IEEE/ACM TRANSACTIONS ON

AUDIO, SPEECH, AND LANGUAGE PROCESSING

A PUBLICATION OF THE IEEE SIGNAL PROCESSING SOCIETY

IEEE Signal Processing Society

Association for Computing Machinery

www.signalprocessingsociety.org

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

PubMed

MEDLINE U.S. National Library of Medicine

2021 VOLUME 29 ITASFA (ISSN 2329-9290)

REGULAR PAPERS

Hierarchical Regulated Iterative Network for Joint Task of Music Detection and Music Relative Loudness Estimation . . . . . *B. Jia, J. Lv, X. Peng, Y. Chen, and S. Yang* 1

Novel Architectures for Unsupervised Information Bottleneck Based Speaker Diarization of Meetings . . . . . *N. Dawalatabad, S. Madikeri, C. C. Sekhar, and H. A. Murthy* 14

Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection . . . . . *M. Yousefi and J. H. L. Hansen* 28

A Deep Adaptation Network for Speech Enhancement: Combining a Relativistic Discriminator With Multi-Kernel Maximum Mean Discrepancy . . . . . *J. Cheng, R. Liang, Z. Liang, L. Zhao, C. Huang, and B. Schuller* 41

Comparison of Artificial Neural Network Types for Infant Vocalization Classification . . . . . *F. Anders, M. Hlawitschka, and M. Fuchs* 54

Harmonic-Temporal Factor Decomposition for Unsupervised Monaural Separation of Harmonic Sounds . . . . . *T. Nakamura and H. Kameoka* 68

Computation of Spherical Harmonic Representations of Source Directivity Based on the Finite-Distance Signature . . . . . *J. Ahrens and S. Bilbao* 83

Improving Automatic Speech Recognition and Speech Translation via Word Embedding Prediction . . . . . *S.-P. Chuang, A. H. Liu, T.-W. Sung, and H.-yi Lee* 93

A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement . . . . . *L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee* 106

Passive Geometry Calibration for Microphone Arrays Based on Distributed Damped Newton Optimization . . . . . *D. Hu, Z. Chen, and F. Yin* 118

An Overview of Voice Conversion and Its Challenges: From Statistical Modeling to Deep Learning . . . . . *B. Sisman, J. Yamagishi, S. King, and H. Li* 132

(Contents Continued on Page i)

IEEE

(Contents Continued from Front Cover)

Steering Study of Linear Differential Microphone Arrays . . . . . *J. Jin, G. Huang, X. Wang, J. Chen, J. Benesty, and I. Cohen* 158
Proportionate Adaptive Filtering Algorithms Derived Using an Iterative Reweighting Framework . . . . . *C.-H. Lee, B. D. Rao, and H. Garudadri* 171
A Novel Approach for Improved Noise Reduction Performance in Feed-Forward Active Noise Control Systems With (Loudspeaker) Saturation Non-Linearity in the Secondary Path . . . . . *S. Ahmed, M. Tufail, M. Rehan, T. Abbas, and A. Majid* 187
Gated Recurrent Fusion With Joint Training Framework for Robust End-to-End Speech Recognition . . . . . *C. Fan, J. Yi, J. Tao, Z. Tian, B. Liu, and Z. Wen* 198
Speech Intelligibility Prediction Using Spectro-Temporal Modulation Analysis . . . . . *A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty* 210
On the Design of Sparse Arrays With Frequency-Invariant Beam Pattern . . . . . *P. L. Son* 226
A Room Compensation Method by Modification of Reverberant Audio Objects . . . . . *D. Menzies, P. Coleman, and F. M. Fazi* 239
Multiple Source Direction of Arrival Estimations Using Relative Sound Pressure Based MUSIC . . . . . *Y. Hu, T. D. Abhayapala, and P. N. Samarasinghe* 253
The Temporal Limits Encoder as a Sound Coding Strategy for Bilateral Cochlear Implants . . . . . *A. Kan and Q. Meng* 265
Exploiting Morphological and Phonological Features to Improve Prosodic Phrasing for Mongolian Speech Synthesis . . . . . *R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li* 274
Multiple Circular Arrays of Vector Sensors for Real-Time Sound Field Analysis . . . . . *F. Ma, T. D. Abhayapala, and W. Zhang* 286
Robust Sound Source Tracking Using SRP-PHAT and 3D Convolutional Neural Networks . . . . . *D. Diaz-Guerra, A. Miguel, and J. R. Beltran* 300
Directly Comparing the Listening Strategies of Humans and Machines . . . . . *V. A. Trinh and M. Mandel* 312
Auxiliary Networks for Joint Speaker Adaptation and Speaker Change Detection . . . . . *L. Sari, M. Hasegawa-Johnson, and S. Thomas* 324
Multiple Acoustic Source Localization in Microphone Array Networks . . . . . *J. Yang, X. Zhong, W. Chen, and W. Wang* 334
Tackling Perception Bias in Unsupervised Phoneme Discovery Using DPGMM-RNN Hybrid Model and Functional Load . . . . . *B. Wu, S. Sakti, J. Zhang, and S. Nakamura* 348
Fast Generation of Sound Zones Using Variable Span Trade-Off Filters in the DFT-Domain . . . . . *T. Lee, L. Shi, J. K. Nielsen, and M. G. Christensen* 363
Multi-Source DOA Estimation in Reverberant Environments by Jointing Detection and Modeling of Time-Frequency Points . . . . . *M. Jia, Y. Wu, C. Bao, and C. Ritz* 379
Speech Enhancement Based on Modulation-Domain Parametric Multichannel Kalman Filtering . . . . . *W. Xue, A. H. Moore, M. Brookes, and P. A. Naylor* 393
A Knowledge Graph Embedding Approach for Metaphor Processing . . . . . *W. Song, J. Guo, R. Fu, T. Liu, and L. Liu* 406
Estimation Reliability Function Assisted Sound Source Localization With Enhanced Steering Vector Phase Difference . . . . . *L. Cheng, X. Sun, D. Yao, J. Li, and Y. Yan* 421
Room Acoustical Parameter Estimation From Room Impulse Responses Using Deep Neural Networks . . . . . *W. Yu and W. B. Kleijn* 436
Affine Projection Algorithm Over Acoustic Sensor Networks for Active Noise Control . . . . . *M. Ferrer, M. de Diego, G. Piñero, and A. Gonzalez* 448
Performance Analysis of the Extended Binaural MVDR Beamformer With Partial Noise Estimation . . . . . *N. Gößling, D. Marquardt, and S. Doclo* 462
Ensemble Bag-of-Audio-Words Representation Improves Paralinguistic Classification Accuracy . . . . . *G. Gosztolya and R. Busa-Fekete* 477
Room Impulse Response Reshaping and Crosstalk Cancellation Using Convex Optimization . . . . . *A. Mertins, M. Maass, and F. Katzberg* 489
Investigating Typed Syntactic Dependencies for Targeted Sentiment Classification Using Graph Attention Neural Network . . . . . *X. Bai, P. Liu, and Y. Zhang* 503
Speech Enhancement via Attention Masking Network (SEAMNET): An End-to-End System for Joint Suppression of Noise and Reverberation . . . . . *B. J. Borgström and M. S. Brandstein* 515
Voice Jitter Estimation Using High-Order Synchrosqueezing Operators . . . . . *J. M. Miramont, M. A. Colominas, and G. Schlotthauer* 527

