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]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Microsoft, STCA, Beijing, China
[3]Microsoft, One Microsoft Way, Redmond, WA, USA

ABSTRACT

Although mask based adaptive beamforming technique benefits speech recognition in far-field, noisy and multi-talker scenarios, it depends on the long time context to estimate target and interference statistics, thus when applied in applications with low latency requirement, its performance usually drops drastically. In contrast, the fixed beamformers do not import time delay but usually have limited capability in acoustic cancellation of interfering source. In this work, we propose a novel multi-channel speech separation system that targets at overlapped speech recognition with low latency processing, which includes four jointly optimized components: a pre-separator, a set of fixed beamformer, an attentional selection module and neural post filtering. With proposed model, low latency processing is achieved by utilizing the known microphone geometry information, while keeps the high quality separation through neural post filtering and end-to-end optimization. In our experiments, we show that the proposed system achieves comparable performance in offline evaluation with the mask based MVDR and speech extraction system, while yield remarkable improvements in the online evaluation.

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

1. INTRODUCTION

Deep learning approaches have brought remarkable progresses to speaker-independent speech separation in the past few years [1, 2, 3, 4] and the consistent improvements of the separated signal quality are reported on the benchmarking dataset such as WSJ0-2mix[2]. However, multi-talker speech recognition is still remained as a challenging problem [5].

Speech separation is a common practice to handle the overlapped speech. Existing efforts in overlapped speech recognition can be roughly categorized into two families: on building a robust separation system as front-end processor for automatic speech recognition (ASR) tasks [6, 7, 8, 9, 10, 11, 12] , or directly train a special multi-talker awared acoustic model [13, 14, 15, 16, 17, 18]. Although better performance is usually expected by the end to end training with acoustic model, the separately optimized front end processing is usually preferable in real world applications such as meeting transcription [19] for two reasons. Firstly, in conversation transcription systems, the individual front end module benefits multiple acoustic component including speech recognition, diarization and speaker verification. Secondly, the industrialized acoustic model is often trained with large amount of data and highly engineering optimized, thus changing the training scheme might potentially introduce performance disturbance and will result in long developing cycle.

In [19], the author introduced the application of speech separation in an advanced conversation transcription system, where a multi-channel separation network, namely speech unmixing network, is trained with permutation-invariant training (PIT) criteria [1]. The speech unmixing continuously separate the input audio stream into two channels, ensuring each channel only contain at most 1 activate speaker. The mask based MVDR is introduced as post processing, for better speech recognition performance.

Although the speech unmixing significantly boosts the transcription accuracy, it requires long processing latency, i.e. more than 1.2s in [19], as MVDR beamformer usually requires long context for accurate estimation of spatial statistics for each frequency. To overcome this, a solution named *unmixing-fixbeam-extraction* (UFE) was introduced in [20]. In UFE, the adaptive beamformer in speech unmixing system was replaced by a set of pre-defined beamformer and a sound source localization (SSL) based beam selection algorithm. To further increase the separation power, the speech extraction model introduced in [9] was applied to post filter the selected beams. The UFE system has been shown to have comparable performance, while significantly reduce the processing latency from 1.20s to 0.38s.

As a subsequent work of [20], in this paper, we propose a novel end-to-end structure of UFE (E2E-UFE) model for robust ASR, which possess the low latency natural of the UFE, but achieved improved performance by an end to end optimization scheme of the whole UFE processing.

The E2E-UFE model follows the same "separation-selection-filtering" scheme as original UFE, with several updates introduced to enable the end to end training. Firstly, as the original sound source localization step was not differentiable, which is replaced by a set of pre-calculated angle feature sampling across the space. Then we propose an attentional selection module between the pre-separation mask embedding, and the pool of fixed beamformer output and angle feature pool. Finally, the PIT training is applied on the filtering output, to enhance the performance of close speaker pairs.

The performance of the E2E-UFE is evaluated in both online and offline mode. The experiments conducted on the simulated and semi-real two-speaker mixtures show comparable results with the UFE system and mask based MVDR beamformer in the offline evaluation while yield significant WER reduction in the online mode, respectively.

