INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

An End-to-end Architecture of Online Multi-channel Speech Separation

Jian Wu[1,2*], *Zhuo Chen*[3], *Jinyu Li*[3], *Takuya Yoshioka*[3], *Zhili Tan*[2], *Ed Lin*[2], *Yi Luo*[3], *Lei Xie*[1†]

[1]Audio, Speech and Language Processing Group (ASLP), School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Microsoft, STCA, Beijing, China
[3]Microsoft, One Microsoft Way, Redmond, WA, USA

{jianwu,lxie}@nwpu-aslp.org, {zhuc,jinyli,tayoshio}@microsoft.com

Abstract

Multi-speaker speech recognition has been one of the key challenges in conversation transcription as it breaks the single active speaker assumption employed by most state-of-the-art speech recognition systems. Speech separation is considered as a remedy to this problem. Previously, we introduced a system, called *unmixing*, *fixed-beamformer* and *extraction* (UFE), that was shown to be effective in addressing the speech overlap problem in conversation transcription. With UFE, an input mixed signal is processed by fixed beamformers, followed by a neural network post filtering. Although promising results were obtained, the system contains multiple individually developed modules, leading potentially sub-optimum performance. In this work, we introduce an end-to-end modeling version of UFE. To enable gradient propagation all the way, an attentional selection module is proposed, where an attentional weight is learnt for each beamformer and spatial feature sampled over space. Experimental results show that the proposed system achieves comparable performance in an offline evaluation with the original separate processing-based pipeline, while producing remarkable improvements in an online evaluation.

Index Terms: multi-channel speech separation, robust speech recognition, speaker extraction, source localization, fixed beamformer

1. Introduction

Deep learning approaches have brought about remarkable progress to speaker-independent speech separation in the past few years [1, 2, 3, 4]. The separated signal quality has been steadily improved on benchmark datasets such as WSJ0-2mix [2]. However, multi-talker speech recognition still remains to be a challenging problem.

Speech separation is a common practice to handle the speech overlaps. Existing efforts in overlapped speech recognition can be roughly categorized into two families: building a robust separation system as a front-end processor to automatic speech recognition (ASR) tasks [5, 6, 7, 8, 9, 10, 11] or developing multi-talker aware ASR models [12, 13, 14, 15, 16, 17, 18]. Although better performance can be expected from the end-to-end training including ASR, the independent front end processing approach is often preferable in real world applications such as meeting transcription [19] for two reasons. Firstly, in the conversation transcription systems, the front end module benefits multiple acoustic processing components, including speech recognition, diarization, and speaker verification. Secondly, commercial ASR models are usually trained with a tremendous amount of data and is highly engineered, making it extremely costly to change the training scheme.

The recent work of [19] applied speech separation to a real-world conversation transcription task, where a multi-channel separation network, namely speech unmixing network, trained with permutation invariant training (PIT) [1] continuously separates the input audio stream into two channels, ensuring each output channel only contains at most one activate speaker. A mask-based adaptive Minimum Variance Distortionless Response (MVDR) beamformer was used for generating enhanced signals. In [20], a fixed beamformer based separation solution was introduced, namely the *unmixing*, *fixed-beamformer* and *extraction* (UFE) system. The mask-based adaptive beamformer of the speech unmixing system is replaced by a process selecting two fixed beamformers from a pre-defined set of beamformers by using a sound source localization (SSL) based beam selection algorithm. This is followed by the speech extraction model introduced in [8] to filter the residual interference in the selected beams. The UFE system has comparable performance with MVDR-based approach, with reduced processing latency.

One limitation of the UFE system lies in its modularized optimization, where each component is individually trained with an indirect objective function. For example, the signal reconstruction objective function used for speech unmixing does not necessarily benefit the accuracy of UFE's beam selection module. As a subsequent work of [20], in this paper, we propose a novel end-to-end structure of UFE (E2E-UFE) model, which utilizes a similar system architecture to UFE, with improved performance thanks to end-to-end optimization. To enable joint training, several updates are implemented on the speech unmixing and extraction networks. We also introduce an attentional module to allow the gradients to propagate though the beam selection module, which was non-differentiable in the original UFE. The performance of the E2E-UFE is evaluated in both block online and offline setups. Our experiments conducted on simulated and semi-real two-speaker mixtures show that E2E-UFE yields comparable results with the original UFE system in the offline evaluation. Significant WER reduction is observed in the block online evaluation.

