symmetry

MDPI

Article

A Pre-Separation and All-Neural Beamformer Framework for Multi-Channel Speech Separation

Wupeng Xie [1,*], Xiaoxiao Xiang [2,3,4], Xiaojuan Zhang [2,3] and Guanghong Liu [1]

[1] Information Science Academy, China Electronics Technology Group Corporation, Beijing 100041, China
[2] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China
[3] Key Laboratory of Electromagnetic Radiation and Sensing Technology, Chinese Academy of Sciences, Beijing 100190, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: xiewupeng15@mails.ucas.ac.cn

Abstract: Thanks to the use of deep neural networks (DNNs), microphone array speech separation methods have achieved impressive performance. However, most existing neural beamforming methods explicitly follow traditional beamformer formulas, which possibly causes sub-optimal performance. In this study, a pre-separation and all-neural beamformer framework is proposed for multi-channel speech separation without following the solutions of the conventional beamformers, such as the minimum variance distortionless response (MVDR) beamformer. More specifically, the proposed framework includes two modules, namely the pre-separation module and the all-neural beamforming module. The pre-separation module is used to obtain pre-separated speech and interference, which are further utilized by the all-neural beamforming module to obtain frame-level beamforming weights without computing the spatial covariance matrices. The evaluation results of the multi-channel speech separation tasks, including speech enhancement subtasks and speaker separation subtasks, demonstrate that the proposed method is more effective than several advanced baselines. Furthermore, this method can be used for symmetrical stereo speech.

Keywords: multi-channel speech separation; beamforming; pre-separation module; all-neural; speech enhancement

Citation: Xie, W.; Xiang, X.; Zhang, X.; Liu, G. A Pre-Separation and All-Neural Beamformer Framework for Multi-Channel Speech Separation. *Symmetry* **2023**, *15*, 261. https://doi.org/10.3390/sym15020261

Academic Editor: Peng-Yeng Yin

Received: 17 December 2022
Revised: 13 January 2023
Accepted: 15 January 2023
Published: 17 January 2023

1. Introduction

Speech separation can extract target speaker information from speech signals corrupted by interference and reverberation, and it can improve the quality of communication between people. Thanks to the powerful nonlinear modeling capabilities of deep learning, speech separation has received extensive attention and achieved significant performance improvements. Speech can be separated in the time domain [1–3] or the time-frequency domain [4–8]. Since time-frequency domain methods have clearer feature patterns and overall better speech quality; this paper focuses on time-frequency domain methods. According to the number of recording microphones, speech separation can be classified as monaural separation and microphone array separation.

Although single-channel speech separation has achieved impressive performance, it can only use the characteristics of the signal itself. Therefore, single-channel speech separation will inevitably produce speech distortion, especially in a noisy and reverberant far-field environment, which seriously affects the performance of speech separation. On the contrary, multi-channel speech separation tasks can utilize additional spatial information, which is very important to improve the quality of speech after separation in extremely challenging conditions. Therefore, the study of multi-channel speech separation has aroused widespread interest. In traditional multi-channel speech enhancement, the mainstream method is acoustic beamforming [9,10], for example, through generalized

eigenvalue (GEV) beamformers or minimum variance distortionless response (MVDR) beamformers, which enhance signals in specific directions while attenuating signals in other directions. Therefore, early researchers usually combined deep neural networks (DNNs) with spatial filtering methods based on signal-processing theory [11–14], such as GEV beamformers or MVDR beamformers, to design spatial filters, which can reduce the nonlinear distortion introduced by DNN. This process can be briefly described as follows. First, the speech and interference of each channel are independently estimated by the designed single-channel speech separation network. Here the interference is obtained by subtracting the target reverberant signal from the noisy mixture, which contains the other interfering speaker and background noise. Then, the second-order statistics, i.e., speech and interference spatial covariance matrices, are calculated to obtain beamformer weights. However, as the single-channel network is typically trained independently, it may produce unreliable outputs. Another approach is to train the separation network with additional spatial features, such as inter-channel phase and level difference [15–18]. Recent studies regard the trained DNN itself as a nonlinear spectral-spatial filter, which takes the real and imaginary parts of the multi-channel signal as inputs and uses a complex spectral mapping method to generate the real and imaginary spectrograms of the target signal [19–21]. In this way, DNNs can implicitly exploit the direction information contained in the array signals input to the multi-channel.

