INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Multi-input Multi-output Complex Spectral Mapping for Speaker Separation

Hassan Taherian[1], *Ashutosh Pandey*[2], *Daniel Wong*[2], *Buye Xu*[2], and *DeLiang Wang*[1]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Meta Reality Labs Research, USA

taherian.1@osu.edu, {apandey620, ddewong, xub}@meta.com, dwang@cse.ohio-state.edu

Abstract

Current deep learning based multi-channel speaker separation methods produce a monaural estimate of speaker signals captured by a reference microphone. This work presents a new multi-channel complex spectral mapping approach that simultaneously estimates the real and imaginary spectrograms of all speakers at all microphones. The proposed multi-input multi-output (MIMO) separation model uses a location-based training (LBT) criterion to resolve the permutation ambiguity in talker-independent speaker separation across microphones. Experimental results show that the proposed MIMO separation model outperforms a multi-input single-output (MISO) speaker separation model with monaural estimates. We also combine the MIMO separation model with a beamformer and a MISO speech enhancement model to further improve separation performance. The proposed approach achieves the state-of-the-art speaker separation on the open LibriCSS dataset.

Index Terms: MIMO speaker separation, multi-channel complex spectral mapping, location-based training.

1. Introduction

The field of multi-channel speech separation has witnessed substantial progress in recent years thanks to the employment of deep neural networks (DNNs). Early studies on DNN-based multi-channel speaker separation rely on combining monaural speaker separation and conventional beamforming techniques [1, 2, 3, 4]. This approach typically utilizes a DNN to estimate a monaural time-frequency (T-F) mask at each microphone, and the masks are then combined to weigh spatial covariance matrices at the corresponding T-F units or segments. The weighted covariance matrices are used to compute a steering vector for beamforming. Other studies make use of a neural beamformer where the beamforming filters are directly learned through a DNN in either the time domain or frequency domain [5].

Most studies on mask-based beamforming use real-valued masks, which only enhance the magnitude spectrogram of a noisy mixture and leave its phase unchanged. However, for more accurate covariance matrix estimation, Wang et al. employed single-input single-output (SISO) complex spectral mapping to jointly estimate the magnitude and phase of the target speech signal at each microphone independently [6]. The estimated complex spectrograms are then used to compute the spatial covariance matrices directly for beamforming.

This research was supported in part by an National Science Foundation grant (ECCS-2125074), a research contract from Meta Reality Labs, the Ohio Supercomputer Center, and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

Recently, multi-input single-output (MISO) complex spectral mapping has been proposed, and it achieves comparable or better separation performance compared to masking-based beamforming [7, 8, 9]. With MISO complex spectral mapping, a DNN is trained to directly estimate the real and imaginary spectrograms of the target speaker at a reference microphone from those of a multi-channel mixture. A MISO separation model can implicitly learn the spectral and spatial information for a fixed array geometry [7]. In [8] and [10], a MISO model is proposed for binaural speaker separation where the target speech signals at the left and right ears are estimated individually. In another study, MISO complex spectral mapping is integrated with minimum variance distortionless response (MVDR) beamforming and post-filtering to further improve separation [9]. Although MISO complex spectral mapping achieves strong speaker separation performance, it is computationally expensive as the model needs to be applied as many times as the number of microphones for spatial covariance computation. Additionally, this approach requires speaker alignment across microphones, as the outputs at different microphones may have different speaker permutations.

To reduce the computational cost of MISO-based beamforming, one straightforward way is to perform multi-input multi-output (MIMO) speaker separation to estimate the target signal at all microphones simultaneously. Several studies have developed MIMO models for speech dereverberation and speaker separation [11, 12, 13, 14, 15]. Wang et al. integrated a MIMO enhancement model with beamforming and post-filtering for speech dereverberation [11]. The Beam-TasNet, a time-domain separation network, is extended to produce spectro-temporal masks at all microphones and for all speakers in [12]. The separated signals are then iteratively refined by combining the Beam-TasNet and MVDR beamforming. Fu et al. proposed a MIMO separation model to estimate the direction of arrival and beamforming weights for each speaker [13]. Other works have investigated the joint optimization of speaker separation and speech dereverberation for MIMO convolutional beamformers [14, 15].