(Contents Continued on Page ii)

(Contents Continued from Page i)

Speaker Separation Using Speaker Inventories and Estimated Speech ..... *P. Wang, Z. Chen, D. L. Wang, J. Li, and Y. Gong* 537
On the Distribution of Speaker Verification Scores: Generative Models for Unsupervised Calibration ..... *S. Cumani* 547
Acoustic Measure of Vocal Strain Based on Glottal Airflow Periodicity ..... *Y.-R. Chien and J. Guanason* 563
RARS: Recognition of Audio Recording Source Based on Residual Neural Network ..... *X. Shen, X. Shao, Q. Ge, and L. Liu* 575
Learning to Generate Explainable Plots for Neural Story Generation ..... *G. Chen, Y. Liu, H. Luan, M. Zhang, Q. Liu, and M. Sun* 585
A New Class of Differential Beamformers ..... *W. Yang, J. Benesty, G. Huang, and J. Chen* 594
Multichannel Blind Source Separation Based on Evanescent-Region-Aware Non-Negative Tensor Factorization in Spherical Harmonic Domain ..... *Y. Mitsufuji, N. Takamune, S. Koyama, and H. Saruwatari* 607
Robust Constrained MFMVDR Filters for Single-Channel Speech Enhancement Based on Spherical Uncertainty Set ..... *D. Fischer and S. Doclo* 618
Differential Beamforming From the Beampattern Factorization Perspective ..... *X. Zhao, J. Benesty, J. Chen, and G. Huang* 632
Preordering Encoding on Transformer for Translation ..... *Y. Kawara, C. Chu, and Y. Arase* 644
Many-to-Many Voice Transformer Network ..... *H. Kameoka, W.-C. Huang, K. Tanaka, T. Kaneko, N. Hojo, and T. Toda* 656
A Study on Reference Microphone Selection for Multi-Microphone Speech Enhancement ..... *J. Zhang, H. Chen, L.-R. Dai, and R. C. Hendriks* 671
Overview and Evaluation of Sound Event Localization and Detection in DCASE 2019 ..... *A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen* 684
Listening Enhancement in Noisy Environments: Solutions in Time and Frequency Domain ..... *M. Niermann and P. Vary* 699
Gated Recurrent Context: Softmax-Free Attention for Online Encoder-Decoder Speech Recognition ..... *H. Lee, W. H. Kang, S. J. Cheon, H. Kim, and N. S. Kim* 710
On Improved Training of CNN for Acoustic Source Localisation ..... *E. Vargas, J. R. Hopgood, K. Brown, and K. Subr* 720
Deep Normalization for Speaker Vectors ..... *Y. Cai, L. Li, A. Abel, X. Zhu, and D. Wang* 733
Pretraining Techniques for Sequence-to-Sequence Voice Conversion ..... *W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda* 745
Temporal Dynamics of Workplace Acoustic Scenes: Egocentric Analysis and Prediction ..... *A. Jati, A. Nadarajan, R. Peri, K. Mundnich, T. Feng, B. Girault, and S. Narayanan* 756
Modeling Future Cost for Neural Machine Translation ..... *C. Duan, K. Chen, R. Wang, M. Utiyama, E. Sumita, C. Zhu, and T. Zhao* 770
Adaptive Convolution for Semantic Role Labeling ..... *K. Munir, H. Zhao, and Z. Li* 782
Quasi-Periodic Parallel WaveGAN: A Non-Autoregressive Raw Waveform Generative Model With Pitch-Dependent Dilated Convolution Neural Network ..... *Y.-C. Wu, T. Hayashi, T. Okamoto, H. Kawai, and T. Toda* 792
Evolving Multi-Resolution Pooling CNN for Monaural Singing Voice Separation ..... *W. Yuan, B. Dong, S. Wang, M. Unoki, and W. Wang* 807
Generation of Personal Sound Zones With Physical Meaningful Constraints and Conjugate Gradient Method ..... *L. Shi, T. Lee, L. Zhang, J. K. Nielsen, and M. G. Christensen* 823
On the Robustness of the Superdirective Beamformer ..... *X. Chen, J. Benesty, G. Huang, and J. Chen* 838
Generating Images From Spoken Descriptions ..... *X. Wang, T. Qiao, J. Zhu, A. Hanjalic, and O. Scharenborg* 850
Domain-Aware Dialogue State Tracker for Multi-Domain Dialogue Systems ..... *V. Balaraman and B. Magnini* 866
Exemplar-Based Emotive Speech Synthesis ..... *X. Wu, Y. Cao, H. Lu, S. Liu, S. Kang, Z. Wu, X. Liu, and H. Meng* 874
Towards Duration Robust Weakly Supervised Sound Event Detection ..... *H. Dinkel, M. Wu, and K. Yu* 887
Binaural Reproduction Based on Bilateral Ambisonics and Ear-Aligned HRTFs ..... *Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely* 901
Synthesis and Analysis-By-Synthesis of Modulated Diplophonic Glottal Area Waveforms ..... *P. Aichinger and F. Pernkopf* 914
Analysis and Calibration of Lombard Effect and Whisper for Speaker Recognition ..... *F. Kelly and J. H.L. Hansen* 927
Lyric or Dramatic - Vibrato Analysis for Voice Type Classification in Professional Opera Singers ..... *M. Müller, T. Schulz, T. Ermakova, and P. P. Caffier* 943
Incorporating Wireless Communication Parameters Into the E-Model Algorithm ..... *D. Z. Rodríguez, D. Carrillo, M. A. Ramírez, P. H. J. Nardelli, and S. Möller* 956