Much research on the usage of DNNs to directly generate frame-level beamforming weights have been carried out. For example, in [22], the authors used two recurrent neural networks (RNNs) to substitute the covariance matrix inversion and eigenvalue decomposition processes of traditional MVDRs for neural frame-adaptive beamforming. Subsequently, Xu et al. [23] improved it by using one unified RNN-DNN model to calculate beamforming weights directly, which achieves higher speech quality and automatic speech recognition (ASR) accuracy. Moreover, Luo et al. [24] proposed a filter-and-sum network that calculates adaptive beamforming filters for each microphone and sums the filtered outputs of all channels as the final output. However, most current methods obtain limited performance improvement and lack of adequate robustness. For example, FasNet+TAC obtains a slight performance improvement under extreme conditions, i.e., a high overlap ratio or small speaker angle.

In this paper, a pre-separation and all-neural beamformer framework for multi-channel speech separation is proposed which is called PsBf and consists of the pre-separation module and the all-neural beamforming module. Note that the proposed PsBf is evaluated on two subtasks, including reverberant speech enhancement and speaker separation tasks. The contributions of this paper can be summarised as follows. (i) The pre-separation module is designed to output multiple filter weights, and it is constructed by combining a multi-scale aggregation block (MSAB) with a dual-path recurrent neural network (DPRNN). The introduction of MSAB makes it possible for the network to better obtain the contextual information of different scales. The local and global features can be modeled by DPRNN at the same time. (ii) A new all-neural beamforming module that directly learns frame-wise beamforming weights from estimated speech and interference, without following the form of conventional beamforming solutions, is proposed. In this way, the neural network can learn better filter weights in a data-driven manner. (iii) Experimental results show that in terms of microphone array speech enhancement and speaker separation tasks, the PsBf proposed in this paper shows significantly better performance than several other baseline methods. Furthermore, even under extremely challenging conditions, the proposed PsBf still achieves acceptable separation performance. In addition, this approach can be used for symmetrical stereo speech.

The remainder of the paper is organized as follows. Sections 2 and 3 introduce the problem definition and model architecture, respectively. The experimental setup is given in Section 4. The experimental results and analysis are presented in Section 5. Section 6 gives the concluding remarks.

2. Problem Formulation

The signals containing reverb noise and input into the P-channel microphone array can be expressed in the short-time Fourier transform (STFT) domain:

$$\mathbf{Y}(t,f) = \mathbf{S}(t,f) + \mathbf{N}(t,f) = \mathbf{d}_{ref}(f)S_{ref}(t,f) + \mathbf{N}(t,f), \quad (1)$$

where $\mathbf{Y}(t,f)$, $\mathbf{S}(t,f)$, and $\mathbf{N}(t,f)$ represent the reverberant mixture, reverberant speech, and reverberant interference at the time t and frequency f, respectively. S_{ref} denotes the STFT of the target speech signal in the reference microphone, and $\mathbf{d}_{ref}(f)$ is the relative transfer function of the target speech to the reference channel. The spatial information contained in the receiving signal between multiple channels will not change due to arbitrary selection of the reference channel, and the first channel is selected as a reference channel in this paper. The purpose of microphone array speech separation is to extract the target signal S_{ref} from the noisy and reverberant signals. It should be pointed out that the objective of this study is to separate the target speech rather than dereverberation, hence the reverberated pure speech as the training label.

3. Model Description

Figure 1 shows the structure diagram of PsBf, which contains a pre-separation module and an all-neural beamforming module. The pre-separation module is a variant of the TPRNN proposed previously [25], which is used to obtain pre-separated speech and interference. The all-neural beamforming module simulates conventional beamformers to calculate frame-level beamforming weights. Next, these two modules are described in detail.

