LOW-LATENCY SPEAKER-INDEPENDENT CONTINUOUS SPEECH SEPARATION

Takuya Yoshioka[†]*, Zhuo Chen*[†]*, Changliang Liu, Xiong Xiao, Hakan Erdogan, Dimitrios Dimitriadis*

Microsoft, One Microsoft Way, Redmond, WA, USA

ABSTRACT

Speaker independent continuous speech separation (SI-CSS) is a task of converting a continuous audio stream, which may contain overlapping voices of unknown speakers, into a fixed number of continuous signals each of which contains no overlapping speech segment. A separated, or cleaned, version of each utterance is generated from one of SI-CSS's output channels nondeterministically without being split up and distributed to multiple channels. A typical application scenario is transcribing multi-party conversations, such as meetings, recorded with microphone arrays. The output signals can be simply sent to a speech recognition engine because they do not include speech overlaps. The previous SI-CSS method uses a neural network trained with permutation invariant training and a data-driven beamformer and thus requires much processing latency. This paper proposes a low-latency SI-CSS method whose performance is comparable to that of the previous method in a microphone array-based meeting transcription task. This is achieved (1) by using a new speech separation network architecture combined with a double buffering scheme and (2) by performing enhancement with a set of fixed beamformers followed by a neural post-filter.

Index Terms— Meeting transcription, continuous speech separation, speaker-independent speech separation, microphone arrays

1. INTRODUCTION

Overlapping speech is omnipresent in natural human-to-human conversations. Yet it presents a significant challenge to the current speech recognition systems, which assume an input acoustic signal to consist of up to one speaker's voice at every time instance. This work investigates the problem of recognizing human-to-human conversations which may include overlapping voices by using a meeting transcription task. We assume a microphone array to be used for audio capturing. The number of conversation participants is not known in advance.

Speech separation, whose goal is to untangle a mixture of co-occuring speech signals, could potentially solve the overlapping speech problem in far-field conversation transcription. A variety of speech separation methods have been proposed in the past quater-century, ranging from independent component or vector analysis [1], nonnegative matrix or tensor factorization [2], time-frequency (TF) bin clustering [3] to deep learning [4, 5]. While considerable progress has been made, a far-field conversation transcription system that can handle speech overlaps has yet to be realized. Almost all existing speech separation methods operate on pre-segmented utterances. This requires yet another problem to be solved: speech segmentation, the goal of which is to trim each utterance from an input audio stream even when the utterance is overlapped by other voices. Many separation methods further assume the number of active speakers to be known beforehand, which does not hold in practice.

[†] Equally major contributions.

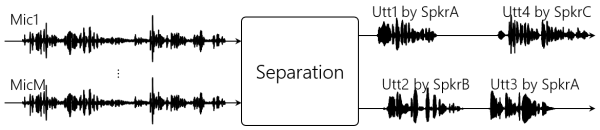

Fig. 1. Speaker-independent continuous speech separation.

Speaker-independent continuous speech separation (SI-CSS)[1] was proposed in [6] to avoid these problems. The idea is that, given a continuous audio stream, we want to generate a fixed number of time-synchronous separated signals as illustrated in Fig. 1. Each utterance constituting the input audio "spurts" from one of the output channels. When the number of active speakers is fewer than that of the output channels, the extra channels generate zero-valued signals. Thus, by performing speech recognition for each separated signal, a word transcription of the entire input conversation is obtained whether it contains speech overlaps or not. This approach was shown to work well for meeting audio [6], outperforming the state-of-the-art data-driven beamformer using neural mask estimation [7, 8].

This paper proposes a new SI-CSS method that runs with lower latency than the previous method. Two new components are introduced to achieve the low latency processing. Firstly, instead of a previous bidirectional model, we employ a new separation network architecture that has recurrent connections in the forward direction and performs fixed-length look-ahead using dilated convolution. Secondly, the segment-based data-driven beamformer of the previous method is replaced by a set of fixed beamformers followed by neural post-filtering. The post-filter removes interfering voices that remain in the beamformed signal. This is necessary as the fixed beamformers cannot precisely filter out interfering point-source signals, or other speakers' voices. The new method is shown to work comparably to the method of [6] in a meeting transcription task while requiring much lower processing latency. A novel sound source localization (SSL) method based on a complex angular central Gaussian (cACG) distribution [9] is also described.

