blaszczak, Marzena Zygis

Investigating the Perception Production Link Through Perceptual Adaptation and Phonetic Convergence 3103
Lena-Marie Huttner, Noël Nguyen, Martin J. Pickering

Emotion Prompting for Speech Emotion Recognition 3108
Xingfa Zhou, Min Li, Lan Yang, Rui Sun, Xin Wang, Huayi Zhan

Speech-In-Speech Recognition is Modulated by Familiarity to Dialect 3113
Jessica L. L. Chin, Elena Talevska, Mark Antoniou

BASEN: Time-Domain Brain-Assisted Speech Enhancement Network with Convolutional Cross Attention in Multi-Talker Conditions 3117
Jie Zhang, Qingtian Xu, Qiu-Shi Zhu, Zhen-Hua Ling

Are Retroflex-To-Dental Sibilant Substitutions in Polish Children's Speech an Example of a Covert Contrast? a Preliminary Acoustic Study 3122
Zuzanna Miodonska, Claartje Levelt, Natalia Mocko, Michal Krecichwost, Agata Sage, Pawel Badura

SPEAKER AND LANGUAGE IDENTIFICATION 2

Reversible Neural Networks for Memory-Efficient Speaker Verification 3127
Bei Liu, Yanmin Qian

ECAPA++: Fine-Grained Deep Embedding Learning for TDNN Based Speaker Verification 3132
Bei Liu, Yanmin Qian

TO-Rawnet: Improving RawNet with TCN and Orthogonal Regularization for Fake Audio Detection 3137
Chenglong Wang, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, Shuai Zhang, Ruibo Fu, Xun Chen

Fooling Speaker Identification Systems with Adversarial Background Music 3142
Chu-Xiao Zuo, Jia-Yi Leng, Wu-Jun Li

Mutual Information-Based Embedding Decoupling for Generalizable Speaker Verification 3147
Jianchen Li, Jiqing Han, Shiwen Deng, Tieran Zheng, Yongjun He, Guibin Zheng

Target Active Speaker Detection with Audio-Visual Cues 3152
Yidi Jiang, Ruijie Tao, Zexu Pan, Haizhou Li

Improving End-To-End Neural Diarization Using Conversational Summary Representations 3157
Samuel J. Broughton, Lahiru Samarakoon

Phase Perturbation Improves Channel Robustness for Speech Spoofing Countermeasures 3162
Yongyi Zang, You Zhang, Zhiyao Duan

Improving Training Datasets for Resource-Constrained Speaker Recognition Neural Networks 3167
Pierre-Michel Bousquet, Mickael Rouvier

Instance-Based Temporal Normalization for Speaker Verification 3172
Thanathai Lertpetchpun, Ekapol Chuangsuwanich

On the Robustness of Wav2vec 2.0 Based Speaker Recognition Systems 3177
Sergey Novoselov, Galina Lavrentyeva, Anastasia Avdeeva, Vladimir Volokhov, Nikita Khmelev, Artem Akulov, Polina Leonteva

P-Vectors: A Parallel-Coupled TDNN/Transformer Network for Speaker Verification 3182
Xiyuan Wang, Fangyuan Wang, Bo Xu, Liang Xu, Jing Xiao

Group GMM-ResNet for Detection of Synthetic Speech Attacks 3187
Zhenchun Lei, Yan Wen, Yingen Yang, Changhong Liu, Minglei Ma

Robust Training for Speaker Verification Against Noisy Labels 3192
Zhihua Fang, Liang He, Hanhan Ma, Xiaochen Guo, Lin Li

Self-Distillation into Self-Attention Heads for Improving Transformer-Based End-To-End Neural Speaker Diarization 3197
Ye-Rin Jeoung, Jeong-Hwan Choi, Ju-Seok Seong, Jehyun Kyung, Joon-Hyuk Chang

Build a SRE Challenge System: Lessons from VoxSRC 2022 and CNSRC 2022 3202
Zhengyang Chen, Bing Han, Xu Xiang, Houjun Huang, Bei Liu, Yanmin Qian

Describing the Phonetics in the Underlying Speech Attributes for Deep and Interpretable Speaker Recognition 3207
Imen Ben-Amor, Jean-François Bonastre, Benjamin O'Brien, Pierre-Michel Bousquet

Range-Based Equal Error Rate for Spoof Localization 3212
Lin Zhang, Xin Wang, Erica Cooper, Nicholas Evans, Junichi Yamagishi

Exploring the English Accent-Independent Features for Speech Emotion Recognition Using Filter and Wrapper-Based Methods for Feature Selection 3217
Nowshin Tabassum, Tasfia Tabassum, Fardin Saad, Tahiya Sultana Safa, Hasan Mahmud, Md. Kamrul Hasan

Powerset Multi-Class Cross Entropy Loss for Neural Speaker Diarization 3222
Alexis Plaquet, Hervé Bredin

A Method of Audio-Visual Person Verification by Mining Connections Between Time Series 3227
Peiwen Sun, Shanshan Zhang, Zishan Liu, Yougen Yuan, Taotao Zhang, Honggang Zhang, Pengfei Hu

SPEECH RECOGNITION: ARCHITECTURE, SEARCH, AND LINGUISTIC COMPONENTS 3

A Model for Every User and Budget: Label-Free and Personalized Mixed-Precision Quantization 3232
Edward Fish, Umberto Michieli, Mete Ozay

Modeling Dependent Structure for Utterances in ASR Evaluation 3237
Zhe Liu, Fuchun Peng

ASR for Low Resource and Multilingual Noisy Code-Mixed Speech 3242
Tushar Verma, Atul Shree, Ashutosh Modi

Accurate and Reliable Confidence Estimation Based on Non-Autoregressive End-To-End Speech Recognition System 3247
Xian Shi, Haoneng Luo, Zhifu Gao, Shiliang Zhang, Zhijie Yan

Combining Multilingual Resources and Models to Develop State-Of-The-Art E2E ASR for Swedish 3252
Lukas Mateju, Jan Nouza, Petr Cerva, Jindrich Zdansky, Frantisek Kynych

Two Stage Contextual Word Filtering for Context Bias in Unified Streaming and Non-Streaming Transducer 3257
Zhanheng Yang, Sining Sun, Xiong Wang, Yike Zhang, Long Ma, Lei Xie

Towards Continually Learning New Languages 3262
Ngoc-Quan Pham, Jan Niehues, Alex Waibel

N-Best T5: Robust ASR Error Correction Using Multiple Input Hypotheses and Constrained Decoding Space 3267
Rao Ma, Mark J. F. Gales, Kate M. Knill, Mengjie Qian

SememeASR: Boosting Performance of End-To-End Speech Recognition Against Domain and Long-Tailed Data Shift with Sememe Semantic Knowledge 3272
Jiaxu Zhu, Changhe Song, Zhiyong Wu, Helen Meng

MiniStreamer: Enhancing Small Conformer with Chunked-Context Masking for Streaming ASR Applications on the Edge 3277
Haris Gulzar, Monikka Roslianna Busto, Takeharu Eda, Katsutoshi Itoyama, Kazuhiro Nakadai

CoMFLP: Correlation Measure Based Fast Search on ASR Layer Pruning 3282
Wei Liu, Zhiyuan Peng, Tan Lee

Exploration on HuBERT with Multiple Resolution 3287
Jiatong Shi, Yun Tang, Hirofumi Inaguma, Hongyu Gong, Juan Pino, Shinji Watanabe

Quantization-Aware and Tensor-Compressed Training of Transformers for Natural Language Understanding 3292
Zi Yang, Samridhi Choudhary, Siegfried Kunzmann, Zheng Zhang

Word-Level Confidence Estimation for CTC Models 3297
Burin Naowarat, Thananchai Kongthaworn, Ekapol Chuangsuwanich

Multilingual Contextual Adapters to Improve Custom Word Recognition in Low-Resource Languages 3302
Devang Kulshreshtha, Saket Dingliwal, Brady Houston, Sravan Bodapati

Unsupervised Active Learning: Optimizing Labeling Cost-Effectiveness for Automatic Speech Recognition 3307
Zhisheng Zheng, Ziyang Ma, Yu Wang, Xie Chen

4D ASR: Joint Modeling of CTC, Attention, Transducer, and Mask-Predict Decoders 3312
Yui Sudo, Shakeel Muhammad, Brian Yan, Jiatong Shi, Shinji Watanabe

Neural Model Reprogramming with Similarity Based Mapping for Low-Resource Spoken Command Recognition 3317
Hao Yen, Pin-Jui Ku, Chao-Han Huck Yang, Hu Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, Yu Tsao

Language-Specific Boundary Learning for Improving Mandarin-English Code-Switching Speech Recognition 3322
Zhiyun Fan, Linhao Dong, Chen Shen, Zhenlin Liang, Jun Zhang, Lu Lu, Zejun Ma

Mixture-Of-Expert Conformer for Streaming Multilingual ASR 3327
Ke Hu, Bo Li, Tara Sainath, Yu Zhang, Françoise Beaufays

Lossless 4-Bit Quantization of Architecture Compressed Conformer ASR Systems on the 300-Hr Switchboard Corpus 3332
Zhaoqing Li, Tianzi Wang, Jiajun Deng, Junhao Xu, Shoukang Hu, Xunying Liu

Compressed MoE ASR Model Based on Knowledge Distillation and Quantization 3337
Yuping Yuan, Zhao You, Shulin Feng, Dan Su, Yanchun Liang, Xiaohu Shi, Dong Yu

ACOUSTIC MODEL ADAPTATION FOR ASR

Factorised Speaker-Environment Adaptive Training of Conformer Speech Recognition Systems 3342
Jiajun Deng, Guinan Li, Xurong Xie, Zengrui Jin, Mingyu Cui, Tianzi Wang, Shujie Hu, Mengzhe Geng, Xunying Liu

Text Only Domain Adaptation with Phoneme Guided Data Splicing for End-To-End Speech Recognition 3347
Wei Wang, Xun Gong, Hang Shao, Dongning Yang, Yanmin Qian

Cross-Lingual Cross-Age Adaptation for Low-Resource Elderly Speech Emotion Recognition..... 3352
Samuel Cahyawijaya, Holy Lovenia, Willy Chung, Rita Frieske, Zihan Liu, Pascale Fung

Modular Domain Adaptation for Conformer-Based Streaming ASR 3357
Qiujia Li, Bo Li, Dongseong Hwang, Tara Sainath, Pedro M. Mengibar

Don't Stop Self-Supervision: Accent Adaptation of Speech Representations Via Residual Adapters..... 3362
Anshu Bhatia, Sanchit Sinha, Saket Dingliwal, Karthik Gopalakrishnan, Sravan Bodapati, Katrin Kirchhoff

SGEM: Test-Time Adaptation for Automatic Speech Recognition Via Sequential-Level Generalized Entropy Minimization 3367
Changhun Kim, Joonhyung Park, Hajin Shim, Eunho Yang

SPEECH SYNTHESIS: EXPRESSIVITY

A Generative Framework for Conversational Laughter: Its 'Language Model' and Laughter Sound Synthesis..... 3372
Hiroki Mori, Shunya Kimura

Towards Spontaneous Style Modeling with Semi-Supervised Pre-Training for Conversational Text-To-Speech Synthesis..... 3377
Weiqin Li, Shun Lei, Qiaochu Huang, Yixuan Zhou, Zhiyong Wu, Shiyin Kang, Helen Meng

Beyond Style: Synthesizing Speech with Pragmatic Functions..... 3382
Harm Lameris, Joakim Gustafson, Éva Székely

ECat: An End-To-End Model for Multi-Speaker TTS & Many-To-Many Fine-Grained Prosody Transfer 3387
Ammar Abbas, Sri Karlapati, Bastian Schnell, Penny Karanasou, Marcel Granero Moya, Amith Nagaraj, Ayman Boustati, Nicole Peinelt, Alexis Moinet, Thomas Drugman

MULTI-MODAL SYSTEMS

BeAts: Bengali Speech Acts Recognition Using Multimodal Attention Fusion..... 3392
Ahana Deb, Sayan Nag, Ayan Mahapatra, Soumitri Chattopadhyay, Aritra Marik, Pijush Kanti Gayen, Shankha Sanyal, Archi Banerjee, Samir Karmakar

Improving the Gap in Visual Speech Recognition Between Normal and Silent Speech Based on Metric Learning..... 3397
Sara Kashiwagi, Keitaro Tanaka, Qi Feng, Shigeo Morishima

Whistle-To-Text: Automatic Recognition of the Silbo Gomero Whistled Language 3402
Agata Jakubiak

A Novel Interpretable and Generalizable Re-Synchronization Model for Cued Speech Based on a Multi-Cuer Corpus 3407
Lufei Gao, Shan Huang, Li Liu

Visually Grounded Few-Shot Word Acquisition with Fewer Shots 3412
Leanne Nortje, Benjamin Van Niekerk, Herman Kamper

JAMFN: Joint Attention Multi-Scale Fusion Network for Depression Detection 3417
Li Zhou, Zhenyu Liu, Zixuan Shangguan, Xiaoyan Yuan, Yutong Li, Bin Hu

