Iteratively Refined Multi-Channel Speech Separation

Xu Zhang , Changchun Bao * , Xue Yang , Jing Zhou

Posted Date: 22 May 2024

doi: 10.20944/preprints202405.1434.v1

Keywords: speech separation; microphone array; minimum variance distortionless response (MVDR); beamforming; iterative separation

 doi:10.20944/preprints202405.1434.v1

Article

Iteratively Refined Multi-Channel Speech Separation

Xu Zhang, Changchun Bao*, Xue Yan and Jing Zhou

Institute of Speech and Audio Information Processing, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China
* Correspondence: baochch@bjut.edu.cn

Abstract: The combination of neural network and beamforming has been proved to be very effective in multi-channel speech separation. But its performance faces a challenge in complex environment. In this paper, an iteratively refined multi-channel speech separation method is proposed for the challenge, where the proposed is composed of initial separation and iterative separation. In initial separation, the time-frequency domain dual-path recurrent neural network neural network (TFDPRNN), minimum variance distortionless response (MVDR) beamformer and post-separation (also TFDPRNN) are cascaded for obtaining the first additional input in iterative separation. In iterative separation, the MVDR beamformer and post-separation are iteratively used, where the output of the MVDR beamformer is used as an additional input of the post-separation network and the final output comes from post-separation module. This iteration of the beamformer and post-separation is fully employed for promoting their individual optimization, which ultimately improves the overall performance of speech separation in multi-speaker scenarios. Experiments on the spatialized version of the WSJ0-2mix corpus show that our proposed method is significantly better than the current popular methods. In addition, the method also has a good effect on the dereverberation task.

Keywords: speech separation; microphone array; minimum variance distortionless response (MVDR); beamforming; iterative separation

1. Introduction

Currently, the speech separation technology is playing a crucial role in human-computer interaction, audio processing and communication systems [1,2]. With the advancement of technology, especially in deep learning, the significant progress has been made in achieving efficient speech separation, especially in the fields of single-channel based speech separation [3–6]. Although single-channel based methods perform well in some environments, their effectiveness is limited in the case of more complex acoustic environments. Therefore, in order to overcome these limitations and further improve the performance of speech separation, the multi-channel speech separation methods [7] have been explored. Traditional multi-channel speech separation methods, such as delay-and-sum beamforming [8], work well in certain situations. However, in the reverberant environments, with the increase of sound sources, it is often difficult for the traditional methods to effectively separate speech signals. The difficulty is mainly focused on the fact that traditional beamforming techniques rely on relatively simple signal processing strategies, which are not ideal in dealing with dynamic changes of sound source positions or in dealing with complex and variable acoustic environments.

Due to the limitation of conventional multi-channel methods, we particularly pay attention to the approach based on the neural beamforming, because it combines the powerful nonlinear modelling capabilities of neural network with the beamformer. In classical neural beamforming, the neural network is used to obtain an initial speech separation. Subsequently, this initially separated speech and original speech are together used into the beamformer to compute spatial covariance matrices (SCM) [9]. In addition, the post filter cascaded to the beamformer is incorporated to further optimize the quality and intelligibility of the separated speech [10]. Compared with the conventional beamforming methods, the neural beamforming has been demonstrated to have significant advantages in dealing with complex acoustic environments, which has made it a mainstream method in research and application in recent years [9,11–13].

 doi:10.20944/preprints202405.1434.v1

For example, a masking-based neural beamforming method was developed in [11] and [13], in which multiple single-channel long short-term memory (LSTM) networks were first used to estimate the masks of the speakers, then these masks were used to estimate the SCM of speech and noise used for the MVDR beamformer. This kind of method shows the significant improvement in the performance of speech separation compared to conventional beamforming methods. In addition, a signal-based neural beamforming method was proposed in [9], in which a time-domain audio separation network (TasNet) was used to pre-separate the speech, and the separated speech was used to calculates the SCM used in the MVDR beamformer. This method achieves a better performance than using ideal ratio mask in the MVDR beamformer [9]. The research work in [14] indicated that reverberation has a significant impact on the separation while using the TasNet. This observation inspired us to explore a speech separation method with the anti-reverberation ability, aiming at achieving more accurate SCM used in MVDR beamformer for improving beamforming performance. Consequently, in our previous work [15], a time-frequency domain dual-path recurrent neural network (TFDPRNN) has been proposed for getting better performance of speech separation in reverberant environment. A significant performance improvement was achieved by combining the MVDR beamformer and TFDPRNN (called Beam-TFDPRNN). Although these neural beamforming methods have a good performance, they are still restricted to the linear filtering operation and the performance is limited. Therefore, we will explore other ways to improve the performance of neural beamforming in this paper.