2. Overview of UFE System

The outline of the UFE pipeline is depicted in Figure. 1, which consists four major components, the fixed beamformer, mask based sound source localization (SSL), speech unmixing network and location based speech extraction network.

In UFE, the M-channel short-time Fourier transform (STFT) of the input speech mixture $\mathbf{Y}_{0,\cdots,M-1} = \{\mathbf{Y}_0, \cdots, \mathbf{Y}_{M-1}\}$ is firstly processed by the speech unmixing module, where a time-frequency mask (TF mask) is estimated

* Work done during internship at Microsoft STCA Beijing.
† corresponding author

http://dx.doi.org/10.21437/Interspeech.2020-1981

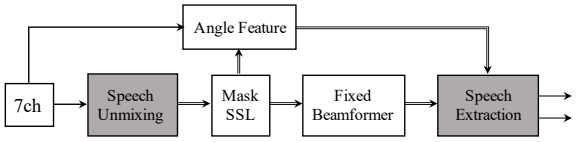

Figure 1: *Overview of the UFE system. The grey block is an neural network trained independently.*

for each participating speaker. In this work, we set the maximum number for simultaneously talking speakers to two so two masks $\mathbf{M}_{0,1} \in \mathbb{R}^{T\times F}$ are generated by unmixing network. The speech unmixing module is trained with permutation invariant training (PIT) criteria with scaled-invariant signal-to-noise ratio (Si-SNR) [21] objective function:

$$\mathcal{L} = -\max_{\phi\in\mathcal{P}} \sum_{(i,j)\in\phi} \text{Si-SNR}(\mathbf{s}_i, \mathbf{x}_j), \tag{1}$$

where $\mathcal{P}$ refers all possible permutations, $\mathbf{x}_j$ is the clean reference of speaker j, and $\mathbf{s}_i$ refers the separated signal of speaker i, which is obtained via inverse short-time Fourier transform (iSTFT):

$$\mathbf{s}_i = \text{iSTFT}\left(\mathbf{M}_i \odot \mathbf{Y}_0\right). \tag{2}$$

Then the sound source localization module is applied to estimate the spatial angle for each separated source with weighted maximum likelihood estimation [20]. The direction of the i-th speaker is estimated via finding a discrete angle θ sampled from 0° to 360° that maximizes the following function:

$$\mathcal{D}_{\theta,i} = -\sum_{t,f} \mathbf{M}_{i,tf} \log\left(1 - \frac{|\mathbf{y}_{t,f}^H \mathbf{h}_{\theta,f}|^2}{1+\epsilon}\right) \tag{3}$$

where $\mathbf{h}_{\theta,f}$ is the normalized steer vector on each frequency band f for source direction θ, ϵ refers a small flooring value, and t denotes frame index in STFT.

With the estimated direction, one beamformer is then selected for each source from a set of pre-defined beamformer, defined as $\mathbf{w}_{n,f} \in \mathbb{C}^{M\times 1}$, where n indexes the beam and each beam has an center angle that is sampled uniformly across the space, and the beamformed signal on each time-frequency bin is obtained by Eqn. 4

$$b_{i,t,f} = \mathbf{w}_{i,f}^H \mathbf{y}_{t,f}, \tag{4}$$

where $\mathbf{y}_{t,f} = [\mathbf{Y}_{0,tf}, \cdots, \mathbf{Y}_{M-1,tf}]^T$.

Finally, the location based speech extraction [8] is applied on each selected beam, and estimates the TF mask based on the input of the beam spectrogram, the inter-microphone phase difference (IPD) and the angle feature [8, 22]. The angle feature on frequency band f is computed as

$$\mathbf{a}_{\theta,f} = \frac{1}{P} \sum_{i,j\in\psi} \cos(\mathbf{o}_{ij,f} - \Delta_{\theta,ij,f}), \tag{5}$$

where ψ contains P microphone pairs and and $\mathbf{o}_{ij,f} = \angle\mathbf{y}_{i,f} - \angle\mathbf{y}_{j,f}$ represents the observed IPD between channel i and j. $\Delta_{\theta,ij,f}$ is the ground truth phase difference given the direction of arrival θ and array geometry. The final output signal is obtained via applying the TF masking on corresponding selected beam, followed by iSTFT.

As fixed beamformer doesn't need to estimate filter coefficients based on input data, it has the potential achieve low latency processing and more robust performance in challenge acoustic environments. And the speech extraction network compensates the limitation in spacial discrimination in fixed beamformer.