The rest of the paper is organized as follows. In Section 2, we briefly reviewed the cascaded UFE system. In Section 3, the E2E-UFE structure is proposed to joint model source localization, unmixing and extraction network. The experimental settings, training details, evaluation scheme, performance results and analysis will be

* Work done during internship at Microsoft STCA Beijing

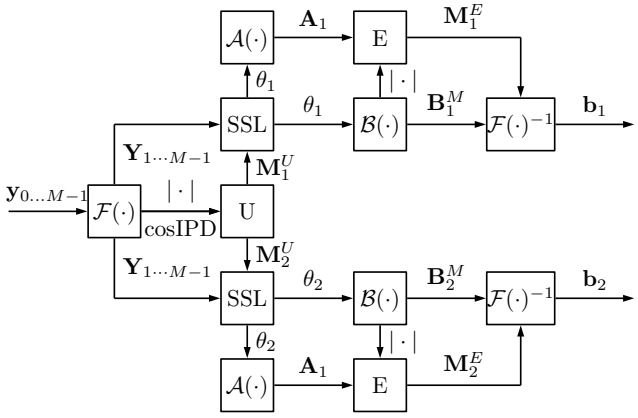

Fig. 1. Overview of the UFE system. $\mathcal{F}$, $\mathcal{B}$, $\mathcal{A}$ and SSL denote short-time Fourier transform (STFT), fixed beamforming, angle feature computation and SSL algorithm, respectively. $\mathbf{M}_i^U$ and $\mathbf{M}_i^E$ represent the TF-masks of the i-th speaker generated by *unmixing* (U) and *extraction* (E) network. $\mathbf{A}_i$ and $\mathbf{B}_i^M$ denote the angle feature and the selected beam given the speaker direction θ_i. The *unmixing* and *extraction* model are trained independently.

discussed in Section 4.

2. OVERVIEW OF UFE SYSTEM

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

The $X_{c,tf}$ is firstly processed by the speech unmixing module, resulting two speaker masks, denoted as $\mathbf{M}_{i,tf}$, where i indexes the speaker, c refers the channel index, t, f index the time and frequency axes.

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

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

where $\mathbf{h}_{\theta,f}$ is the normalized steer vector on each frequency band f given for each angle θ.

Meanwhile, a set of fixed beamformer $w_{n,f}$ is pre-defined, where n index the beam, the beam center is sampled uniformly across the space. For each speaker, the closest beam is selected from the estimated angle, and the beamfored signal is obtained by equation xxx

$$\mathcal{B}(\theta, f) = \mathbf{w}_{\theta,f}^H \mathbf{y}_{t,f}, \quad (2)$$

where $\mathbf{w}_{\theta,f} \in \mathbf{C}^{M\times 1}$ represents the beam weight that covers the look direction θ and $\mathbf{y}_{t,f} \in \mathbf{C}^{M\times 1}$ denotes STFT of the mixture signal $\mathbf{y}_M$ at frame t.

Then, the location based speech extraction[9] is applied on each selected beam, which takes spectrogram of selected beam, the inter-microphone phase difference(eqn xx) (IPD) among all channels and angle feature(eqnxx) as input, where xxx The angle feature, also known as location based bias, differentiates the target and interfering speaker, avoiding the permutation problem in multi-speaker system. Note that the speech extraction is applied on each beam independently. Finally each mask estimated by the speech extraction module is applied on corresponding beam, to obtain the final enhanced speech, as shown in eqn.xxx.

$$\mathcal{A}(\theta, f) = \sum_{i,j\in\psi} \cos(\mathbf{o}_{ij,f} - \mathbf{r}_{\theta,ij,f}). \quad (3)$$

$$\mathcal{A}(\theta, f) = \sum_{i,j\in\psi} \cos(\mathbf{o}_{ij,f} - \mathbf{r}_{\theta,ij,f}). \quad (4)$$

3. END-TO-END UFE

To enable the end-to-end training and improve the speech separation quality, four updates to the original UFE model are proposed to each component, while the whole processing sequence operates similarly. The system workflow is depicted in figure xx. In e2e frame work, the whole process is largely simplified, where the input of xx. Inside, a similar four processing blocks can be still roughly divided.