In recent years, a notable trend of speech separation is focused on the iteratively refined structures. In the fields of single-channel speech separation, the demonstrated results in [16–18] show that the accuracy of speech separation can be significantly improved through iterative optimization. In the fields of multi-channel speech separation, the introduction of an iterative structure also demonstrated great potential. For example, a MVDR beamformer and TasNet was used as the iterative structure in [19], a time-domain real-valued generalized Wiener filter (TD-GWF) and TasNet was used as the iterative structure in [14] , and a time-domain dilated convolutional neural network (TDCN) and multi-channel Wiener filter (MCWF) was used as the iterative structure in [20]. The performance of these methods has been improved a lot compared with the no-iteration version. Therefore, in this paper, we also consider to use iterative structure to improve the performance of the neural beamforming. In this paper, we greatly extend the iterative version of our previous work [15], in which a new neural beamforming structure, that is, improved Beam-TFDPRNN (iBeam-TFDPRNN) is proposed. The main contributions of this paper are summarized as follows:

- The structure of the original neural beamforming is revised. Specifically, two main changes are made. Frist, the number of the time-frequency domain path scanning block in the neural network is reduced to 3 from original 6. This simplification improves the training efficiency and inference speed of the model, while reducing the complexity and resource consumption of the model. Second, an iteratively refined separation method is proposed, which combines the initially separated speech with the original mixed signal as an auxiliary input for the iterative network. By repeating this process in N iteration stages, the MVDR beamformer and post-separation network are mutually promoted. As a result, the separation results are effectively improved.
- The proposed method not only evaluates each stage of the multi-stage iterative processes, but also uses more evaluation metrics to get a more comprehensive evaluation. The experimental results show that the proposed method works well on the spatialized version of the WSJ0-2mix data corpus, and outperforms the current popular methods greatly. In addition, it is noted that our proposed method also performs well in the dereverberation task.
- The rest of this paper is organized as follows: Section 2 presents the details of our proposed method. Section 3 describes the experimental setup and analysis. Finally, Section 4 concludes the paper.

2. Proposed Method

 doi:10.20944/preprints202405.1434.v1

In this section, the proposed iBeam-TFDPRNN will be introduced. First, the signal model that provides a foundation for the subsequent discussions will be described. Then, the architecture of iBeam-TFDPRNN and the loss function used in the proposed method will be given.

2.1. Signal Model

In this paper, the far-field signal model with Q speakers in time-domain is considered as follows:

$$x_c = \sum_{q=1}^{Q} y_{q,c} = \sum_{q=1}^{Q} s_q * h_{q,c} \tag{1}$$

where x_c denotes the received signal by the c^{th} microphone, $1 \le c < C$, C denotes the total number of the microphones, $y_{q,c}$ denotes the signal captured by the c^{th} microphone corresponding to the q^{th} speaker, s_q denotes the original source signal of the q^{th} speaker, $h_{q,c}$ denotes the room impulse response (RIR) from the q^{th} speaker to the c^{th} microphone.

2.2. Initial Separation

In the Figure 1, the structure of initial separation is shown. It is composed of three parts, i.e., the TFDPRNN used for pre-separation with one input (mixed signal), MVDR beamformer and post-separation consisted of same TFDPRNN with two inputs (mixed signal and the output of MVDR beamformer). In this structure, the mixed signal coming from multiple microphones is first fed into this pre-separation network. Then, the pre-separated signal and the mixed speech signal are used to obtain the statistical information of the MVDR beamformer. Finally, the post-separation network is arranged at the backend of beamformer to get the finally refined initial separation speech. The following is a detailed introduction to the initial separation.

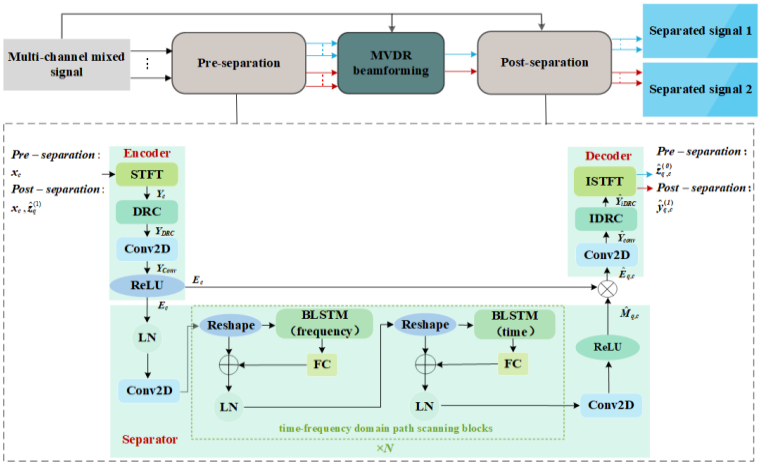

Figure 1. The overall structure of the initial separation.

During the pre-separation, the collected signals by each microphone are individually fed into respective TFDPRNN module, where each network adopts a classical encoder-separator-decoder structure for the processing.

1. In the encoder section, firstly, the mixed signal x_c is transformed into time-frequency representation Y_c by short-term Fourier transform (STFT). Then, this representation Y_c is applied into dynamic range compression (DRC) module to obtain Y_{DRC}. Subsequently, the local features Y_{Conv} are extracted from Y_{DRC} through a 2D convolutional (Conv2D) layer. Finally, these features Y_{Conv} are passed through a rectified linear unit (ReLU) activation function to obtain the encoded feature E_c. The whole encoder section can be expressed as:

$$E_c = \text{Encoder}\{x_c\}_c \tag{2}$$

where $\text{Encoder}\{\cdot\}_c$ corresponds to the encoder of the c^{th} microphone, E_c denotes the representation of encoder of the c^{th} microphone.