(Contents Continued on Page iii)

(Contents Continued from Page ii)

Non-Linear-Echo Based Anti-Collusion Mechanism for Audio Signals . . . . . *T. Zong, Y. Xiang, I. Natgunanathan, L. Gao, G. Hua, and W. Zhou* 969
CTNet: Conversational Transformer Network for Emotion Recognition . . . . . *Z. Lian, B. Liu, and J. Tao* 985
Neural Machine Translation With Explicit Phrase Alignment . . . . . *J. Zhang, H. Luan, M. Sun, F. Zhai, J. Xu, and Y. Liu* 1001
Cognitive Load Estimation From Speech Commands to Simulated Aircraft . . . . . *M. Vukovic, M. Stolar, and M. Lech* 1011
Geometry Calibration for Acoustic Transceiver Networks Based on Network Newton Distributed Optimization . . . . . *D. Hu, Z. Chen, and F. Yin* 1023
Perceptual-Similarity-Aware Deep Speaker Representation Learning for Multi-Speaker Generative Modeling . . . . . *Y. Saito, S. Takamichi, and H. Saruwatari* 1033
Vocal Tract Length Estimation Using Accumulated Means of Formants and Its Effects on Speaker-Normalization . . . . . *T. Sakata, N. Ikeda, Y. Ueda, and A. Watanabe* 1049
Modified Magnitude-Phase Spectrum Information for Spoofing Detection . . . . . *J. Yang, H. Wang, R. K. Das, and Y. Qian* 1065
Audio-Visual Deep Neural Network for Robust Person Verification . . . . . *Y. Qian, Z. Chen, and S. Wang* 1079
Deep Selective Memory Network With Selective Attention and Inter-Aspect Modeling for Aspect Level Sentiment Classification . . . . . *P. Lin, M. Yang, and J. Lai* 1093
Improved Acoustic Word Embeddings for Zero-Resource Languages Using Multilingual Transfer . . . . . *H. Kamper, Y. Matusevych, and S. Goldwater* 1107
Audio-Based Piano Performance Evaluation for Beginners With Convolutional Neural Network and Attention Mechanism . . . . . *W. Wang, J. Pan, H. Yi, Z. Song, and M. Li* 1119
Quasi-Periodic WaveNet: An Autoregressive Raw Waveform Generative Model With Pitch-Dependent Dilated Convolution Neural Network . . . . . *Y.-C. Wu, T. Hayashi, P. L. Tobing, K. Kobayashi, and T. Toda* 1134
Late-Reverberation Synthesis Using Interleaved Velvet-Noise Sequences . . . . . *V. Välimäki and K. Prawda* 1149
Multi-Turn Dialogue Reading Comprehension With Pivot Turns and Knowledge . . . . . *Z. Zhang, J. Li, and H. Zhao* 1161
Sparsity-Based Audio Declipping Methods: Selected Overview, New Algorithms, and Large-Scale Evaluation . . . . . *C. Gaultier, S. Kitić, R. Gribonval, and N. Bertin* 1174
Mixed Source Sound Field Translation for Virtual Binaural Application With Perceptual Validation . . . . . *L. Birnie, T. Abhayapala, V. Tourbabin, and P. Samarasinghe* 1188
Meta-Learning With Latent Space Clustering in Generative Adversarial Network for Speaker Diarization . . . . . *M. Pal, M. Kumar, R. Peri, T. J. Park, S. H. Kim, C. Lord, S. Bishop, and S. Narayanan* 1204
Sensor Selection for Relative Acoustic Transfer Function Steered Linearly-Constrained Beamformers . . . . . *J. Zhang, J. Du, and L.-R. Dai* 1220
Zero-Shot Audio Classification Via Semantic Embeddings . . . . . *H. Xie and T. Virtanen* 1233
Phoneme-Unit-Specific Time-Delay Neural Network for Speaker Verification . . . . . *X. Chen and C. Bao* 1243
Optimal Output-Constrained Active Noise Control Based on Inverse Adaptive Modeling Leak Factor Estimate . . . . . *D. Shi, W.-S. Gan, B. Lam, S. Wen, and X. Shen* 1256
Dense CNN With Self-Attention for Time-Domain Speech Enhancement . . . . . *A. Pandey and D. L. Wang* 1270
Knowing Where to Leverage: Context-Aware Graph Convolutional Network With an Adaptive Fusion Layer for Contextual Spoken Language Understanding . . . . . *L. Qin, W. Che, M. Ni, Y. Li, and T. Liu* 1280
Transfer Learning From Speech Synthesis to Voice Conversion With Non-Parallel Training Data . . . . . *M. Zhang, Y. Zhou, L. Zhao, and H. Li* 1290
Neural Network Adaptation and Data Augmentation for Multi-Speaker Direction-of-Arrival Estimation . . . . . *W. He, P. Motlicek, and J.-M. Odobez* 1303
Improving Skip-Gram Embeddings Using BERT . . . . . *Y. Wang, L. Cui, and Y. Zhang* 1318
Deep Graph-Based Character-Level Chinese Dependency Parsing . . . . . *L. Wu and M. Zhang* 1329
Integrating Knowledge Into End-to-End Speech Recognition From External Text-Only Data . . . . . *Y. Bai, J. Yi, J. Tao, Z. Wen, Z. Tian, and S. Zhang* 1340
Convolutional Maximum-Likelihood Distortionless Response Beamforming With Steering Vector Estimation for Robust Speech Recognition . . . . . *B. J. Cho and H.-M. Park* 1352
An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation . . . . . *D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen* 1368
On the Design of Differential Kronecker Product Beamformers . . . . . *G. Itzhak, J. Benesty, and I. Cohen* 1397
Partially Matching Projection Decoding Method Evaluation Under Different Playback Conditions . . . . . *Z. Ge, L. Li, and T. Qu* 1411
Analyzing Multimodal Sentiment Via Acoustic- and Visual-LSTM With Channel-Aware Temporal Convolution Network . . . . . *S. Mai, S. Xing, and H. Hu* 1424

(Contents Continued on Page iv)

(Contents Continued from Page iii)