3.1. Speech unmixing

In E2E frame work,

3.2. Feature and beamformer

We include fixed beamformer (subnet B) and angle feature extractor (subnet A) as a part of the network, as depicted in Fig.2.The complex multiplication of the fixed beamformer in Eqn.2 can be easily implemented using the multiplication of two real matrices [21], while the linear parameters are initialized with the fixed beam weights. Different from UFE system, the fixed beamformer in E2E-UFE always generate N_B beams. STFT and iSTFT are realized using 1D convolutional operations. A is a special network that is responsible for computing N_D candidate angle feature matrices on N_D directions sampled from 0° to 360°, following Eqn.4. It this work, we freeze the beamformer weight during training and $N_B = 18, N_D = 36$. However, as a neural network layer, the beamformer can be also jointly learnt during training, as suggested in [22].

3.3. Attention based angle selection

Since the SSL based beam selection in original UFE model is non-differentiable, to enable the joint optimization, we introduce an attention module in the E2E-UFE which selects the beam based on similarity metric. Firstly, two sets of embedding $\mathbf{E}_i^U, \mathbf{E}_d^B \in \mathbf{R}^{T\times D}$ for each unmixing output $\mathbf{M}_i^U$ and fixed beam $\mathbf{B}_d$ are estimated through linear mapping

$$\mathbf{E}_i^U = \mathbf{M}_i^U \mathbf{W}_m, \quad (5)$$

$$\mathbf{E}_g^B = |\mathbf{B}_g| \mathbf{W}_b, \quad (6)$$

where i, g denote speaker and beam index, respectively. D is the embedding size and $\mathbf{W}_m, \mathbf{W}_b \in \mathbf{R}^{F\times D}$ are transform matrices. We add absolution in Eqn.5 because $\mathbf{B}_g$ is a complex-valued matrix. Then the pair-wise dot-product distance between each mask

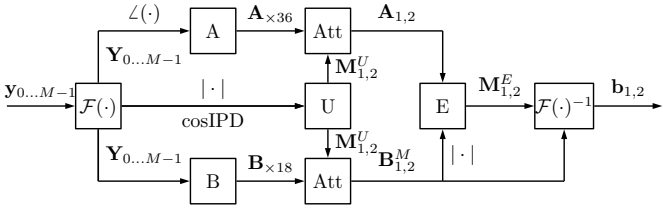

Fig. 2. Overview of the E2E-UFE structure. A and B are two special network, using for angle feature extraction and fixed beamforming. The whole network including U, E and attention network can be jointly optimized.

embedding and beam embedding are computed to form the similarity matrix:

$$s_{i,g,t} = (\sqrt{D})^{-1} \left(\mathbf{E}^U_{i,t}\right)^T \mathbf{E}^B_{g,t} \tag{7}$$

The similarity matrix is then averaged along time axis, followed by a softmax activation, to generate the weight $w_{i,g}$ for each acoustic source

$$\hat{s}_{i,g} = (T)^{-1} \sum_t s_{i,g,t}, \tag{8}$$

$$w_{i,g} = \text{softmax}_g(\hat{s}_{i,g}) \tag{9}$$

The output beam $\mathbf{B}^M_i$ is calculated by the weighed average with each beam and its corresponding attention weights for each source, as shown in Eqn.9

$$\mathbf{B}^M_i = \sum_g w_{i,g} \mathbf{B}_i \tag{10}$$

The similar progress is also applied to obtain the angle features $\mathbf{A}_{1,2}$ as depicted in Fig.2, replacing the SSL based angle feature generation in original UFE. And the weighted combination of angle feature is used in the neural filtering step.