Previous MIMO separation models typically use permutation-invariant training (PIT) [16] to address the permutation ambiguity problem in talker-independent speaker separation. However, as we demonstrate in this paper, a PIT-based MIMO separation model underperforms the corresponding MISO model. Furthermore, these models have mostly been evaluated in simulated environments, leaving their generalization to realistic recordings uncertain. In this study, we propose a MIMO complex spectral mapping approach for both speaker separation and speech dereverberation. Our approach predicts the direct-path complex spectrograms of all speakers at all microphones simultaneously. We train the MIMO

10.21437/Interspeech.2023-318

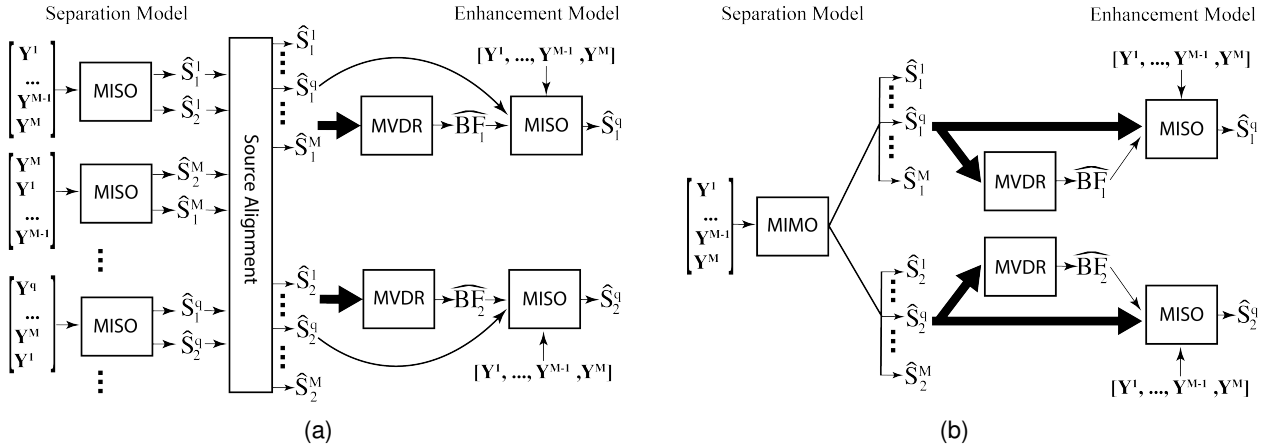

Figure 1: *Schematic diagram of (a) MISO-BF-MISO and (b) MIMO-BF-MISO systems for two-speaker separation and dereverberation.*

separation model using location-based training (LBT) [17], which significantly improves separation performance over the widely-used PIT criterion for MIMO complex spectral mapping. Our proposed MIMO separation model outperforms the MISO model and achieves the state-of-the-art results on the recorded LibriCSS dataset [18]. Compared to MISO, MIMO complex spectral mapping is conceptually and computationally simpler. It also eliminates the need for source alignment across different microphones and is expected to preserve inter-channel cues better, hence facilitating downstream multi-channel speech processing tasks such as localization.

2. System Description

In this section, we first describe the MISO-BF-MISO system [9], which previously achieved the best results on the LibriCSS dataset. Afterward, we will introduce our proposed approach, MIMO complex spectral mapping, for joint speech separation and dereverberation.

2.1. MISO-BF-MISO

Figure 1a depicts the MISO-BF-MISO system, which comprises a MISO separation model, a MVDR beamformer, and a MISO enhancement model for post-filtering. Given a M-channel mixture signal $\boldsymbol{Y} = [Y^1, \ldots, Y^M]$ in short-time Fourier transform (STFT) domain, the MISO separation model estimates the direct-path complex spectrograms of N speakers $\hat{S}_n^q, n \in [1, \ldots, N]$ at the reference microphone q. The MISO separation model is trained with the utterance-level PIT criterion [16] to resolve the permutation ambiguity.

