Guest editorial: Special issue on advances in deep learning based speech processing

Xiaolei Zhang, Lei Xie, Eric Fosler-Lussier, Emmanuel Vincent

Guest Editorial: Special Issue on Advances in Deep Learning Based Speech Processing

Xiao-Lei Zhang[a], Lei Xie[b], Eric Fosler-Lussier[c], Emmanuel Vincent[d]

[a]*School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.*
[b]*Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.*
[c]*Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA.*
[d]*Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France.*

Deep learning has triggered a big revolution on speech processing. The revolution started from the successful application of deep neural networks to automatic speech recognition, and was quickly spread to other topics of speech processing, including speech analysis, speech enhancement and separation, speaker and language recognition, speech synthesis, and spoken language understanding. Such tremendous success is achieved by the long-term evolution of neural network technologies as well as the big explosion of speech data and fast development of computing power.

Although such a big success has been made, deep learning based speech processing still has many challenges for real-world wide deployment. For example, when the distance between a speaker and a microphone array is larger than 10 meters, the word error rate of a speech recognizer may be as high as over 50%; end-to-end deep learning based speech processing systems have shown potential advantages over hybrid systems, however, they still have a high requirement to large-scale labelled speech data; deep-learning-based speech synthesis has been highly competitive with human-sounding speech to traditional methods, however, the models are not stable, lacks controllability and are still too large and slow to be able to put into mobile and IoT devices, etc.

Accordingly, new theoretical methods in deep learning and speech processing are required to tackle the above challenges, as well as to yield novel insights into new directions and problems.

This Special Issue provides a collection of state-of-the-art research works focusing on recent advances of deep learning based speech processing, which presents novel contributions addressing theoretical and practical aspects of deep learning related speech processing techniques. After a rigorous review of 66 high-quality articles that were submitted, 26 articles were selected for inclusion in this Special Issue. They cover major topics of speech processing, including speech enhancement and separation (8 articles), speech recognition (5 articles), speaker and language recognition (5 articles), speech synthesis (3 articles), and speech emotion recognition (2 articles), respectively. Three articles related to general topics of speech processing are incorporated as well. A brief summary of these articles is provided herein.

Speech enhancement and separation is a topic of recovering clean speech from noisy speech that may include various kinds of interference signals. It is conventional formulated as a signal processing problem. Recently, deep learning made breakthroughs to speech enhancement and separation, particularly in adverse acoustic environments, which quickly becomes a new research paradigm. However, it still faces many challenges, even in the deep learning era. A creative work was made by Zhang *et al.* [1] for the classic Active Noise Control (ANC) problem. They formulated ANC as a supervised learning problem, and proposed a deep learning approach, called deep ANC. Unlike other deep learning techniques for speech enhancement which use deep networks to predict waveforms of speech or their alternatives, they employed deep learning to encode the optimal control parameters corresponding to different noises and environments, and hence made deep learning applicable to ANC.

Monaural speech enhancement and separation is a historically difficult topic, particularly in the presence of reverberation. A good deep architecture that incorporates sufficient prior knowledge is important for real-world applications. In this Special Issue, several advanced deep architectures were proposed, and new state-of-the-art performance was reported. Xian *et al.* [2] proposed a convolutional fusion network for monaural speech enhancement, which fully exploits cross-band information. Chen *et al.* [3] proposed dual-stream deep attractor networks with multi-domain learning to efficiently perform both dereverberation and separation tasks. Li *et al.* [4] proposed to generate speech spectra via a new type of adversarial training framework, named μ-law spectrum generative adversarial network, for speech separation. Borgstrom *et al.* [5] proposed an end-to-end neural network architecture based on self-attention for hearing aid. Huang *et al.* [6] proposed a lightweight speaker extraction model, named TinyWASE, by compressing speaker extraction models with ultra-low precision quantization and knowledge distillation, which is able run on resource-constrained devices.