QUESTION ANSWERING FROM SPEECH

Prompt Guided Copy Mechanism for Conversational Question Answering 3422
Yong Zhang, Zhitao Li, Jianzong Wang, Yiming Gao, Ning Cheng, Fengying Yu, Jing Xiao

Composing Spoken Hints for Follow-On Question Suggestion in Voice Assistants 3427
Pedro Faustini, Besnik Fetahu, Giuseppe Castellucci, Anjie Fang, Oleg Rokhlenko, Shervin Malmasi

On Monotonic Aggregation for Open-Domain QA 3432
Sang-Eun Han, Yeonseok Jeong, Seung-Won Hwang, Kyungjae Lee

Question-Context Alignment and Answer-Context Dependencies for Effective Answer Sentence Selection 3437
Minh Van Nguyen, Kishan KC, Toan Nguyen, Thien Huu Nguyen, Ankit Chadha, Thuy Vu

Multi-Scale Attention for Audio Question Answering 3442
Guangyao Li, Yixin Xu, Di Hu

Enhancing Visual Question Answering Via Deconstructing Questions and Explicating Answers 3447
Feilong Chen, Minglun Han, Jing Shi, Shuang Xu, Bo Xu

MULTI-TALKER METHODS IN SPEECH PROCESSING

SEF-Net: Speaker Embedding Free Target Speaker Extraction Network 3452
Bang Zeng, Suo Hongbin, Yulong Wan, Ming Li

Cascaded Encoders for Fine-Tuning ASR Models on Overlapped Speech 3457
Richard Rose, Oscar Chang, Olivier Siohan

TokenSplit: Using Discrete Speech Representations for Direct, Refined, and Transcript-Conditioned Speech Separation and Recognition 3462
Hakan Erdogan, Scott Wisdom, Xuankai Chang, Zalán Borsos, Marco Tagliasacchi, Neil Zeghidour, John R. Hershey

Unified Modeling of Multi-Talker Overlapped Speech Recognition and Diarization with a Sidecar Separator 3467
Lingwei Meng, Jiawen Kang, Mingyu Cui, Haibin Wu, Xixin Wu, Helen Meng

Time-Domain Transformer-Based Audiovisual Speaker Separation 3472
Vahid Ahmadi Kalkhorani, Anurag Kumar, Ke Tan, Buye Xu, Deliang Wang

Multi-Stream Extension of Variational Bayesian HMM Clustering (MS-VBx) for Combined End-To-End and Vector Clustering-Based Diarization 3477
Marc Delcroix, Naohiro Tawara, Mireia Diez, Federico Landini, Anna Silnova, Atsunori Ogawa, Tomohiro Nakatani, Lukáš Burget, Shoko Araki

VOLUME 6

Unsupervised Adaptation with Quality-Aware Masking to Improve Target-Speaker Voice Activity Detection for Speaker Diarization 3482
Shutong Niu, Jun Du, Maokui He, Chin-Hui Lee, Baoxiang Li, Jiakui Li

BA-SOT: Boundary-Aware Serialized Output Training for Multi-Talker ASR 3487
Yuhao Liang, Fan Yu, Yangze Li, Pengcheng Guo, Shiliang Zhang, Qian Chen, Lei Xie

Improving Label Assignments Learning by Dynamic Sample Dropout Combined with Layer-Wise Optimization in Speech Separation 3492
Chenyang Gao, Yue Gu, Ivan Marsic

Joint Compensation of Multi-Talker Noise and Reverberation for Speech Enhancement with Cochlear Implants Using One Or More Microphones 3497
Clément Gaultier, Tobias Goehring

Speaker Diarization for ASR Output with T-Vectors: A Sequence Classification Approach 3502
Midia Yousefi, Naoyuki Kanda, Dongmei Wang, Zhuo Chen, Xiaofei Wang, Takuya Yoshioka

GPU-Accelerated Guided Source Separation for Meeting Transcription 3507
Desh Raj, Daniel Povey, Sanjeev Khudanpur

Overlap Aware Continuous Speech Separation Without Permutation Invariant Training 3512
Linfeng Yu, Wangyou Zhang, Chenda Li, Yanmin Qian

Weakly-Supervised Speech Pre-Training: A Case Study on Target Speech Recognition 3517
Wangyou Zhang, Yanmin Qian

Directional Speech Recognition for Speaker Disambiguation and Cross-Talk Suppression 3522
Ju Lin, Niko Moritz, Ruiming Xie, Kaustubh Kalgaonkar, Christian Fuegen, Frank Seide

Mixture Encoder for Joint Speech Separation and Recognition 3527
Simon Berger, Peter Vieting, Christoph Boeddeker, Ralf Schlüter, Reinhold Haeb-Umbach

SOCIOPHONETICS

Aberystwyth English Pre-Aspiration in Apparent Time 3532
Míša Michaela Hejná, Adèle Jatteau

Speech Entrainment in Chinese Story-Style Talk Shows: The Interaction Between Gender and Role 3537
Yanting Sun, Hongwei Ding

Sociodemographic and Attitudinal Effects on Dialect Speakers' Articulation of the Standard Language: Evidence from German-Speaking Switzerland 3542
Carina Steiner, Dieter Studer-Joho, Corinne Lanthemann, Andrin Büchler, Adrian Leemann

Vowel Normalisation in Latent Space for Sociolinguistics 3547
James Burridge

SPEAKER AND LANGUAGE DIARIZATION

Attention-Based Encoder-Decoder Network for End-To-End Neural Speaker Diarization with Target Speaker Attractor 3552
Zhengyang Chen, Bing Han, Shuai Wang, Yanmin Qian

Robust Self Supervised Speech Embeddings for Child-Adult Classification in Interactions Involving Children with Autism 3557
Rimita Lahiri, Tiantian Feng, Rajat Hebbar, Catherine Lord, So Hyun Kim, Shrikanth Narayanan

The DISPLACE Challenge 2023 - DIarization of SPeaker and LAnguage in Conversational Environments 3562
Shikha Baghel, Shreyas Ramoji, Sidharth, Ranjana H, Prachi Singh, Somil Jain, Pratik Roy Chowdhuri, Kaustubh Kulkarni, Swapnil Padhi, Deepu Vijayasenan, Sriram Ganapathy

Lexical Speaker Error Correction: Leveraging Language Models for Speaker Diarization Error Correction 3567
Rohit Paturi, Sundararajan Srinivasan, Xiang Li

The SpeeD--ZevoTech Submission at DISPLACE 2023 3572
Gabriel Pirlogeanu, Dan Oneata, Alexandru-Lucian Georgescu, Horia Cucu

End-To-End Neural Speaker Diarization with Absolute Speaker Loss 3577
Chao Wang, Jie Li, Xiang Fang, Jian Kang, Yongxiang Li

SPEECH EMOTION RECOGNITION 2

A Context-Constrained Sentence Modeling for Deception Detection in Real Interrogation 3582
Ya-Tse Wu, Yuan-Ting Chang, Shao-Hao Lu, Jing-Yi Chuang, Chi-Chun Lee

MetricAug: A Distortion Metric-Lead Augmentation Strategy for Training Noise-Robust Speech Emotion Recognizer 3587
Ya-Tse Wu, Chi-Chun Lee

The Co-Use of Laughter and Head Gestures Across Speech Styles 3592
Bogdan Ludusan, Marin Schröer, Martina Rossi, Petra Wagner

EmotionNAS: Two-Stream Neural Architecture Search for Speech Emotion Recognition 3597
Haiyang Sun, Zheng Lian, Bin Liu, Ying Li, Jianhua Tao, Licai Sun, Cong Cai, Meng Wang, Yuan Cheng

Pre-Finetuning for Few-Shot Emotional Speech Recognition 3602
Maximillian Chen, Zhou Yu

Integrating Emotion Recognition with Speech Recognition and Speaker Diarisation for Conversations 3607
Wen Wu, Chao Zhang, Philip C. Woodland

Utility-Preserving Privacy-Enabled Speech Embeddings for Emotion Detection 3612
Chandrashekhar Lavania, Sanjiv Das, Xin Huang, Kyu J. Han

Node-Weighted Graph Convolutional Network for Depression Detection in Transcribed Clinical Interviews 3617
Sergio Burdisso, Esaú Villatoro-Tello, Srikanth Madikeri, Petr Motlicek

Laughter in Task-Based Settings: Whom We Talk to Affects How, When, and How Often We Laugh 3622
Catarina Branco, Isabel Trancoso, Paulo Infante, Khiet P. Truong

Exploring Downstream Transfer of Self-Supervised Features for Speech Emotion Recognition 3627
Yuanbo Fang, Xiaofen Xing, Xiangmin Xu, Weibin Zhang

Leveraging Semantic Information for Efficient Self-Supervised Emotion Recognition with Audio-Textual Distilled Models 3632
Danilo De Oliveira, Navin Raj Prabhu, Timo Gerkmann

Two-Stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining 3637
Yuan Gao, Chenhui Chu, Tatsuya Kawahara

Investigating Acoustic Cues for Multilingual Abuse Detection 3642
Yash Thakran, Vinayak Abrol

A Novel Frequency Warping Scale for Speech Emotion Recognition 3647
Premjeet Singh, Goutam Saha

Multi-Scale Temporal Transformer for Speech Emotion Recognition 3652
Zhipeng Li, Xiaofen Xing, Yuanbo Fang, Weibin Zhang, Hengsheng Fan, Xiangmin Xu

Distant Speech Emotion Recognition in an Indoor Human-Robot Interaction Scenario..... 3657
Nicolás Grágeda, Eduardo Alvarado, Rodrigo Mahu, Carlos Busso, Néstor Becerra Yoma

A Study on Prosodic Entrainment in Relation to Therapist Empathy in Counseling Conversation 3662
Dehua Tao, Tan Lee, Harold Chui, Sarah Luk

SHOW AND TELL: LANGUAGE LEARNING AND EDUCATIONAL RESOURCES

A Unified Framework to Improve Learners' Skills of Perception and Production Based on Speech Shadowing and Overlapping 3667
Nobuaki Minematsu, Noriko Nakanishi, Yingxiang Gao, Haitong Sun

Speak & Improve: L2 English Speaking Practice Tool 3669
Diane Nicholls, Kate M. Knill, Mark J. F. Gales, Anton Ragni, Paul Ricketts

Measuring Prosody in Child Speech Using SoapBox Fluency API..... 3671
Mauro Nicolao, Brenda McGuirk, Declan Moore, Niall Mullally, Lora Lynn O'Mahony, Emma O'Neill, Amelia C. Kelly

Teaching Non-Native Sound Contrasts Using Visual Biofeedback..... 3673
Shawn Nissen

Large-Scale Automatic Audiobook Creation..... 3675
Brendan Walsh, Mark Hamilton, Greg Newby, Xi Wang, Serena Ruan, Sheng Zhao, Lei He, Shaofei Zhang, Eric Dettinger, William T. Freeman, Markus Weimer

QVoice: Arabic Speech Pronunciation Learning Application..... 3677
Yassine El Kheir, Fouad Khnaisser, Shammur Absar Chowdhury, Hamdy Mubarak, Shazia Afzal, Ahmed M. Ali

Asking Questions: An Innovative Way to Interact with Oral History Archives 3679
Jan Švec, Martin Bulín, Adam Frémund, Filip Polák

DisfluencyFixer: A Tool to Enhance Language Learning Through Speech to Speech Disfluency Correction 3681
Vineet Bhat, Preethi Jyothi, Pushpak Bhattacharyya

Technology Pipeline for Large Scale Cross-Lingual Dubbing of Lecture Videos into Multiple Indian Languages 3683
Anusha Prakash, Arun Kumar, Ashish Seth, Bhagyashree Mukherjee, Ishika Gupta, Jom Kuriakose, Jordan F, K V Vikram, Mano R Kumar M, Metilda Sagaya Mary, Mohammad Wajahat, Mohana N, Mudit Batra, Navina K, Nihal John George, Nithya Ravi, Pruthwik Mishra, Sudhanshu Srivastava, Vasista Sai Lodagala, Vandan Mujadia, Kada Sai Venkata Vineeth, Vrunda N. Sukhadia, Dipti Sharma, Hema Murthy, Pushpak Bhattacharyya, S Umesh, Rajeev Sangal

MyVoice: Arabic Speech Resource Collaboration Platform 3685
Yousseif Elshahawy, Yassine El Kheir, Shammur Absar Chowdhury, Ahmed M. Ali

Personal Primer Prototype 1: Invitation to Make Your Own Embooked Speech-Based Educational Artifact 3687
Daniel D. Hromada, Hyungjoong Kim

ANALYSIS OF SPEECH AND AUDIO SIGNALS 3

Time-Frequency Domain Filter-And-Sum Network for Multi-Channel Speech Separation 3689
Zhewen Deng, Yi Zhou, Hongqing Liu

Audio-Visual Fusion Using Multiscale Temporal Convolutional Attention for Time-Domain Speech Separation 3694
Debang Liu, Tianqi Zhang, Mads Græsbøll Christensen, Ying Wei, Zeliang An

An Efficient Speech Separation Network Based on Recurrent Fusion Dilated Convolution and Channel Attention 3699
Junyu Wang