 doi:10.20944/preprints202405.1434.v1

2. In the separator section, firstly, the encoded feature E_c is sent to the layer normalization (LN) for standardization, followed by a Conv2D layer to obtain the feature $\hat{Y}_{conv}$. Subsequently, the feature $\hat{Y}_{conv}$ is sent into N time-frequency domain scanning blocks using a time-frequency scanning mechanism [21,22]. Each scanning block consists of two recurrent modules, where the first recurrent module utilizes a Bi-directional LSTM (BLSTM) network layer along the frequency axis, while the second recurrent module utilizes BLSTM along the time axis. Both modules all include reshaping, LN, and fully connected (FC) operations. Finally, after processing through these modules, the features are further refined through a Conv2D layer and a ReLU activation function, resulting in the separated mask $\hat{M}_{q,c}$. The whole separator section can be expressed as:

$$\hat{M}_{q,c} = \text{Separator}\left\{E_c\right\}_c \tag{3}$$

3. where Separator{.}$_c$ denotes the separator corresponding to the signal of the c^{th} microphone and $\hat{M}_{q,c}$ denotes the mask of the q^{th} speaker in the c^{th} microphone.
4. Thus, the separated masks are element-wise multiplied with the encoded feature E_c to obtain the separated feature representation $\hat{E}_{q,c}$.

$$\hat{E}_{q,c} = \hat{M}_{q,c} \odot E_c \tag{4}$$

where E_c denotes the encoded feature representation in the c^{th} microphone, $\odot$ denotes Hadamard product.

5. In the decoder section, the separated feature $\hat{E}_{q,c}$ passes through a Conv2D layer, inverse DRC (IDRC) and inverse STFT (ISFFT) to obtain the finally separated waveforms $\hat{y}_{q,c}^{(0)}$, where the superscript (0) denotes the first stage. The whole decoder section can be expressed as:

$$\hat{y}_{q,c}^{(0)} = \text{Decoder}\{\hat{E}_{q,c}\}_q \tag{5}$$

where Decoder{.}$_q$ denotes the decoder of the q^{th} speaker, $\hat{y}_{q,c}^{(0)}$ denotes the separated waveform of the q^{th} speaker in the c^{th} microphone.

During the MVDR beamforming, these separated waveforms will be employed in the computation of the SCM concerned in the MVDR beamformer, i.e.,

$$\hat{\mathbf{\Phi}}_f^{\text{Target}} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{Y}}_{q,t,f}\hat{\mathbf{Y}}_{q,t,f}^{H} \tag{6}$$

$$\hat{\mathbf{\Phi}}_f^{\text{Interfer}} = \frac{1}{T}\sum_{t=1}^{T}(\mathbf{Y}_{t,f} - \hat{\mathbf{Y}}_{q,t,f})(\mathbf{Y}_{t,f} - \hat{\mathbf{Y}}_{q,t,f})^{H} \tag{7}$$

where $\hat{\mathbf{\Phi}}_f^{\text{Target}} \in \mathbb{C}^{C\times C}$ and $\hat{\mathbf{\Phi}}_f^{\text{Interfer}} \in \mathbb{C}^{C\times C}$ represent the SCMs of the speech and interference sources, $\hat{\mathbf{Y}}_{q,t,f} \in \mathbb{C}^{C\times 1}$ is the estimated clean signal vector composed of the STFT coefficients of C microphones at time-frequency bins, which is computed from the output signals of multiple TFDPRNN modules for the q^{th} speaker. $\hat{\mathbf{Y}}_{t,f} \in \mathbb{C}^{C\times 1}$ is the multi-channel signal vector, which also consists of the STFT coefficients of C microphones at time-frequency bins. The notation H denotes the Hermitian transpose operation. Based on these SCMs, the MVDR beamformer's weights can be obtained as follows:

$$\mathbf{w}_f = \frac{\left(\hat{\mathbf{\Phi}}_f^{\text{Interfer}}\right)^{-1}\hat{\mathbf{\Phi}}_f^{\text{Target}}}{\text{Tr}\left\{\left(\hat{\mathbf{\Phi}}_f^{\text{Interfer}}\right)^{-1}\hat{\mathbf{\Phi}}_f^{\text{Target}}\right\}}\mathbf{u} \tag{8}$$

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 May 2024 doi:10.20944/preprints202405.1434.v1

where $^{-1}$ denotes the inverse of a matrix, Tr{ } denotes trace of a matrix, which is the sum of the diagonal elements of the matrix and **u** denotes a one-hot vector representing the reference microphone.

Subsequently, the separated signal by the MVDR beamformer is given by

$$\hat{z}_q^{(1)} = \mathrm{ISTFT}\left\{\mathbf{w}_f^H \mathbf{Y}_{t,f}\right\} \tag{9}$$

where $\hat{z}_{q,c}^{(1)}$ denotes the estimated signal of the q^{th} speaker in the initial separation and ISTFT{ } denotes the ISTFT operation.

During the post-separation process, the output $\hat{z}_q^{(1)}$ is regarded as additional input along with the original mixed signal xc and sent to the post-separation network to obtain the initial separation output $\hat{y}_{q,c}^{(1)}$. Specially, the post-separation network has the same structure as the pre-separation network.

2.3. Iterative separation

The overall structure of the iBeam-TFDPRNN is shown in Figure 2. This new structure is divided into two stages, i.e., initial separation and iterative separation. The initial separation has been described in section 2.2. The iterative separation contains a MVDR beamformer and a post-separation network. For convenience, the pre-separation in initial separation is called as stage 0, the first combination of MVDR beamformer and post-separation after stage 0 is called stage 1, in the subsequent iteration separation, the first iteration is called stage 2, the second iteration is called stage 3 and so on.

