EFFICIENT INTEGRATION OF FIXED BEAMFORMERS AND SPEECH SEPARATION NETWORKS FOR MULTI-CHANNEL FAR-FIELD SPEECH SEPARATION

Zhuo Chen, Takuya Yoshioka, Xiong Xiao, Jinyu Li, Michael L. Seltzer, Yifan Gong

Microsoft AI & Research, One Microsoft Way, Redmond, WA, USA

ABSTRACT

Speech separation research has significantly progressed in recent years thanks to the rapid advances in deep learning technology. However the performance of recently proposed single-channel neural network-based speech separation methods is still limited especially in reverberant environments. To push the performance limit, we recently developed a method of integrating beamforming and single-channel speech separation approaches. This paper proposes a novel architecture that integrates multi-channel beamforming and speech separation in a much more efficient way than our previous method. The proposed architecture comprises a set of fixed beamformers, a beam prediction network, and a speech separation network based on permutation invariant training (PIT). The beam prediction network takes in the beamformed audio signals and estimates the best beam for each speaker constituting the input mixture. Two variants of PIT-based speech separation networks are proposed. Our approach is evaluated on reverberant speech mixtures under three different mixing conditions, covering cases where speakers partially overlap or one speaker's utterance is very short. The experimental results show that the proposed system significantly outperforms the conventional single-channel PIT system, producing the same performance as a single-channel system using oracle masks.

Index Terms— Cocktail party problem, permutation invariant training, acoustic beamforming, beam prediction, speech separation

1. INTRODUCTION

Thanks to the recent advances in deep learning, machine realization of the cocktail party effect has been significantly progressed. Given an audio recording of multiple speakers talking at the same time, the cocktail party problem is defined as separating and recognizing the spoken content of each speaker [1]. The cocktail party problem is considered as one of the most challenging problems in speech signal processing for decades because the systems must track more than one speakers unlike in many other speech processing tasks. The huge acoustic variations that are often seen in natural multi-talker scenarios, caused by reverberation, additive noise, and speaker variability, also need to be handled.

Two classes of neural network-based algorithms were proposed to address the speaker-independent speech separation problem: an embedding approach and a direct approach. With the former class of algorithms, each time-frequency (TF) point of the observed mixture signal is embedded into a fixed-dimensional space by a neural network. The embedding features, each associated with a certain TF point, are clustered, where each cluster is assumed to correspond to one of the speakers participating in the input mixture. The neural network is trained so that the embedding features of the same speaker are close to each other. Deep clustering (DC) [2, 3] and deep attractor networks [4] are two representative embedding-based methods. The direct approach mostly follows the single-speaker mask-based speech enhancement framework [5, 6, 7], where a neural network is asked to output spectral masks to be applied to the microphone signal to enhance the speech. The masks are trained to minimize the difference between the enhanced speech and the clean speech. Permutation invariant training (PIT) [8, 9] can be regarded as an extension of this to multi-speaker scenarios. In the standard WSJ0-derived task, which uses fully overlapped speech signals, PIT was shown to perform comparably to the embedding based method [8]. PIT is preferable in practice because it allows for simpler and more efficient implementation.

While the neural network-based approaches have brought significant advances to the speech separation research, their separation performance is still limited for real world applications. In addition, as shown in [10], single-channel systems are vulnerable to acoustic variations resulting from the presence of reverberation. Integrating multi-channel processing with speech separation algorithms is one possible way for improving the separation performance in such far-field environments. The use of multiple microphones allows us to exploit spatial information in addition to spectral features. One method of integrating multi-channel processing and DC is described in [11], where clustering is performed using DC-derived embedding features and spatial features. While the method does not require knowledge of the microphone array geometry, it necessitates a rather complex clustering algorithm to jointly deal with the two types of features. In [10], a set of fixed beamformers, each with a distinct directivity pattern, is first applied to an observed multi-channel signal to yield a set of beamformed audio. Then a separation network is applied to each beamformed signal. Out of the speech separation results for all the beams, the best separated signal is selected for each speaker by a post selection algorithm. This system achieved significant improvement over the single-channel system for different types of mixtures including 2, 3 and 4 speakers. However, performing speech separation on each of the beamformed signals is computationally very expensive and not tolerable for many applications.