3.4. Joint speech extraction

In E2E-UFE model, two beams are jointly optimized. Therefore the post filtering network takes both combined beams and angle features as input, outputting the two beams simultaneously. We use the Si-SNR [4] as the loss function to optimize the network. Representing the expected beam signal as $\mathbf{r}_{1,2}$, the loss function is given in a permutation-free form

$$\mathcal{L}_{\text{beam}} = -\max_{i,j\in\phi}\{\text{Si-SNR}(\mathbf{b}_i, \mathbf{r}_j)\}, \tag{11}$$

where $\mathbf{b}_i = \text{iSTFT}(\mathbf{B}^M_i \odot \mathbf{M}^E_i)$ and ϕ is a set of permutations over two speakers. To better direct the training of the unmixing network, we add an additional loss for unmixing part as a regular term:

$$\mathcal{L}_{\text{ch0}} = -\max_{i,j\in\phi}\{\text{Si-SNR}(\mathbf{c}_i, \mathbf{s}_j)\}, \tag{12}$$

where $\mathbf{c}_i = \text{iSTFT}(\mathbf{Y}_{\text{ch0}} \odot \mathbf{M}^U_i)$ and $\mathbf{s}_{1,2}$ denotes the clean signal at the reference channel (0 used in our experiments). The final objective function is a combination of those two forms with a weight α

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{\text{ch0}} + \alpha\mathcal{L}_{\text{beam}}. \tag{13}$$

Note that, although we still use signal reconstruction as neural network training criteria, it is straightforward to jointly optimize with the acoustic model, as suggested in [22].

4. EXPERIMENTS

4.1. Dataset

We create a large-scale close-talk dataset including Librispeech [23] and three internal dataset from Microsoft for far-field mixture data simulation. The sampling rate is 16kHz and the training mixture utterances are simulated and segmented into 4s chunks on-the-fly. The simulation details of each epoch is configured by a specific file independently. Geometry of the microphone array and the fixed beamformer are some as the work [20], and the room impulse responses (RIRs), isotropic noise are generated in advance. The T60 value, signal-to-noise ratio (SNR) and signal-to-interference ratio (SIR) for each utterance is randomly sampled from the range [0.1, 0.5] s, [10, 20] dB and [-5, 5] dB, respectively. The two speakers are put at least 20° apart from each other and 1 meter away from the center of microphones. The overlapping ratio is controlled in 50$\sim$100% during training.

For model evaluation we create two types of the test set, *simu* and *semi-real* and each set contains two subsets which overlapping ratios are distributed in the range of 20$\sim$50% and 50$\sim$100%, denoted as *S* and *L*, respectively. The *simu* set uses the close-talk data from the *dev* set of the Librispeech with simulated RIRs that is different from the training stage and *semi-real* is mixed using single-speaker real recordings. Each subset in *simu* contains 3000 utterances and *semi-real* contains 2000 utterances. Additional noise is only added in *simu* set as the *semi-real* already contains real-recording background noise. Similar with the training settings, the speakers are kept at least 20° apart from each other in each dataset.

4.2. Baseline systems

The mask based MVDR beamformer we used follows the equation

$$\mathbf{w}^{\text{MVDR}}_f = \frac{(\mathbf{R}^n_f)^{-1}\mathbf{d}_f}{\mathbf{d}^H_f(\mathbf{R}^n_f)^{-1}\mathbf{d}_f} \tag{14}$$

and the steer vector $\mathbf{d}_f$ is estimated using the principal eigenvector of the $\mathbf{R}^s_f$. Two spatial correlation matrices $\mathbf{R}^k_f, k \in \{s, n\}$ is estimated via

$$\mathbf{R}^k_f = \frac{\sum_t m^k_{t,f}\mathbf{y}^H_{t,f}\mathbf{y}_{t,f}}{\sum_t m^k_{t,f}}. \tag{15}$$

As the unmixing network do not predict noise masks here, we use $m^n_{t,f} = 1 - m^s_{t,f}$ instead.

During experiment, we found that the UFE trained with PIT criteria, i.e. joint optimizing two beams, yields significantly better results than original UFE. Therefore we use PIT-UFE instead of separately trained components as the baseline to the proposed E2E-UFE model. SSL from Eqn.1 was used to estimate the DoA for the PIT-UFE system and we tested it with the mask from both unidirectional and bidirectional unmixing network. An oracle DoA result was also calculated for reference and a layer trajectory LSTM acoustic model introduced in [24] with a trigram language model was used to generate recognition result.