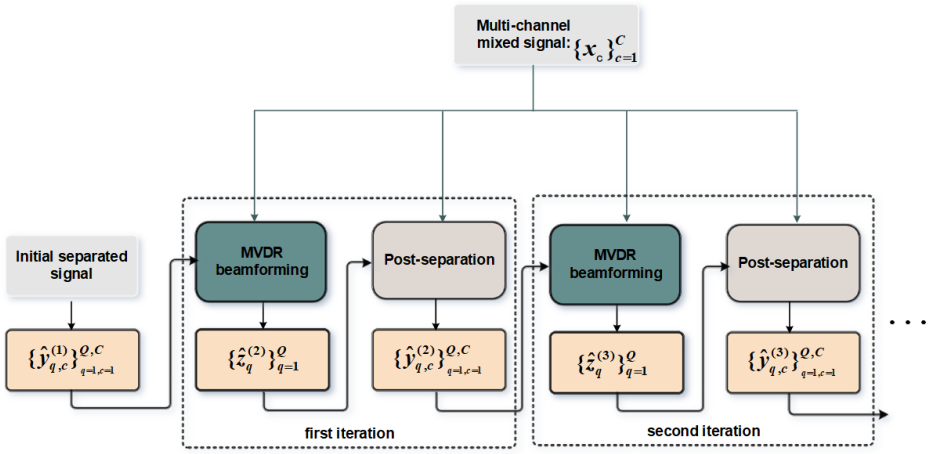

Figure 2. The structure of iBeam-TFDPRNN.

Specifically, in the first iteration, the initially separated signals $\{\hat{y}_{q,c}^{(1)}\}_{q=1,c=1}^{Q,C}$ of Q speakers on all C microphones and multi-channel mixed signals $\{x_c\}_{c=1}^{C}$ are sent to the MVDR beamformer to obtain output $\{\hat{z}_q^{(2)}\}_{q=1}^{Q}$, then, the output $\{\hat{z}_q^{(2)}\}_{q=1}^{Q}$ together with the original mixed signal $\{x_c\}_{c=1}^{C}$ are sent to the post-separation network to obtain output $\{\hat{y}_{q,c}^{(2)}\}_{q=1,c=1}^{Q,C}$. In the second iterative, the output $\{\hat{y}_{q,c}^{(2)}\}_{q=1,c=1}^{Q,C}$ and mixed signals $\{x_c\}_{c=1}^{C}$ are sent to MVDR beamformer of the next stage to obtain output $\{\hat{z}_q^{(3)}\}_{q=1}^{Q}$, then, the output $\{\hat{z}_q^{(3)}\}_{q=1}^{Q}$ together with the original mixed signal $\{x_c\}_{c=1}^{C}$ are sent to the post-separation network of the next stage to obtain output $\{\hat{y}_{q,c}^{(3)}\}_{q=1,c=1}^{Q,C}$, and continue to repeat this process N times. From this iterative separation, we can see that the output of post-separation will servers as an additional input of the MVDR beamformer of next stage. Through this iterative loop, the results of MVDR beamformer and post-separation is fully employed for promoting their individual optimization, and ultimately improves the overall performance.

 doi:10.20944/preprints202405.1434.v1

2.4. Loss Funtion

The loss function is used to calculate the signal-to-distortion ratio (SDR) between the separated signal and the original signal in the first stage and each subsequent iteration, and then adds these SDRs to form the total loss. The joint loss function can be expressed as follows:

$$Loss = -SDR(\hat{y}_{q,c}^{(0)}, y_{q,c}) - SDR(\hat{y}_{q,c}^{(1)}, y_{q,c}) - SDR(\hat{y}_{q,c}^{(2)}, y_{q,c}) \quad (10)$$

where

$$SDR(y,s) = 10log_{10}(\frac{\|s\|^2}{\|s-y\|^2}), \quad (11)$$

denotes the *SDR* calculation operator, s and y denote the reference and the separated speech signals, respectively, $y_{q,c}$ denotes the original clean signal, $\hat{y}_{q,c}^{(0)}$ denotes the pre-separated signal , $\hat{y}_{q,c}^{(1)}$ denotes the initially separated signal and $\hat{y}_{q,c}^{(2)}$ denotes the output signal of post-separtion after the first iteration.

Since each iteration can improve the last iteration result, the same loss function can be reapplied at each iteration. Here, a group of the pre-determined loss functions are used to reduce complexity during the training process for making the model easier to train. Therefore, in this paper, only a three-stage loss function is used.

3. Experimental

3.1. Datasets and Microphone Structure

6. The effectiveness of our proposed method is evaluated using the spatialized version of the WSJ0-2mix dataset given in [23]. This dataset contains 20,000 training data (about 30 hours), 5,000 validation data (about 10 hours) and 3,000 test data (about 5 hours), respectively. All utterances in the training and validation sets are either expanded or truncated to four seconds, and the sampling rate for all audio data is set to 8 kHz. This dataset comprises "min" and "max" versions: in the "min" version, the speech is truncated to match the duration of the shorter utterance, while in the "max" version, the speech is extended to match the longer utterance. The "min" version is used for training and validation, while the "max" version is used for testing to maintain consistency with baseline methods. When mixing speech from two speakers, the signal-to-interference ratio (SIR) of the speech signals varies from -5dB to +5dB. Then, these adjusted speech signals are convolved with the RIRs to simulate the reverberation effect in real environments. The RIRs are simulated by the image method proposed by Allen and Berkley in 1979 [24]. During the simulation, the length and width of the room varies from 5m to 10m, the height varies from 3m to 4m, the reverberation time varies from 0.2s to 0.6s and the positions of the microphones and speakers are all randomly selected. The microphone array consists of 8 omnidirectional microphones placed inside a virtual sphere. The center of this sphere is roughly located at the center of the room and the radius of the sphere randomly selected from 7.5m to 12.5m. The first four microphones are used for training and validation: two are symmetrically positioned on the surface of the sphere, while the other two are randomly positioned inside the sphere. The last four microphones are used for testing, and they are randomly positioned within the area defined by the first two microphones, which can evaluate the performance of the model in unseen microphone configurations.