Next, the estimated complex spectrograms are used to compute target $\hat{\boldsymbol{\Phi}}_{S_n}$ and non-target $\hat{\boldsymbol{\Phi}}_{V_n}$ covariance matrices for MVDR beamforming [9]:

$$\hat{\boldsymbol{\Phi}}_{S_n}(f) = \frac{1}{T}\sum_t \hat{\boldsymbol{S}}_n(t,f)\hat{\boldsymbol{S}}_n(t,f)^H$$
$$\hat{\boldsymbol{\Phi}}_{V_n}(f) = \frac{1}{T}\sum_t \hat{\boldsymbol{V}}_n(t,f)\hat{\boldsymbol{V}}_n(t,f)^H, \quad (1)$$

where $\hat{\boldsymbol{S}}_n = [\hat{S}_n^1, \ldots, \hat{S}_n^M]$ and $\hat{\boldsymbol{V}}_n(t,f) = \boldsymbol{Y}(t,f) - \hat{\boldsymbol{S}}_n(t,f)$ are the complex STFT vectors of the estimated target and interference signals for speaker n at time t and frequency f, respectively. Symbol H denotes the Hermitian operator and T is the total number of frames. To derive covariance matrices, complex spectrograms should be estimated at all microphones. For this purpose, separation is performed M times by circularly shifting the microphone order to predict the direct-path signal for all speakers at each microphone. Note that the microphone rotation method only works for uniform circular arrays. For non-circular arrays, a dedicated MISO model needs to be trained at each microphone. Furthermore, the outputs need to be aligned across all microphones. Source alignment is done by aligning the outputs at each non-reference microphone to the outputs at the reference microphone based on their magnitude distance [9].

In the last stage, the beamforming results $\widehat{BF}_n$ are stacked with the multi-channel mixture signal and the estimated target signal for each speaker at the reference microphone, i.e. $[\widehat{BF}_n, \boldsymbol{Y}, \hat{S}_n^q]$, for further enhancement. The second MISO model only performs enhancement and does not need to resolve the permutation ambiguity problem, as the problem has already been resolved in the separation stage.

2.2. DNN Architecture

We employ the Dense-UNet architecture [19] for both MISO separation and enhancement models. The real and imaginary components of the input signals are stacked and fed to the Dense-UNet. The architecture comprises four downsampling layers and four upsampling layers interleaved with nine densely-connected convolutional neural network blocks. Each dense block contains five convolutional layers with $C = 76$ channels, a kernel size of 3×3 and a stride of 1×1. After the last dense block, we use a 1×1 convolutional layer with $O = 2 \times N$ and $O = 2$ channels to produce complex spectrogram estimates for the MISO separation and enhancement models, respectively. For more information about the Dense-UNet architecture, please refer to [19]. The MISO models are trained with ℓ_1 norm loss of real and imaginary spectrograms of estimated and target speech with an additional magnitude loss term [20]:

$$\mathcal{L}(\hat{S}, S) = \left\| \Re(\hat{S}) - \Re(S) \right\|_1 + \left\| \Im(\hat{S}) - \Im(S) \right\|_1 + \left\| |\hat{S}| - |S| \right\|_1, \quad (2)$$

where $\Re(.)$ and $\Im(.)$ extract real and imaginary parts, $|\ .\ |$ computes magnitude and $\|\ .\ \|_1$ computes ℓ_1 norm.

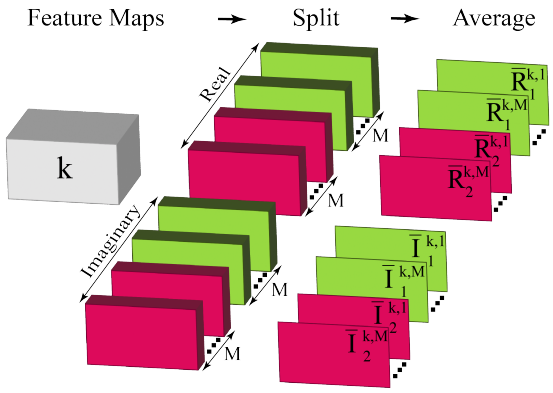

Figure 2: *Generating low-resolution complex spectrograms based on decoder layer k feature maps for $N = 2$ speakers.*