3.1. Pre-Separation Module

TPRNN, which has been proposed previously, is suitable for multi-channel speech separation on distributed microphone arrays. Here, the input features of the model are transformed from the time domain to the time-frequency domain, as it is empirically found that the time-frequency domain speech separation method can obtain better performance than the time-domain speech separation method. The encoder is composed of two layers with an MSAB inserted into between them, and it is used to input the reverb mixture $\mathbf{Y}(t,f)$. MSAB can obtain richer contextual information. As shown in Figure 2, MSAB mainly consists of three modules, namely the input layer, the UNet-shaped multi-scale feature extraction layer and the output layer. By utilizing MSAB, the local and global information of speech can be effectively captured. The first layer of the encoder consists of 1×1 convolutions, layer normalization [26] (LN) and a parameter-rectified linear unit [27] (PReLU) activation, which increases the number of channels to $C = 64$, remains unchanged in subsequent modules. The second layer uses 1×3 convolutions with a stride of 2 for halving the frequency dimensions size, followed by LN and PReLU.

The input of the stacked DPRNN block is the output of the encoder. DPRNN mainly includes two-stage processing, in which local and global processing is performed iteratively and alternately. The local processing consists of bi-directional gate recurrent units [28] (BGRU), a fully connected layer (FC), and a global LN (gLN). As for global processing, BGRU and gLN are correspondingly replaced by GRU and cumulative LN (cLN) to guarantee strict causality. There are X units in each direction of each BGRU layer. The outputs of all DPRNN blocks are fed into the adaptive feature fusion block to aggregate diverse intermediate features. In addition to replacing the convolution in the encoder with sub-pixel convolution, the decoder is a reverse version of the encoder. Sub-pixel convolution can replace transposed convolution to reconstruct compressed features, and its upsampling ratio is 2. The advantage of adding skip connections between the encoder and the decoder is that the flow of information in the network is improved. The last layer of the decoder is a 1×1 convolution, and linear activation layer is used due to the need to generate an unbounded mask $\tilde{M}$. The mask $\tilde{M}$ has 2P channels, where P indicates the number of micro-

phones, and the number 2 represents the real and imaginary two parts, so pre-separated speech $\tilde{S}$ and interference $\tilde{N}$ can be obtained.