3.2. Model Configuration

7. Here are the basic model configuration for the experiment, most of which are the same as the original settings [22]. The window settings of STFT are a 32ms frame length and a 16ms hop size. In the encoder section, the kernel size of the Conv2D layer is set to (7, 7) in order to extract local feature, while in other sections, the kernel size of Conv2D layer is set to (1, 1). In the separator section, the number of the time-frequency domain path scanning blocks in this paper is reduced to 3 from original 6 for simplifying the model. Each block contains two BLSTM layers, with each

 doi:10.20944/preprints202405.1434.v1

BLSTM layer consisting of 128 hidden units. In addition, a parameter sharing strategy is adopted in this model, meaning the same parameters are used during the iteration separation. This strategy reduces the total number of the parameters in the model, thus decreasing the computational requirements.

3.3. Training Configuration

8. In the model training, the batch size is set to 1. Utterance-level permutation invariant training (uPIT) is applied to address the source permutation problem. The Adam optimizer is utilized and the learning rate is set to 1×10^{-3}. Additionally, the maximum norm value of gradient clipping is set to 5. The networks and the comparison network are trained for 150 epochs to ensure fairness of the experiments.

3.4. Evaluation Metrics

9. The SDR of blind source separation evaluation (BSS-Eval) [25] and scale-invariant signal-to-distortion ratio (SI-SNR) are chosen as the main objective measures of separation accuracy. Furthermore, perceptual evaluation of speech quality (PESQ), short-time objective intelligibility (STOI) and SIR are used to further evaluate the accuracy of separated speech. It is worth noting that during the evaluation process, the first microphone is selected as the reference by default.

4. Results

4.1. Analysis of Iterative Results

10. The separation results of our proposed method at different stages are shown in Table 1. Here, $\hat{y}_{q,1}^{(n)}$ denotes the speech signal of each speaker on the first microphone after the TFDPRNN module at the n^{th} stage, and $\hat{z}_q^{(n)}$ denotes the speech signal of each speaker after the MVDR beamformer at the n^{th} stage.
11. From Table 1, we can see that after the first iteration, the SDR performance of $\hat{z}_q^{(n)}$ is increased by 17.46% and the SI-SDR performance of $\hat{z}_q^{(n)}$ is increased by 22.66%. This shows that the first iteration can bring high gains to our model. However, after the second iteration, although some improvement is still observed with each iteration, the performance improvements of $\hat{z}_q^{(n)}$ in SDR and SI-SDR are very slow, with an increase of less than 1%. This indicates that as the number of iterations increases, the estimation of the SCM becomes more accurate and gradually approaches the inherent performance upper limit of MVDR beamforming. On the other hand, $\hat{y}_{q,1}^{(1)}$ is improved by about 51% over $\hat{y}_{q,1}^{(0)}$ in SDR and SI-SDR, and $\hat{y}_{q,1}^{(2)}$ is improved by about 11% over $\hat{y}_{q,1}^{(1)}$ on these two metrics as well. In the subsequent stages, although the performance is continually improved, the rate of improvement is less than 2%. The performance of $\hat{y}_{q,1}^{(n)}$ and $\hat{z}_q^{(n)}$ on the SIR and PESQ are improved after each iteration. The STOI performance of $\hat{y}_{q,1}^{(n)}$ and $\hat{z}_q^{(n)}$ are remained at about 0.99. Overall, the performance of our model is improved obviously in the iterative process.
12. In order to describe the results of the method more graphically, Figure 3 shows a comparison of the spectrograms for a single speaker. They include original clean speech signal, the original reverberant speech signal, the mixed speech signal, the output signals of neural network at different stages, and the output signals of the beamformer at different stages, respectively. Observing the output spectrograms of neural network in the left column, it is clear when the processing stage increases, the clarity and quality of the spectrograms gradually are improved. For example, compared to the spectrogram of original reverberant signal, it can be observed that certain spectral components in the spectrogram of the first neural network are reduced (e.g., the red circles). After iterative processing, these signal components are gradually recovered in the spectrogram of the second neural network. In the speech spectrogram in the right column, the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 22 May 2024 doi:10.20944/preprints202405.1434.v1

separation effect of the beamformer at different stages is observed, and there are no significant changes or improvements. When comparing the speech spectrograms in the left and right columns, we can see that the output spectrograms processed by the neural network exhibit relatively higher clarity compared to those processed by the beamformer. This comparison reveals the potential advantages of neural networks in processing speech signals.

13. In general, the iteration of the MVDR and post-separation is fully employed for promoting their individual optimization, which ultimately improves the overall performance. In addition, both the outputs of MVDR and post-separation can be used as the final output of the model. However, the performance of the output of post-separation is better than the MVDR beamformer from Table 1, so the former is used as the final output of the model. In addition, each additional processing will lead to an increase in the real-time factor (RTF) of the model, thereby increasing the data processing time. Considering this, the output of post-separation after the first iteration is used as the evaluation results of our model for comparison with other methods.