2.3. MIMO Complex Spectral Mapping

The MIMO complex spectral mapping method allows a DNN to estimate the complex spectrogram of target speech from all microphones using a multi-channel noisy mixture. The MIMO architecture remains the same, except for an increase in the number of channels in the output layer to $O = 2 \times M \times N$. This extension adds only a negligible increase in the number of parameters of the MIMO model by $2 \times N \times C \times (M-1)$ compared to a MISO model. To train the MIMO separation model, we generalize LBT [17] to incorporate the complex spectrogram estimates from all microphones. The generalized LBT loss function is defined as follows:

$$\mathcal{L}_{\text{LBT}}(\hat{\boldsymbol{S}}, \boldsymbol{S}) = \frac{1}{NM} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathcal{L}(\hat{S}_n^m, S_{\lambda_n}^m), \tag{3}$$

where $\lambda_1, \lambda_2, \ldots, \lambda_N \in [1, \ldots, N]$ are speaker indices sorted in ascending order based on speaker azimuths or distances relative to the microphone array [17]. In this study, only the azimuth criterion is considered for LBT.

Additionally, we propose an extension to the multi-resolution LBT (MR-LBT) loss function [21] for MIMO separation models. With MR-LBT, we estimate the complex spectrograms from low to high time and frequency resolution in decoder layers. Specifically, the channel dimension of the output feature maps for every decoder layer is divided into $N \times M$ real groups and $N \times M$ imaginary groups, each representing a real or imaginary component of a microphone for a speaker. The feature maps within each group are then averaged across the channel dimension. At decoder layer k, a low-resolution estimate of target speech in the STFT domain is created at all microphones:

$$\hat{\boldsymbol{S}}_n^k = [\bar{R}_n^{k,1} + i\bar{I}_n^{k,1}, \ldots, \bar{R}_n^{k,M} + i\bar{I}_n^{k,M}] \tag{4}$$

where $\bar{R}_n^{k,m}$ and $\bar{I}_n^{k,m}$ are the averaged feature maps from the real and imaginary groups of microphone m and speaker n. Symbol i denotes the imaginary unit. Figure 2 illustrates the low-resolution estimation of complex spectrograms for $N = 2$ speakers. Complementary LBT losses between lower-resolution estimates and clean signals are calculated for every decoder layer and added to the Eq. (3) loss:

$$\mathcal{L}_{\text{LBT}}(\hat{\boldsymbol{S}}, \boldsymbol{S}) + \sum_{k=1}^{K_d - 1} \mathcal{L}_{\text{LBT}}(\hat{\boldsymbol{S}}^k, g^{K_d - k}(\boldsymbol{S})) \tag{5}$$

where K_d is the total number of decoder layers, and $g^x(.)$ is a 2D average pooling function with a kernel size of 2×2 and a stride of 2×2, applied recursively for x iterations.

We also combine the MIMO separation model with a MVDR beamformer and a MISO enhancement model to create a MIMO-BF-MISO system. The proposed system is illustrated in Figure 1b. The MISO enhancement model uses $[\widehat{BF_n}, \boldsymbol{Y}, \hat{\boldsymbol{S}}_n]$ as inputs to predict $\hat{S}_n^q$. All MISO enhancement models are trained using the loss function in Eq. (2).

3. Experimental Setup

We validate the proposed separation models for conversational speech recognition task in unmatched reverberant conditions. The evaluation is performed using the LibriCSS corpus [18], which contains 10 hours of partially overlapped utterances. The utterances are taken from the Librispeech development set and retransmitted with loudspeakers to capture real room reverberation. The recording device is a circular array with $M = 7$ microphones and $4.25\,\text{cm}$ radius. The LibriCSS corpus is divided into 6 sessions with different overlap ratios: 0S (no overlap with a 0.1-0.5 s pause between utterances), 0L (no overlap with a 2.9-3.0 s pause between utterances), 10%, 20%, 30% and 40% overlaps.

We processed the LibriCSS recordings using the continuous speech separation (CSS) framework [18]. Each recording was segmented using a sliding window of $2.4\,\text{s}$ with a segment shift of $1.2\,\text{s}$. The CSS framework can handle any number of speakers, with the assumption that there are at most $N = 2$ speakers within a segment. For segments without overlapped speech, the model only performs dereverberation. In this case, the input is mapped to the first output, and a zero signal is assigned to the second output. Finally, the processed segments are concatenated using the stitching algorithm proposed by [18].

