Introduction to the Issue on Far-Field Speech Processing in the Era of Deep Learning: Speech Enhancement, Separation, and Recognition

I. Introduction

FAR-FIELD speech processing has become an active field of research due to recent scientific advancements and its widespread use in commercial products. This field of research deals with speech enhancement and recognition using one or more microphones placed at a distance from one or more speakers. Although the topic has been studied for a long time, recent successful applications starting with the Amazon Echo and challenge activities including CHiME and REVERB projects greatly accelerated progress in this field. Concurrently, deep learning has created a new paradigm that has led to major breakthroughs both in front-end signal enhancement, extraction, and separation, as well as in back-end automatic speech recognition (ASR). Furthermore, more deep learning provides a means of jointly optimizing all components of far-field speech processing in an end-to-end fashion. This special issue is a forum to gather the latest findings in this very active field of research. The special issue is highly relevant to the audio and acoustics, speech and language, and machine learning for signal processing communities.

II. Overviews

We had 15 submissions for this special issue, and finally accepted 7 articles covering the state-of-the-art boosted by deep neural networks in this area. It covers multichannel speech enhancement based on beamforming techniques, multichannel feature extraction, overlapped speech analysis, and new dataset for far-field speech processing. This editorial first describes the technical achievements of this special issue with respect to multichannel speech enhancement based on deep learning by Zmolikova *et al*., Drude *et al.*, Chakrabarty *et al.*, and Sun *et al.* in Secion II-A. Then, it describes multichannel feature extraction by Rodomagoulakis *et al.* in Section II-B. These achievements are essential technologies to improve the performance of far-field speech processing applications. Section II-C describes the outcome of overlapped speech analysis by Andrei *et al.* by performing perception studies of human abilities for speaker counting and overlapped speech detection. Section II-D also describes the introduction of a new dataset of real room impulse responses (RIR), background noise and re-transmitted speech data by Szöke *et al*. The database facilitates far-field speech processing studies by providing further variations of far-field speech/acoustic recordings. The following subsections provide a detailed summary of each method.

A. Multichannel Speech Enhancement

One of the major outcomes of this special issue is to establish multichannel speech enhancement techniques by combining masking-based approaches and array signal processing including beamforming. The mask is usually represented as (speaker-dependent/target) speech and noise presence probabilities, which are estimated based on a statistical model including a complex Gaussian mixture model (CGMM) or its variants and a deep neural network. This mask information is used to provide accurate spatial statistics for noise and (speaker-dependent) speech signals, which enables the estimation of robust beamforming filters.

In "Time-frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks," Chakrabarty *et al.* first uses a layer-wise combination of recurrent neural network and convolutional neural network (CNN) to accurately estimate the noise and speech masks for each time-frequency bin. These mask estimates are then employed either as a real valued gain to obtain the desired signal or as an indicator for the recursive update of power spectral density matrices used for obtaining accurate beamformers. The method implements online multichannel speech enhancement and shows improvement in terms of several speech enhancement metrics (fwSNR, STOI, and PESQ) compared with conventional beamforming techniques.

In "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," Zmolikova *et al.* tackles a multi-speaker speech enhancement problem by recovering one target speaker from a mixture instead of performing speech separation for all speakers in a mixture. This approach does not suffer from the label permutation problem and can also be applied to the situation where the number of speakers in a mixture is unknown. The approach first estimates the target speaker masks given his/her enrolled speech segments based on a neural network inspired by an acoustic model adaptation technique. The mask can be used to estimate beamformers similar to the other multichannel speech enhancement to make a beam pattern for a specific target speaker. Speaker-Beam achieved significant improvement from conventional deep-learning-based speech separation techniques in terms of both speech enhancement and recognition metrics.

Digital Object Identifier 10.1109/JSTSP.2019.2925640

In "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," Drude *et al.* provides a unified view for blind source separation schemes based on multichannel speech separation and enhancement within an encoder, latent model, and decoder framework. Based on this framework, the proposed method integrates deep clustering or a deep attracter network as an encoder, complex angular central GMM as a latent model, and mask-based beamforming as a decoder. From the experiments on both artificially mixed speech and true recording of speech mixtures, the proposed integration method consistently outperforms the individual components.