Binaural Sound Localization in Noisy Environments Using Frequency-Based Audio Vision Transformer (FAViT) 3704
Waradon Phokhinanan, Nicolas Obin, Sylvain Argentieri

Contrastive Learning Based Deep Latent Masking for Music Source Separation 3709
Jihyun Kim, Hong-Goo Kang

Speaker Extraction with Detection of Presence and Absence of Target Speakers 3714
Ke Zhang, Marvin Borsdorf, Zexu Pan, Haizhou Li, Yangjie Wei, Yi Wang

PIAVE: A Pose-Invariant Audio-Visual Speaker Extraction Network 3719
Qinghua Liu, Meng Ge, Zhizheng Wu, Haizhou Li

Spatial LibriSpeech: An Augmented Dataset for Spatial Audio Learning 3724
Miguel Sarabia, Elena Menyaylenko, Alessandro Toso, Skyler Seto, Zakaria Aldeneh, Shadi Pirhosseinloo, Luca Zappella, Barry-John Theobald, Nicholas Apostoloff, Jonathan Sheaffer

Image-Driven Audio-Visual Universal Source Separation 3729
Chenxing Li, Ye Bai, Yang Wang, Feng Deng, Yuanyuan Zhao, Zhuo Zhang, Xiaorui Wang

Joint Blind Source Separation and Dereverberation for Automatic Speech Recognition Using Delayed-Subsource MNMF with Localization Prior 3734
Mieszko Fras, Marcin Witkowski, Konrad Kowalczyk

SDNet: Stream-Attention and Dual-Feature Learning Network for Ad-Hoc Array Speech Separation 3739
Honglong Wang, Chengyun Deng, Yanjie Fu, Meng Ge, Longbiao Wang, Gaoyan Zhang, Jianwu Dang, Fei Wang

Deeply Supervised Curriculum Learning for Deep Neural Network-Based Sound Source Localization 3744
Min-Sang Baek, Joon-Young Yang, Joon-Hyuk Chang

Multi-Channel Separation of Dynamic Speech and Sound Events 3749
Takuya Fujimura, Robin Scheibler

Rethinking the Visual Cues in Audio-Visual Speaker Extraction 3754
Junjie Li, Meng Ge, Zexu Pan, Rui Cao, Longbiao Wang, Jianwu Dang, Shiliang Zhang

Using Semi-Supervised Learning for Monaural Time-Domain Speech Separation with a Self-Supervised Learning-Based SI-SNR Estimator 3759
Shaoxiang Dang, Tetsuya Matsumoto, Yoshinori Takeuchi, Hiroaki Kudo

Investigation of Training Mute-Expressive End-To-End Speech Separation Networks for an Unknown Number of Speakers 3764
Younggwan Kim, Hyungjun Lim, Kiho Yeom, Eunjoo Seo, Hoodong Lee, Stanley Jungkyu Choi, Honglak Lee

SR-SRP: Super-Resolution Based SRP-PHAT for Sound Source Localization and Tracking 3769
Jae-Heung Cho, Joon-Hyuk Chang

Dual-Memory Multi-Modal Learning for Continual Spoken Keyword Spotting with Confidence Selection and Diversity Enhancement 3774
Zhao Yang, Dianwen Ng, Xizhe Li, Chong Zhang, Rui Jiang, Wei Xi, Yukun Ma, Chongjia Ni, Jizhong Zhao, Bin Ma, Eng Siong Chng

FN-SSL: Full-Band and Narrow-Band Fusion for Sound Source Localization 3779
Yabo Wang, Bing Yang, Xiaofei Li

A Neural State-Space Modeling Approach to Efficient Speech Separation 3784
Chen Chen, Chao-Han Huck Yang, Kai Li, Yuchen Hu, Pin-Jui Ku, Eng Siong Chng

Locate and Beamform: Two-Dimensional Locating All-Neural Beamformer for Multi-Channel Speech Separation 3789
Yanjie Fu, Meng Ge, Honglong Wang, Nan Li, Haoran Yin, Longbiao Wang, Gaoyan Zhang, Jianwu Dang, Chengyun Deng, Fei Wang

Monaural Speech Separation Method Based on Recurrent Attention with Parallel Branches 3794
Xue Yang, Changchun Bao, Xu Zhang, Xianhong Chen

Ontology-Aware Learning and Evaluation for Audio Tagging 3799
Haohe Liu, Qiuqiang Kong, Xubo Liu, Xinhao Mei, Wenwu Wang, Mark D. Plumbley

SPEECH CODING AND ENHANCEMENT 3

Multi-Dataset Co-Training with Sharpness-Aware Optimization for Audio Anti-Spoofing 3804
Hye-Jin Shim, Jee-Weon Jung, Tomi Kinnunen

Reducing the Prior Mismatch of Stochastic Differential Equations for Diffusion-Based Speech Enhancement 3809
Bunlong Lay, Simon Welker, Julius Richter, Timo Gerkmann

Complex-Valued Neural Networks for Voice Anti-Spoofing 3814
Nicolas M. Müller, Philip Sperl, Konstantin Böttinger

DeepVQE: Real Time Deep Voice Quality Enhancement for Joint Acoustic Echo Cancellation, Noise Suppression and Dereverberation 3819
Nicolae Catalin Ristea, Evgenii Indenbom, Ando Saabas, Tanel Pärnamaa, Jegor Guzhvin, Ross Cutler

Diffiner: A Versatile Diffusion-Based Generative Refiner for Speech Enhancement 3824
Ryosuke Sawata, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Takashi Shibuya, Shusuke Takahashi, Yuki Mitsufuji

HD-DEMUCS: General Speech Restoration with Heterogeneous Decoders 3829
Doyeon Kim, Soo-Whan Chung, Hyewon Han, Youna Ji, Hong-Goo Kang

MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra 3834
Ye-Xin Lu, Yang Ai, Zhen-Hua Ling

TridentSE: Guiding Speech Enhancement with 32 Global Tokens 3839
Dacheng Yin, Zhiyuan Zhao, Chuanxin Tang, Zhiwei Xiong, Chong Luo

Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features 3844
Chenglong Wang, Jiangyan Yi, Jianhua Tao, Chu Yuan Zhang, Shuai Zhang, Xun Chen

Self-Supervised Learning with Diffusion-Based Multichannel Speech Enhancement for Speaker Verification Under Noisy Conditions 3849
Sandipana Dowerah, Ajinkya Kulkarni, Romain Serizel, Denis Jouvet

Two-Stage Voice Anonymization for Enhanced Privacy 3854
Francesco Nespoli, Daniel Barreda, Jöerg Bitzer, Patrick A. Naylor

Personalized Dereverberation of Speech 3859
Ruilin Xu, Gurunandan Krishnan, Changxi Zheng, Shree K. Nayar

Weighted Von Mises Distribution-Based Loss Function for Real-Time STFT Phase Reconstruction Using DNN 3864
Nguyen Binh Thien, Yukoh Wakabayashi, Yuting Geng, Kenta Iwai, Takanobu Nishiura

Deep Multi-Frame Filtering for Hearing Aids 3869
Hendrik Schröter, Tobias Rosenkranz, Alberto N. Escalante-B., Andreas Maier

Aligning Speech Enhancement for Improving Downstream Classification Performance 3874
Yan Xiong, Visar Berisha, Chaitali Chakrabarti

DNN-Based Parameter Estimation for MVDR Beamforming and Post-Filtering 3879
Minseung Kim, Sein Cheong, Jong Won Shin

FRA-RIR: Fast Random Approximation of the Image-Source Method 3884
Yi Luo, Jianwei Yu

Rethinking Complex-Valued Deep Neural Networks for Monaural Speech Enhancement 3889
Haibin Wu, Ke Tan, Buye Xu, Anurag Kumar, Daniel Wong

Harmonic Enhancement Using Learnable Comb Filter for Light-Weight Full-Band Speech Enhancement Model 3894
Xiaohuai Le, Tong Lei, Li Chen, Yiqing Guo, Chao He, Cheng Chen, Xianjun Xia, Hua Gao, Yijian Xiao, Piao Ding, Shenyi Song, Jing Lu

SPOKEN LANGUAGE TRANSLATION, INFORMATION RETRIEVAL, SUMMARIZATION, RESOURCES, AND EVALUATION 3

How Does Pretraining Improve Discourse-Aware Translation? 3899
Zhihong Huang, Longyue Wang, Siyou Liu, Derek F. Wong

PATCorrect: Non-Autoregressive Phoneme-Augmented Transformer for ASR Error Correction 3904
Ziji Zhang, Zhehui Wang, Rajesh Kamma, Sharanya Eswaran, Narayanan Sadagopan

Model-Assisted Lexical Tone Evaluation of Three-Year-Old Chinese-Speaking Children by Also Considering Segment Production 3909
Shu-Chuan Tseng, Yi-Fen Liu, Xiang-Li Lu

Sentence Embedder Guided Utterance Encoder (SEGUE) for Spoken Language Understanding 3914
Yi Xuan Tan, Navonil Majumder, Soujanya Poria

Joint Time and Frequency Transformer for Chinese Opera Classification 3919
Qiang Li, Beibei Hu

AdaMS: Deep Metric Learning with Adaptive Margin and Adaptive Scale for Acoustic Word Discrimination 3924
Myunghun Jung, Hoirin Kim

Investigating Reproducibility at Interspeech Conferences: A Longitudinal and Comparative Perspective 3929
Mohammad Arvan, A. Seza Dogruöz, Natalie Parde

An Efficient Approach for the Automated Segmentation and Transcription of the People's Speech Sorpus 3939
Astik Biswas, Abdelmoumene Boumadane, Stephane Peillon, Gildas Bleas

Diverse Feature Mapping and Fusion Via Multitask Learning for Multilingual Speech Emotion Recognition 3944
Shi-Wook Lee

Take the Hint: Improving Arabic Diacritization with Partially-Diacritized Text 3949
Parnia Bahar, Mattia Di Gangi, Nick Rossenbach, Mohammad Zeineldeen

Low-Resource Cross-Lingual Adaptive Training for Nigerian Pidgin 3954
Pin-Jie Lin, Muhammed Saeed, Ernie Chang, Merel Scholman

Efficient Adaptation of Spoken Language Understanding Based on End-To-End Automatic Speech Recognition 3959
Eesung Kim, Aditya Jajodia, Cindy Tseng, Divya Neelagiri, Taeyeon Ki, Vijendra Raj Apsingekar

PhonMatchNet: Phoneme-Guided Zero-Shot Keyword Spotting for User-Defined Keywords 3964
Yong-Hyeok Lee, Namhyun Cho

Mix Before Align: Towards Zero-Shot Cross-Lingual Sentiment Analysis Via Soft-Mix and Multi-View Learning 3969
Zhihong Zhu, Xuxin Cheng, Dongsheng Chen, Zhiqi Huang, Hongxiang Li, Yuexian Zou

AlignAtt: Using Attention-Based Audio-Translation Alignments as a Guide for Simultaneous Speech Translation 3974
Sara Papi, Marco Turchi, Matteo Negri

Incremental Blockwise Beam Search for Simultaneous Speech Translation with Controllable Quality-Latency Tradeoff 3979
Peter Polák, Brian Yan, Shinji Watanabe, Alex Waibel, Ondrej Bojar

Zambezi Voice: A Multilingual Speech Corpus for Zambian Languages 3984
Claytone Sikasote, Kalinda Siaminwe, Stanly Mwape, Bangiwe Zulu, Mofya Phiri, Martin Phiri, David Zulu, Mayumbo Nyirenda, Antonios Anastasopoulos

ANTI-SPOOFING FOR SPEAKER VERIFICATION

Towards Single Integrated Spoofing-Aware Speaker Verification Embeddings 3989
Sung Hwan Mun, Hye-Jin Shim, Hemlata Tak, Xin Wang, Xuechen Liu, Md Sahidullah, Myeonghun Jeong, Min Hyun Han, Massimiliano Todisco, Kong Aik Lee, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, Nam Soo Kim, Jee-Weon Jung

Pseudo-Siamese Network Based Timbre-Reserved Black-Box Adversarial Attack in Speaker Identification 3994
Qing Wang, Jixun Yao, Ziqian Wang, Pengcheng Guo, Lei Xie

Betray Oneself: A Novel Audio DeepFake Detection Model Via Mono-To-Stereo Conversion 3999
Rui Liu, Jinhua Zhang, Guanglai Gao, Haizhou Li

Robust Audio Anti-Spoofing Countermeasure with Joint Training of Front-End and Back-End Models 4004
Xingming Wang, Bang Zeng, Suo Hongbin, Yulong Wan, Ming Li

Improved DeepFake Detection Using Whisper Features 4009
Piotr Kawa, Marcin Plata, Michal Czuba, Piotr Szymanski, Piotr Syga

DoubleDeceiver: Deceiving the Speaker Verification System Protected by Spoofing Countermeasures 4014
Mengao Zhang, Ke Xu, Hao Li, Lei Wang, Chengfang Fang, Jie Shi

SPEECH CODING: INTELLIGIBILITY

On Training a Neural Residual Acoustic Echo Suppressor for Improved ASR 4019
Sankaran Panchapagesan, Turaj Zakizadeh Shabestary, Arun Narayanan