We used the default automatic speech recognition (ASR) backend provided with the LibriCSS corpus [18]. The LibriCSS corpus contains two ASR evaluation scenarios: utterance-wise and continuous evaluations. In the utterance-wise scenario, the ground-truth utterance boundaries are provided. The ASR backend scores each separated signal independently, and the one with the lower word error rate (WER) is considered. In the continuous evaluation, the utterance boundaries are unknown, with 8-10 utterances in each recording. The decoding results from both separated signals are combined to compute the final WER.

To generate the training and validation data, we followed the setup described in [9] and used simulated room impulse responses (RIRs) [23, 24]. We created 192K two-speaker mixtures with different overlap ratios from the LibriSpeech dataset. Each mixture was convolved with 7-channel microphone array RIRs with the same array geometry as the LibriCSS recording device. To generate RIRs, we positioned the sources in rectangular rooms with random length, width, and height dimensions ranging from 5×5×3 to 10×10×4 meters. The microphone array was placed in the center of the room, and the source positions were uniformly sampled from 360 candidate azimuth angles in the range of -180°to 180° with a 1° resolution. The reverberation time (T60) was randomly sampled between 0.2 and 0.6 s.

We trained the separation and enhancement networks sequentially, starting with a learning rate of 0.001, which gradually decayed using a cosine annealing learning rate scheduler. For MISO models, we designated the first microphone as the reference microphone ($q = 1$). For MIMO models, we used the estimated complex spectrograms of the first microphone for

Table 1: *WER results (in %) of comparison systems for utterance-wise and continuous evaluation with 7-channel array on LibriCSS. 'SC' refers to a frame-wise speaker counter which corrects separation errors for non-overlapped utterances. MIMO models introduce a slight computation overhead. Our MISO and MIMO models require 195.21G and 195.32G multiply-accumulate (MAC) operations, respectively, to process a 2.4-second segment.*

			Utterance-wise						**Continuous**					
	#Parameters	Criterion	0S	0L	10%	20%	30%	40%	0S	0L	10%	20%	30%	40%
Unprocessed	–	–	11.8	11.7	18.8	27.2	35.6	43.3	15.4	11.5	21.7	27.0	34.3	40.5
BLSTM [18]	21.8M	PIT	8.3	8.4	11.6	16.0	18.4	21.6	11.9	9.7	13.4	15.1	19.7	22.0
Conformer [22]	58.7M	PIT	7.2	7.5	9.6	11.3	13.7	15.1	11.0	8.7	12.6	13.5	17.6	19.6
MISO [9]	6.9M	PIT	7.7	7.5	7.9	9.6	11.3	13.0	10.7	10.5	10.9	11.5	13.8	15.3
MISO-BF-MISO+SC [9]	13.8M	PIT	5.8	5.8	5.9	6.5	7.7	8.3	7.7	7.5	7.4	8.4	9.7	11.3
MISO	6.9M	MR-LBT	7.3	7.9	7.6	9.2	10.8	11.4	9.2	9.8	9.1	10.5	12.2	12.8
+SISO	13.7M		6.9	7.1	7.1	8.5	9.5	10.6	10.7	10.3	9.4	9.8	11.7	12.5
+MISO	13.8M		6.5	7.0	6.5	7.7	8.4	9.0	9.2	9.1	8.2	9.2	9.8	9.9
+BF-MISO	13.8M		6.5	6.8	6.5	7.4	8.4	8.8	9.6	10.0	8.8	8.6	10.4	10.2
MISO Large	14.6M	MR-LBT	7.5	7.2	7.4	8.6	9.8	11.5	9.8	9.1	9.2	10.3	11.8	13.8
MIMO	6.9M	PIT	10.5	11.8	10.9	12.8	13.7	16.1	16.5	17.7	15.5	16.3	17.8	20.2
MIMO+SC	6.9M	PIT	8.1	9.4	8.6	10.9	12.5	15.4	8.5	8.3	8.9	10.3	12.7	14.7
MIMO	6.9M	LBT	8.3	9.5	8.1	9.4	10.7	11.8	10.5	10.6	9.5	10.6	12.4	12.6
MIMO Large	14.6M	PIT	6.4	7.0	7.8	9.6	11.7	13.5	9.5	8.8	10.5	11.4	13.6	16.0
MIMO Large+SC	14.6M	PIT	6.4	6.8	7.5	9.6	11.7	14.0	7.9	7.3	8.7	10.0	12.7	14.6
MIMO Large	14.6M	LBT	7.1	7.7	7.7	9.1	10.4	11.6	10.3	10.5	10.6	11.4	13.0	13.4
MIMO Large	14.6M	MR-LBT	7.6	7.1	7.3	8.6	10.1	11.6	8.6	7.9	8.8	9.7	11.9	12.9
+MISO	21.5M		6.5	6.3	6.0	7.1	8.0	9.1	7.8	7.8	7.6	7.7	9.4	10.2
+BF-MISO	21.5M		6.3	6.1	6.0	6.8	7.4	8.5	7.4	7.5	7.2	7.4	8.8	9.6