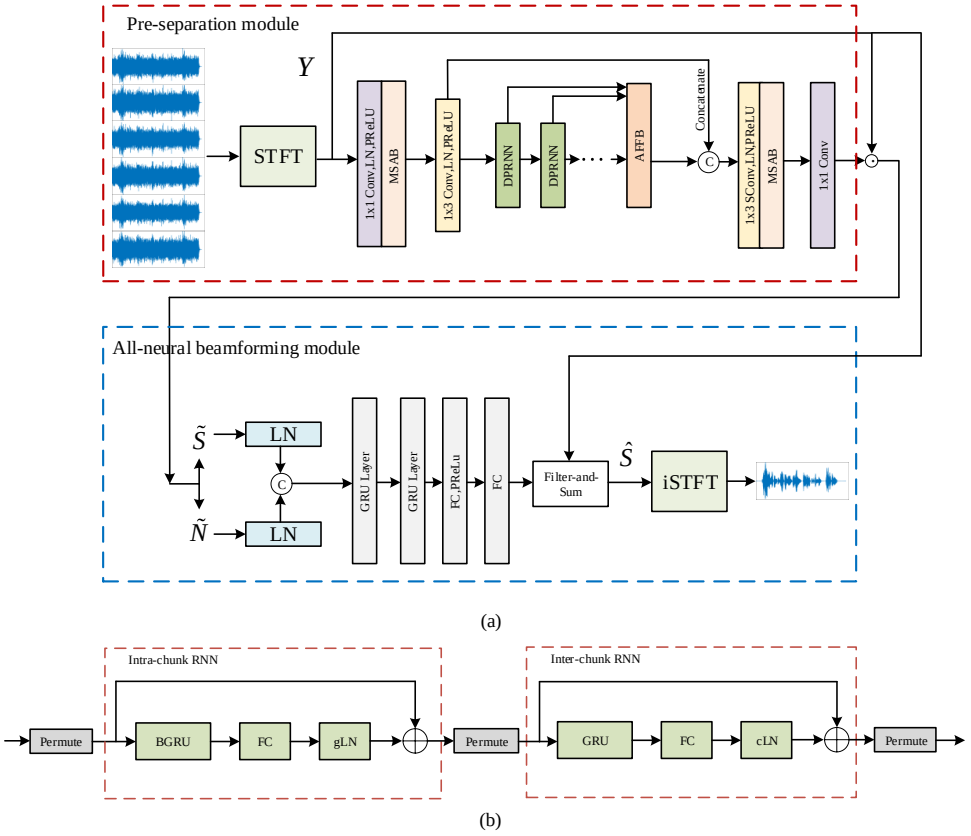


Figure 1. The entire architecture of the PsBf. (**a**) PsBf architecture. (**b**) DPRNN block. "gLN" represents global layer normalization, "cLN" represents cumulative layer normalization, "SConv" represents sub-pixel convolution, "AFFB" means adaptive feature fusion block. ⊙ indicates the complex-domain multiplication, and filter-and-sum denotes the operations expressed in Equation (4).

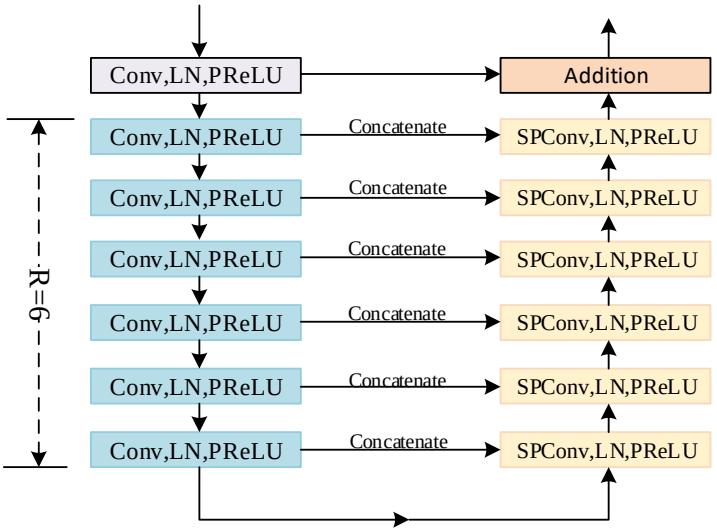


Figure 2. The diagram of MSAB. *R* represents the number of convolutional layers in the multi-scale feature extraction layer.

3.2. All-Neural Beamforming Module

After estimating speech and interference, the all-neural beamforming module is applied to calculate frame-level beamforming weights. As opposed to the previous all-neural beamformer, which needs to follow the form of conventional beamformer solutions, i.e., by computing the spatial covariance matrix and its inverse, the estimated speech and interference are directly used to calculate the beamforming weight, which can improve the performance of the system. In this module, LN is first employed to normalize speech and interference. The normalized speech and interference are then concatenated along the channel dimension and fed into two GRU layers with 128 hidden nodes to simulate a frame-by-frame update beamforming process. Two FC layers are used for estimating the weights, one of which is followed by PReLU; the other is followed by linear output. The following formula gives the overall process.

$$\tilde{\mathbf{w}}_{bf}(t,f) = \mathrm{GRU}([LN(\tilde{\mathbf{S}}), LN(\tilde{\mathbf{N}})]) \tag{2}$$

$$\mathbf{w}_{bf}(t,f) = FC(\mathrm{PReLU}(FC(\tilde{\mathbf{w}}_{bf}(t,f)))) \tag{3}$$

$$\hat{S} = (\mathbf{w}_{bf}(t,f))^{H}\mathbf{Y}(t,f) \tag{4}$$

where $\mathbf{w}_{bf}(t,f) \in \mathbb{C}^{P}$ is the frame-wise weight. The time-domain enhanced signal $\hat{s}$ can be obtained by performing an inverse STFT (iSTFT) on $\hat{S}$.

It should be pointed out here that PsBf is inspired by TPRNN and has been changed. First, the network operates in the time-frequency domain, which has more explicit spectral features than the time domain. Second, rather than introducing a third path RNN to explicitly model spatial information, the method proposed implicitly exploits spatial information encoded in the multi-channel input signal. Third, the all-neural beamforming module is additionally added, which simulates the traditional spatial filtering beamformer and can effectively restore speech quality while reducing distortion.