3. End-to-end UFE

The proposed end-to-end UFE system is depicted in Figure 2. The overall system workflow is similar to original UFE, while the E2E framework largely simplifies the whole process. The proposed system takes the multi-channel recording as input, and directly outputs two separated speech. A single objective function on the top of the network is used to optimize all parameters.

In original UFE, three components are non-differentiable, which are SSL module, beam selection module and angle feature extraction. To ensure the joint training, we introduce updates to each component.

3.1. Pre-separation layer

In E2E framework, the permutation ambiguity is handled in the final objective function, the unmixing module in UFE reduces to a stack of pre-separation layers. Same as the original UFE, the network takes the IPD and spectrogram of first channel recording as input feature. The pre-separation layers consists of a stack of recurrent layers, followed by H linear projection layers. Here we use $H = 2$ as we consider at most 2 speakers in this paper. After processed by pre-separation layers, an intermediate representation $\mathbf{E} \in \mathbb{R}^{H\times T\times K}$ is formed where K denotes the embedding dimension. We refer $\mathbf{E}$ as the "pre-separation mask" in later context.

3.2. Attentional selection

To avoid the hard angle selection in sound source module, i.e. Eqn. 3. An attention module is applied in E2E-UFE system, which consists of a pool of beamformed signal and angle feature, followed by an attentional selection to estimate the location based bias for final extraction layers.

3.2.1. Spatial feature pool

The spatial feature pool is formed by stacking the spatial feature pointing to different directions. Two pools are formed, one for fixed beamforming and the other for angle feature. For beam pool $\mathbf{B} \in \mathbb{C}^{N_b\times T\times F}$, we calculate the spectrogram of signal obtained though all pre-defined fixed beamformer. In this work, we use $N_b = 18$ beamformers to scan the horizontal space, i.e., 20 degree is covered by each beamformer. The angle feature pool $\mathbf{A} \in \mathbb{R}^{N_a\times T\times F}$ are formed similarly, with $N_a = 36$ directions.

Note that in original UFE, only the beam and angle feature corresponding to the selected angle are calculated, while the E2E UFE calculates beamformed signal and angle feature from all directions beforehand, resulting in an increased computation burden. But this also open the possibility of jointly optimize the beamformer and the angle feature, as suggested in [23], as they are now part of the network. In this work, we freeze the beamformer filter coefficients and angle feature representation. The complex operation in beamforming is implemented using the multiplication of two real matrices [24].

3.2.2. Attentional beam & angle selection

With pre-separation mask $\mathbf{E}$, beam pool $\mathbf{B}$ and angle feature set $\mathbf{A}$ as input, an attention selection module is implemented to

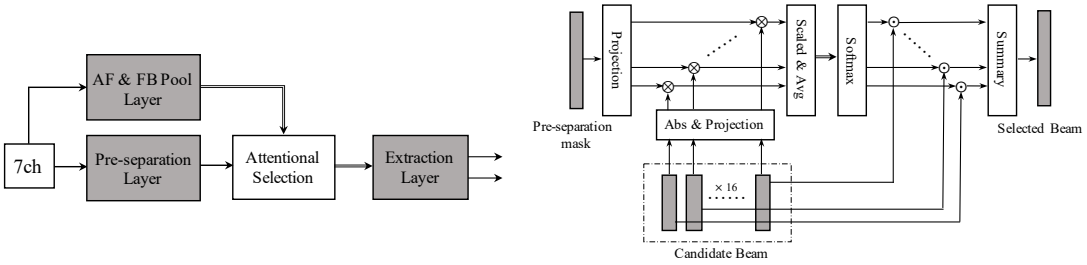

Figure 2: *Overview of the E2E-UFE system (left) and scheme of the attentional beam selection (right). AF and FB are abbreviation of the angle feature and fixed beamformer, respectively. The extraction layers accept both of the weighted beam and angle feature.*

form the location based acoustic bias for each source. The intuition for the attention selection is straightforward, where one attention weight is estimated for each beam and angle feature, based on their learnt similarity with pre-separation representations, followed by a weighted sum to form the final beam and angle feature that are sent to the final extraction layers. In more detail, the attention module is operated in four steps. We use the beam attention as example for illustration, while the selection of angle feature operates in the same manner. The corresponding scheme is depicted in Figure. 2.