Extending DNN-Based Multiplicative Masking to Deep Subband Filtering for Improved Dereverberation 4024
Jean-Marie Lemercier, Julian Tobergte, Timo Gerkmann

UnSE: Unsupervised Speech Enhancement Using Optimal Transport 4029
Wenbin Jiang, Fei Wen, Yifan Zhang, Kai Yu

MC-SpEx: Towards Effective Speaker Extraction with Multi-Scale Interfusion and Conditional Speaker Modulation 4034
Jun Chen, Wei Rao, Zilin Wang, Jiuxin Lin, Yukai Ju, Shulin He, Yannan Wang, Zhiyong Wu

Causal Signal-Based DCCRN with Overlapped-Frame Prediction for Online Speech Enhancement 4039
Julitta Bartolewska, Stanislaw Kacprzak, Konrad Kowalczyk

Gesper: A Restoration-Enhancement Framework for General Speech Reconstruction 4044
Wenzhe Liu, Yupeng Shi, Jun Chen, Wei Rao, Shulin He, Andong Li, Yannan Wang, Zhiyong Wu

RESOURCES FOR SPOKEN LANGUAGE PROCESSING

Multimodal Personality Traits Assessment (MuPTA) Corpus: The Impact of Spontaneous and Read Speech 4049
Elena Ryumina, Dmitry Ryumin, Maxim Markitantov, Heysem Kaya, Alexey Karpov

MOCKS 1.0: Multilingual Open Custom Keyword Spotting Testset 4054
Mikolaj Pudo, Mateusz Wosik, Adam Cieslak, Justyna Krzywdziak, Bozena Lukasiak, Artur Janicki

MD3: The Multi-Dialect Dataset of Dialogues 4059
Jacob Eisenstein, Vinodkumar Prabhakaran, Clara Rivera, Dorottya Demszky, Devyani Sharma

MuAViC: A Multilingual Audio-Visual Corpus for Robust Speech Recognition and Robust Speech-To-Text Translation 4064
Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, Changhan Wang

Thai Dialect Corpus and Transfer-Based Curriculum Learning Investigation for Dialect Automatic Speech Recognition 4069
Artit Suwanbandit, Burin Naowarat, Orathai Sangpetch, Ekapol Chuangsuwanich

HK-LegiCoST: Leveraging Non-Verbatim Transcripts for Speech Translation 4074
Cihan Xiao, Henry Li Xinyuan, Jinyi Yang, Dongji Gao, Matthew Wiesner, Kevin Duh, Sanjeev Khudanpur

NEW COMPUTATIONAL STRATEGIES FOR ASR TRAINING AND INFERENCE

A Metric-Driven Approach to Conformer Layer Pruning for Efficient ASR Inference 4079
Dhanush Bekal, Karthik Gopalakrishnan, Karel Mundnich, Srikanth Ronanki, Sravan Bodapati, Katrin Kirchhoff

Distillation Strategies for Discriminative Speech Recognition Rescoring 4084
Prashanth Gurunath Shivakumar, Jari Kolehmainen, Yile Gu, Ankur Gandhe, Ariya Rastrow, Ivan Bulyko

Another Point of View on Visual Speech Recognition 4089
Baptiste Pouthier, Laurent Pilati, Giacomo Valenti, Charles Bouveyron, Frederic Precioso

RASR2: The RWTH ASR Toolkit for Generic Sequence-To-Sequence Speech Recognition 4094
Wei Zhou, Eugen Beck, Simon Berger, Ralf Schlüter, Hermann Ney

Streaming Speech-To-Confusion Network Speech Recognition 4099
Denis Filimonov, Prabhat Pandey, Ariya Rastrow, Ankur Gandhe, Andreas Stolcke

Accurate and Structured Pruning for Efficient Automatic Speech Recognition 4104
Huiqiang Jiang, Li Lyna Zhang, Yuang Li, Yu Wu, Shijie Cao, Ting Cao, Yuqing Yang, Jinyu Li, Mao Yang, Lili Qiu

MERLION CCS CHALLENGE: MULTILINGUAL EVERYDAY RECORDINGS - LANGUAGE IDENTIFICATION ON CODE-SWITCHED CHILD-DIRECTED SPEECH

MERLIon CCS Challenge: A English-Mandarin Code-Switching Child-Directed Speech Corpus for Language Identification and Diarization 4109
Victoria Y. H. Chua, Hexin Liu, Leibny Paola Garcia, Fei Ting Woon, Jinyi Wong, Xiangyu Zhang, Sanjeev Khudanpur, Andy W. H. Khong, Justin Dauwels, Suzy J. Styles

Spoken Language Identification System for English-Mandarin Code-Switching Child-Directed Speech 4114
Shashi Kant Gupta, Sushant Hiray, Prashant Kukde

Improving Wav2vec2-Based Spoken Language Identification by Learning Phonological Features 4119
Mostafa Shahin, Zheng Nan, Vidhyasaharan Sethu, Beena Ahmed

Language Identification Networks for Multilingual Everyday Recordings 4124
Kiran Praveen, Balaji Radhakrishnan, Kamini Sabu, Abhishek Pandey, Mahaboob Ali Basha Shaik

Investigating Model Performance in Language Identification: Beyond Simple Error Statistics 4129
Suzy J. Styles, Victoria Y. H. Chua, Fei Ting Woon, Hexin Liu, Leibny Paola Garcia, Sanjeev Khudanpur, Andy W. H. Khong, Justin Dauwels

HEALTH-RELATED SPEECH ANALYSIS

Classification of Vocal Intensity Category from Speech Using the Wav2vec2 and Whisper Embeddings 4134
Manila Kodali, Sudarsana Reddy Kadiri, Paavo Alku

The Effect of Clinical Intervention on the Speech of Individuals with PTSD: Features and Recognition Performances 4139
Alexander Kathan, Andreas Triantafyllopoulos, Shahin Amiriparian, Sabrina Milkus, Alexander Gebhard, Jonas Hohmann, Pauline Muderlak, Jürgen Schottdorf, Björn W. Schuller, Richard Musil

Analysis and Automatic Prediction of Exertion from Speech: Contrasting Objective and Subjective Measures Collected While Running 4144
Andreas Triantafyllopoulos, Alexander Gebhard, Alexander Kathan, Maurice Gerczuk, Shahin Amiriparian, Björn W. Schuller

The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection 4149
Fuxiang Tao, Anna Esposito, Alessandro Vinciarelli

Comparing Hand-Crafted Features to Spectrograms for Autism Severity Estimation 4154
Marina Eni, Ilan Dinstein, Yaniv Zigel

Acoustic Characteristics of Depression in Older Adults' Speech: The Role of Covariates 4159
Carmen Mijnders, Esther Janse, Paul Naarding, Khiet P. Truong

AUTOMATIC AUDIO CLASSIFICATION AND AUDIO CAPTIONING

Dual Transformer Decoder Based Features Fusion Network for Automated Audio Captioning 4164
Jianyuan Sun, Xubo Liu, Xinhao Mei, Volkan Kiliç, Mark D. Plumbley, Wenwu Wang

Adapting a ConvNeXt Model to Audio Classification on AudioSet 4169
Thomas Pellegrini, Ismail Khalfaoui-Hassani, Etienne Labbé, Timothée Masquelier

Few-Shot Class-Incremental Audio Classification Using Stochastic Classifier 4174
Yanxiong Li, Wenchang Cao, Jialong Li, Wei Xie, Qianhua He

VOLUME 7

Enhance Temporal Relations in Audio Captioning with Sound Event Detection 4179
Zeyu Xie, Xuenan Xu, Mengyue Wu, Kai Yu

SPEECH PERCEPTION, PRODUCTION, AND ACQUISITION 2

First Language Effects on Second Language Perception: Evidence from English Low-Vowel Nasal Sequences Perceived by L1 Mandarin Chinese Listeners 4184
Sijia Zhang

Motor Control Similarity Between Speakers Saying "A Souk" Using Inverse Atlas Tongue Modeling 4189
Ursa Maity, Fangxu Xing, Jerry Prince, Maureen Stone, El Fakhri Georges, Jonghye Woo, Sidney Fels

Assessing Phrase Break of ESL Speech with Pre-Trained Language Models and Large Language Models 4194
Zhiyi Wang, Shaoguang Mao, Wenshan Wu, Yan Xia, Yan Deng, Jonathan Tien

A Relationship Between Vocal Fold Vibration and Droplet Production 4199
Tsukasa Yoshinaga, Takayuki Arai, Akiyoshi Iida

Audio, Visual and Audiovisual Intelligibility of Vowels Produced in Noise 4204
Maeva Garnier

Optimal Control of Speech with Context-Dependent Articulatory Targets 4209
Benjamin Elie, Juraj Šimko, Alice Turk

Computational Modeling of Auditory Brainstem Responses Derived from Modified Speech 4214
Tzu-Han Zoe Cheng, Paul Calamia

Leveraging Label Information for Multimodal Emotion Recognition 4219
Peiying Wang, Sunlu Zeng, Junqing Chen, Lu Fan, Meng Chen, Youzheng Wu, Xiaodong He

Improving End-To-End Modeling for Mandarin-English Code-Switching Using Lightweight Switch-Routing Mixture-Of-Experts 4224
Fengyun Tan, Chaofeng Feng, Tao Wei, Shuai Gong, Jinqiang Leng, Wei Chu, Jun Ma, Shaojun Wang, Jing Xiao

Frequency Patterns of Individual Speaker Characteristics at Higher and Lower Spectral Ranges 4229
Zhao Zhang, Ju Zhang, Ziyu Zhu, Yujie Chi, Kiyoshi Honda, Jianguo Wei

Adaptation to Predictive Prosodic Cues in Non-Native Standard Dialect 4234
Sabine Gosselke Berthelsen

Head Movements in Two- And Four-Person Interactive Conversational Tasks in Noisy and Moderately Reverberant Conditions 4239
Alan Archer-Boyd, Rainer Martin

Second Language Identification of Vietnamese Tones by Native Mandarin Learners 4244
Juqiang Chen, Ailing Qin, Hui Chang, Hua Chen

Nasal Vowel Production and Grammatical Processing in French-Speaking Children with Cochlear Implants and Normal-Hearing Peers. 4249
Sophie Fagniart, Véronique Delvaux, Brigitte Charlier, Bernard Harmegnies, Anne Huberlant, Myriam Piccaluga, Kathy Huet

Emotion Classification with EEG Responses Evoked by Emotional Prosody of Speech 4254
Zechen Zhang, Xihong Wu, Jing Chen

L2-Mandarin Regional Accent Variability During Mandarin Tone-Word Training Facilitates English Listeners' Subsequent Tone Categorizations 4259
Yanping Li, Michael D. Tyler, Denis Burnham, Catherine T. Best

HumanDiffusion: Diffusion Model Using Perceptual Gradients 4264
Yota Ueda, Shinnosuke Takamichi, Yuki Saito, Norihiro Takamune, Hiroshi Saruwatari

Queer Events, Relationships, and Sports: Does Topic Influence Speakers' Acoustic Expression of Sexual Orientation? 4269
Sven Kachel, Manuel Pöhlmann, Christine Nussbaum

SPEECH SYNTHESIS

Epoch-Based Spectrum Estimation for Speech 4274
Jón Guðnason, Guolin Fang, Mike Brookes

OverFlow: Putting Flows on Top of Neural Transducers for Better TTS 4279
Shivam Mehta, Ambika Kirkland, Harm Lameris, Jonas Beskow, Éva Székely, Gustav Eje Henter

ADAPTERMIX: Exploring the Efficacy of Mixture of Adapters for Low-Resource TTS Adaptation 4284
Ambuj Mehrish, Abhinav Ramesh Kashyap, Li Yingting, Navonil Majumder, Soujanya Poria

Prior-Free Guided TTS: An Improved and Efficient Diffusion-Based Text-Guided Speech Synthesis 4289
Won-Gook Choi, So-Jeong Kim, Taeho Kim, Joon-Hyuk Chang

UnDiff: Unsupervised Voice Restoration with Unconditional Diffusion Model 4294
Anastasiia Iashchenko, Pavel Andreev, Ivan Shchekotov, Nicholas Babaev, Dmitry Vetrov

Pruning Self-Attention for Zero-Shot Multi-Speaker Text-To-Speech 4299
Hyungchan Yoon, Changhwan Kim, Eunwoo Song, Hyun-Wook Yoon, Hong-Goo Kang

Interpretable Style Transfer for Text-To-Speech with ControlVAE and Diffusion Bridge 4304
Wenhao Guan, Tao Li, Yishuang Li, Hukai Huang, Qingyang Hong, Lin Li

Towards Robust FastSpeech 2 by Modelling Residual Multimodality 4309
Fabian Kögel, Bac Nguyen, Fabien Cardinaux

Real Time Spectrogram Inversion on Mobile Phone 4314
Oleg Rybakov, Marco Tagliasacchi, Yunpeng Li, Liyang Jiang, Xia Zhang, Fadi Biadsy

Automatic Tuning of Loss Trade-Offs Without Hyper-Parameter Search in End-To-End Zero-Shot Speech Synthesis 4319
Seongyeon Park, Bohyung Kim, Tae-Hyun Oh