It should also be pointed out that most previous studies used fully overlapped utterances for evaluation, which are rarely seen in actual usage scenarios. In reality, utterances by different speakers usually only partially overlap. Another typical cause of speech overlap is back-channels, where one speaker's utterance is very short.

In this paper, we propose a multi-channel speech separation method that efficiently integrates fixed beamformers and neural network speech separation systsems. As with [10], a set of pre-defined beamformers is applied to an input multi-channel signal. Then a beam selection network is trained to predict the best beam to use for each speaker. The outputs from the selected beams are fed into a PIT network for further separation. This architecture has several desirable properties as discussed in Section 2.4. The proposed method

is evaluated using test sets which encompass three different mixing conditions.

The rest of the paper is organized as follows. Section 2 explains the proposed system. Section 3 describes our experimental setup, followed by results and discussion in Section 4. Section 5 concludes this paper.

2. PROPOSED ARCHITECTURE

The basic design idea behind our proposed architeture is to perform multi-channel signal processing using pre-defined well-engineered beamformers, followed by additional single-channel processing. The multi-channel processing part enhances individual speakers' signals by using pre-defined beamformers. Our single channel processing further performs speech separation on the beamformed signals, which relies on spectral properties of the speech.

Figure 1 shows a diagram of the proposed architecture. First, the input multi-channel signal is processed with a set of fixed beamformers. The beamformer directivity patterns are optimally desinged in advance for a target microphone array geometry. Then a beam prediction network is applied to the beamformed speech signals. For each speaker, this network predicts the beam that provides the maximum signal-to-distortion ratio (SDR) for the speaker. Finally, the signals from the selected beams are fed to a PIT-trained speech separation network to obtain the final separation result.

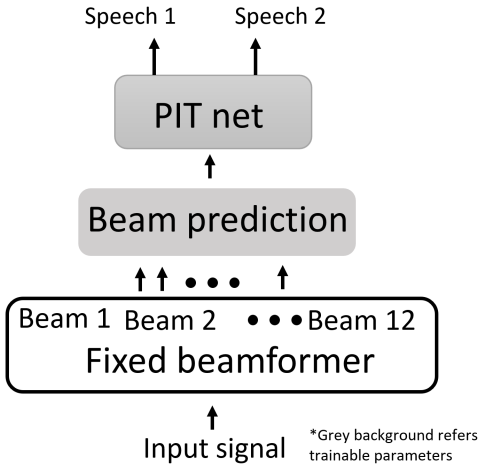

Fig. 1. Proposed speech separation system architecture.

2.1. Fixed beamformers

A set of pre-defined beamformers is firstly applied to the multi-channel input signal. The idea here is to uniformly 'sample' the space of direction of arrivals (DOAs) by using a set of fixed beamformers that have different look directions. Therefore, while this paper presents one realization of the proposed framework by using a specific set of beamformers, we expect that the framework can be applied to arbitrary arrays.

As with our prior work in [10], we make use of differential beamforming to define the beamformer set. One advantage of differential beamformers is that they have acoustic nulls, which allows spatially isolated sound sources to be effectively removed. Our beamformers were engineered by following [12]. In our experiments, we simply used 12 differential beamformers, whose look directions were separated by 30 degrees, covering the whole 360 degrees. The beam directivity patterns were hand-crafted.