A Joint Model for Named Entity Recognition With Sentence-Level Entity Type Attentions . . . . . *T. Qian, M. Zhang, Y. Lou, and D. Hua* 1438
Ambisonic Signal Processing DNNs Guaranteeing Rotation, Scale and Time Translation Equivariance . . . . . *R. Sato, K. Niwa, and K. Kobayashi* 1449
Iterative Echo Labeling Algorithm With Convex Hull Expansion for Room Geometry Estimation . . . . . *S. Park and J.-W. Choi* 1463
Overlapping Speaker Segmentation Using Multiple Hypothesis Tracking of Fundamental Frequency . . . . . *A. O. T. Hogg, C. Evers, A. H. Moore, and P. A. Naylor* 1479
Controlling Elevation and Azimuth Beamwidths With Concentric Circular Microphone Arrays . . . . . *R. Sharma, I. Cohen, and B. Berdugo* 1491
A Multiple-Integration Encoder for Multi-Turn Text-to-SQL Semantic Parsing . . . . . *R.-Z. Wang, Z.-H. Ling, J.-B. Zhou, and Y. Hu* 1503
Bayesian Learning of LF-MMI Trained Time Delay Neural Networks for Speech Recognition . . . . . *S. Hu, X. Xie, S. Liu, J. Yu, Z. Ye, M. Geng, X. Liu, and H. Meng* 1514
Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence . . . . . *M. Torcoli, T. Kastner, and J. Herre* 1530
Voice Activity Detection in the Wild: A Data-Driven Approach Using Teacher-Student Training . . . . . *H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu* 1542
Detection of Multiple Steganography Methods in Compressed Speech Based on Code Element Embedding, Bi-LSTM and CNN With Attention Mechanisms . . . . . *S. Li, J. Wang, P. Liu, M. Wei, and Q. Yan* 1556
Deformable Self-Attention for Text Classification . . . . . *Q. Ma, J. Yan, Z. Lin, L. Yu, and Z. Chen* 1570
Extracting and Predicting Word-Level Style Variations for Speech Synthesis . . . . . *Y.-J. Zhang and Z.-H. Ling* 1582
Exploiting Temporal Context in CNN Based Multisource DOA Estimation . . . . . *A. Bohlender, A. Spriet, W. Tirry, and N. Madhu* 1594
Determined BSS Based on Time-Frequency Masking and Its Application to Harmonic Vector Analysis . . . . . *K. Yatabe and D. Kitamura* 1609
TutorNet: Towards Flexible Knowledge Distillation for End-to-End Speech Recognition . . . . . *J. W. Yoon, H. Lee, H. Y. Kim, W. I. Cho, and N. S. Kim* 1626
Self-Supervised Representation Learning With Path Integral Clustering for Speaker Diarization . . . . . *P. Singh and S. Ganapathy* 1639
A Graph-to-Sequence Learning Framework for Summarizing Opinionated Texts . . . . . *P. Wei, J. Zhao, and W. Mao* 1650
Near-Field Superdirectivity: An Analytical Perspective . . . . . *D. Y. Levin, S. Markovich-Golan, and S. Gannot* 1661
Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations . . . . . *J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen* 1675
Time-Domain Audio Source Separation With Neural Networks Based on Multiresolution Analysis . . . . . *T. Nakamura, S. Kozuka, and H. Saruwatari* 1687
FSPRM: A Feature Subsequence Based Probability Representation Model for Chinese Word Embedding . . . . . *Y. Zhang, Y. Liu, J. Zhu, and X. Wu* 1702
Any-to-Many Voice Conversion With Location-Relative Sequence-to-Sequence Modeling . . . . . *S. Liu, Y. Cao, D. Wang, X. Wu, X. Liu, and H. Meng* 1717
An Optimal Envelope-Based Noise Reduction Method for Cochlear Implants: An Upper Bound Performance Investigation . . . . . *R. A. Chiea, M. H. Costa, and J. A. Cordioli* 1729
Adaptive Adapters: An Efficient Way to Incorporate BERT Into Neural Machine Translation . . . . . *J. Guo, Z. Zhang, L. Xu, B. Chen, and E. Chen* 1740
Group Communication With Context Codec for Lightweight Source Separation . . . . . *Y. Luo, C. Han, and N. Mesgarani* 1752
Hierarchical Neighbor Propagation With Bidirectional Graph Attention Network for Relation Prediction . . . . . *Z. Xie, R. Zhu, J. Liu, G. Zhou, and J. X. Huang* 1762
Beamforming with Cube Microphone Arrays Via Kronecker Product Decompositions . . . . . *X. Wang, J. Benesty, J. Chen, G. Huang, and I. Cohen* 1774
Towards Model Compression for Deep Learning Based Speech Enhancement . . . . . *K. Tan and D. L. Wang* 1785
Nonlinear Spatial Filtering in Multichannel Speech Enhancement . . . . . *K. Tesch and T. Gerkmann* 1795
Expressive TTS Training With Frame and Style Reconstruction Loss . . . . . *R. Liu, B. Sisman, G. Gao, and H. Li* 1806
Chinese Lexical Simplification . . . . . *J. Qiang, X. Lu, Y. Li, Y. Yuan, and X. W.* 1819
Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement . . . . . *A. Li, W. Liu, C. Zheng, C. Fan, and X. Li* 1829

(Contents Continued on Page v)

(Contents Continued from Page iv)