2. SPEAKER-INDEPENDENT CONTINUOUS SPEECH SEPARATION

This section defines the SI-CSS task and briefly reviews the method proposed in [6]. The goal of SI-CSS is to transform an input signal, which may last for hours, into a fixed number of signals so that each output signal does not have overlapping speech segments. In this paper, we set the number of the output channels to two because three or more people rarely speak simultaneouly in meetings except

[1] [6] referred to CSS as unmixing transduction. In this paper, we use the term CSS as we feel it is more intuitive.

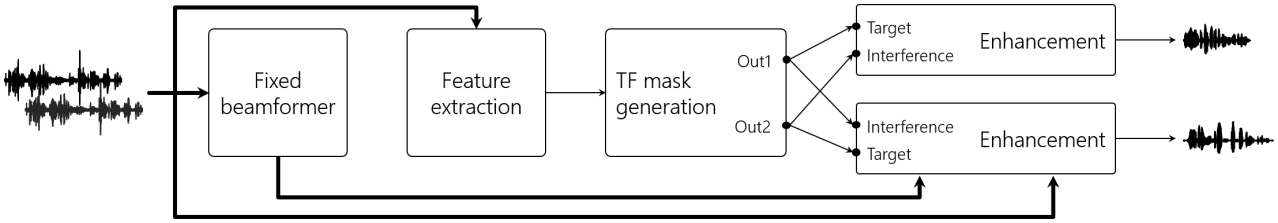

Fig. 2. Processing flow diagram of proposed method. Upper enhancement module also receives microphone array and beamformed signals as input. Thick lines represent multi-channel signals.

for laughter segments [10]. A rigorous definition of the task can be found in [6]. SI-CSS greatly facilitates transcribing conversations that include speech overlaps because we only have to perform speech recognition for each separated signal.

The method proposed in [6] achieves SI-CSS as follows. Firstly, single- and multi-channel features are extracted from an input microphone array signal. The features include magnitude spectra of an arbitrarily chosen reference microphone and inter-microphone phase differences (IPDs) [11, 12]. The stream of the feature vectors are chopped up into short segments by using a T_{W}-second sliding window with a constant shift of T_{S} seconds. For each segment, the extracted feature vectors are passed to a speech separation neural network that generates three TF masks: two for speech, one for noise. Such a network can be trained with permutation invariant training (PIT) [13]. The generated TF masks are used to construct two MVDR beamformers, each yielding a distinct separated signal. The beamformers are constructed by using the TF masks in a data dependent way [7, 14]. In order for the beamformers to make use of a certain amount of future acoustic context so that the separation performance does not degrade at the end of each segment, the last T_{M} second-part of each segment is discarded. Finally, the order of the separated signals are flipped if necessary to keep the output signal order consistent across segments.

The processing latency of the method of [6] is $T_{\mathrm{S}} + T_{\mathrm{M}} + \alpha T_{\mathrm{W}}$ seconds, where α is the real time factor (RTF) to process each T_{W}-second segment. While future hardware and algorithmic improvements may reduce the RTF factor, $\alpha(< 1)$, to some extent, the fixed cost of $T_{\mathrm{S}} + T_{\mathrm{M}}$ shall inevitably remain. Our latest experimental configuration sets T_{S}, T_{M}, and T_{W} at 0.8, 0.4, and 2.4, respectively, which reasonably balances the separation performance and the computational cost.

3. PROPOSED METHOD

Figure 2 illustrates the processing flow of the proposed method. Firstly, magnitude spectra and IPD features are extracted from an input multi-channel signal. They are fed to a TF mask generation module, which is implemented by using a neural network trained with a mean squared error (MSE) PIT loss as with the previous method (see [6] for details). The TF mask generation module *continuously* yields two sets of TF masks with a small time lag. While the TF masks may be applied directly to the input signal, direct masking tends to end up with degrading speech recognition performance due to speech distortion. Thus, we fed them to another system component, referred to as an enhancement module in Fig. 2, which utilizes fixed beamformers and a neural network-based post-filter. The rest of this section details each component other than the enhancement module, which we elaborate on in the next section.