Email addresses: xiaolei.zhang@nwpu.edu.cn (Xiao-Lei Zhang), lxie@nwpu.edu.cn (Lei Xie), fosler-lussier.1@osu.edu (Eric Fosler-Lussier), emmanuel.vincent@inria.fr (Emmanuel Vincent)

Multichannel or multimodal speech enhancement and separation extends the monaural case by incorporating additional information and important sources to further improve the performance. In this Special Issue, two works are included. Li *et al.* [7] proposed a novel dual-channel deep-neural-network-based Generalized Sidelobe Canceller (GSC) structure, called nnGSC. The core idea is to make each module of the traditional GSC fully learnable, and use an acoustic model to perform joint optimization of speech recognition with GSC. Chen *et al.* [8] proposed a visual embedding approach to improve embedding aware speech enhancement by synchronizing visual lip frames at the phone and place of articulation levels.

Automatic Speech Recognition (ASR) is a task of transcribing speech into text. It was the first big breakthrough of deep learning in applications. However, low-resource ASR is still challenging in real-world applications of ASR. This Special Issue contains several works on this topic with specific applications. Iranzo-Sanchez *et al.* [9] presented a state-of-the-art streaming speech translation system, in which neural-based models integrated in the ASR and machine translation components are carefully adapted in terms of their training and decoding procedures. They addressed the low-resource and streaming problems of ASR simultaneously. Yang *et al.* [10] proposed an unsupervised pre-training approach to utilize the speech data of two native languages (the learner's native and target languages) together, for the data sparsity problem of the non-native mispronunciation recognition. Zhao *et al.* [11] proposed an end-to-end keyword spotting system in which they introduced an attention mechanism and a novel energy scorer to make decisions with the locations of the keywords. Their experimental results on four low resource conditions demonstrate the effectiveness of the system. Liu *et al.* [12] proposed incremental training with revised loss function, data augmentation, and fine-grained training, which is able to improve the accuracy for the low-resource or even unseen user-defined keywords while maintaining high accuracy for pre-defined keywords.

Besides the low resource problem, how to train an ASR system with large amount of data is also dramatically important. Haider *et al.* [13] presented a novel Natural Gradient and hessian-Free (NGHF) optimisation framework for neural network training that can operate efficiently in a distributed manner. Their experiments show that NGHF not only achieves larger word error rate reductions than standard stochastic gradient descent or Adam, but also requires orders of magnitude fewer parameter updates.

Speaker recognition is a topic of recognizing the identity of a speaker. Deep learning based speaker recognition dominates the topic at present. In this Special Issue, an overview article of deep-learning-based speaker recognition was included [14]. It summarizes the subtasks of speaker recognition, including speaker identification, speaker verification, speaker diarization, robust speaker recognition. Regarding the research paradigms of speaker recognition, it contains DNN/i-vector, x-vector, stage-wise diarization, and end-to-end diarization. Many acoustic features and datasets were summarized as well.

From the above overview, one see that different parts of speech contribute unequally to the voiceprint of a speaker. Therefore, attention mechanism is helpful. For this respect, Miao *et al.* [15] improved conventional attention mechanism by a novel mixed-order attention for low frame-level speech features, and a nonlocal attention mechanism and a dilated residual structure to balance fine grained local information and multi-scale long-range information, which achieves a much wider context than purely local attention. Shi *et ai.* [16] proposed a frame-level encoder and attention to the segments of speech, as well as another segment level attention to construct an utterance representation.

Given that there is empirically no unique outstanding acoustic features or models suitable for various test scenarios, another thought for improving the performance of speaker recognition is to combine complementary acoustic features or models. Sun *et al.* [17] proposed a c-vector method by combining multiple sets of complementary d-vectors derived from systems with different neural network components, including 2-dimensional self-attentive, gated additive, bilinear pooling structures, etc. *Language recognition* shares quite a similar research trend with speaker recognition. Li *et al.* [18] investigated the efficiency of integrating multiple acoustic features for language recognition, and further explored two kinds of training constraints to integrate the features. One option introduced auxiliary classification constraints with adaptive weights for loss functions in feature encoder sub-networks, and the other option introduced the canonical correlation analysis constraint to maximize the correlation of different feature representations.