ASR evaluation. Following [9], we included the spectral magnitude of the mixture signal $\boldsymbol{Y}$ from the first microphone for all MISO and MIMO models. As a comparison baseline, we also report the results for a MIMO separation model trained with the PIT criterion:

$$\mathcal{L}_{\text{PIT}}(\hat{\boldsymbol{S}}, \boldsymbol{S}) = \frac{1}{NM} \min_{\psi \in \Psi} \sum_{n=1}^{N} \sum_{m=1}^{M} \mathcal{L}(\hat{S}_n^m, S_{\psi(n)}^m), \quad (6)$$

where symbol Ψ is the set of all permutations of N speakers with ψ referring to one permutation.

4. Evaluation Results

Table 1 compares our MISO and MIMO separation models with other competitive methods for utterance-wise and continuous evaluations on LibriCSS. The proposed systems in [18] and [22] use BLSTM and conformer architectures, respectively, to estimate real-valued masks. The systems in [9] employ MISO complex spectral mapping using a UNet model with a temporal convolutional network. To correct separation errors for segments without speaker overlaps, a dedicated speaker counting (SC) network is used in [9]. It is observed that when a model is trained on two-speaker mixtures, it sometimes emits an intelligible residual signal in the second output for single speaker utterances, leading to more decoding errors. The SC network counts the number of speakers at each frame and merges the separated outputs for non-overlapped frames.

We observe that the MISO separation model trained with MR-LBT obtains substantially better results than BLSTM, conformer and PIT-based MISO. To further improve the WER scores, we combine the MISO outputs with the reference microphone mixture signal, i.e. $[Y^1, \hat{S}_n]$ to train a SISO enhancement model. As shown in the table, the MISO-SISO model achieves only marginal improvements over MISO. On the other hand, including the multi-channel mixture and beamformer signals for post-filtering in the MISO-BF-MISO system substantially improves its performance. The MISO-MISO system yields comparable results to MISO-BF-MISO, indicating that the MISO enhancement model is essential for further WER reduction.

Our MIMO separation model trained with the PIT criterion performs significantly worse than MISO separation models. To investigate further, we incorporated a SC network to correct the separation errors in the PIT-based MIMO model. The results indicated a significant improvement in WER. However, even with the SC network, the PIT-based MIMO model still produced worse estimates of target speech than MISO models, especially in higher overlap ratios. We also experimented with a larger number of convolutional kernels ($C = 112$) in the PIT-based MIMO model, and the results suggest that increasing the number of parameters can improve its performance. In contrast, increasing the number of parameters for LBT-based MISO and MIMO models resulted in smaller improvements in performance.

Our best results are achieved with a large MIMO model trained using the MR-LBT criterion, which exhibits excellent speech recognition performance and outperforms even MISO models. Despite being trained only on simulated RIRs, the MIMO model trained with the MR-LBT criterion generalizes well to real microphone array recordings. By combining the MIMO separation model with the MVDR beamformer and the MISO enhancement model, we achieve state-of-the-art results on the LibriCSS dataset without the need for a SC network. For instance, our MIMO-BF-MISO system achieves a WER of 9.6% in continuous evaluation on a 40% overlap condition, which is better than the MISO-BF-MISO system and does not require circular shifting or source alignment.