Firstly, the pre-separation mask and beam pool are projected using two linear layers:

$$\mathbf{V}^P = \mathbf{E}\mathbf{W}_p, \tag{6}$$

$$\mathbf{V}^B = |\mathbf{B}|\mathbf{W}_b, \tag{7}$$

where $\mathbf{W}_p \in \mathbb{R}^{K\times D}$ and $\mathbf{W}_b \in \mathbb{R}^{F\times D}$ are projection layer weights that convert the pre-separation mask and beam pool into the same dimension D, resulting updated embedding matrices $\mathbf{V}^P \in \mathbb{R}^{H\times T\times D}$ and $\mathbf{V}^B \in \mathbb{R}^{N_b\times T\times D}$.

Then a pair-wised similarity matrix is defined between each frame in $\mathbf{V}^P$ and $\mathbf{V}^B$ using dot product distance, scaled by $(\sqrt{D})^{-1}$. Averaging the similarity matrix along the time axis resulted in beam selection for different time resolution, which is passed by the softmax function to generate the final weight. In Eqn. 8, $s_{h,b,t}$ is the similarity score between h-th pre-separation mask and b-th beam at time t, $\hat{s}_{h,b}$ refers the time averaged weights and $w_{h,b}$ is the final attention weight for each beam. Finally, the weight average operation is performed in order to get the combined beam $\hat{\mathbf{B}}_h$ for h-th speaker, as shown in the Eqn. 11.

$$s_{h,b,t} = (\sqrt{D})^{-1}\left(\mathbf{V}^P_{h,t}\right)^T \mathbf{V}^B_{b,t} \tag{8}$$

$$\hat{s}_{h,b} = (T)^{-1}\sum_t s_{h,b,t}, \tag{9}$$

$$w_{h,b} = \text{softmax}_b(\hat{s}_{h,b}), \tag{10}$$

$$\hat{\mathbf{B}}_h = \sum_b w_{h,b}\mathbf{B}_b. \tag{11}$$

The combined angle feature $\hat{\mathbf{A}}_h$ can be calculated with the same mechanism. The proposed attention module connects the special feature, pre-separation and the later extraction step, ensuring the gradient can be passed in an end-to-end optimization scheme. Note that, the averaging step in Eqn. 9 can be adjusted according to different application scenarios. For offline processing, averaging over entire utterance usually leads to more robust estimation, assuming the position of the speaker is not changed. While averaging only based on past information is more desirable for online processing. The same mechanism can be applied with the other information as well, e.g., speaker inventory [25] or visual clues, etc.

3.3. Joint speech extraction

The combined beam and angle feature estimated via the attentional selection module are processed by the extraction layers. The extraction layers have essentially the same structure as the original UFE, except that the PIT training criteria is required as the permutation ambiguity is not disentangled by unmixing module in E2E framework. We use the clean source from the ground truth beam selection as training target, so both beam selection and wave reconstruction will be optimized with one objective function. Denoting $\mathbf{r}_i$ as the training target for the speaker-i, the objective function is given in a permutation-free form:

$$\mathcal{L} = -\max_{\phi\in\mathcal{P}}\sum_{(i,j)\in\phi}\text{Si-SNR}(\mathbf{s}_i, \mathbf{r}_j), \tag{12}$$

where the $\mathbf{s}_i$ is the network's estimation of the speaker-i.

4. Experiments

4.1. Dataset

The proposed system was trained with multi-channel artificially mixed speech. A total of 1000 hours of training speech data was generated. Source clean speech signals were taken from publicly available datasets, including LibriSpeech[1] [26], Common Voice[2], as well as Microsoft internal recordings. Seven-channel signals were simulated by convolving clean speech signals with artificial room impulse responses (RIRs) generated with the image method [27]. We used the same microphone array geometry as the one used in [20]. The T60 reverberation times were uniformly sampled from [0.1, 0.5] s with a room size of [2,20] m in length and width and [2,5] m in height. The speaker and microphone locations were randomly determined in the simulated rooms. Simulated isotropic noise [28] was added to each mixing utterance at an SNR sampled from [10, 20] dB. We made sure each speech mixture contained one or two speakers, with the mixing SNR between [-5, 5] dB and an average overlapping ratio of 50%. All the data had a sampling rate of 16 kHz.