Weighted Orthogonal Vector Rejection Method for Loudspeaker-Based Binaural Audio Reproduction ..... E. C. Hamdan and F. M. Fazi 1844
Deep Learning Based Real-Time Speech Enhancement for Dual-Microphone Mobile Phones ..... K. Tan, X. Zhang, and D. L. Wang 1853
Indoor Multi-Speaker Localization Based on Bayesian Nonparametrics in the Circular Harmonic Domain ..... K. SongGong, H. Chen, and W. Wang 1864
Double-Cross-Correlation Processing for Blind Sampling-Rate and Time-Offset Estimation ..... A. Chinaev, P. Thüne, and G. Enzner 1881
Fast End-to-End Speech Recognition Via Non-Autoregressive Models and Cross-Modal Knowledge Transferring From BERT ..... Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang 1897
Multimodal Representations for Synchronized Speech and Real-Time MRI Video Processing ..... Ö.ykü. D. Köse and M. Saraçlar 1912
The Detection of Parkinson's Disease From Speech Using Voice Source Information ..... N. P. Narendra, B. Schuller, and P. Alku 1925
SNR-Based Features and Diverse Training Data for Robust DNN-Based Speech Enhancement ..... R. Rehr and T. Gerkmann 1937
A Joint Diagonalization Based Efficient Approach to Underdetermined Blind Audio Source Separation Using the Multichannel Wiener Filter ..... N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani 1950
Second-Order Semantic Role Labeling With Global Structural Refinement ..... H. Fei, S. Wu, Y. Ren, and D. Ji 1966
F0 Perturbation Due to Articulatory Movements: Filtering, Characterization and Applications ..... H. M. Torres, M. Güemes, J. A. Gurlekian, and D. A. Evin 1977
Receptive Field Regularization Techniques for Audio Classification and Tagging With Deep Convolutional Neural Networks ..... K. Koutini, H. Eghbal-zadeh, and G. Widmer 1987
Multi-microphone Complex Spectral Mapping for Utterance-wise and Continuous Speech Separation ..... Z.-Q. Wang, P. Wang, and D. L. Wang 2001
Relation Extraction in Dialogues: A Deep Learning Model Based on the Generality and Specialty of Dialogue Text ..... M. Zhou, D. Ji, and F. Li 2015
Domain-Shift Conditioning Using Adaptable Filtering Via Hierarchical Embeddings for Robust Chinese Spell Check ..... M. Nguyen, G. H. Ngo, and N. F. Chen 2027
The Effect of Partial Time-Frequency Masking of the Direct Sound on the Perception of Reverberant Speech ..... L. Madmoni, S. Tibor, I. Nelken, and B. Rafaely 2037
Corpus-Aware Graph Aggregation Network for Sequence Labeling ..... H. Chen, Q. Ma, L. Yu, Z. Lin, and J. Yan 2048
Towards Robust Speech Super-Resolution ..... H. Wang and D. L. Wang 2058
Audio-Visual Multi-Channel Integration and Recognition of Overlapped Speech ..... J. Yu, S.-X. Zhang, B. Wu, S. Liu, S. Hu, M. Geng, X. Liu, H. Meng, and D. Yu 2067
Conditioned Source Separation for Musical Instrument Performances ..... O. Slizovskaia, G. Haro, and E. Gómez 2083
Bayesian Learning for Deep Neural Network Adaptation ..... X. Xie, X. Liu, T. Lee, and L. Wang 2096
Nearest Kronecker Product Decomposition Based Linear-in-The-Parameters Nonlinear Filters ..... S. S. Bhattacharjee and N. V. George 2111
A Multi-Agent Communication Based Model for Nested Named Entity Recognition ..... C. Li, G. Wang, J. Cao, and Y. Cai 2123
Blind Separation for Multiple Moving Sources With Labeled Random Finite Sets .... J. Ong, B. T. Vo, and S. Nordholm 2137
PROTOTYPE-TO-STYLE: Dialogue Generation With Style-Aware Editing on Retrieval Memory ..... Y. Su, Y. Wang, D. Cai, S. Baker, A. Korhonen, and N. Collier 2152
A Wave Digital Newton-Raphson Method for Virtual Analog Modeling of Audio Circuits with Multiple One-Port Nonlinearities ..... A. Bernardini, E. Bozzo, F. Fontana, and A. Sarti 2162
Proximal Normalized Subband Adaptive Filtering for Acoustic Echo Cancellation ..... G. Guo, Y. Yu, R. C. de Lamare, Z. Zheng, L. Lu, and Q. Cai 2174
Audibility of Group-Delay Equalization ..... J. Liski, A. Mäkivirta, and V. Välimäki 2189
Corruption Is Not All Bad: Incorporating Discourse Structure Into Pre-Training via Corruption for Essay Scoring ..... F. S. Mim, N. Inoue, P. Reisert, H. Ouchi, and K. Inui 2202
Graph-Based Clustering of Dolphin Whistles ..... D. Kipnis and R. Diamant 2216
Language-Independent Approach for Automatic Computation of Vowel Articulation Features in Dysarthric Speech Assessment ..... Y. Liu, N. Penttilä, T. Ihalainen, J. Lintula, R. Convey, and O. Räsänen 2228
Impulsive Noise Detection for Speech Enhancement in HHT Domain ..... C. Medina, R. Coelho, and L. Zão 2244

(Contents Continued on Page vi)

(Contents Continued from Page v)