Speech synthesis, a.k.a. Text-To-Speech (TTS), is a topic of generating speech from text. A challenge problem is how to generate speech efficiently and vividly. Liu *et al.* [19] proposed a TTS model, named FastTalker, for high-quality speech synthesis at low computational cost. The core idea is to use a non-autoregressive context decoder to generate acoustic features efficiently, and then add a shallow autoregressive acoustic decoder on top of the non-autoregressive context decoder to retrieve the temporal information of the acoustic signal. Dahmani *et al.* [20] first presented an expressive audiovisual corpus, and then proposed to learn emotional latent representation with a conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis. In Nallanthighal *et al.*'s study [21], they emphasized the importance of respiratory voice during TTS by exploring techniques for sensing breathing signal and breathing parameters from speech using deep learning architectures. The conclusion of the study may also help us understand the respiratory health using one's speech.

Speech emotion recognition aims to identify the feeling of a speaker through his/her voice. It can be categorized into discrete speech emotion recognition where the emotion states are discrete values, and continuous dimensional speech recognition where the emotion states are in a continuous space. Regarding the discrete speech emotion recognition, Zhao *et al.* [22] presented an efficient self-attention residual dilated network incorporating Connectionist Temporal Classification (CTC) loss to address the challenging issue of modelling long-term temporal dependencies of emotional speech. Regarding the continuous dimensional speech emotion recognition, Peng *et al.* [23] investigated multi-resolution representations of an auditory

perception model, and proposed a novel feature called multi-resolution modulation-filtered cochleagram for predicting valence and arousal values of emotional primitives.

Some new concepts that are out of the above topics are accommodated in this large Special Issue as well. For example, Guizzo *et al.* [24] proposed a novel anti-transfer learning strategy to make a deep model focus on its targeted task, where anti-transfer learning avoids the learning of representations that have been learned for an orthogonal task, i.e., one that is not relevant and potentially confounding for the target task, such as speaker identity for speech recognition or speech content for emotion recognition. This extends the potential use of pre-trained models that have become increasingly available.

Some interesting applications of speech processing technologies were included as well. [25] proposed two neural network architectures for modeling unsupervised lexical learning from raw acoustic inputs, named ciwGAN (Categorical InfoWave Generative Adversarial Networks) and fiwGAN (Featural InfoWaveGAN), which combine Deep Convolutional GAN architecture for audio data with the information theoretic extension of GAN and propose a new latent space structure that can model featural learning simultaneously with a higher level classification. [26] proposed a novel Residual Network (ResNet)-based technique with short-duration speech segments as the input to improve the performance and applicability of detecting impaired speech.

The Guest Editors would like to thank the authors for their high-quality contributions, and the Editor-in-Chiefs for their support throughout the process and realization of this Special Issue. The Guest Editors are also grateful for all the reviewers who helped ensure the quality of the articles included in this issue.

References

[1] H. Zhang, D. Wang, Deep anc: A deep learning approach to active noise control, Neural Networks 141 (2021) 1–10.

[2] Y. Xian, Y. Sun, W. Wang, S. M. Naqvi, Convolutional fusion network for monaural speech enhancement, Neural Networks 143 (2021) 97–107.

[3] H. Chen, P. Zhang, A dual-stream deep attractor network with multi-domain learning for speech dereverberation and separation, Neural Networks 141 (2021) 238–248.

[4] H. Li, Y. Xu, D. Ke, K. Su, μ-law sgan for generating spectra with more details in speech enhancement, Neural Networks 136 (2021) 17–27.

[5] B. J. Borgström, M. S. Brandstein, G. A. Ciccarelli, T. F. Quatieri, C. J. Smalt, Speaker separation in realistic noise environments with applications to a cognitively-controlled hearing aid, Neural Networks 140 (2021) 136–147.