4.3. Training details

Our experiments take log magnitude spectrum as spectral features and the chunk-level mean-variance normalization is used during training. The frame length in STFT is 32ms long with 16ms hop size. For unmixing network, cosIPDs between three microphone pairs $(1, 4)$, $(2, 5)$, $(3, 6)$ are concatenated as spatial features and δ is set as 2 in uni-directional structure. The angle feature used in

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

Method	DoA	*simu (S/L)*	*semi-real (S/L)*
Mixed Beam	Oracle	67.40/52.40	70.92/57.63
Clean Beam	Oracle	10.67/10.56	20.34/19.71
IRM + MVDR	×	21.69/20.04	28.52/26.75
PIT-UFE	Oracle	16.15/18.22	31.22/33.38
	bid-U + SSL	**16.41/18.43**	**32.91/36.02**
	uni-U + SSL	16.44/18.55	35.60/37.54
E2E-UFE	×	16.85/18.98	**33.89/35.92**
bid-U + MVDR	×	**23.43/21.52**	**35.89/34.72**
uni-U + MVDR	×	24.65/23.39	40.29/38.27

extraction network is summarized over six microphone pairs $(p, 0)$ where $p \in [1, 6]$. We use Adam optimizer and train both network for a maximum of 80 epochs with a weight decay value of $1e^{-5}$ and the early stopping strategy. Initial learning rate is set to $1e^{-3}$ and will halve if no validation improvement in 2 consecutive epochs.

Both unmixing and extraction network contain 3 layers, each with 512 units and a dropout rate of 0.2. For the better convergence, we use the well-trained unmixing and extraction network to initialize the corresponding parts in E2E-UFE and train the network with a smaller learning rate $1e^{-4}$ for 20 epochs. The training data and other configurations are kept same as UFE. α in Eqn.12 was adopt 0.8 in our experiments.

4.4. Evaluation scheme

The proposed system and baseline were evaluated in both offline and online setup. In offline evaluation, the MVDR, SSL and attentional selection weight was averaged over the whole utterance. And in online evaluation, we use the *double buffering* scheme which is similar to [20]. Each time the network processes T_c seconds audio segment with a look-back length of T_b seconds. The T_b seconds buffer is used to initialize the inner states of the unidirectional recurrent network as well as to align the output orders. The look-back buffer in the SSL aims to create a smooth estimation of the directions. Spatial correlation matrices $\mathbf{R}^k_{b,f}$ used in online MVDR are recursively estimated in each segment b with a forget factor β:

$$\mathbf{R}^k_{b,f} = \beta \mathbf{R}^k_{b-1,f} + (1-\beta)\frac{\sum_{t\in b} m^k_{t,f}\mathbf{y}^H_{t,f}\mathbf{y}_{t,f}}{\sum_{t\in b} m^k_{t,f}}, k \in \{n, s\} \quad (16)$$

We used $T_c = 2s$, $\beta = 0.8$ and the unidirectional unmixing layers to reduce processing latency in the experiments.

4.5. Results

The offline evaluation result are shown in Table 1. The fixed beamformer (Mixed Beam) brings a high WER even using the oracle DoA and the result of the clean beam shows the upper bound performance of the UFE system. The unidirectional unmixing structure achieved similar performance with bidirectional model for both MVDR and PIT-UFE baseline on *simu* set, as the spatial pattern is usually more stable for simulated dataset, while for *semi-real* set, the bidirectional unmixing model showed clear performance advantages. In comparison, although using unidirectional unmixing layers, the E2E-UPE achieved a comparable performance with bidirectional MVDR and PIT-UFE and a significantly better performance than unidirectional baselines, showing the efficacy of the end-to-end training scheme.