2.2. Beam prediction network

The 12 beamformed signals obtained as described above are fed into a beam selection network, which predicts the best beam to use for each speaker. Assuming the stationarity of the speaker locations, i.e., each speaker is supposed to stay in one of the 12 "beam areas", we propose to use a long short-term memory (LSTM) network for beam prediction. Input to the beam prediction network consists of the magnitude spectra of all the 12 beamformed signals, concatenated with inter-microphone phase difference (IPD) features. The IPD features were calculated by using the first microphone signal as reference.

For each speaker, the best beam is defined as the one that maximizes the SDR for the speaker. In a single-speaker scenario, this usually corresponds to the beam that points to the speaker's direction. But, this is not the case in the multi-speaker scenario because it is possible that the beam with a look direction different from the target speaker direction may better enhance the target speech signal when the null of the beam is pointing at an interfering speaker. Therefore, when training the beam prediction network, we calculated the SDR of each speaker for each of the 12 beams and picked up the best one as the prediction target instead of relying on direction information. For each time frame, a N-hot vector is formed as reference, where N refers the number of active speaker for that frame. And the objective is the binary cross entropy between the network output and the reference.

2.3. Speech separation network

After the beam selection step, the output signals of the selected beams are passed to a speech separation network. In this work, PIT is used to train the speech separation network. PIT uses a neural network that generates spectral mask for each participating speaker. The fundamental difficulty in training such networks is that we do not know which output mask should be associated with which speaker during training. To take account of this ambiguity, PIT examines all possible speaker permutations. The permutation that yields the minimum loss is selected to invoke backpropagation learning. Thus, the PIT loss function can be written as equation (1), where X,m and Y refers the clean speech, estimated mask and mixed speech, respectively:

$$\mathcal{L} = \min_{J \in perm(I)} \sum_{i=1}^{I} \sum_{t \in T} \|m_{j_i,t} * Y_t - X_{i,t}\|^2, \;\; J = (j_1, \cdots, j_I). \tag{1}$$

I and T represent the number of speakers and that of time frames.

Unlike in the conventional single-channel PIT scenario, we may use multiple beamformed signals as input to the PIT network, where each beamformed signal may correspond to different speakers. Thus, in this paper, we investigate two different speech separation networks, which we call speech enhancement PIT (SE-PIT) and multi-view PIT (MV-PIT), as shown in Fig. 2.

2.3.1. Speech enhancement PIT

In each of the beams selected by the beam prediction network, it is expected that one speaker (which we call a target speaker here) is

more dominant than others. Therefore, it may be sufficient for the speech separation network to enhance only the target speaker's signal instead of trying to restore all speakers' signals in all the selected beams. Thus, with the speech enhancement PIT scheme proposed here, we use only the reconstruction errors for the target speaker as shown in equation (2), where source index i is removed since the speaker is uniquely determined for the selected beam:

$$\mathcal{L} = \min_{J \in perm(I)} \sum_{t \in T} \left\| m_{j,t} * Y_t - \tilde{X}_t \right\|^2 . \quad (2)$$

Note that $\tilde{X}$ denotes a beamformed signal. While this appears to be similar to the conventional single-speaker mask-based speech enhancement methods, the PIT approach, which searches for the best permutation, is still used here because the separation performance of the fixed beamformers is limited and thus the permutation ambiguity may still exist.

2.3.2. Multi-view PIT

With multi-view PIT, we propose to use all the selected beams as input to the PIT network. In this scheme, the PIT network needs to produce the spectral masks for all participating speakers. To be more precise, the input feature vector is the concatenation of the magnitude spectrogram of the selected beamformed signals. The network is trained to estimate the masks for each speaker, where the masks for each speaker are applied to the corresponding beamformed signal. In this paper, we assume there is at most 2 speakers in the mixture, however the proposed model can easily generalized to more speaker mixtures. Since the target speaker information in one beam can be exploited to cancel that speaker's signal when enhancing other speakers, the multiple selected beams can provide complementary information. Equation (3) shows the multi-view PIT loss function.