[6] Y. Huang, Y. Hao, J. Xu, B. Xu, Compressing speaker extraction model with ultra-low precision quantization and knowledge distillation, Neural Networks 154 (2022) 13–21.

[7] G. Li, S. Liang, S. Nie, W. Liu, Z. Yang, Deep neural network-based generalized sidelobe canceller for dual-channel far-field speech recognition, Neural Networks 141 (2021) 225–237.

[8] H. Chen, J. Du, Y. Hu, L.-R. Dai, B.-C. Yin, C.-H. Lee, Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement, Neural Networks 143 (2021) 171–182.

[9] J. Iranzo-Sánchez, J. Jorge, P. Baquero-Arnal, J. A. Silvestre-Cerdà, A. Giménez, J. Civera, A. Sanchis, A. Juan, Streaming cascade-based speech translation leveraged by a direct segmentation model, Neural Networks 142 (2021) 303–315.

[10] L. Yang, K. Fu, J. Zhang, T. Shinozaki, Non-native acoustic modeling for mispronunciation verification based on language adversarial representation learning, Neural Networks 142 (2021) 597–607.

[11] Z. Zhao, W.-Q. Zhang, End-to-end keyword search system based on attention mechanism and energy scorer for low resource languages, Neural Networks 139 (2021) 326–334.

[12] L. Liu, M. Yang, X. Gao, Q. Liu, Z. Yuan, J. Zhou, Keyword spotting techniques to improve the recognition accuracy of user-defined keywords, Neural Networks 139 (2021) 237–245.

[13] A. Haider, C. Zhang, F. L. Kreyssig, P. C. Woodland, A distributed optimisation framework combining natural gradient with hessian-free for discriminative sequence training, Neural Networks 143 (2021) 537–549.

[14] Z. Bai, X.-L. Zhang, Speaker recognition based on deep learning: An overview, Neural Networks 140 (2021) 65–99.

[15] X. Miao, I. McLoughlin, W. Wang, P. Zhang, D-mona: A dilated mixed-order non-local attention network for speaker and language recognition, Neural Networks 139 (2021) 201–211.

[16] Y. Shi, Q. Huang, T. Hain, H-vectors: Improving the robustness in utterance-level speaker embeddings using a hierarchical attention model, Neural Networks 142 (2021) 329–339.

[17] G. Sun, C. Zhang, P. C. Woodland, Combination of deep speaker embeddings for diarisation, Neural Networks 141 (2021) 372–384.

[18] L. Li, Z. Li, Y. Liu, Q. Hong, Deep joint learning for language recognition, Neural Networks 141 (2021) 72–86.

[19] R. Liu, B. Sisman, Y. Lin, H. Li, Fasttalker: A neural text-to-speech architecture with shallow and group autoregression, Neural Networks 141 (2021) 306–314.

[20] S. Dahmani, V. Colotte, V. Girard, S. Ouni, Learning emotions latent representation with cvae for text-driven expressive audiovisual speech synthesis, Neural Networks 141 (2021) 315–329.

[21] V. S. Nallanthighal, Z. Mostaani, A. Härmä, H. Strik, M. Magimai-Doss, Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings, Neural Networks 141 (2021) 211–224.

[22] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, B. W. Schuller, Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition, Neural Networks 141 (2021) 52–60.

[23] Z. Peng, J. Dang, M. Unoki, M. Akagi, Multi-resolution modulation-filtered cochleagram feature for lstm-based dimensional emotion recognition from speech, Neural Networks 140 (2021) 261–273.

[24] E. Guizzo, T. Weyde, G. Tarroni, Anti-transfer learning for task invariance in convolutional neural networks for speech processing, Neural Networks 142 (2021) 238–251.

[25] G. Beguš, Ciwgan and fiwgan: Encoding information in acoustic data to model lexical learning with generative adversarial networks, Neural Networks 139 (2021) 305–325.

[26] S. Gupta, A. T. Patil, M. Purohit, M. Parmar, M. Patel, H. A. Patil, R. C. Guido, Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments, Neural Networks 139 (2021) 105–117.