Two test sets were created for model evaluation. The first test set was created by using the same generation pipeline as the one for the training data, denoted as the *simu* test set, which amounts to 3000 utterances. The speakers were sampled from the *test-clean* set in LibriSpeech. There was no shared speakers

[1] http://www.openslr.org/12/

[2] https://voice.mozilla.org/en

Table 1: *WER (%) performance in the offline evaluation.*

Method	*simu*		*semi-real*	
	OV35	OV75	OV35	OV75
Mixed Beam	67.40	52.40	70.92	57.63
Clean Beam	10.67	10.56	20.34	19.71
UFE	16.44	18.55	35.60	37.54
E2E-UFE	**16.85**	**18.98**	**33.89**	**35.92**

in the training and test sets. The second test set was generated by directly mixing our internal real recorded multi-channel single speaker signals. 2000 mixed utterances were created with the same mixing strategy as in the training set, except that no scaling was applied on the source signals. We refer to this set as the *semi-real* test set. For each set, we created two overlapping conditions, whose overlap ratio ranged from 20—50% or 50–100%. We denote these two condition as OV35 and OV75, respectively.

4.2. Baseline systems

The original UFE system served as the baseline of the proposed E2E architecture. We observed that when trained with the PIT criterion, the extraction model of UFE yielded significantly better results. Therefore, we used PIT-trained extraction in our UFE baseline. For reference, we included the results obtained with the fixed beamforming system applied directly to speech mixtures (Mixed Beam) and those obtained by applying the same beamformers to the clean utterances (Clean Beam), where the beams were selected based on oracle direction of arrival information.

4.3. Training scheme

In the proposed E2E-UFE framework, both extraction and unmixing layers consisted of three contextual LSTM layers [29], each with 512 nodes and a dropout rate of 0.2. For better convergence, the unmxing and extraction networks were pre-trained individually before joint optimization. The same model architecture for unmixing and extraction was used for the UFE baseline. The log magnitude spectrum with an FFT size of 512 and a hop of 256 samples was used as spectral features for all networks. For the unmixing network, cosIPDs between three microphone pairs $(1, 4)$, $(2, 5)$, $(3, 6)$ were extracted.

We used Adam optimizer and train both the networks for a maximum of 80 epochs with a weight decay value of $1e^{-5}$. The early stopping strategy was used to avoid over-fitting. Initial learning rate was set to $1e^{-3}$ and halved if no validation improvement was observed for two consecutive epochs. For joint training in E2E-UFE, a smaller learning rate $1e^{-4}$ was used for fine tuning.

4.4. Evaluation scheme

All systems were evaluated in offline and block online setups. In the offline evaluation, the system was allowed to use the information from an entire utterance. That is, SSL and attentional selection, i.e., Eqn. 3 and 8, respectively, were performed by using averages over the whole utterance. In the block online processing, a double buffering [20] scheme was applied, where each system estimated the output block-wisely through time. Each evaluation block contained a two second window, with additional two or four second history information. The hop between two evaluation block was two seconds, resulting in an average latency of one second.

Table 2: *WER (%) performance in the online evaluation.*

Method (history)	*simu*		*semi-real*	
	OV35	OV70	OV35	OV70
UFE (2s)	24.10	31.40	44.05	45.13
UFE (4s)	23.66	28.85	43.49	44.06
E2E-UFE (2s)	17.50	19.43	38.64	39.98
E2E-UFE (4s)	**17.09**	**19.10**	**36.67**	**39.11**

The word error rate (WER) was used as a performance metric. The ASR pipeline we used for decoding included a tri-gram language model and an acoustic model consisting of six layers of 512-element layer trajectory LSTM [30]. The acoustic model was trained with maximum mutual information (MMI) [31] on 30k hours of noise-corrupted data.

4.5. Results

The offline evaluation results are shown in Table 1. The simple fixed beamforming (Mixed Beam) yielded a high WER even though it used the oracle DoA. The result of the clean beam sets the upper bound to the UFE performance. The proposed E2E-UFE system achieved comparable performance as the original UFE for the simulated data set, while demonstrating a clear performance advantage in *semi-real* the semi-real set, showing the efficacy of the end-to-end training scheme. Overall, E2E-UFE achieved 4.8% and 4.3% relative WER reduction over the UFE system on OV35 and OV75 of the *semi-real*, respectively, reaching 33.89% and 35.92% WERs.

Table 2 shows the block online evaluation results. E2E-UFE shows robustness for different look-back configurations (a 2s or 4s history context), achieving slightly worse results than for the offline evaluation on both datasets. On the *simu* set, E2E-UFE showed no significant degradation compared with the offline performance. It achieved lower WERs than the original UFE. On the *semi-real* set, it brought about a 12.47% average relative WER reduction compared with the UFE system using a 2 s history context, while on the *simu* set, the relative reduction increases to 29.71% . By contrast, the original UFE resulted in a much larger performance degradation for the online evaluation, degrading from 16.44/18.55% to 24.10/31.40% on the *simu* set and 35.60/37.54% to 44.05/43.15% on the *semi-real* set. One hypothesis for the robustness of E2E-UFE is that, during training, the E2E-UFE model already optimized for wrong beam selections, while for the original UFE, only the correct beams were selected as input. Another potential reason could be that the sparification trick in [20] was not applied in either UFE or E2E-UFE, which might result in more energy leakage for UFE system, while E2E-UFE system doesn't suffer from this problem as all modules are jointly optimized.