$$\mathcal{L} = \min_{J \in perm(I)} \sum_{t \in T} \left\| m_{j,t} * Y_{j,t} - \tilde{X}_{i,t} \right\|^2 , \quad (3)$$

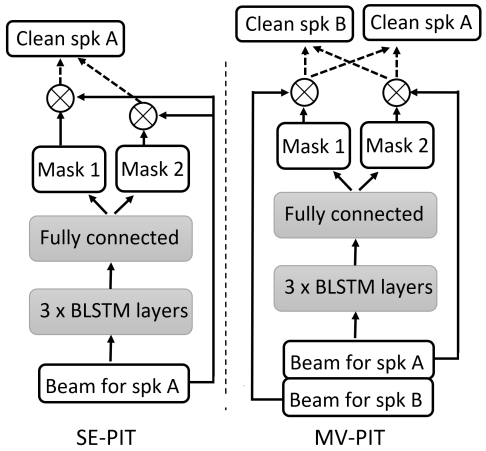

Fig. 2. Left: Speech enhancement PIT. Right: Multi-view PIT.

2.4. Comparison with other approaches

The proposed model leverages both spectral and spatial information by combining beamforming and PIT-based speech separation, which leads to a conceptually simple yet effective solution. Our approach has several advantages compared with other solutions that were proposed to address the neural network-based multi-channel speech separation problem.

Firstly, compared with our prior work [10], which is based on a late beam-selection strategy (see the discussion in Section 1), the proposed early beam-selection approach enables reduction in computational cost. With late beam-selection, single-channel speech separation is performed for each of the 12 beamformed signals, resulting in 24 output signals in the two-speaker mixture case. Then, the best signal is selected for each speaker. While late selection seems easier than selecting the best beams to use before performing neural network-based separation, this architecture requires a lot of computational cost and thus inappropriate for many usage scenarios. For example, the transmission of 12 beamformed waves to cloud is usually not feasible in real world applications. By contrast, with the proposed early beam-selection strategy, neural network-based speech separation needs to be run only once for each mixture. It is also straightforward to perform joint training between the two network, enabling a more end-to-end optimization.

Secondly, the proposed approach involves neither clustering nor adaptive beamforming, which significantly reduces processing latency. To perform clustering and adaptive beamforming, such as generalized eigenvalue filtering, as in [13, 14, 15, 16, 17], sufficient statistics need to be computed at test time. This means that the mixed speech must be observed for a certain duration before separation takes place, and the movement of the speaker usually cause significant turbulence in the accumulated statistics. The use of predefined beamformers and a direct separation approach like PIT allows to avoid such difficulty. Though in this paper, we applied Bi directional LSTM in the experiment, it is straightforward to replace it with LSTM network when the latency is the main concern.

Lastly, using fixed beamformers is much simpler and more efficient by exploiting prior knowledge of the array geometry.

3. EXPERIMENTAL SETUP

3.1. Data

Three different test sets were generated to evaluate the proposed systems under different mixing conditions. We targeted at two speaker mixing senarios in experiment session. Figure 3 illustrates the three mixing configurations considered in our experiments, which are referred to as full overlap (FO), partial overlap (PO) and single dominant (SD). For each mixing configuration, a one-hour test set was created by artificially reverberating anechoic speech signals and mixing them. The clean speech was sampled from our internal collection of utterances spoken by 44 speakers. The image method was used to create the room impulse response for the simulation, where the room dimensions(2 20m), the room T60 time(0.1s 0.9s), and the microphone and speaker locations were randomly determined for each speech mixture. The mixing signal levels and the amount of overlap (for PO and SD) were also randomly chosen. A 40-hour training set was also generated in the same way, where the clean speech was sampled from the Wall Street Journal (WSJ) SI-284 set.The microphone array consists of 7 sensors, i.e. six microphone equally distributed in a circle with radius of 4.25cm, and a center microphone. The audio signals were sampled at 16,000 Hz.