3.1. Fixed beamformers

For real time applications, beamformers designed for a specific microphone array geometry are more advantageous than the data-driven beamforming approach [7, 14]. It is noteworthy that, as demonstrated in [8], a well-designed fixed beamformer is as effective at reducing background noise as the state-of-the-art data-driven beamformer.

We designed a set of 18 fixed beamformers, each with a distinct focus direction, for the seven-channel circular microphone array that we used for our data collection. The focus directions of neighboring beamformers are separated by 20 degrees. The beam pattern for each direction was optimized to maximize the output signal-to-noise ratio for simulated environments.

3.2. Feature extraction

Multiple independent reports [11, 12, 15] show IPD feature's effectiveness for neural speech separation. In this work, we make use of both the IPDs and the magnitude spectrum of the signal of the first, or reference, microphone. The IPDs are computed between the reference microphone and each of the other microphones.

3.3. Time-frequency mask generation

A neural network trained with PIT generates TF masks for speech separation from the features computed as described above. The most prominent advantage of PIT over other speech separation mask estimation schemes, such as spatial clustering [16, 17], deep clustering [4], and deep attractor networks [18], is that it does not require prior knowledge of the number of active speakers. When only one speaker is active, the PIT-trained network yields zero-valued masks from extra output channels. This is desirable for SI-CSS because we always generate a fixed number of output signals.

3.3.1. Network architecture

Prior work on PIT often utilized bidirectional models. A neural network trained with PIT can not only separate speech signals for each short time frame but also keep the order of the output signals consistent across short time frames. This is possible largely because the network is penalized if it changes the output signal order at some middle point of an utterance during training. On the other hand, for the network to be able to consistently assign an output channel to each separated signal frame, it is also beneficial for the network to take account of some future acoustic context [13]. Therefore, bidirectional models are inherently advantageous while their use hinders low latency processing.

In this paper, we propose to use a hybrid of a unidirectional recurrent neural network (RNN) and a convolutional neural network (CNN). Figure 3 depicts the architecture of our RNN-CNN hybrid

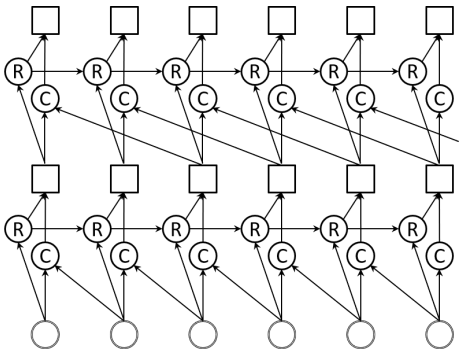

Fig. 3. RNN-CNN hybrid model. "R" and "C" circles represent recurrent (LSTM) and convolution nodes, respectively. Square nodes perform splicing. Double circles represent input nodes.

model. The temporal acoustic dependency in the forward direction is modeled by the RNN, or more specifically a long short term memory (LSTM) network. On the other hand, the CNN captures the backward acoustic dependency. Dilated convolution [19] is used as shown in Fig. 3 to efficiently cover a fixed length of future acoustic context. Our experimental system consists of a projection layer with 1024 units, two RNN-CNN hybrid layers, and two parallel fully connected layers with sigmoid nonlinearity. The final layer's activations are used as TF masks for speech separation. With the two RNN-CNN hybrid layers, our model utilizes four ($= N_{\mathrm{LF}}$) future frames, where our frame shift is 0.016 seconds.

3.3.2. Double buffering

While the PIT-trained network is designed to assign an output channel to each separated speech frame consistently across short time frames, we cannot simply keep feeding the network with the feature vectors for a long time. Firstly, the speech separation network is trained on mixed speech segments of up to $T_{\mathrm{TR}}(= 10)$ seconds during the learning phase. The resultant model does not necessarily keep the output order consistent beyond T_{TR} seconds. In addition, RNN's state values tend to saturate after a while when it is exposed to a long feature vector stream [20]. Therefore, the state values need to be refreshed at some interval in such a way that keeps the output order consistent.