5. Conclusions

In this study, we have proposed a MIMO complex spectral mapping approach for joint speech dereverberation and speaker separation. We have also extended multi-resolution LBT to MIMO separation. Our experimental results demonstrate the importance of using the LBT criterion for training MIMO speaker separation models. With lower computational complexity, our proposed MIMO-BF-MISO system achieves the state-of-the-art results on the LibriCSS dataset.

6. References

[1] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 196–200.

[2] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. ICASSP*, 2016, pp. 5210–5214.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. L. Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[4] R. Gu, S.-X. Zhang, Y. Zou, and D. Yu, "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," *IEEE Signal Processing Letters*, vol. 28, pp. 1370–1374, 2021.

[5] B. Li, T. N. Sainath, R. J. Weiss, K. W. Wilson, and M. Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition," in *Proc. Interspeech*, 2016, pp. 1976–1980.

[6] Z.-Q. Wang, P. Wang, and D. L. Wang, "Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR," *IEEE/ACM Trans. Audio Speech Lang. Process.*, pp. 1778–1787, 2020.

[7] K. Tan, Z.-Q. Wang, and D. L. Wang, "Neural spectrospatial filtering," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 605–621, 2022.

[8] K. Tan, B. Xu, A. Kumar, E. Nachmani, and Y. Adi, "SAGRNN: Self-attentive gated RNN for binaural speaker separation with interaural cue preservation," *IEEE Signal Processing Letters*, vol. 28, pp. 26–30, 2021.

[9] Z.-Q. Wang, P. Wang, and D. L. Wang, "Multi-microphone complex spectral mapping for utterance-wise and continuous speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 2001–2014, 2021.

[10] C. Han, Y. Luo, and N. Mesgarani, "Real-time binaural speech separation with preserved spatial cues," in *Proc. ICASSP*, 2020, pp. 6404–6408.

[11] Z.-Q. Wang and D. L. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *Proc. ICASSP*, 2020, pp. 486–490.

[12] H. Chen, Y. Yang, F. Dang, and P. Zhang, "Beam-Guided TasNet: An iterative speech separation framework with multi-channel output," in *Proc. Interspeech*, 2022, pp. 866–870.

[13] Y. Fu, H. Yin, M. Ge, L. Wang, G. Zhang, J. Dang, C. Deng, and F. Wang, "MIMO-DBnet: Multi-channel input and multiple outputs DOA-aware beamforming network for speech separation," *arXiv:2212.03401*, 2022.

[14] M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Multichannel speech dereverberation and separation with optimized combination of linear and non-linear filtering," in *Proc. ICASSP*, 2012, pp. 4057–4060.

[15] T. Nakatani, R. Ikeshita, K. Kinoshita, H. Sawada, and S. Araki, "Computationally efficient and versatile framework for joint optimization of blind speech separation and dereverberation." in *Proc. Interspeech*, 2020, pp. 91–95.

[16] 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 Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[17] H. Taherian, K. Tan, and D. L. Wang, "Multi-channel talker-independent speaker separation through location-based training," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2791–2800, 2022.

[18] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo, J. Wu, X. Xiao, and J. Li, "Continuous speech separation: Dataset and analysis," in *Proc. ICASSP*, 2020, pp. 7284–7288.

[19] Y. Liu and D. L. Wang, "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[20] Z.-Q. Wang and D. L. Wang, "Deep learning based target cancellation for speech dereverberation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 941–950, 2020.

[21] H. Taherian and D. L. Wang, "Multi-resolution location-based training for multi-channel continuous speech separation," *arXiv:2301.06458, Proc. ICASSP, in press*, 2023.

[22] S. Chen, Y. Wu, Z. Chen, J. Wu, J. Li, T. Yoshioka, C. Wang, S. Liu, and M. Zhou, "Continuous speech separation with conformer," in *Proc. ICASSP*, 2021, pp. 5749–5753.

[23] 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.

[24] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, 2018, pp. 351–355.