3.2. Proposed and reference systems

Bi-directional LSTM (BLSTM) networks [18] were used for both beam prediction and PIT-based speech separation. Our beam pre-

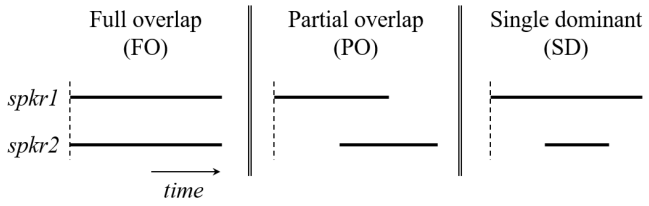

Fig. 3. Three mixing configurations considered in our experiments.

diction network consisted of three BLSTM layers, each with 300 forward cells and 300 backward cells. On top of the BLSTM layers, there was a fully connected layer of size 12, i.e., the number of predefined beamformers. In addition, we also used the segment-based batch normalization layer proposed in [13], which we found to be crucial under reverberant conditions. The beam prediction network was trained to minimize the binary cross entropy between the network outputs and the "ground-truth" beams that maximize the SDRs for the training speakers.

Our speech separation network had three BLSTM layers, each with 1024 forward cells and 1024 backward cells. As with the beam prediction network, we used the segment-based batch normalization layer before the BLSTM layers. The top layer consisted of two parallel fully connected layers with sigmoid activation for generating spectral masks. Magnitude spectra of the beamformed signals were used as input to the separation network, which were obtained by using 32-ms window with a shift of 16 ms. The separation network was trained with PIT.

Four reference systems were built for benchmarking purposes, three of which used oracle information. One reference system performed speech separation by using ideal ratio masks (IRM), which were calculated based on the reverberated source signals. The IRM system reveals the performance upper bound of any single channel based separation systems. Another system (IRM-OB) performed IRM-based speech separation for the beamformed signals, where the beams were chosen to maximize the SDRs, i.e., oracle beam selection. The IRM-OB performance shows the upper bound of the proposed approach combining multiple beamformers and single-channel speech separation. A minimum variance distortionless response beamformer using oracle spatial covariances was also included (OMVDR) for comparison, which shows the performance limit for linear beamforming algorithms. To calculate the oracle beamformer, we firstly applied the ideal ratio masks, calculated from reverberant source signals, to the multi-channel mixture signal. Then an MVDR beamformer was estimated based on the spatial covariance matrix from the masked speech signals. Lastly, we included the results of the conventional single channel PIT by using the same network architecture. We also report the SDR results for the unprocessed speech (ORI), oracle beams (OB) and predicted beams (PB).

To measure the impact that each component of the proposed system has on the separation performance, we evaluated the proposed system in two different ways. Firstly, we evaluated the performance of the system that applied single-channel separation to the oracle beams (SE-PIT-OB, MV-PIT-OB). Secondly, we evaluated the system that performed speech separation on the beams selected by our beam prediction network (SE-PIT-PB, MV-PIT-PB), which made use of no oracle information.

All systems were evaluated in terms of SDR, estimated with the bss_eval toolbox [19], where a high SDR indicates better separation performance. For each test mixture, the SDRs were calculated for all possible speaker permutations and then the permutation that produced the highest SDRs was selected. The SDRs reported below were obtained by averaging the results over all the test samples.

Table 1. SDRs of different separation systems for different mixing conditions.