To address this problem, we propose a double buffering scheme as illustrated in Fig 4. We feed feature vectors to the network for $T_{\mathrm{W}}(= 2.4)$ seconds. Because the model uses a fixed length of future context, the output TF masks can be obtained with a limited processing latency. Halfway through processing the first buffer, we start a new buffer from fresh RNN state values. The new buffer is processed for another T_{W} seconds. By using the TF masks generated for the first $T_{\mathrm{W}}/2$-second half, we determine the best output order for the second buffer. The order is determined so that the MSE can be minimized between the separated signals obtained for the last half of the previous buffer and those for the first half of the current buffer. By using two buffers in this way, the TF masks can be continuously generated for a long stream of audio in real time.

4. TARGET SPEECH ENHANCEMENT

Given two TF masks, one for a target speaker and one for an interfering speaker, and multiple beamformed signals, the enhancement module generates a signal where the target speaker is enhanced

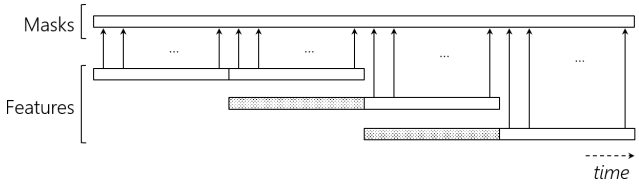

Fig. 4. Double buffering for real-time CSS. TF masks calculated for shaded blocks are used only for ordering output channels.

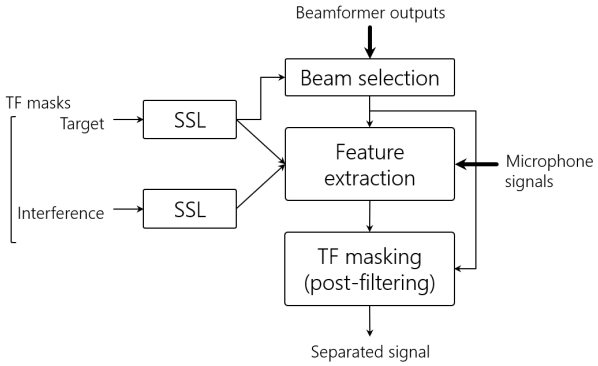

Fig. 5. Processing flow diagram of target speech enhancement.

against the interfering speaker and background noise. As shown in Fig. 5, this is performed by first selecting the beamformer channel pointing at the target speaker direction and then post-filtering the signal with TF masks derived from a post-filtering neural network. Unlike the separation network, the post-filtering network receives the target and interference angles as input in addition to the microphone and beamformed signals in order to enhance only the target speaker's voice. Our network model does not use any future data frames.

4.1. Sound source localization

The enhancement processing starts with performing SSL for each of the target and interference speakers. The estimated directions are used both for selecting the beamformer channel and as an input to the post-filtering network.

For computational efficiency, the target and interference directions are estimated every N_{S} frames, or $0.016N_{\mathrm{S}}$ seconds. For each of the target and interference, SSL is performed by using the input multi-channel audio and the TF masks in frames $(n-N_{\mathrm{W}}, n]$, where n refers to the current frame index. The estimated directions are used for processing the frames in $(n-N_{\mathrm{M}}-N_{\mathrm{S}}, n-N_{\mathrm{M}}]$, resulting in delay of N_{M} frames. The "margin" of length N_{M} is introduced so that SSL leverages a small amount of future context. In our experiments, N_{M}, N_{S}, and N_{W} are set at 20, 10, and 50, respectively.

SSL is achieved with maximum likelihood estimation using the TF masks as observation weights. We hypothesize that each magnitude-normalized multi-channel observation vector, $\boldsymbol{z}_{t,f}$, follows a cACG distribution [9] as follows:

$$p(\boldsymbol{z}_{t,f}|\omega) = 0.5\pi^{-M}(M-1)!|\boldsymbol{B}_{f,\omega}|^{-1}\left(\boldsymbol{z}_{t,f}^{H}\boldsymbol{B}_{f,\omega}^{-1}\boldsymbol{z}_{t,f}\right)^{-M}, \quad (1)$$

where ω denotes an incident angle, M the number of microphones, and $\boldsymbol{B}_{f,\omega} = (\boldsymbol{h}_{f,\omega}\boldsymbol{h}_{f,\omega}^{H} + \varepsilon\boldsymbol{I})$ with $\boldsymbol{h}_{f,\omega}$, $\boldsymbol{I}$, and ε being the steering vector for angle ω, the M-dimensional identify matrix, and a small