A Low-Resource Pipeline for Text-To-Speech from Found Data with Application to Scottish Gaelic 4324
Dan Wells, Korin Richmond, William Lamb

Self-Supervised Solution to the Control Problem of Articulatory Synthesis 4329
Paul K. Krug, Peter Birkholz, Branislav Gerazov, Daniel R. Van Niekerk, Anqi Xu, Yi Xu

Hierarchical Timbre-Cadence Speaker Encoder for Zero-Shot Speech Synthesis 4334
Joun Yeop Lee, Jae-Sung Bae, Seongkyu Mun, Jihwan Lee, Ji-Hyun Lee, Hoon-Young Cho, Chanwoo Kim

ZET-Speech: Zero-Shot Adaptive Emotion-Controllable Text-To-Speech Synthesis with Diffusion and Style-Based Models 4339
Minki Kang, Wooseok Han, Sung Ju Hwang, Eunho Yang

Improving WaveRNN with Heuristic Dynamic Blending for Fast and High-Quality GPU Vocoding 4344
Muyang Du, Chuan Liu, Jiaxing Qi, Junjie Lai

Intelligible Lip-To-Speech Synthesis with Speech Units 4349
Jeongsoo Choi, Minsu Kim, Yong Man Ro

Parameter-Efficient Learning for Text-To-Speech Accent Adaptation 4354
Li-Jen Yang, Chao-Han Huck Yang, Jen-Tzung Chien

Controlling Formant Frequencies with Neural Text-To-Speech for the Manipulation of Perceived Speaker Age 4359
Ziya Khan, Lovisa Wihlborg, Cassia Valentini-Botinhao, Oliver Watts

FastFit: Towards Real-Time Iterative Neural Vocoder by Replacing U-Net Encoder with Multiple STFTs 4364
Won Jang, Dan Lim, Heayoung Park

ISTFTNet2: Faster and More Lightweight iSTFT-Based Neural Vocoder Using 1D-2D CNN 4369
Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, Shogo Seki

VITS2: Improving Quality and Efficiency of Single-Stage Text-To-Speech with Adversarial Learning and Architecture Design 4374
Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, Sangjin Kim

Controlling Multi-Class Human Vocalization Generation Via a Simple Segment-Based Labeling Scheme 4379
Hieu-Thi Luong, Junichi Yamagishi

SPEECH RECOGNITION: SIGNAL PROCESSING, ACOUSTIC MODELING, ROBUSTNESS, ADAPTATION 4

Vistaar: Diverse Benchmarks and Training Sets for Indian Language ASR 4384
Kaushal Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M. Khapra, Pratyush Kumar

Domain Adaptive Self-Supervised Training of Automatic Speech Recognition 4389
Cong-Thanh Do, Rama Doddipatla, Mohan Li, Thomas Hain

There is More than One Kind of Robustness: Fooling Whisper with Adversarial Examples 4394
Raphael Olivier, Bhiksha Raj

MT-SLVR: Multi-Task Self-Supervised Learning for Transformation In(Variant) Representations 4399
Calum Heggan, Tim Hospedales, Sam Budgett, Mehrdad Yaghoobi

Reducing Barriers to Self-Supervised Learning: HuBERT Pre-Training with Academic Compute 4404
William Chen, Xuankai Chang, Yifan Peng, Zhaoheng Ni, Soumi Maiti, Shinji Watanabe

Blank-Regularized CTC for Frame Skipping in Neural Transducer 4409
Yifan Yang, Xiaoyu Yang, Liyong Guo, Zengwei Yao, Wei Kang, Fangjun Kuang, Long Lin, Xie Chen, Daniel Povey

The Tag-Team Approach: Leveraging CLS and Language Tagging for Enhancing Multilingual ASR 4414
Kaousheik Jayakumar, Vrunda N. Sukhadia, A Arunkumar, S Umesh

Improving RNN-Transducers with Acoustic LookAhead 4419
Vinit S. Unni, Ashish Mittal, Preethi Jyothi, Sunita Sarawagi

Everyone Has an Accent 4424
Nina Markl, Catherine Lai

Some Voices Are Too Common: Building Fair Speech Recognition Systems Using the CommonVoice Dataset 4428
Lucas Maison, Yannick Estève

Information Magnitude Based Dynamic Sub-Sampling for Speech-To-Text 4433
Yuhao Zhang, Chenghao Gao, Kaiqi Kou, Chen Xu, Tong Xiao, Jingbo Zhu

KEYNOTE 3

What's in a Rise? the Relevance of Intonation for Attention Orienting 4438
Martine Grice

SPEECH SYNTHESIS: CONTROLLABILITY AND ADAPTATION

HierVST: Hierarchical Adaptive Zero-Shot Voice Style Transfer 4439
Sang-Hoon Lee, Ha-Yeong Choi, Hyung-Seok Oh, Seong-Whan Lee

VISinger2: High-Fidelity End-To-End Singing Voice Synthesis Enhanced by Digital Signal Processing Synthesizer 4444
Yongmao Zhang, Heyang Xue, Hanzhao Li, Lei Xie, Tingwei Guo, Ruixiong Zhang, Caixia Gong

EdenTTS: A Simple and Efficient Parallel Text-To-Speech Architecture with Collaborative Duration-Alignment Learning 4449
Youneng Ma, Junyi He, Meimei Wu, Guangyue Hu, Haojun Fei

Generalizable Zero-Shot Speaker Adaptive Speech Synthesis with Disentangled Representations 4454
Wenbin Wang, Yang Song, Sanjay Jha

Speech Inpainting: Context-Based Speech Synthesis Guided by Video 4459
Juan Felipe Montesinos, Daniel Michelsanti, Gloria Haro, Zheng-Hua Tan, Jesper Jensen

STEN-TTS: Improving Zero-Shot Cross-Lingual Transfer for Multi-Lingual TTS with Style-Enhanced Normalization Diffusion Framework 4464
Chung Tran, Chi Mai Luong, Sakriani Sakti

SEARCH METHODS AND DECODING ALGORITHMS FOR ASR

Average Token Delay: A Latency Metric for Simultaneous Translation 4469
Yasumasa Kano, Katsuhito Sudoh, Satoshi Nakamura

Automatic Speech Recognition Transformer with Global Contextual Information Decoder 4474
Yukun Qian, Xuyi Zhuang, Mingjiang Wang

Time-Synchronous One-Pass Beam Search for Parallel Online and Offline Transducers with Dynamic Block Training 4479
Yui Sudo, Shakeel Muhammad, Yifan Peng, Shinji Watanabe

Prefix Search Decoding for RNN Transducers 4484
Kiran Praveen, Advait Vinay Dhopeshwarkar, Abhishek Pandey, Balaji Radhakrishnan

WhisperX: Time-Accurate Speech Transcription of Long-Form Audio 4489
Max Bain, Jaesung Huh, Tengda Han, Andrew Zisserman

Implementing Contextual Biasing in GPU Decoder for Online ASR 4494
Iuliia Nigmatulina, Srikanth Madikeri, Esaú Villatoro-Tello, Petr Motlicek, Juan Zuluaga-Gomez, Karthik Pandia, Aravind Ganapathiraju

SPEECH SIGNAL ANALYSIS

MF-PAM: Accurate Pitch Estimation Through Periodicity Analysis and Multi-Level Feature Fusion 4499
Woo-Jin Chung, Doyeon Kim, Soo-Whan Chung, Hong-Goo Kang

Enhancing Speech Articulation Analysis Using a Geometric Transformation of the X-Ray Microbeam Dataset 4504
Ahmed Adel Attia, Mark Tiede, Carol Espy-Wilson

Matching Acoustic and Perceptual Measures of Phonation Assessment in Disordered Speech - A Case Study 4508
Melanie Jouaiti, Pippa Kirby, Ravi Vaidyanathan

Improved Contextualized Speech Representations for Tonal Analysis 4513
Jiahong Yuan, Xingyu Cai, Kenneth Church

A Study on the Importance of Formant Transitions for Stop-Consonant Classification in VCV Sequence 4518
Siddarth Chandrasekar, Arvind Ramesh, Tilak Purohit, Prasanta Kumar Ghosh

FusedF0: Improving DNN-Based F0 Estimation by Fusion of Summary-Correlograms and Raw Waveform Representations of Speech Signals 4523
Eray Eren, Lee Ngee Tan, Abeer Alwan

SPEECH EMOTION RECOGNITION 3

Improving Joint Speech and Emotion Recognition Using Global Style Tokens 4528
Jehyun Kyung, Ju-Seok Seong, Jeong-Hwan Choi, Ye-Rin Jeoung, Joon-Hyuk Chang

Speech Emotion Recognition by Estimating Emotional Label Sequences with Phoneme Class Attribute 4533
Ryotaro Nagase, Takahiro Fukumori, Yoichi Yamashita

Unsupervised Transfer Components Learning for Cross-Domain Speech Emotion Recognition 4538
Shenjie Jiang, Peng Song, Shaokai Li, Keke Zhao, Wenming Zheng

Dual Memory Fusion for Multimodal Speech Emotion Recognition 4543
Darshana Prisayad, Tharindu Fernando, Sridha Sridharan, Simon Denman, Clinton Fookes

Hybrid Dataset for Speech Emotion Recognition in Russian Language 4548
Vladimir Kondratenko, Nikolay Karpov, Artem Sokolov, Nikita Savushkin, Oleg Kutuzov, Fyodor Minkin

Speech Emotion Recognition Using Decomposed Speech Via Multi-Task Learning 4553
Jia-Hao Hsu, Chung-Hsien Wu, Yu-Hung Wei

CONNECTING SPEECH-SCIENCE AND SPEECH-TECHNOLOGY FOR CHILDREN'S SPEECH

Prospective Validation of Motor-Based Intervention with Automated Mispronunciation Detection of Rhotics in Residual Speech Sound Disorders 4558
Nina R Benway, Jonathan L Preston

Classifying Rhoticity of /ɹ/ in Speech Sound Disorder Using Age-And-Sex Normalized Formants 4563
Nina R Benway, Jonathan L Preston, Asif Salekin, Yi Xiao, Harshit Sharma, Tara McAllister

Acoustic-To-Articulatory Speech Inversion Features for Mispronunciation Detection of /ɹ/ in Child Speech Sound Disorders 4568
Nina R Benway, Yashish M Siriwardena, Jonathan L Preston, Elaine Hitchcock, Tara McAllister, Carol Espy-Wilson

Using Commercial ASR Solutions to Assess Reading Skills in Children: A Case Report 4573
Timothy Piton, Enno Hermann, Angela Pasqualotto, Marjolaine Cohen, Mathew Magimai.-Doss, Daphné Bavelier

Exploiting Diversity of Automatic Transcripts from Distinct Speech Recognition Techniques for Children's Speech 4578
Christopher Gebauer, Lars Rumberg, Hanna Ehlert, Ulrike Lüdtke, Joern Ostermann

Uncertainty Estimation for Connectionist Temporal Classification Based Automatic Speech Recognition 4583
Lars Rumberg, Christopher Gebauer, Hanna Ehlert, Maren Wallbaum, Ulrike Lüdtke, Joern Ostermann

BabySLM: Language-Acquisition-Friendly Benchmark of Self-Supervised Spoken Language Models 4588
Marvin Lavechin, Yaya Sy, Hadrien Titeux, María Andrea Cruz Blandón, Okko Räsänen, Hervé Bredin, Emmanuel Dupoux, Alejandrina Cristia

Data Augmentation for Children ASR and Child-Adult Speaker Classification Using Voice Conversion Methods 4593
Shuyang Zhao, Mittul Singh, Abraham Woubie, Reima Karhila

Developmental Articulatory and Acoustic Features for Six to Ten Year Old Children 4598
Vishwas M. Shetty, Steven M. Lulich, Abeer Alwan

Automatically Predicting Perceived Conversation Quality in a Pediatric Sample Enriched for Autism 4603
Yahan Yang, Sunghye Cho, Maxine Covello, Azia Knox, Osbert Bastani, James Weimer, Edgar Dobriban, Robert Schultz, Insup Lee, Julia Parish-Morris

An Equitable Framework for Automatically Assessing Children's Oral Narrative Language Abilities 4608
Alexander Johnson, Hariram Veeramani, Natarajan Balaji Shankar, Abeer Alwan

An Analysis of Goodness of Pronunciation for Child Speech 4613
Xinwei Cao, Zijian Fan, Torbjørn Svendsen, Giampiero Salvi

Measuring Language Development from Child-Centered Recordings 4618
Yaya Sy, William N. Havard, Marvin Lavechin, Emmanuel Dupoux, Alejandrina Cristia

Speaking Clearly, Understanding Better: Predicting the L2 Narrative Comprehension of Chinese Bilingual Kindergarten Children Based on Speech Intelligibility Using a Machine Learning Approach 4623
Hiuching Hung, Paula A. Pérez-Toro, Tomás Arias-Vergara, Andreas Maier, Elmar Nöth

Speech Breathing Behavior During Pauses in Children 4628
Delphine Charuau, Béatrice Vaxelaire, Rudolph Sock

Understanding Spoken Language Development of Children with ASD Using Pre-Trained Speech Embeddings 4633
Anfeng Xu, Rajat Hebbar, Rimita Lahiri, Tiantian Feng, Lindsay Butler, Lue Shen, Helen Tager-Flusberg, Shrikanth Narayanan