	FO	PO	SD
PIT	3.82	2.83	2.34
IRM	6.66	6.97	6.79
IRM-OB	10.53	10.87	10.78
OMVDR	9.86	9.66	9.46
MV-PIT-OB	8.00	9.27	8.5
SE-PIT-OB	6.78	7.96	7.31
MV-PIT-PB	6.19	7.33	6.5
SE-PIT-PB	5.99	7.28	6.38
OB	4.2	4.01	4.00
PB	2.72	2.57	2.39
ORI	-1.56	-1.45	0.04

4. RESULTS AND DISCUSSION

Table 1 lists the results of all the speech separation systems considered. We can seen that both SE-PIT and MV-PIT methods significantly improved the SDRs compared with the unprocessed speech (ORI) in all mixing conditions, which confirms the efficacy of the proposed architecture. While the use of the automatically predicted beams degraded the SDR by $\sim$2 dB compared with using the oracle beams (MV-PIT-OB vs. MV-PIT-PB or SE-PIT-OB vs. SE-PIT-PB), our proposed neural network-based beam selection worked well. This can be seen from the fact that MV-PIT-PB and SE-PIT-PB significantly outperformed the single-channel PIT.

As regards the comparison of the two speech separation schemes, speech-enhancement PIT and multi-view PIT, the latter consistently showed superior performance. This confirms our speculation that all the selected beams provide complementary information.

It is also noteworthy that the proposed system, MV-PIT-PB, was comparable even with the single-channel IRM system. This demonstrates the usefulness of the spatial information obtained from multiple microphones and confirms our belief that it is important to further investigate multi-channel speech separation to realize speech separation systems usable in practical application scenarios.

Finally, the IRM-OB and OMVDR showed better performance than the proposed system, thanks to the access to the oracle data. The IRM-OB leads to around 3dB improvement to the MV-PIT-OB, indicating that there is still room for improvement.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed an efficient framework for integrating a set of fixed beamformers and neural network-based speech separation. A beam prediction network was proposed to estimate the best acoustic beam for each speaker participating in the input speech mixture. The use of fixed beamformers and PIT separation allows for efficient implementation and potentially low latency processing compared with previously proposed clustering-based adaptive algorithms. The proposed system was evaluated in a far-field speech separation task and shown to be able to separate reverberant speech signals under different mixing conditions.

In the future, we will extend this work to multi-talker speech recognition, and jointly train all the components in a progressive way which has been shown more effective than the simple joint training [20].

6. REFERENCES

[1] E Colin Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

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

[3] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, 2016.

[4] Zhuo Chen, Yi Luo, and Nima Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 246–250.

[5] Yuxuan Wang, Narayanan Arun, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[6] Zhuo Chen, Yan Huang, Jinyu Li, and Yifan Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," 2017.

[7] Yuxuan Wang and DeLiang Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[8] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 241–245.

[9] 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*, vol. 25, no. 10, pp. 1901–1913, 2017.

[10] Zhuo Chen, Jinyu Li, Takuya Yohioka, Jesper Jensen, Xiong Xiao, Huaming Wang, Zhenghao Wang, and Yifan Gong, "Cracking the cocktail party problem by multi-beam deep attractor network," in *Automatic Speech Recognition and Understanding Workshop, 2017*. IEEE, 2017.

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

[12] Dipl-Ing Hannes Pessentheiner, "Differential microphone arrays," 2013.

[13] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 196–200.

[14] Emanuël Anco Peter Habets, Jacob Benesty, Israel Cohen, Sharon Gannot, and Jacek Dmochowski, "New insights into the mvdr beamformer in room acoustics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 158–170, 2010.

[15] Mehrez Souden, Jacob Benesty, and Sofiène Affes, "A study of the lcmv and mvdr noise reduction filters," *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4925–4935, 2010.

[16] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5210–5214.

[17] Takuya Higuchi, Nobutaka Ito, Shoko Araki, Takuya Yoshioka, Marc Delcroix, Tomohiro Nakatani, Takuya Higuchi, Nobutaka Ito, Shoko Araki, Takuya Yoshioka, et al., "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 780–793, 2017.

[18] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[19] Cédric Févotte, Rémi Gribonval, and Emmanuel Vincent, "Bss_eval toolbox user guide–revision 2.0," 2005.

[20] Zhehuai Chen, Jasha Droppo, Jinyu Li, and Wayne Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 26, no. 1, pp. 184–196, 2018.