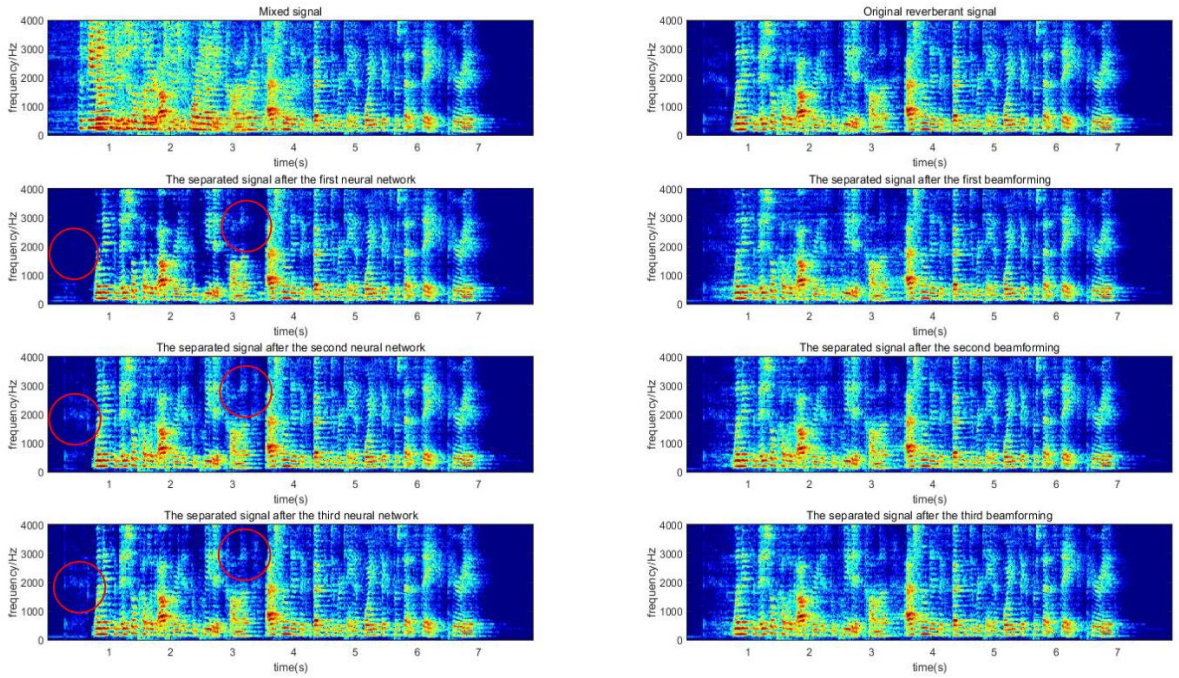

Figure 3. Spectrogram comparison of the separated speech at different stages.

Table 1. Comparison of speech separation results at different stages.

stage	RTF	SDR		SI-SDR		SIR		PESQ		STOI	
		$\hat{z}_q^{(n)}$	$\hat{y}_{q,1}^{(n)}$	$\hat{z}_q^{(n)}$	$\hat{y}_{q,1}^{(n)}$	$\hat{z}_q^{(n)}$	$\hat{y}_{q,1}^{(n)}$	$\hat{z}_q^{(n)}$	$\hat{y}_{q,1}^{(n)}$	$\hat{z}_q^{(n)}$	$\hat{y}_{q,1}^{(n)}$
0	0.024	-	14.41	-	13.91	-	30.67	-	4.13	-	0.98
1	0.049	18.79	21.84	17.08	21.13	27.21	30.67	3.93	4.13	0.99	0.98
2	**0.074**	22.07	**24.17**	20.95	**23.48**	33.21	**33.21**	3.93	**4.23**	0.99	**0.99**
3	0.104	22.10	24.48	21.07	23.84	33.23	33.93	3.99	4.25	0.99	0.99
4	0.125	22.21	**24.91**	21.20	**24.26**	33.80	**34.36**	3.98	**4.26**	0.99	**0.99**
5	0.148	22.31	24.60	21.34	23.98	34.05	34.24	4.00	4.25	0.99	0.99

4.2. Comparison with Reference Methods

At present, there are two kinds of mainstream methods to address the multi-channel speech separation task, one is frequency-domain based separation method, another one is time-domain based separation method. The most popular multi-channel speech separation methods are listed as the follows:

 doi:10.20944/preprints202405.1434.v1

14. Filter-and-Sum Network (FaSNet) [26] is a time-domain method that uses a neural network to implement beamforming technology. This method utilizes deep learning to automatically learn and optimize the weights and parameters of the beamformer. The core advantage of this method is its adaptability, allowing the network to adjust according to the complexity and diversity of speech signal;
15. Narrow-band (NB) -BLSTM [27] is a frequency-method using the BLSTM network, which is specially focused on narrow-band frequency processing and is trained by full-band methods to improve its performance. By processing each narrow-band frequency component separately in frequency domain, this method can effectively identify and separate individual speakers for the overlapped speech;
16. Beam-TasNet [9] is a classical speech separation method that combines time-domain and frequency-domain approaches. First, the time-domain neural network is used for pre-separation. Subsequently, these pre-separated speech signals are used to calculate the SCM of the beamformer. Finally, the separated signal is obtained by the beamformer;
17. Beam-Guided TasNet [19] is a two-stage speech separation method that also combines both time-domain and frequency-domain approaches. In the first stage, the initial speech separation is performed using the Beam-TasNet. In the second stage, the network structure remains the same as the Beam-TasNet, but the input includes the output from the first stage. This iterative process helps to further refine the separation of the initial speech.
18. Beam-TFDPRNN [15] is our previously proposed time-frequency speech separation method, which also uses a neural beamforming structure like Beam-TasNet. This method has more advantages in the reverberant environment, because it uses a time-frequency domain network with more anti-reverberant ability for the pre-separation.