Measuring Phonological Precision in Children with Cleft Lip and Palate 4638
Tomás Arias-Vergara, Elizabeth Londoño-Mora, Paula A. Pérez-Toro, Maria Schuster, Elmar Nöth, Juan Rafael Orozco-Arroyave, Andreas Maier

A Study on Using Duration and Formant Features in Automatic Detection of Speech Sound Disorder in Children 4643
Si-Ioi Ng, Cymie Wing-Yee Ng, Tan Lee

Influence of Utterance and Speaker Characteristics on the Classification of Children with Cleft Lip and Palate 4648
Ilja Baumann, Dominik Wagner, Franziska Braun, Sebastian P. Bayerl, Elmar Nöth, Korbinian Riedhammer, Tobias Bocklet

DIALOG MANAGEMENT

Parameter-Efficient Low-Resource Dialogue State Tracking by Prompt Tuning 4653
Mingyu Derek Ma, Jiun-Yu Kao, Shuyang Gao, Arpit Gupta, Di Jin, Tagyoung Chung, Nanyun Peng

An Autoregressive Conversational Dynamics Model for Dialogue Systems 4658
Matthew McNeill, Rivka Levitan

Style-Transfer Based Speech and Audio-Visual Scene Understanding for Robot Action Sequence Acquisition from Videos 4663
Chiori Hori, Puyuan Peng, David Harwath, Xinyu Liu, Kei Ota, Siddarth Jain, Radu Corcodel, Devesh Jha, Diego Romeres, Jonathan Le Roux

Speech Aware Dialog System Technology Challenge (DSTC11) 4668
Hagen Soltau, Izhak Shafran, Mingqiu Wang, Abhinav Rastogi, Jeffrey Zhao, Ye Jia, Wei Han, Yuan Cao, Aramys Miranda

Knowledge-Retrieval Task-Oriented Dialog Systems with Semi-Supervision 4673
Yucheng Cai, Hong Liu, Zhijian Ou, Yi Huang, Junlan Feng

Tracking Must Go on : Dialogue State Tracking with Verified Self-Training 4678
Jihyun Lee, Chaebin Lee, Yunsu Kim, Gary Geunbae Lee

SPEAKER RECOGNITION 2

Ordered and Binary Speaker Embedding 4683
Jiaying Wang, Xianglong Wang, Namin Wang, Lantian Li, Dong Wang

Self-FiLM: Conditioning GANs with Self-Supervised Representations for Bandwidth Extension Based Speaker Recognition 4688
Saurabh Kataria, Jesús Villalba, Laureano Moro-Velazquez, Thomas Thebaud, Najim Dehak

Curriculum Learning for Self-Supervised Speaker Verification 4693
Hee-Soo Heo, Jee-Weon Jung, Jingu Kang, Young-Ki Kwon, Bong-Jin Lee, You Jin Kim, Joon Son Chung

Introducing Self-Supervised Phonetic Information for Text-Independent Speaker Verification 4698
Ziyang Zhang, Wu Guo, Bin Gu

A Teacher-Student Approach for Extracting Informative Speaker Embeddings from Speech Mixtures 4703
Tobias Cord-Landwehr, Christoph Boeddeker, Catalin Zorila, Rama Doddipatla, Reinhold Haeb-Umbach

Experimenting with Additive Margins for Contrastive Self-Supervised Speaker Verification 4708
Theo Lepage, Reda Dehak

PHONETICS, PHONOLOGY, AND PROSODY 2

Nonbinary American English Speakers Encode Gender in Vowel Acoustics 4713
Maxwell Hope, Charlotte Ward, Jason Lilley

Coarticulation of Sibe Vowels and Dorsal Fricatives in Spontaneous Speech: An Acoustic Study 4718
Jared Sharp, Matthew Faytak, Hasutai Fei Xiong Liu

Using Speech Synthesis to Explain Automatic Speaker Recognition: A New Application of Synthetic Speech 4723
Georgina Brown, Christin Kirchhübel, Ramiz Cuthbert

Same F0, Different Tones: A Multidimensional Investigation of Zhangzhou Tones 4728
Yishan Huang

Discovering Phonetic Feature Event Patterns in Transformer Embeddings 4733
Patrick Cormac English, John D. Kelleher, Julie Carson-Berndsen

A System for Generating Voice Source Signals that Implements the Transformed LF-Model Parameter Control..... 4738
Zihan Wang, Christer Gobl

Speaker-Independent Speech Inversion for Estimation of Nasalance 4743
Yashish M Siriwardena, Carol Espy-Wilson, Suzanne Boyce, Mark Tiede, Liran Oren

Effects of Tonal Coarticulation and Prosodic Positions on Tonal Contours of Low Rising Tones: In the Case of Xiamen Dialect..... 4748
Yiying Hu, Hui Feng, Qinghua Zhao, Aijun Li

Durational and Non-Durational Correlates of Lexical and Derived Geminates in Arabic 4753
Amel Issa

Mapping Phonemes to Acoustic Symbols and Codes Using Synchrony in Speech Modulation Vectors Estimated by the Travellingwave Filter Bank..... 4758
Ashwin Rao

Rhythmic Characteristics of L2 German Speech by Advanced Chinese Learners 4763
Lindun Ge, Min Xu, Hongwei Ding

(Dis)agreement and Preference Structure Are Reflected in Matching Along Distinct Acoustic-Prosodic Features 4768
Anneliese Kelterer, Margaret Zellers, Barbara Schuppler

Vowel Reduction by Greek-Speaking Children: The Effect of Stress and Word Length..... 4773
Polychronia Christodoulidou, Katerina Nicolaidis, Dimitrios Stamovlasis

Pitch Distributions in a Very Large Corpus of Spontaneous Finnish Speech 4778
Mietta Lennes, Minnaleena Toivola

Speech Enhancement Patterns in Human-Robot Interaction: A Cross-Linguistic Perspective..... 4783
Jacek Kudera, Katharina Zahner-Ritter, Jakob Engel, Nathalie Elsässer, Philipp Hutmacher, Carolin Worstbrock

SPEECH SYNTHESIS: EXPRESSIVITY

Controllable Generation of Artificial Speaker Embeddings Through Discovery of Principal Directions 4788
Florian Lux, Pascal Tilli, Sarina Meyer, Ngoc Thang Vu

Dual Audio Encoders Based Mandarin Prosodic Boundary Prediction by Using Multi-Granularity Prosodic Representations..... 4793
Ruishan Li, Yingming Gao, Yanlu Xie, Dengfeng Ke, Jinsong Zhang

NoreSpeech: Knowledge Distillation Based Conditional Diffusion Model for Noise-Robust Expressive TTS 4798
Dongchao Yang, Songxiang Liu, Helin Wang, Jianwei Yu, Chao Weng, Yuexian Zou

MaskedSpeech: Context-Aware Speech Synthesis with Masking Strategy..... 4803
Ya-Jie Zhang, Wei Song, Yanghao Yue, Zhengchen Zhang, Youzheng Wu, Xiaodong He

Narrator Or Character: Voice Modulation in an Expressive Multi-Speaker TTS 4808
Tankala Pavan Kalyan, Preeti Rao, Preethi Jyothi, Pushpak Bhattacharyya

CASEIN: Cascading Explicit and Implicit Control for Fine-Grained Emotion Intensity Regulation 4813
Yuhao Cui, Xiongwei Wang, Zhongzhou Zhao, Wei Zhou, Haiqing Chen

Semi-Supervised Learning for Continuous Emotional Intensity Controllable Speech Synthesis with Disentangled Representations..... 4818
Yoori Oh, Juheon Lee, Yoseob Han, Kyogu Lee

Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis..... 4823
Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, Emmanuel Dupoux

ComedicSpeech: Text to Speech for Stand-Up Comedies in Low-Resource Scenarios..... 4828
Yuyue Wang, Huan Xiao, Yihan Wu, Ruihua Song

Neural Speech Synthesis with Enriched Phrase Boundaries 4833
Marie Kunešová, Jindrich Matoušek

Cross-Lingual Prosody Transfer for Expressive Machine Dubbing 4838
Jakub Swiatkowski, Duo Wang, Mikolaj Babianski, Patrick Lumban Tobing, Ravichander Vipperla, Vincent Pollet

Synthesis After a Couple PINTs: Investigating the Role of Pause-Internal Phonetic Particles in Speech Synthesis and Perception 4843
Mikey Elmers, Johannah O'Mahony, Éva Székely

Accentor: An Explicit Lexical Stress Model for TTS Systems 4848
Diana Geneva, Georgi Shopov, Kostadin Garov, Maria Todorova, Stefan Gerdjikov, Stoyan Mihov

A Neural TTS System with Parallel Prosody Transfer from Unseen Speakers 4853
Slava Shechtman, Raul Fernandez

Diverse and Expressive Speech Prosody Prediction with Denoising Diffusion Probabilistic Model..... 4858
Xiang Li, Songxiang Liu, Max W. Y. Lam, Zhiyong Wu, Chao Weng, Helen Meng

Prosody Modeling with 3D Visual Information for Expressive Video Dubbing 4863
Zhihan Yang, Shansong Liu, Xu Li, Haozhe Wu, Zhiyong Wu, Ying Shan, Jia Jia

LightClone: Speaker-Guided Parallel Subnet Selection for Few-Shot Voice Cloning 4868
Jie Wu, Jian Luan, Yujun Wang

VOLUME 8

EE-TTS: Emphatic Expressive TTS with Linguistic Information..... 4873
Yi Zhong, Chen Zhang, Xule Liu, Chenxi Sun, Weishan Deng, Haifeng Hu, Zhongqian Sun

Stochastic Pitch Prediction Improves the Diversity and Naturalness of Speech in Glow-TTS 4878
Sewade Ogun, Vincent Colotte, Emmanuel Vincent

ContextSpeech: Expressive and Efficient Text-To-Speech for Paragraph Reading..... 4883
Yujia Xiao, Shaofei Zhang, Xi Wang, Xu Tan, Lei He, Sheng Zhao, Frank K. Soong, Tan Lee

PromptStyle: Controllable Style Transfer for Text-To-Speech with Natural Language Descriptions 4888
Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Lei Xie, Zhifei Li

Creating Personalized Synthetic Voices from Post-Glossectomy Speech with Guided Diffusion Models 4893
Yusheng Tian, Guangyan Zhang, Tan Lee

SPEECH RECOGNITION: SIGNAL PROCESSING, ACOUSTIC MODELING, ROBUSTNESS, ADAPTATION 5

Towards Multi-Task Learning of Speech and Speaker Recognition 4898
Nik Vaessen, David A. Van Leeuwen

Regarding Topology and Variant Frame Rates for Differentiable WFST-Based End-To-End ASR 4903
Zeyu Zhao, Peter Bell

2-Bit Conformer Quantization for Automatic Speech Recognition 4908
Oleg Rybakov, Phoenix Meadowlark, Shaojin Ding, David Qiu, Jian Li, David Rim, Yanzhang He

Time-Domain Speech Enhancement for Robust Automatic Speech Recognition 4913
Yufeng Yang, Ashutosh Pandey, Deliang Wang

Multi-Channel Multi-Speaker Transformer for Speech Recognition 4918
Guo Yifan, Tian Yao, Suo Hongbin, Wan Yulong

Fake the Real: Backdoor Attack on Deep Speech Classification Via Voice Conversion 4923
Zhe Ye, Terui Mao, Li Dong, Diqun Yan

Dialect Speech Recognition Modeling Using Corpus of Japanese Dialects and Self-Supervised Learning-Based Model XLSR 4928
Shogo Miwa, Atsuhiko Kai

Contextualized End-To-End Speech Recognition with Contextual Phrase Prediction Network 4933
Kaixun Huang, Ao Zhang, Zhanheng Yang, Pengcheng Guo, Bingshen Mu, Tianyi Xu, Lei Xie

Competitive and Resource Efficient Factored Hybrid HMM Systems Are Simpler than You Think 4938
Tina Raissi, Christoph Lüscher, Moritz Gunz, Ralf Schlüter, Hermann Ney

MMSpeech: Multi-Modal Multi-Task Encoder-Decoder Pre-Training for Speech Recognition 4943
Xiaohuan Zhou, Jiaming Wang, Zeyu Cui, Shiliang Zhang, Zhijie Yan, Jingren Zhou, Chang Zhou

Biased Self-Supervised Learning for ASR 4948
Florian L. Kreyssig, Yangyang Shi, Jinxi Guo, Leda Sari, Abdel-Rahman Mohamed, Philip C. Woodland

A Unified Recognition and Correction Model Under Noisy and Accent Speech Conditions 4953
Zhao Yang, Dianwen Ng, Chong Zhang, Rui Jiang, Wei Xi, Yukun Ma, Chongjia Ni, Jizhong Zhao, Bin Ma, Eng Siong Chng

Wav2vec 2.0 ASR for Cantonese-Speaking Older Adults in a Clinical Setting 4958
Ranzo Huang, Brian Mak

BAT: Boundary Aware Transducer for Memory-Efficient and Low-Latency ASR 4963
Keyu An, Xian Shi, Shiliang Zhang