flooring value. Given a set of observations, $Z = \{\boldsymbol{z}_{t,f}\}$, we want to maximize the following log likelihood function with respect to ω:

$$L(\omega) = \sum_{t,f} m_{t,f} \log p(\boldsymbol{z}_{t,f}|\omega), \tag{2}$$

where ω can take a discrete value in $[0, 360)$ and $m_{t,f}$ denotes the TF mask provided by the separation network. It can be shown that the log likelihood function reduces to the following simple form:

$$L(\omega) = -\sum_{t,f} m_{t,f} \log\big(1 - ||\boldsymbol{z}_{t,f}^H \boldsymbol{h}_{f,\omega}||^2/(1+\varepsilon)\big). \tag{3}$$

$L(\omega)$ is computed for every possible discrete angle value. The ω value that gives the highest score is picked as a direction estimate. Further analysis of the cACG-based SSL method will be conducted in a separate paper.

4.2. Neural post-filtering

The beamformer signal selected based on the estimated target speaker's direction is further processed with TF masking. The aim is to cancel the interfering speaker's voice that has been left to the beamformed signal. This post-filtering is indispensable because fixed beamformers are usually designed to remove diffuse noise and thus cannot remove interfering speech signals effectively.

For this purpose, we employ the direction-informed target speech extraction method proposed in [21]. The method uses a neural network that accepts features computed based on the target and interference directions to focus on the target direction and give less attention to the interference direction. The network generates TF masks that can extract only the target speaker component from the input beamformed audio. The directional feature is calculated for each TF bin as a sparsified version of the cosine distance between the target direction's steering vector and the microphone array signal. The IPD features and the magnitude spectrum of the beamformed signal are also fed to the network. The model consists of four uni-directional LSTM layers, each with 600 units, and is trained to minimize the MSE of clean and TF mask-processed signals. We refer the reader to [21] for further details.

In summary, the minimum processing latency required for executing the proposed method is $N_{\mathrm{LF}} + N_{\mathrm{M}}$ frames, where the frame shift is 0.016 seconds. In our experiments, the look-ahead size, N_{LF}, of the RNN-CNN hybrid model is four while N_{M} is set at 20. This is much smaller than the lower-bound latency of the previous method, i.e., $T_{\mathrm{S}} + T_{\mathrm{M}}$ seconds.

5. EXPERIMENTS

We conducted meeting speech recognition experiments to evaluate the effectiveness of the proposed SI-CSS method. We performed SI-CSS on multi-microphone meeting recordings and sent the separated signals to a speech recognition engine to obtain word transcriptions. The results were scored with asclite tool [22], which aligns multiple (two for our work) hypotheses against multiple speaker-specific reference transcriptions to generate word error rate (WER) estimates.

We recorded and transcribed six meetings at our Speech Group. Both headset microphones and a seven-channel circular microphone array were used. Our meetings were conducted at multiple conference rooms. The number of the meeting attendees varied from four to eleven as shown in Table 1.

Our separation network was trained on 600 hours of artificially reverberated and mixed speech signals while the post-filter network was trained on 1.5K hours of data. See [6, 21] for our simulation and training procedures. Multi-channel dereverberation is performed prior to SI-CSS in real time by using the weighted prediction error (WPE) method [23]. Our acoustic model was sequence-trained on 33K hours of audio, including artificially contaminated speech. Decoding was performed with a trigram language model.

Table 1. %WER of different methods for meeting transcription. Numbers of meeting attendees shown in parentheses. FBF: fixed beamformer; PF: post-filter.

System	S1	S2	S3 (Proposed)
Sep. model	BLSTM	R/CNN hybrid	R/CNN hybrid
Enh. method	MVDR	MVDR	FBF-PF
MTG0 (4)	22.1	23.7	20.7
MTG1 (6)	17.0	18.1	18.0
MTG2 (6)	20.6	21.8	22.0
MTG3 (8)	28.0	27.8	28.4
MTG4 (4)	28.5	29.8	29.5
MTG5 (11)	21.0	22.9	20.5
Overall	21.5	22.7	21.7

Table 2. Impact of SSL window configurations.