The experimental results on the spatialized version of the WSJ0-2mix dataset for the proposed method and the popular methods are shown in Table 2. It should be emphasized that the results of Beam-TasNet are directly cited from the original paper, while the result of Beam-Guided TasNet is obtained by our replicate test, the difference is about 0.2dB from the original result.

Table 2. Comparison with reference methods on the spatialized version of the WSJ0-2mix dataset.

Method	Param	SDR	SI-SDR	PESQ	SIR	STOI
FaSNet	2.8 M	11.96	11.69	3.16	18.97	0.93
NB-BLSTM	1.2 M	8.22	6.90	2.44	12.13	0.83
Beam-TasNet	5.4M	17.40	-	-	-	-
Beam-Guided TasNet	5.5M	20.52	19.49	3.88	27.49	0.98
Beam-TFDPRNN	2.7 M	17.20	16.80	3.68	26.77	0.96
iBeam-TFDPRNN	2.8M	24.17	23.48	4.23	33.21	0.99

From Table 2, we can see that the performance of the FaSNet and NB-BLSTM is not-satisfactory. In comparison, the Beam-TasNet and Beam-TFDPRNN demonstrate good separation performance. The Beam-Guided TasNet further improves the performance of Beam-TasNet by employing an iteratively refined structure. The proposed method, iBeam-TFDPRNN, significantly outperforms the other methods. In addition, our proposed model has only 2.8M parameters, which is smaller than the parameters of most other methods. In conclusion, the proposed method in this paper has excellent performance compared to the reference methods.

4.3. Dereverberation Results

In the previous section, we only discuss the performance of the proposed method in the task of the reverberation. However, in order to comprehensively evaluate the performance of the proposed method and explore its performance in different tasks, this section will explore the separation performance of the proposed method on the dereverberation task.

 doi:10.20944/preprints202405.1434.v1

The experimental results in Table 3 exhibit the performance of our proposed method and reference methods on the dereverberation task. We can see that our proposed method has significant advantage than reference methods. Specifically, the proposed method achieves an SDR of 20.2dB, much better than Beam-TasNet and exceeds Beam-Guided TasNet by 2.1dB. Additionally, it exceeds the oracle mask-based MVDR by 8.2dB, and narrows the gap to just 0.9dB with the oracle signal-based MVDR. These results highlight the effectiveness of our proposed method in dereverberation tasks, demonstrating its potential for real-world applications.

Table 3. Dereverberation performance on the spatialized version of the WSJ0-2mix dataset.

Method	SDR	
	$\hat{y}_{q,1}^{(n)}$	$\hat{z}_{q,1}^{(n)}$
Beam TasNet	10.8	14.6
Beam-Guided TasNet	16.5	17.1
iBeam- TFDPRNN	**20.2**	19.7
Oracle mask-based MVDR	11.4	12.0
Oracle signal-based MVDR	∞	21.1

5. Conclusions and Discussion

In this paper, an iteratively refined multi-channel method was proposed to improve the performance of speech separation in complex environments. Benefiting from the strength of neural beamforming and the multi-stage iteratively refined structure, the proposed method achieved outstanding performance. The experiments on the spatialized version of the WSJ0-2mix corpus show that the proposed method not only has a good separation performance in a reverberant environment, but also has significant advantages compared to current popular speech separation methods. In addition, the model shows a promising ability on dereverberation task. However, the dataset in this paper does not contain noise components. Therefore, exploring speech separation problems in noisy environments will be our future research direction, and the effectiveness of this method in realistic environments will be further validated by using noisy datasets such as LibriCSS [28] and WHAMR! [29].

Author Contributions: Conceptualization, X.Z, C.C.B; methodology, X.Z.; software, X.Z.; validation, X.Z, X.Y., J.Z.; formal analysis, X.Z, X.Y., J.Z.; investigation, X.Z.; resources, X.Z.; data curation, X.Z.; writing—original draft preparation, X.Z.; writing—review and editing, X.Z., C.C.B, X.Y, and J.Z.; visualization, X.Z.; supervision, C.C.B.; project administration, C.C.B; funding acquisition, C.C.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 61831019.

Acknowledgments: The authors are grateful to the thorough reviewers.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, Z.; Li, J.; Xiao, X.; Yoshioka, T.; Wang, H.; Wang, Z.; Gong, Y. Cracking the Cocktail Party Problem by Multi-Beam Deep Attractor Network. In Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); December 2017; pp. 437–444.
2. Qian, Y.; Weng, C.; Chang, X.; Wang, S.; Yu, D. Past Review, Current Progress, and Challenges Ahead on the Cocktail Party Problem. *Frontiers Inf Technol Electronic Eng* **2018**, *19*, 40–63, doi:10.1631/FITEE.1700814.
3. Chen, J.; Mao, Q.; Liu, D. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation 2020.