Bayes Risk Transducer: Transducer with Controllable Alignment Prediction 4968
Jinchuan Tian, Jianwei Yu, Hangting Chen, Brian Yan, Chao Weng, Dong Yu, Shinji Watanabe

Multi-View Frequency-Attention Alternative to CNN Frontends for Automatic Speech Recognition 4973
Belen Alastruey, Lukas Drude, Jahn Heymann, Simon Wiesler

SPEECH, VOICE, AND HEARING DISORDERS 2

Investigating the Dynamics of Hand and Lips in French Cued Speech Using Attention Mechanisms and CTC-Based Decoding 4978
Sanjana Sankar, Denis Beautemps, Frédéric Elisei, Olivier Perrotin, Thomas Hueber

Hearing Loss Affects Emotion Perception in Older Adults: Evidence from a Prosody-Semantics Stroop Task 4983
Yingyang Wang, Min Xu, Jing Shao, Lan Wang, Nan Yan

Cochlear-Implant Listeners Listening to Cochlear-Implant Simulated Speech 4988
Fanhui Kong, Nengheng Zheng, Xianren Wang, Hao He, Jan W. H. Schnupp, Qinglin Meng

Validation of a Task-Independent Cepstral Peak Prominence Measure with Voice Activity Detection 4993
Olivia M. Murton, Abigail E. Haenssler, Marc F. Maffei, Kathryn P. Connaghan, Jordan Green

Score-Balanced Loss for Multi-Aspect Pronunciation Assessment 4998
Heejin Do, Yunsu Kim, Gary Geunbae Lee

Federated Learning for Secure Development of AI Models for Parkinson's Disease Detection Using Speech from Different Languages 5003
Soroosh Tayebi Arasteh, Cristian David Ríos-Urrego, Elmar Nöth, Andreas Maier, Seung Hee Yang, Jan Rusz, Juan Rafael Orozco-Arroyave

F0inTFS: A Lightweight Periodicity Enhancement Strategy for Cochlear Implants 5008
Huali Zhou, Fanhui Kong, Nengheng Zheng, Qinglin Meng

Differentiating Acoustic and Physiological Features in Speech for Hypoxia Detection 5013
Benjamin O'Brien, Adrien Gresse, Jean-Baptise Billaud, Guilhem Belda, Jean-François Bonastre

Mandarin Electrolaryngeal Speech Voice Conversion Using Cross-Domain Features 5018
Hsin-Hao Chen, Yung-Lun Chien, Ming-Chi Yen, Shu-Wei Tsai, Tai-Shih Chi, Hsin-Min Wang, Yu Tsao

Audio-Visual Mandarin Electrolaryngeal Speech Voice Conversion 5023
Yung-Lun Chien, Hsin-Hao Chen, Ming-Chi Yen, Shu-Wei Tsai, Hsin-Min Wang, Yu Tsao, Tai-Shih Chi

Which Aspects of Motor Speech Disorder Are Captured by Mel Frequency Cepstral Coefficients? Evidence from the Change in STN-DBS Conditions in Parkinson's Disease 5027
Vojtech Illner, Petr Krýže, Jan Švihlík, Mário Sousa, Paul Krack, Elina Tripoliti, Robert Jech, Jan Rusz

Detecting Manifest Huntington's Disease Using Vocal Data 5032
Vinod Subramanian, Namhee Kwon, Raymond Brueckner, Nate Blaylock, Henry O'Connell, Luis Sierra, Clementina Ullman, Karen Hildebrand, Simon Laganiere

Exploring Multi-Task Learning and Data Augmentation in Dementia Detection with Self-Supervised Pretrained Models 5037
Minchuan Chen, Chenfeng Miao, Jun Ma, Shaojun Wang, Jing Xiao

SPEECH ACTIVITY DETECTION AND MODELING

GL-SSD: Global and Local Speech Style Disentanglement by Vector Quantization for Robust Sentence Boundary Detection in Speech Stream 5042
Kuncai Zhang, Wei Zhou, Pengcheng Zhu, Haiqing Chen

Semantic VAD: Low-Latency Voice Activity Detection for Speech Interaction 5047
Mohan Shi, Yuchun Shu, Lingyun Zuo, Qian Chen, Shiliang Zhang, Jie Zhang, Li-Rong Dai

Dynamic Encoder RNN for Online Voice Activity Detection in Adverse Noise Conditions 5052
Prithvi R. R. Gudepu, Jayesh M. Koroth, Kamini Sabu, Mahaboob Ali Basha Shaik

Point to the Hidden: Exposing Speech Audio Splicing Via Signal Pointer Nets 5057
Denise Moussa, Germans Hirsch, Sebastian Wankerl, Christian Riess

Real-Time Causal Spectro-Temporal Voice Activity Detection Based on Convolutional Encoding and Residual Decoding 5062
Jingyuan Wang, Jie Zhang, Li-Rong Dai

SVVAD: Personal Voice Activity Detection for Speaker Verification 5067
Zuheng Kang, Jianzong Wang, Junqing Peng, Jing Xiao

MULTILINGUAL MODELS FOR ASR

Learning Cross-Lingual Mappings for Data Augmentation to Improve Low-Resource Speech Recognition 5072
Muhammad Umar Farooq, Thomas Hain

AfriNames: Most ASR Models "Butcher" African Names 5077
Tobi Olatunji, Tejumade Afonja, Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Chris Chinenye Emezue, Amina Mardiyyah Rufai, Sahib Singh

Towards Dialect-Inclusive Recognition in a Low-Resource Language: Are Balanced Corpora the Answer? 5082
Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, Ailbhe Ní Chasaide

Svarah: Evaluating English ASR Systems on Indian Accents 5087
Tahir Javed, Sakshi Joshi, Vignesh Nagarajan, Sai Sundaresan, Janki Nawale, Abhigyan Raman, Kaushal Bhogale, Pratyush Kumar, Mitesh M. Khapra

N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition 5092
Bashar Talafha, Abdul Waheed, Muhammad Abdul-Mageed

The MALACH Corpus: Results with End-To-End Architectures and Pretraining 5097
Michael Picheny, Qin Yang, Daiheng Zhang, Lining Zhang

SPEECH ENHANCEMENT AND BANDWIDTH EXPANSION

Unsupervised Speech Enhancement with Deep Dynamical Generative Speech and Noise Models 5102
Xiaoyu Lin, Simon Leglaive, Laurent Girin, Xavier Alameda-Pineda

Noise-Robust Bandwidth Expansion for 8K Speech Recordings 5107
Yin-Tse Lin, Bo-Hao Su, Chi-Han Lin, Shih-Chan Kuo, Jyh-Shing Roger Jang, Chi-Chun Lee

MdctGAN: Taming Transformer-Based GAN for Speech Super-Resolution with Modified DCT Spectra 5112
Chenhao Shuai, Chaohua Shi, Lu Gan, Hongqing Liu

Zoneformer: On-Device Neural Beamformer for In-Car Multi-Zone Speech Separation, Enhancement and Echo Cancellation 5117
Yong Xu, Vinay Kothapally, Meng Yu, Shixiong Zhang, Dong Yu

Low-Complexity Broadband Beampattern Synthesis Using Array Response Control 5122
Jiayi Xu, Jian Li, Weixin Meng, Xiaodong Li, Chengshi Zheng

A GAN Speech Inpainting Model for Audio Editing Software 5127
Haixin Zhao

ARTICULATION

Deep Speech Synthesis from MRI-Based Articulatory Representations 5132
Peter Wu, Tingle Li, Yijing Lu, Yubin Zhang, Jiachen Lian, Alan W Black, Louis Goldstein, Shinji Watanabe, Gopala K. Anumanchipalli

Learning to Compute the Articulatory Representations of Speech with the MIRRORNET 5137
Yashish M Siriwardena, Carol Espy-Wilson, Shihab Shamma

Generating High-Resolution 3D Real-Time MRI of the Vocal Tract 5142
Martin Strauch, Antoine Serrurier

Exploring a Classification Approach Using Quantised Articulatory Movements for Acoustic to Articulatory Inversion 5147
Jesuraj Bandekar, Sathvik Udupa, Prasanta Kumar Ghosh

NEURAL PROCESSING OF SPEECH AND LANGUAGE: ENCODING AND DECODING THE DIVERSE AUDITORY BRAIN

MEG Encoding Using Word Context Semantics in Listening Stories 5152
Subba Reddy Oota, Nathan Trouvain, Frederic Alexandre, Xavier Hinaut

Investigating the Cortical Tracking of Speech and Music with Sung Speech 5157
Giorgia Cantisani, Amirhossein Chalehchaleh, Giovanni Di Liberto, Shihab Shamma

Coherence Estimation Tracks Auditory Attention in Listeners with Hearing Impairment 5162
Oskar Keding, Emina Alickovic, Martin A. Skoglund, Maria Sandsten

Speech Taskonomy: Which Speech Tasks Are the Most Predictive of fMRI Brain Activity? 5167
Subba Reddy Oota, Veeral Agarwal, Mounika Marreddy, Manish Gupta, Raju Bapi

Exploring Auditory Attention Decoding Using Speaker Features 5172
Zelin Qiu, Jianjun Gu, Dingding Yao, Junfeng Li

Enhancing the EEG Speech Match Mismatch Tasks with Word Boundaries 5177
Akshara Soman, Vidhi Sinha, Sriram Ganapathy

Similar Hierarchical Representation of Speech and Other Complex Sounds in the Brain and Deep Residual Networks: An MEG Study 5182
Tzu-Han Zoe Cheng, Kuan-Lin Chen, Juliane Schubert, Ya-Ping Chen, Tim Brown, John Iversen

Effects of Spectral Degradation on the Cortical Tracking of the Speech Envelope 5187
Alexis Deighton Macintyre, Tobias Goehring

Effects of Spectral and Temporal Modulation Degradation on Intelligibility and Cortical Tracking of Speech Signals 5192
Ignacio Calderon De Palma, Laura S. Lopez, Alejandro Lopez Valdes

PERCEPTION OF PARALINGUISTICS

Transfer Learning for Personality Perception Via Speech Emotion Recognition 5197
Yuanchao Li, Peter Bell, Catherine Lai

A Stimulus-Organism-Response Model of Willingness to Buy from Advertising Speech Using Voice Quality 5202
Mizuki Nagano, Yusuke Ijima, Sadao Hiroya

Voice Passing : A Non-Binary Voice Gender Prediction System for Evaluating Transgender Voice Transition 5207
David Doukhan, Simon Devauchelle, Lucile Girard-Monneron, Mía Chávez Ruz, V. Chaddouk, Isabelle Wagner, Albert Rilliard

Influence of Personal Traits on Impressions of One's Own Voice 5212
Hikaru Yanagida, Yusuke Ijima, Naohiro Tawara

Pardon My Disfluency: The Impact of Disfluency Effects on the Perception of Speaker Competence and Confidence 5217
Ambika Kirkland, Joakim Gustafson, Éva Székely

Cross-Linguistic Emotion Perception in Human and TTS Voices 5222
Iona Gessinger, Michelle Cohn, Benjamin R. Cowan, Georgia Zellou, Bernd Möbius

TECHNOLOGIES FOR CHILD SPEECH PROCESSING

Joint Learning Feature and Model Adaptation for Unsupervised Acoustic Modelling of Child Speech 5227
Richeng Duan

Automatic Assessment of Oral Reading Accuracy for Reading Diagnostics 5232
Bo Molenaar, Cristian Tejedor-Garcia, Catia Cucchiarini, Helmer Strik

An ASR-Enabled Reading Tutor: Investigating Feedback to Optimize Interaction for Learning to Read 5237
Yu Bai, Ferdy Hubers, Catia Cucchiarini, Roeland Van Hout, Helmer Strik

Adaptation of Whisper Models to Child Speech Recognition 5242
Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, Horia Cucu

SHOW AND TELL: MEDIA AND COMMERCIAL APPLICATIONS

Let's Give a Voice to Conversational Agents in Virtual Reality 5247
Michele Yin, Gabriel Roccabruna, Abhinav Azad, Giuseppe Riccardi

FOOCTTS: Generating Arabic Speech with Acoustic Environment for Football Commentator 5249
Massa Baali, Ahmed M. Ali

Video Summarization Leveraging Multimodal Information for Presentations..... 5251
Hanchao Liu, Dapeng Chen, Rongjun Li, Wenyuan Xue, Wei Peng

What Questions Are My Customers Asking?: Towards Actionable Insights from Customer Questions in Contact Center Calls..... 5253
Varun Nathan, Devashish Deshpande, Ayush Kumar, Cijo George, Jithendra Vepa

COnVoy: A Contact Center Operated Pipeline for Voice of Customer Discovery..... 5255
Rishabh Tripathi, Digvijay Anil Ingle, Ayush Kumar, Cijo George, Jithendra Vepa

NeMo Forced Aligner and Its Application to Word Alignment for Subtitle Generation..... 5257
Elena Rastorgueva, Vitaly Lavrukhin, Boris Ginsburg

CauSE: Causal Search Engine for Understanding Contact-Center Conversations..... 5259
Anup Pattnaik, Tanay Narshana, Aashraya Sachdeva, Cijo George, Jithendra Vepa