A Novel Loss Function and Training Strategy for Noise-Robust Keyword Spotting . . . . . *I. López-Espejo, Z.-H. Tan, and J. Jensen* 2254
Recent Progress in the CUHK Dysarthric Speech Recognition System . . . . . *S. Liu, M. Geng, S. Hu, X. Xie, M. Cui, J. Yu, X. Liu, and H. Meng* 2267
Desynchronization Attacks Resilient Watermarking Method Based on Frequency Singular Value Coefficient Modification . . . . . *J. Zhao, T. Zong, Y. Xiang, L. Gao, W. Zhou, and G. Beliakov* 2282
Sparse Representations With Legendre Kernels for DOA Estimation and Acoustic Source Separation . . . . . *M. B. Çöteli and H. Hacıhabiboğlu* 2296
DNN-Based Mask Estimation for Distributed Speech Enhancement in Spatially Unconstrained Microphone Arrays . . . . . *N. Furnon, R. Serizel, S. Essid, and I. Illina* 2310
Hybrid Speech and Text Analysis Methods for Speaker Change Detection . . . . . *O. H. Anidjar, I. Lapidot, C. Hajaj, A. Dvir, and I. G.* 2324
Multi-Task Sequence Tagging for Emotion-Cause Pair Extraction Via Tag Distribution Refinement . . . . . *C. Fan, C. Yuan, L. Gui, Y. Zhang, and R. Xu* 2339
TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech . . . . . *A. T. Liu, S.-W. Li, and H.-yi Lee* 2351
Converting Foreign Accent Speech Without a Reference . . . . . *G. Zhao, S. Ding, and R. Gutierrez-Osuna* 2367
Phoneme Level Lyrics Alignment and Text-Informed Singing Voice Separation . . . . . *K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau* 2382
High-Order Pair-Wise Aspect and Opinion Terms Extraction With Edge-Enhanced Syntactic Graph Convolution . . . . . *S. Wu, H. Fei, Y. Ren, B. Li, F. Li, and D. Ji* 2396
Sentiment Time Series Calibration for Event Detection . . . . . *J. Wu, L. Shang, and X. Gao* 2407
Learning Context-Aware Convolutional Filters for Implicit Discourse Relation Classification . . . . . *K. Munir, H. Zhao, and Z. Li* 2421
Editorial: Special Issue on the Eighth Dialog System Technology Challenge . . . . . *S. Kim, H. Schulz, C. Gunasekara, C. Hori, A. Rastogi, and L. F. D'Haro* 2434
Randomly Wired Network Based on RoBERTa and Dialog History Attention for Response Selection . . . . . *B. Kim, J. Seo, and M.-W. Koo* 2437
Deep Contextualized Utterance Representations for Response Selection and Dialogue Analysis . . . . . *J.-C. Gu, T. Li, Z.-H. Ling, Q. Liu, Z. Su, Y.-P. Ruan, and X. Zhu* 2443
End-to-End Recurrent Cross-Modality Attention for Video Dialogue . . . . . *Y.-W. Chu, K.-Y. Lin, C.-C. Hsu, and L.-W. Ku* 2456
Conversational Semantic Role Labeling . . . . . *K. Xu, H. Wu, L. Song, H. Zhang, L. Song, and D. Yu* 2465
Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog . . . . . *Z. Li, Z. Li, J. Zhang, Y. Feng, and J. Zhou* 2476
GRTr: Generative-Retrieval Transformers for Data-Efficient Dialogue Domain Adaptation . . . . . *I. Shalyminov, A. Sordoni, A. Atkinson, and H. Schulz* 2484
Domain Adaptive Meta-Learning for Dialogue State Tracking . . . . . *J. Zeng, Y. Yin, Y. Liu, Y. Ge, and J. Su* 2493
D-Score: Holistic Dialogue Evaluation Without Reference . . . . . *C. Zhang, G. Lee, L. F. D'Haro, and H. Li* 2502
HDRS: Hindi Dialogue Restaurant Search Corpus for Dialogue State Tracking in Task-Oriented Environment . . . . . *S. Malviya, R. Mishra, S. K. Barnwal, and U. S. Tiwary* 2517
Overview of the Eighth Dialog System Technology Challenge: DSTC8 . . . . . *S. Kim, M. Galley, C. Gunasekara, S. Lee, A. Atkinson, B. Peng, H. Schulz, J. Gao, J. Li, M. Adada, M. Huang, L. Lastras, J. K. Kummerfeld, W. S. Lasecki, C. Hori, A. Cherian, T. K. Marks, A. Rastogi, X. Zang, S. Sunkara, and R. Gupta* 2529
Label and Context Augmentation for Response Selection at DSTC8 . . . . . *M. Jeong, S. Choi, J. Yeo, and S.-won Hwang* 2541
History Reuse and Bag-of-Words Loss for Long Summary Generation . . . . . *Q. Liu, L. Chen, Y. Yuan, and H. Wu* 2551
PhaseDCN: A Phase-Enhanced Dual-Path Dilated Convolutional Network for Single-Channel Speech Enhancement . . . . . *L. Zhang, M. Wang, Q. Zhang, X. Wang, and M. Liu* 2561
Guided Generative Adversarial Neural Network for Representation Learning and Audio Generation Using Fewer Labelled Audio Data . . . . . *K. N. Haque, R. Rana, J. Liu, J. H. L. Hansen, N. Cummins, C. Busso, and B. W. Schuller* 2575
Gamma Boltzmann Machine for Audio Modeling . . . . . *T. Nakashika and K. Yatabe* 2591
Attending From Foresight: A Novel Attention Mechanism for Neural Machine Translation . . . . . *X. Li, L. Liu, Z. Tu, G. Li, S. Shi, and M. Q.-H. Meng* 2606
Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition . . . . . *H. Zhou, J. Du, Y. Zhang, Q. Wang, Q.-F. Liu, and C.-H. Lee* 2617
Learning Cross-Lingual Mappings in Imperfectly Isomorphic Embedding Spaces . . . . . *Y. Li, K. Yu, and Y. Zhang* 2630

(Contents Continued on Page vii)

(Contents Continued from Page vi)