In "A speaker-dependent approach to separation of far-field multi-talker microphone array speech for front-end processing in the CHiME-5 challenge," Sun *et al.* focuses on reporting their multichannel speech enhancement component of their whole multichannel ASR system that achieved the best performance in the CHiME-5 challenge. Similarly to the other techniques, mask-based beamforming techniques are essential components of the proposed system, but it combines single- and multi-channel speech enhancement iteratively to refine speaker-dependent masks or beamformed signals. With this approach, the method significantly reduces the WER compared with the official CHiME-5 speech enhancement baseline.

B. Multichannel Feature Extraction

In "Improved frequency modulation features for multichannel distant speech recognition," Rodomagoulakis *et al.* investigates the use of robust frequency modulation features in a multichannel far-field speech recognition scenario. The paper proposes to use a multichannel demodulation technique to improve the demodulation of speech resonances and enable more accurate estimation of instantaneous modulations. The modulation features are enriched based on compressed instantaneous frequencies or hierarchical deep bottleneck features, and fed into state-of-the-art acoustic models based on DNNs. The effectiveness of the proposed method is evaluated across various far-field speech recognition corpora including the DIRHA-English, AMI, and CHiME-4 corpora, showing consistent improvement across them.

C. Overlapped Speech Analysis

In "Overlapped speech detection and competing speaker counting humans vs. deep learning," Andrei *et al.* first designs a perception study to evaluate the human ability of speaker counting, which estimates the maximum number of speakers given a recording, and overlapped speech detection in a single channel audio file. Compared to previous studies, this investigation is on a much larger scale. It includes a perception study in three sessions with 31, 38 and 80 volunteers respectively. It provides a more reliable statistical analysis of human speaker counting and overlapping speech detection. Given this investigation, the rest of this paper compares their proposed automatic speaker counting and overlapped speech detection based on deep learning with the human performance. The deep learning based automatic method outperforms the human level performance, and especially a CNN-based approach can perform speaker counting and overlapped speech detection with shorter spans of input signal.

D. Database

In "Building and evaluation of a real room impulse response dataset," Szöke *et al.* presents BUT ReverbDB, which is a new dataset of real room impulse responses (RIR), background noise and re-transmitted speech collected by Brno University of Technology. In addition to the details of the dataset, this paper also provides a detailed description (know-how) of RIR collection (hardware, software, post-processing). The rest of the paper discusses an ASR experiment on this data by augmenting the ASR training data with RIR filtered data, using both real and artificially generated RIRs. In addition, the distribution also includes a Kaldi open source ASR recipe that transforms the AMI close talking data with the RIRs and noise.

III. Summary

This special issue covers various far-field speech processing techniques including speech enhancement, separation and recognition, and their integration. In most of the methods, multichannel speech processing is an essential component to achieve state-of-the-art performance. Deep learning techniques significantly enhances the performance of these techniques. Furthermore, analyses and databases facilitate research and development of such techniques.

Although there has been significant progress in the field of far-field speech processing, there still remain challenging issues including speaker diarization, audio synchronization, dynamic environment (e.g., moving speakers/microphones in a robot scenario). On the other hand, recent progress in end-to-end speech processing and the use of multimodal sensor data could provide alternative solutions for these challenges. We hope this special issue becomes a trigger for the community to tackle these challenging and exciting far-field speech processing technologies.

Shinji Watanabe, *Lead Guest Editor*
Johns Hopkins University
Baltimore, MD 21218 USA

Shoko Araki, *Guest Editor*
NTT Communication Science
Laboratories, Japan
Kyoto 619-0237, Japan

Michiel Bacchiani, *Guest Editor*
Google, Inc., Japan
Minato 106-6126, Japan

Reinhold Haeb-Umbach, *Guest Editor*
Paderborn University
Paderborn 33098, Germany

Michael L. Seltzer, *Guest Editor*
Facebook
Seattle, WA 98109