Tailored Real-Time Call Summarization System for Contact Centers 5261
Aashraya Sachdeva, Sai Nishanth Padala, Anup Pattnaik, Varun Nathan, Cijo George, Ayush Kumar, Jithendra Vepa

Federated Learning Toolkit with Voice-Based User Verification Demo..... 5263
Prathamesh Mandke, Rachel Oberst, Matthias Reisser, Av?it Chakraborty, Christos Louizos, Joseph Soriaga, Daniel Madrigal, Andre Manoel, Nalin Singal, Jeff Omhover, Robert Sim

Learning When to Speak: Latency and Quality Trade-Offs for Simultaneous Speech-To-Speech Translation with Offline Models..... 5265
Liam Dugan, Anshul Wadhawan, Kyle Spence, Chris Callison-Burch, Morgan McGuire, Victor Zordan

Fast Enrollable Streaming Keyword Spotting System: Training and Inference Using a Web Browser 5267
Namhyun Cho, Sunmin Kim, Yoseb Kang, Heeman Kim

Cross-Lingual/Cross-Channel Intent Detection in Contact-Center Conversations..... 5269
Suraj Agrawal, Aashraya Sachdeva, Soumya Jain, Cijo George, Jithendra Vepa

SPEAKER AND LANGUAGE IDENTIFICATION 3

One-Step Knowledge Distillation and Fine-Tuning in Using Large Pre-Trained Self-Supervised Learning Models for Speaker Verification..... 5271
Jungwoo Heo, Chan-Yeong Lim, Ju-Ho Kim, Hyun-Seo Shin, Ha-Jin Yu

Defense Against Adversarial Attacks on Audio DeepFake Detection 5276
Piotr Kawa, Marcin Plata, Piotr Syga

A Conformer-Based Classifier for Variable-Length Utterance Processing in Anti-Spoofing..... 5281
Eros Rosello, Alejandro Gomez-Alanis, Angel M. Gomez, Antonio Peinado

Conformer-Based Language Embedding with Self-Knowledge Distillation for Spoken Language Identification 5286
Feng Wang, Lingyan Huang, Tao Li, Qingyang Hong, Lin Li

CommonAccent: Exploring Large Acoustic Pretrained Models for Accent Classification Based on Common Voice 5291
Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, Cem Subakan

From Adaptive Score Normalization to Adaptive Data Normalization for Speaker Verification Systems..... 5296
Sandro Cumani, Salvatore Sarni

CAM++: A Fast and Efficient Network for Speaker Verification Using Context-Aware Masking 5301
Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, Qian Chen

North Sámi Dialect Identification with Self-Supervised Speech Models..... 5306
Sofoklis Kakouros, Katri Hiovain-Asikainen

Encoder-Decoder Multimodal Speaker Change Detection.....5311
Jee-Weon Jung, Soonshin Seo, Hee-Soo Heo, Geonmin Kim, You Jin Kim, Young-Ki Kwon, Minjae Lee, Bong-Jin Lee

Disentangled Representation Learning for Multilingual Speaker Recognition 5316
Kihyun Nam, Youkyum Kim, Jaesung Huh, Hee-Soo Heo, Jee-Weon Jung, Joon Son Chung

A Compact End-To-End Model with Local and Global Context for Spoken Language Identification 5321
Fei Jia, Nithin Rao Koluguri, Jagadeesh Balam, Boris Ginsburg

On the Robustness of Arabic Speech Dialect Identification..... 5326
Peter Sullivan, Abdelrahim Elmadany, Muhammad Abdul-Mageed

Adaptive Neural Network Quantization for Lightweight Speaker Verification..... 5331
Haoyu Wang, Bei Liu, Yifei Wu, Yanmin Qian

Adversarial Diffusion Probability Model for Cross-Domain Speaker Verification Integrating Contrastive Loss 5336
Xinmei Su, Xiang Xie, Fengrun Zhang, Chenguang Hu

Spoofing Attacker Also Benefits from Self-Supervised Pretrained Model..... 5346
Aoi Ito, Shota Horiguchi

Label Aware Speech Representation Learning for Language Identification..... 5351
Shikhar Vashishth, Shikhar Bharadwaj, Sriram Ganapathy, Ankur Bapna, Min Ma, Wei Han, Vera Axelrod, Partha Talukdar

Exploring the Impact of Back-End Network on Wav2vec 2.0 for Dialect Identification 5356
Qibao Luo, Ruohua Zhou

Improving Speaker Verification with Self-Pretrained Transformer Models 5361
Junyi Peng, Oldrich Plchot, Themos Stafylakis, Ladislav Mosner, Lukáš Burget, Jan "Honza" Cernocký

Handling the Alignment for Wake Word Detection: A Comparison Between Alignment-Based, Alignment-Free and Hybrid Approaches..... 5366
Vinicius Ribeiro, Yiteng Huang, Yuan Shangguan, Zhaojun Yang, Li Wan, Ming Sun

ANALYSIS OF SPEECH AND AUDIO SIGNALS 4

What Do Self-Supervised Speech Representations Encode? An Analysis of Languages, Varieties, Speaking Styles and Speakers 5371
Julian Linke, Mate Kadar, Gergely Dosinszky, Peter Mihajlik, Gernot Kubin, Barbara Schuppler

A Compressed Synthetic Speech Detection Method with Compression Feature Embedding 5376
Jinghong Zhang, Xiaowei Yi, Xianfeng Zhao

Outlier-Aware Inlier Modeling and Multi-Scale Scoring for Anomalous Sound Detection Via Multitask Learning 5381
Yucong Zhang, Suo Hongbin, Yulong Wan, Ming Li

MOSLight: A Lightweight Data-Efficient System for Non-Intrusive Speech Quality Assessment..... 5386
Zitong Li, Wei Li

A Multi-Scale Attentive Transformer for Multi-Instrument Symbolic Music Generation..... 5391
Xipin Wei, Junhui Chen, Zirui Zheng, Li Guo, Lantian Li, Dong Wang

MTANet: Multi-Band Time-Frequency Attention Network for Singing Melody Extraction from Polyphonic Music..... 5396
Yuan Gao, Ying Hu, Liusong Wang, Hao Huang, Liang He

Xiaoicesing 2: A High-Fidelity Singing Voice Synthesizer Based on Generative Adversarial Network..... 5401
Wang Chunhui, Chang Zeng, Xing He

Do Vocal Breath Sounds Encode Gender Cues for Automatic Gender Classification?..... 5406
Mohammad Shaique Solanki, Ashutosh Bharadwaj, Jeevan Kylash, Prasanta Kumar Ghosh

Automatic Exploration of Optimal Data Processing Operations for Sound Data Augmentation Using Improved Differentiable Automatic Data Augmentation 5411
Toki Sugiura, Hiromitsu Nishizaki

A Snoring Sound Dataset for Body Position Recognition: Collection, Annotation, and Analysis 5416
Li Xiao, Xiuping Yang, Xinhong Li, Weiping Tu, Xiong Chen, Weiyan Yi, Jie Lin, Yuhong Yang, Yanzhen Ren

RMVPE: A Robust Model for Vocal Pitch Estimation in Polyphonic Music 5421
Haojie Wei, Xueke Cao, Tangpeng Dan, Yueguo Chen

Spatialization Quality Metric for Binaural Speech..... 5426
Pranay Manocha, Israel Dejene Gebru, Anurag Kumar, Dejan Markovic, Alexander Richard

AsthmaSCELNet: A Lightweight Supervised Contrastive Embedding Learning Framework for Asthma Classification Using Lung Sounds 5431
Arka Roy, Udit Satija

Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification..... 5436
Sangmin Bae, June-Woo Kim, Won-Yang Cho, Hyerim Baek, Soyoun Son, Byungjo Lee, Changwan Ha, Kyongpil Tae, Sungnyun Kim, Se-Young Yun

Remote Assessment for ALS Using Multimodal Dialog Agents: Data Quality, Feasibility and Task Compliance 5441
Vanessa Richter, Michael Neumann, Jordan Green, Brian Richburg, Oliver Roesler, Hardik Kothare, Vikram Ramanarayanan

AudioToken: Adaptation of Text-Conditioned Diffusion Models for Audio-To-Image Generation 5446
Guy Yariv, Itai Gat, Lior Wolf, Yossi Adi, Idan Schwartz

Obstructive Sleep Apnea Screening with Breathing Sounds and Respiratory Effort: A Multimodal Deep Learning Approach 5451
Hector E. Romero, Ning Ma, Guy J. Brown, Sam Johnson

Investigation of Music Emotion Recognition Based on Segmented Semi-Supervised Learning 5456
Yifu Sun, Xulong Zhang, Jianzong Wang, Ning Cheng, Kaiyu Hu, Jing Xiao

SPEECH SYNTHESIS: MULTILINGUALITY; EVALUATION

The Effects of Input Type and Pronunciation Dictionary Usage in Transfer Learning for Low-Resource Text-To-Speech 5461
Phat Do, Matt Coler, Jelske Dijkstra, Esther Klabbers

Resource-Efficient Fine-Tuning Strategies for Automatic MOS Prediction in Text-To-Speech for Low-Resource Languages 5466
Phat Do, Matt Coler, Jelske Dijkstra, Esther Klabbers

Robust Feature Decoupling in Voice Conversion by Using Locality-Based Instance Normalization 5471
Yewei Gu, Xianfeng Zhao, Xiaowei Yi

Zero-Shot Accent Conversion Using Pseudo Siamese Disentanglement Network 5476
Dongya Jia, Qiao Tian, Kainan Peng, Jiaxin Li, Yuanzhe Chen, Mingbo Ma, Yuping Wang, Yuxuan Wang

Automatic Evaluation of Turn-Taking Cues in Conversational Speech Synthesis 5481
Erik Ekstedt, Siyang Wang, Éva Székely, Joakim Gustafson, Gabriel Skantze

GenerTTS: Pronunciation Disentanglement for Timbre and Style Generalization in Cross-Lingual Text-To-Speech 5486
Yahuan Cong, Haoyu Zhang, Haopeng Lin, Shichao Liu, Chunfeng Wang, Yi Ren, Xiang Yin, Zejun Ma

Analysis of Mean Opinion Scores in Subjective Evaluation of Synthetic Speech Based on Tail Probabilities 5491
Yusuke Yasuda, Tomoki Toda

LibriTTS-R: A Restored Multi-Speaker Text-To-Speech Corpus 5496
Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, Ankur Bapna

UniFLG: Unified Facial Landmark Generator from Text Or Speech 5501
Kentaro Mitsui, Yukiya Hono, Kei Sawada

XPhoneBERT: A Pre-Trained Multilingual Model for Phoneme Representations for Text-To-Speech 5506
Linh The Nguyen, Thinh Pham, Dat Quoc Nguyen

ClArTTS: An Open-Source Classical Arabic Text-To-Speech Corpus....5511
Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'Em Mohammad Shatnawi, Hanan Aldarmaki

Diffusion-Based Accent Modelling in Speech Synthesis 5516
Kamil Deja, Georgi Tinchev, Marta Czarnowska, Marius Cotescu, Jasha Droppo

Multilingual Text-To-Speech Synthesis for Turkic Languages Using Transliteration.... 5521
Rustem Yeshpanov, Saida Mussakhojayeva, Yerbolat Khassanov

CVTE-Poly: A New Benchmark for Chinese Polyphone Disambiguation.... 5526
Siheng Zhang, Xingjun Tan, Yanqiang Lei, Xianxiang Wang, Zhizhong Zhang, Yuan Xie

Improving Bilingual TTS Using Language and Phonology Embedding with Embedding Strength Modulator.... 5531
Fengyu Yang, Jian Luan, Meng Meng, Yujun Wang

High-Quality Automatic Voice Over with Accurate Alignment: Supervision Through Self-Supervised Discrete Speech Units.... 5536
Junchen Lu, Berrak Sisman, Mingyang Zhang, Haizhou Li

PronScribe: Highly Accurate Multimodal Phonemic Transcription from Speech and Text 5541
Yang Yu, Matthew Perez, Ankur Bapna, Fadi Haik, Siamak Tazari, Yu Zhang

Expressive Machine Dubbing Through Phrase-Level Cross-Lingual Prosody Transfer 5546
Jakub Swiatkowski, Duo Wang, Mikolaj Babianski, Giuseppe Coccia, Patrick Lumban Tobing, Ravichander Vipperla, Viacheslav Klimkov, Vincent Pollet

Why We Should Report the Details in Subjective Evaluation of TTS More Rigorously 5551
Cheng-Han Chiang, Wei-Ping Huang, Hung-Yi Lee

Speaker-Independent Neural Formant Synthesis.... 5556
Pablo Pérez Zarazaga, Zofia Malisz, Gustav Eje Henter, Lauri Juvela

CALLS: Japanese Empathetic Dialogue Speech Corpus of Complaint Handling and Attentive Listening in Customer Center.... 5561
Yuki Saito, Eiji Iimori, Shinnosuke Takamichi, Kentaro Tachibana, Hiroshi Saruwatari

SASPEECH: A Hebrew Single Speaker Dataset for Text to Speech and Voice Conversion 5566
Orian Sharoni, Roee Shenberg, Erica Cooper

Author Index