 doi:10.20944/preprints202405.1434.v1

4. Luo, Y.; Mesgarani, N. Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1256–1266, doi:10.1109/TASLP.2019.2915167.
5. Subakan, C.; Ravanelli, M.; Cornell, S.; Bronzi, M.; Zhong, J. Attention Is All You Need In Speech Separation. In Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Toronto, ON, Canada, June 6 2021; pp. 21–25.
6. Zhao, S.; Ma, Y.; Ni, C.; Zhang, C.; Wang, H.; Nguyen, T.H.; Zhou, K.; Yip, J.; Ng, D.; Ma, B. MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation 2023.
7. Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 692–730, doi:10/ghncwr.
8. Anguera, X.; Wooters, C.; Hernando, J. Acoustic Beamforming for Speaker Diarization of Meetings. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 2011–2022, doi:10.1109/TASL.2007.902460.
9. Ochiai, T.; Delcroix, M.; Ikeshita, R.; Kinoshita, K.; Nakatani, T.; Araki, S. Beam-TasNet: Time-Domain Audio Separation Network Meets Frequency-Domain Beamformer. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Barcelona, Spain, May 2020; pp. 6384–6388.
10. Zhang, X.; Wang, Z.-Q.; Wang, D. A Speech Enhancement Algorithm by Iterating Single- and Multi-Microphone Processing and Its Application to Robust ASR. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); March 2017; pp. 276–280.
11. Erdogan, H.; Hershey, J.R.; Watanabe, S.; Mandel, M.I.; Roux, J.L. Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks. In Proceedings of the Interspeech 2016; ISCA, September 8 2016; pp. 1981–1985.
12. Gu, R.; Zhang, S.-X.; Zou, Y.; Yu, D. Towards Unified All-Neural Beamforming for Time and Frequency Domain Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 849–862, doi:10.1109/TASLP.2022.3229261.
13. Xiao, X.; Zhao, S.; Jones, D.L.; Chng, E.S.; Li, H. On Time-Frequency Mask Estimation for MVDR Beamforming with Application in Robust Speech Recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: New Orleans, LA, March 2017; pp. 3246–3250.
14. Luo, Y. A Time-Domain Real-Valued Generalized Wiener Filter for Multi-Channel Neural Separation Systems. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 3008–3019, doi:10.1109/TASLP.2022.3205750.
15. Zhang, X.; Bao, C.; Zhou, J.; Yang, X. A Beam-TFDPRNN Based Speech Separation Method in Reverberant Environments. In Proceedings of the 2023 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC); IEEE: ZHENGZHOU, China, November 14 2023; pp. 1–5.
16. Kavalerov, I.; Wisdom, S.; Erdogan, H.; Patton, B.; Wilson, K.; Roux, J.L.; Hershey, J.R. Universal Sound Separation 2019.
17. Tzinis, E.; Wisdom, S.; Hershey, J.R.; Jansen, A.; Ellis, D.P.W. Improving Universal Sound Separation Using Sound Classification. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Barcelona, Spain, May 2020; pp. 96–100.
18. Shi, Z.; Liu, R.; Han, J. LaFurca: Iterative Refined Speech Separation Based on Context-Aware Dual-Path Parallel Bi-LSTM 2020.
19. Chen, H.; Yi, Y.; Feng, D.; Zhang, P. Beam-Guided TasNet: An Iterative Speech Separation Framework with Multi-Channel Output. *arXiv:2102.02998 [eess]* **2022**.
20. Wang, Z.-Q.; Erdogan, H.; Wisdom, S.; Wilson, K.; Raj, D.; Watanabe, S.; Chen, Z.; Hershey, J.R. Sequential Multi-Frame Neural Beamforming for Speech Separation and Enhancement. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT); IEEE: Shenzhen, China, January 19 2021; pp. 905–911.
21. Yang, L.; Liu, W.; Wang, W. TFPSNet: Time-Frequency Domain Path Scanning Network for Speech Separation. 5, doi:10.1109/ICASSP43922.2022.9747554.
22. Yang, X.; Bao, C.; Zhang, X.; Chen, X. Monaural Speech Separation Method Based on Recurrent Attention with Parallel Branches. In Proceedings of the INTERSPEECH 2023; ISCA, August 20 2023; pp. 3794–3798.
23. Wang, Z.-Q.; Le Roux, J.; Hershey, J.R. Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Calgary, AB, April 2018; pp. 1–5.
24. Allen, J.B.; Berkley, D.A. Image Method for Efficiently Simulating Small-room Acoustics. *The Journal of the Acoustical Society of America* **1979**, *65*, 943–950, doi:10.1121/1.382599.
25. Févotte, C.; Gribonval, R.; Vincent, E. BSS_EVAL TOOLBOX USER GUIDE REVISION 2.0. **2011**, 22.
26. Luo, Y.; Ceolini, E.; Han, C.; Liu, S.-C.; Mesgarani, N. FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing. *arXiv:1909.13387 [cs, eess]* **2019**.

 doi:10.20944/preprints202405.1434.v1

27. Quan, C.; Li, X. Multi-Channel Narrow-Band Deep Speech Separation with Full-Band Permutation Invariant Training. In Proceedings of the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Singapore, Singapore, May 23 2022; pp. 541–545.
28. Chen, Z.; Yoshioka, T.; Lu, L.; Zhou, T.; Meng, Z.; Luo, Y.; Wu, J.; Xiao, X.; Li, J. Continuous Speech Separation: Dataset and Analysis 2020.
29. Maciejewski, M.; Wichern, G.; McQuinn, E.; Roux, J.L. WHAMR!: Noisy and Reverberant Single-Channel Speech Separation. In Proceedings of the ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); IEEE: Barcelona, Spain, May 2020; pp. 696–700.