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

Method	T_b(s)	*simu (S/L)*	*semi-real (S/L)*
PIT-UFE	2/2/2	31.40/24.10	44.05/45.13
	2/4/2	29.04/23.78	44.55/44.44
	4/2/2	31.24/24.00	**43.26/43.31**
	4/4/2	**28.85/23.66**	43.49/44.06
E2E-UFE	2	17.50/19.43	38.64/39.98
	4	**17.09/19.10**	**36.67/39.11**
uni-U + MVDR	4	50.99/35.65	53.78/43.87

The online performance is shown in Table 2. The T_b for the UFE system represents the look-back time of the unmixing network, SSL and extraction network, respectively. E2E-UPE shows robust performance for different look-back setup, achieving slightly worse result than the offline evaluation on both dataset. On *simu* set, E2E-UFE shows no significant degradation with the offline mode but the much lower WER than UFE and MVDR method, especially in low overlapping scenario. On *semi-real* set it brings a 12.47% and 22.40% average relative WER reduction compared with the UFE system and the mask based MVDR, respectively. As contrast, both PIT-UFE and MVDR baseline observed a severe performance degradation in online evaluation. One hypothesis for the robustness of E2E-UFE could be that during training, the E2E-UFE model already optimized for the wrong beam selection, while for PIT-UFE, only the correct beams were selected as input.

Online MVDR degrades seriously in low overlapping scenario as the interfering speaker can not be suppressed well in the zero-mask streams, which will increase the insertion errors in the final transcriptions, an additional VAD suggested in [12] could be a potential remedy for this challenge.

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 the joint optimization of the unmixing and extraction network. The experiments are conducted on two 2-speaker dataset (simulated and semi-real mixtures) and the performance is evaluated in both offline and online mode. Experimental results show that E2E-UFE gives comparable performance with UFE and the mask based MVDR in the offline situations but shows 12.47% and 22.40% average relative WER reduction on two test sets in the online mode, respectively.

6. REFERENCES

[1] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper 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] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji 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] Zhong-Qiu Wang, Ke Tan, and DeLiang 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] Yi Luo and Nima 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] Chao Weng, Dong Yu, Michael L Seltzer, and Jasha Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[6] Takuya Higuchi, Nobutaka Ito, Shoko Araki, Takuya Yoshioka, Marc Delcroix, and Tomohiro 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.

[7] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures.," in *Interspeech*, 2017, pp. 2655–2659.

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

[9] Zhuo Chen, Xiong Xiao, Takuya Yoshioka, Hakan Erdogan, Jinyu Li, and Yifan 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.

[10] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the chime-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[11] Jian Wu, Yong Xu, Shi-Xiong Zhang, Lian-Wu Chen, Meng Yu, Lei Xie, and Dong Yu, "Improved speaker-dependent separation for chime-5 challenge," *arXiv preprint arXiv:1904.03792*, 2019.

[12] Takuya Yoshioka, Hakan Erdogan, Zhuo Chen, Xiong Xiao, and Fil Alleva, "Recognizing overlapped speech in meetings: A multichannel separation approach using neural networks," *arXiv preprint arXiv:1810.03655*, 2018.

[13] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[14] Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne 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.

[15] Max WY Lam, Jun Wang, Xunying Liu, Helen Meng, Dan Su, and Dong Yu, "Extract, adapt and recognize: an end-to-end neural network for corrupted monaural speech recognition," *Proc. Interspeech 2019*, pp. 2778–2782, 2019.

[16] Naoyuki Kanda, Shota Horiguchi, Ryoichi Takashima, Yusuke Fujita, Kenji Nagamatsu, and Shinji Watanabe, "Auxiliary interference speaker loss for target-speaker speech recognition," *arXiv preprint arXiv:1906.10876*, 2019.

[17] Xuankai Chang, Yanmin Qian, Kai Yu, and Shinji 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.

[18] Marc Delcroix, Shinji Watanabe, Tsubasa Ochiai, Keisuke Kinoshita, Shigeki Karita, Atsunori Ogawa, and Tomohiro Nakatani, "End-to-end speakerbeam for single channel target speech recognition," *Proc. Interspeech 2019*, pp. 451–455, 2019.

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

[20] Takuya Yoshioka, Zhuo Chen, Changliang Liu, Xiong Xiao, Hakan Erdogan, and Dimitrios 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] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, João Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal, "Deep complex networks," *arXiv preprint arXiv:1705.09792*, 2017.

[22] Wu Minhua, Kenichi Kumatani, Shiva Sundaram, Nikko Ström, and Björn 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.

[23] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev 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.

[24] Jinyu Li, Changliang Liu, and Yifan Gong, "Layer trajectory lstm," *arXiv preprint arXiv:1808.09522*, 2018.