Window size (N_{W})	50	50	70
Margin (N_{M})	20	10	30
%WER	21.7	22.4	21.7

5.1. Results

Table 1 lists the WERs of the previous method (S1) and the proposed method (S3). The performance of a system that yields separated signals by using MVDR and the RNN-CNN hybrid model is also presented (S2). The performance of the proposed method is comparable to that of the previous method. Comparison of S1 and S2 reveals that the use of the RNN-CNN hybrid model slightly degraded the quality of the speech separation masks. The proposed enhancement scheme, combining the fixed beamformers with the post-filter, was less sensitive to the degradation in the TF mask quality. This would be because the separation TF masks are used only for SSL in the proposed method while data-driven MVDR significantly relies on the TF masks.

Table 2 compares the WERs for different SSL window configurations. It can be seen that having a certain number of margin frames has non-negligible impact on the separation performance. A margin of 20 frames, or 0.32 seconds, seems sufficient to achieve the performance on par with the previous method using a bidirectional model and data-driven MVDR beamforming.

6. CONCLUSION

In this paper, we described a novel low-latency SI-CSS method which uses an RNN-CNN hybrid network for generating speech separation TF masks and a set of fixed beamformers followed by a neural post-filter. A double buffering scheme is introduced to continuously generate the TF masks with a short amount of delay. A new maximum likelihood SSL method using a cACG model is also presented. The proposed method achieved comparable meeting transcription accuracy to that of the previously proposed method while significantly reducing the processing latency.

7. REFERENCES

[1] S. Makino, T. W. Lee, and H. Sawada, *Blind speech separation*, Springer, 2007.

[2] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, 2011.

[4] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 31–35.

[5] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings," in *Proc. Interspeech*, 2017, pp. 2650–2654.

[6] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," in *Proc. Interspeech*, 2018, pp. 3038–3042.

[7] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "BLSTM supported GEV beamformer front-end for the 3rd CHiME challenge," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 444–451.

[8] C. Boeddeker, H. Erdogan, T. Yoshioka, and R. Haeb-Umbach, "Exploring practical aspects of neural mask-based beamforming for far-field speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, accepted.

[9] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. Eur. Signal Process. Conf.*, 2016, 1153–1157.

[10] O. Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *Proc. Interspeech*, 2006, pp. 293–296.

[11] T. Yoshioka, H. Erdogan, Z. Chen, and F. Alleva, "Multi-microphone neural speech separation for far-field multi-talker speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5739–5743.

[12] Z.-Q. Wang and D. Wang, "Integrating spectral and spatial features for multi-channel speaker separation," in *Proc. Interspeech*, 2018, pp. 2718–2722.

[13] M. Kolbæk, D. Yu, Z. H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[14] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, Atsunori Ogawa, Keisuke Kinoshita, Masakiyo Fujimoto, Chengzhu Yu, Wojciech J. Fabian, Miquel Espi, Takuya Higuchi, Shoko Araki, and Tomohiro Nakatani, "The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 436–443.

[15] Z.-Q. Wang, J. Le Roux, and J.R. Hershey, "Multi-channel deep clustering: discriminative spectral and spatial embeddings for speaker-independent speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 1–5.

[16] L. Drude, A. Chinaev, D. H. T. Vu, and R. Haeb-Umbach, "Source counting in speech mixtures using a variational em approach for complex watson mixture models," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 6834–6838.

[17] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. Int. Worksh. Acoust. Echo, Noise Contr.*, 2014, pp. 268–272.

[18] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2017, pp. 246–250.

[19] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[20] S.-Y. Chang, B. Li, G. Simko, T. N. Sainath, A. Tripathi, A. van den Oord, and O. Vinyals, "Temporal modeling using dilated convolution and gating for voice-activity-detection," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5549–5553.

[21] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. IEEE Worksh. Spoken Language Tech.*, 2018, to appear.

[22] J. G. Fiscus, J. Ajot, N. Raddle, and C. Laprum, "Multiple dimension Levenshtein edit distance calculations for evaluating automatic speech recognition systems during simulaneous speech," in *Proc. Int. Conf. Language Resources, Evaluation*, 2006, pp. 803–808.

[23] T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 10, pp. 2707–2720, 2012.