UnitNet: A Sequence-to-Sequence Acoustic Model for Concatenative Speech Synthesis . . . . . *X. Zhou, Z.-H. Ling, and L.-R. Dai* 2643
Multi-Tone Phase Coding of Interaural Time Difference for Sound Source Localization With Spiking Neural Networks . . . . . *Z. Pan, M. Zhang, J. Wu, J. Wang, and H. Li* 2656
FifthNet: Structured Compact Neural Networks for Automatic Chord Recognition . . . . . *K. O'Hanlon and M. B. Sandler* 2671
Estimation of Spectral Notches From Pinna Meshes: Insights From a Simple Computational Model . . . . . *S. Spagnol, R. Miccini, M. G. Onofrei, R. Unnthorsson, and S. Serafin* 2683
Target Speaker Verification With Selective Auditory Attention for Single and Multi-Talker Speech . . . . . *C. Xu, W. Rao, J. Wu, and H. Li* 2696
Minimum Processing Beamforming . . . . . *A. Zahedi, M. S. Pedersen, J. Østergaard, T. U. Christiansen, L. Bramsløw, and J. Jensen* 2710
Multichannel Iterative Noise Reduction Filters in the Short-Time-Fourier-Transform Domain Based on Kronecker Product Decomposition . . . . . *X. Wang, J. Chen, X. Chen, J. Guo, and Q. Xiang* 2725
Robust Q-Gradient Subband Adaptive Filter for Nonlinear Active Noise Control . . . . . *K.-L. Yin, Y.-F. Pu, and L. Lu* 2741
Monaural Speech Separation Using Speaker Embedding From Preliminary Separation . . . . . *J. Byun and J. W. Shin* 2753
On the Design of 3D Steerable Beamformers With Uniform Concentric Circular Microphone Arrays . . . . . *X. Zhao, G. Huang, J. Chen, and J. Benesty* 2764
A Unified Target-Oriented Sequence-to-Sequence Model for Emotion-Cause Pair Extraction . . . . . *Z. Cheng, Z. Jiang, Y. Yin, N. Li, and Q. Gu* 2779
Robust Voice Feature Selection Using Interval Type-2 Fuzzy AHP for Automated Diagnosis of Parkinson's Disease . . . . . *H. Azadi, M.-R. Akbarzadeh-T, H.-R. Kobravi, and A. Shoeibi* 2792
Sinsy: A Deep Neural Network-Based Singing Voice Synthesis System . . . . . *Y. Hono, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda* 2803
Multi-Granularity Sequence Alignment Mapping for Encoder-Decoder Based End-to-End ASR . . . . . *J. Tang, J. Zhang, Y. Song, I. McLoughlin, and L.-R. Dai* 2816
Exploiting Translation Model for Parallel Corpus Mining . . . . . *C. Leong, X. Liu, D. F. Wong, and L. S. Chao* 2829
Wavesplit: End-to-End Speech Separation by Speaker Clustering . . . . . *N. Zeghidour and D. Grangier* 2840
Learning Waveform-Based Acoustic Models Using Deep Variational Convolutional Neural Networks . . . . . *D. Oglic, Z. Cvetkovic, and P. Sollich* 2850
Privacy-Preserving Audio Classification Using Variational Information Feature Extraction . . . . . *A. Nelus and R. Martin* 2864
Recurrent Neural Networks and Acoustic Features for Frame-Level Signal-to-Noise Ratio Estimation . . . . . *H. Li, D. L. Wang, X. Zhang, and G. Gao* 2878
Generating Responses With a Given Syntactic Pattern in Chinese Dialogues . . . . . *Y. Zhou, X. Zheng, and X. Huang* 2888
Binaural Auralization of Microphone Array Room Impulse Responses Using Causal Wiener Filtering . . . . . *V. Gunnarsson and M. Sternad* 2899
Sensor Imperfection Tolerance Analysis of Robust Linear Differential Microphone Arrays . . . . . *Z. Chen, H. Chen, and Q. Tu* 2915
CSS-LM: A Contrastive Framework for Semi-Supervised Fine-Tuning of Pre-Trained Language Models . . . . . *Y. Su, X. Han, Y. Lin, Z. Zhang, Z. Liu, P. Li, J. Zhou, and M. Sun* 2930
A Causality-Constrained Frequency-Domain Least-Squares Filter Design Method for Crosstalk Cancellation . . . . . *T. Kabzinski and P. Jax* 2942
CTC-Based Learning of Chroma Features for Score–Audio Music Retrieval . . . . . *F. Zalkow and M. Müller* 2957
Multi-Branch Convolutional Macaron net for Sound Event Detection . . . . . *T. K. Chan and C. S. Chin* 2972
FluentNet: End-to-End Detection of Stuttered Speech Disfluencies With Deep Learning . . . . . *T. Kourkounakis, A. Hajavi, and A. Etemad* 2986
Multi-Metric Optimization Using Generative Adversarial Networks for Near-End Speech Intelligibility Enhancement . . . . . *H. Li and J. Yamagishi* 3000
Predict-Then-Decide: A Predictive Approach for Wait or Answer Task in Dialogue Systems . . . . . *Z. Lin, S. Cui, G. Li, X. Kang, F. Ji, F. Li, Z. Zhao, H. Chen, and Y. Zhang* 3012
Localization Based on Enhanced Low Frequency Interaural Level Difference . . . . . *M. Calis, S. van de Par, R. Heusdens, and R. C. Hendriks* 3025
Native-Nonnative Voice Conversion by Residual Warping in a Sparse, Anchor-Based Representation . . . . . *C. Liberatore* 3040
Spatial Active Noise Control Based on Kernel Interpolation of Sound Field . . . . . *S. Koyama, J. Brunnström, H. Ito, N. Ueno, and H. Saruwatari* 3052
LSBert: Lexical Simplification Based on BERT . . . . . *J. Qiang, Y. Li, Y. Zhu, Y. Yuan, Y. Shi, and X. Wu* 3064

(Contents Continued on Page viii)

(Contents Continued from Page vii)