5. Conclusion

In this paper, we proposed an end-to-end structure of multi-channel speech separation, named E2E-UFE, for robust ASR. It replaces the SSL module in the previously proposed UFE system with a small attention network and enables joint optimization of the unmixing and extraction networks. The experiments were conducted on two 2-speaker datasets (simulated and semi-real mixtures) and the performance was evaluated for both offline and online settings. The experimental results showed that E2E-UFE provided comparable performance with the UFE system in the offline situations and yielded an average relative WER reduction of 12.47% on block online processing.

6. References

[1] 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 Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 10, pp. 1901–1913, 2017.

[2] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[3] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 71–75.

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[5] T. Higuchi, N. Ito, S. Araki, T. Yoshioka, M. Delcroix, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[6] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures." in *Interspeech*, 2017, pp. 2655–2659.

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

[8] 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 *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 558–565.

[9] C. Boeddeker, J. Heitkaemper, J. Schmalenstroeer, L. Drude, J. Heymann, and R. Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[10] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, and D. Yu, "Improved speaker-dependent separation for chime-5 challenge," *arXiv preprint arXiv:1904.03792*, 2019.

[11] T. Yoshioka, H. Erdogan, Z. Chen, X. Xiao, and F. Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," *arXiv preprint arXiv:1810.03655*, 2018.

[12] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[13] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 184–196, 2018.

[14] M. W. Lam, J. Wang, X. Liu, H. Meng, D. Su, and D. Yu, "Extract, adapt and recognize: an end-to-end neural network for corrupted monaural speech recognition," *Proc. Interspeech 2019*, pp. 2778–2782, 2019.

[15] N. Kanda, S. Horiguchi, R. Takashima, Y. Fujita, K. Nagamatsu, and S. Watanabe, "Auxiliary interference speaker loss for target-speaker speech recognition," *arXiv preprint arXiv:1906.10876*, 2019.

[16] X. Chang, Y. Qian, K. Yu, and S. Watanabe, "End-to-end monaural multi-speaker asr system without pretraining," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6256–6260.

[17] M. Delcroix, S. Watanabe, T. Ochiai, K. Kinoshita, S. Karita, A. Ogawa, and T. Nakatani, "End-to-end speakerbeam for single channel target speech recognition," *Proc. Interspeech 2019*, pp. 451–455, 2019.

[18] N. Kanda, Y. Gaur, X. Wang, Z. Meng, and T. Yoshioka, "Serialized output training for end-to-end overlapped speech recognition," *arXiv preprint arXiv:2003.12687*, 2020.

[19] T. Yoshioka, I. Abramovski *et al.*, "Advances in online audio-visual meeting transcription," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2019.

[20] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitriadis, "Low-latency speaker-independent continuous speech separation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6980–6984.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[22] Z.-Q. Wang and D. Wang, "On spatial features for supervised speech separation and its application to beamforming and robust asr," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5709–5713.

[23] W. Minhua, K. Kumatani, S. Sundaram, N. Ström, and B. Hoffmeister, "Frequency domain multi-channel acoustic modeling for distant speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6640–6644.

[24] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, "Deep complex networks," *arXiv preprint arXiv:1705.09792*, 2017.

[25] P. Wang, Z. Chen, X. Xiao, Z. Meng, T. Yoshioka, T. Zhou, L. Lu, and J. Li, "Speech separation using speaker inventory," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 230–236.

[26] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[27] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[28] E. A. Habets and S. Gannot, "Generating sensor signals in isotropic noise fields," *The Journal of the Acoustical Society of America*, vol. 122, no. 6, pp. 3464–3470, 2007.

[29] J. Li, L. Lu, C. Liu, and Y. Gong, "Improving layer trajectory lstm with future context frames," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6550–6554.

[30] J. Li, C. Liu, and Y. Gong, "Layer trajectory lstm," *arXiv preprint arXiv:1808.09522*, 2018.

[31] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks." in *Interspeech*, vol. 2013, 2013, pp. 2345–2349.