Contrastive Information Extraction With Generative Transformer ..... *N. Zhang, H. Ye, S. Deng, C. Tan, M. Chen, S. Huang, F. Huang, and H. Chen* 3077
Minimum-Volume Multichannel Nonnegative Matrix Factorization for Blind Audio Source Separation ..... *J. Wang, S. Guan, S. Liu, and X.-L. Zhang* 3089
A Room Impulse Response Measurement Method Robust Towards Nonlinearities Based on Orthogonal Periodic Sequences ..... *A. Carini, S. Cecchi, A. Terenzi, and S. Orcioni* 3104
Quantization-Aware Binaural MWF Based Noise Reduction Incorporating External Wireless Devices ..... *J. Zhang and C. Li* 3118
Knowledge Enhanced Fact Checking and Verification ..... *B. Zhu, X. Zhang, M. Gu, and Y. Deng* 3132
A Superfast Toeplitz Matrix Inversion Method for Single- and Multi-Channel Inverse Filters and Its Application to Room Equalization ..... *M. A. Poletti and P. D. Teal* 3144
Detecting Source Contextual Barriers for Understanding Neural Machine Translation ..... *G. Li, L. Liu, C. Zhu, R. Wang, T. Zhao, and S. Shi* 3158
Audio-Aware Spoken Multiple-Choice Question Answering With Pre-Trained Language Models ..... *C.-C. Kuo, K.-Y. Chen, and S.-B. Luo* 3170
Addressing Extraction and Generation Separately: Keyphrase Prediction With Pre-Trained Language Models ..... *R. Liu, Z. Lin, and W. Wang* 3180
Sarcasm Detection with Commonsense Knowledge ..... *J. Li, H. Pan, Z. Lin, P. Fu, and W. Wang* 3192
Keyword Search Using Attention-Based End-to-End ASR and Frame-Synchronous Phoneme Alignments ..... *R. Yang, G. Cheng, H. Miao, T. Li, P. Zhang, and Y. Yan* 3202
Flexibly Focusing on Supporting Facts, Using Bridge Links, and Jointly Training Specialized Modules for Multi-Hop Question Answering ..... *T. Alkhaldi, C. Chu, and S. Kurohashi* 3216
An Efficient Filter Bank Structure for Adaptive Notch Filtering and Applications ..... *W. Wu, Y. Xiao, J. Lin, L. Ma, and K. Khorasani* 3226
Synthesizing Spoken Descriptions of Images ..... *X. Wang, J. van der Hout, J. Zhu, M. Hasegawa-Johnson, and O. Scharenborg* 3242
Enhancement of Noisy Reverberant Speech Using Polynomial Matrix Eigenvalue Decomposition ..... *V. W. Neo, C. Evers, and P. A. Naylor* 3255
Wave Digital Modeling and Implementation of Nonlinear Audio Circuits With Nullors ..... *R. Giampiccolo, M. G. de Bari, A. Bernardini, and A. Sarti* 3267
Speech Emotion Recognition Using Sequential Capsule Networks ..... *X. Wu, Y. Cao, H. Lu, S. Liu, D. Wang, Z. Wu, X. Liu, and H. Meng* 3280
PSLA: Improving Audio Tagging With Pretraining, Sampling, Labeling, and Aggregation ..... *Y. Gong, Y.-A. Chung, and J. Glass* 3292
Review and Arrange: Curriculum Learning for Natural Language Understanding ..... *L. Zhang, Z. Mao, B. Xu, Q. Wang, and Y. Zhang* 3307
Automatic Detection of Affective Flattening in Schizophrenia: Acoustic Correlates to Sound Waves and Auditory Perception ..... *F. He, L. He, J. Zhang, Y. Li, and X. Xiong* 3321
Efficient Learning Approach for Pronominal Anaphora and Ellipsis Identification and Resolution in Arabic Texts ..... *S. Mathlouthi Bouzid and C. Ben Othmane Zribi* 3335
Semantic Change Detection With Gaussian Word Embeddings ..... *A. Yüksel, B. Uğurlu, and A. Koç* 3349
Medical Term and Status Generation From Chinese Clinical Dialogue With Multi-Granularity Transformer ..... *M. Li, L. Xiang, X. Kang, Y. Zhao, Y. Zhou, and C. Zong* 3362
F_0-Noise-Robust Glottal Source and Vocal Tract Analysis Based on ARX-LF Model ..... *Y. Li, J. Tao, D. Erickson, B. Liu, and M. Akagi* 3375
Efficient Estimate of Sentence's Representation Based on the Difference Semantics Model ..... *X. Liao, Y. Huang, Y. Wei, C. Zhang, F. Wang, and Y. Wang* 3384
TAU-Net: Temporal Activation U-Net Shared With Nonnegative Matrix Factorization for Speech Enhancement in Unseen Noise Environments ..... *K. M. Jeon, G. W. Lee, N. K. Kim, and H. K. Kim* 3400
Robustness of Speech Spoofing Detectors Against Adversarial Post-Processing of Voice Conversion ..... *Y.-Y. Ding, H.-J. Lin, L.-J. Liu, Z.-H. Ling, and Y. Hu* 3415
Language Agnostic Speaker Embedding for Cross-Lingual Personalized Speech Generation ..... *Y. Zhou, X. Tian, and H. Li* 3427
Speech Enhancement Using Multi-Stage Self-Attentive Temporal Convolutional Networks ..... *J. Lin, A. J. de L. van Wijngaarden, K.-C. Wang, and M. C. Smith* 3440

(Contents Continued on Page ix)

(Contents Continued from Page viii)

HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units ..... *W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed* 3451
Time-Frequency-Bin-Wise Linear Combination of Beamformers for Distortionless Signal Enhancement ..... *K. Yamaoka, N. Ono, and S. Makino* 3461
Convolutive Prediction for Monaural Speech Dereverberation and Noisy-Reverberant Speaker Separation ..... *Z.-Q. Wang, G. Wichern, and J. L. Roux* 3476
Learning Deep Direct-Path Relative Transfer Function for Binaural Sound Source Localization ..... *B. Yang, H. Liu, and X. Li* 3491
Pre-Training With Whole Word Masking for Chinese BERT ..... *Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang* 3504
Counterfactually Fair Automatic Speech Recognition ..... *L. Sarı, M. Hasegawa-Johnson, and C. D. Yoo* 3515
Multi-Channel Multi-Frame ADL-MVDR for Target Speech Separation ..... *Z. Zhang, Y. Xu, M. Yu, S.-X. Zhang, L. Chen, D. S. Williamson, and D. Yu* 3526
Reduction of Subjective Listening Effort for TV Broadcast Signals With Recurrent Neural Networks ..... *N. L. Westhausen, R. Huber, H. Baumgartner, R. Sinha, J. Rennies, and B. T. Meyer* 3541
Subword-Based Compact Reconstruction for Open-Vocabulary Neural Word Embeddings ..... *S. Sasaki, J. Suzuki, and K. Inui* 3551
Asynchronous Decentralized Distributed Training of Acoustic Models ..... *X. Cui, W. Zhang, A. Kayi, M. Liu, U. Finkler, B. Kingsbury, G. Saon, and D. Kung* 3565
Spatial Active Noise Control in Rooms Using Higher Order Sources ..... *J. Zhang, W. Zhang, J. A. Zhang, T. D. Abhayapala, and L. Zhang* 3577
Multimodal Emotion Recognition With Temporal and Semantic Consistency ..... *B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang* 3592
Regularized Phrase-Based Topic Model for Automatic Question Classification With Domain-Agnostic Class Labels ..... *S. Supraja, A. W. H. Khong, and S. Tatinati* 3604
Sound Field Reproduction With a Cylindrical Loudspeaker Array Using First Order Wall Reflections ..... *N. Maeda, F. M. Fazi, and F.-M. Hoffmann* 3617
Coupling a Generative Model With a Discriminative Learning Framework for Speaker Verification ..... *X. Lu, P. Shen, Y. Tsao, and H. Kawai* 3631
Effects of Additive Noise in Binaural Rendering of Spherical Microphone Array Signals ..... *H. Helmholz, D. Lou Alon, S. V. A. Garí, and J. Ahrens* 3642
Speech Reconstruction With Reminiscent Sound Via Visual Voice Memory ..... *J. Hong, M. Kim, S. J. Park, and Y. M. Ro* 3654
Robustness of Acoustic Rake Filters in Minimum Variance Beamforming ..... *R. Weisman, T. Shlomo, V. Tourbabin, P. Calamia, and B. Rafaely* 3668
Mixed Precision Low-Bit Quantization of Neural Network Language Models for Speech Recognition ..... *J. Xu, J. Yu, S. Hu, X. Liu, and H. Meng* 3679
Learning Fine-Grained Fact-Article Correspondence in Legal Cases ..... *J. Ge, Y. Huang, X. Shen, C. Li, and W. Hu* 3694
High-Resolution Piano Transcription With Pedals by Regressing Onset and Offset Times ..... *Q. Kong, B. Li, X. Song, Y. Wan, and Y. Wang* 3707