3.3. Loss Function

The phase-constrained magnitude [29] (PCM) loss is adopted to train the speech enhancement model. PCM is an improved version of STFT magnitude loss, which can eliminate an unknown artifact due to the magnitude loss of STFT.

$$L_{PCM}(s,\hat{s}) = \frac{1}{2}\cdot L_{SM}(s,\hat{s}) + \frac{1}{2}\cdot L_{SM}(n,\hat{n}) \tag{5}$$

where s and n represent a clean speech signal and a interference signal, respectively. $\hat{s}$ and $\hat{n}$ indicate the separated speech and interference, respectively. L_{SM} is the magnitude loss of STFT, and its expression is

$$\begin{aligned} L_{SM}(s,\hat{s}) = \tfrac{1}{T\cdot F}\sum_{t=0}^{T-1}\sum_{f=0}^{F-1} &|(|S_r(t,f)| + |S_i(t,f)|) \\ &-(|\hat{S}_r(t,f)| + |\hat{S}_i(t,f)|)| \end{aligned} \tag{6}$$

where $S(t,f)$ and $\hat{S}(t,f)$ are the STFTs of s and $\hat{s}$ in the time t and frequency f, respectively. S_r and S_i represent the real and the imaginary part of the complex spectrogram S, respectively. $L_{SM}(n,\hat{n})$ is also defined in this way.

For the speaker separation task, the scale-invariant signal-to-noise ratio (SI-SNR) is adopted to train the network. SI-SNR is defined as

$$\mathrm{Si-SNR} := 10\log_{10}\left(\frac{||\alpha s||^2}{||\alpha s - \hat{s}||^2}\right) \tag{7}$$

where $\alpha = \frac{\langle \hat{s}, s \rangle}{\langle s, s \rangle}$ is just a scalar. $\langle \cdot, \cdot \rangle$ and $|| \cdot ||$ represent the inner product and Euclidean norm, respectively. Utternce-level permutation-invariant training [30] is utilized for solving the source permutation problem during the training process. Note that instead of being provided to two modules explicitly, the supervision is only available for the final estimate. Therefore, the entire network can be trained in an end-to-end manner.

4. Experimental Setup

4.1. Datasets

The speech enhancement task is evaluated on the Librispeech dataset [31], and the image method is employed to generate multi-channel mixture speech signals. Two different microphone arrays were placed in the room. One is a circular array of six microphones; these were evenly distributed in a circle with a radius of 5 cm. The other is a line array of six microphones, and the distance between these microphones was set to 4 cm. The room size is randomly sampled from 3 m × 3 m × 2.5 m (*length* × *width* × *height*) to 10 m × 10 m × 4 m, which can cover most of the actual scenes. The reverberation time (T60) is sampled from 0.1 to 0.5 s at random. When generating the training set and validation set, the clean speech signal comes from train-clean-100 and dev-clean, respectively, and the noise signal comes from DNS challenge dataset [32]. The relative SNR between speech and noise is randomly sampled among {−5 dB, −4 dB, −3 dB, −2 dB, −1 dB, 0 dB}. For each microphone array, 40,000 training utterances and 5000 validation utterances are generated. For the test set, in order to test the generalization of the model to unseen noise, two challenging noises from NOISEX-92 [33], namely babble and factory1, are selected. Multi-channel mixture speech signals under three SNR conditions are generated, namely −5 dB, 0 dB, and 5 dB. For each SNR condition, 600 noisy utterances are generated. The experiments are conducted under the circular array if not stated otherwise.

The same dataset is used for [34] as for the two-speaker separation task, which contains training set, validation set and test set. During each mixing process, two different speakers at various SNRs between 0 dB and 5 dB are randomly selected. The overlap ratio between the two speakers is uniformly sampled between 0% and 100%.

4.2. Experimental Configurations

The sampling rate of generated utterances is set to 16 kHz. A 32 ms Hanning window is used to segment the signal into frames, and the frame shift is 16 ms. A 512-point STFT is performed on the framed signal, which generates 257-D spectral vectors. The Adam [35] optimizer is adopted to train the entire network, and the initialized learning rate and batch size are set to 0.001 and 16, respectively. In each mini-batch, the sentence length is fixed to 4s. The loss on the validation set affects the value of the learning rate and the learning behavior of the model. If the best model is not found for three consecutive times, the learning rate is halved, and if the best model is not found for five consecutive times, the model training is terminated early. To avoid exploding gradients, gradient clipping with a maximum L2-norm of 5 is applied. The model for the speech enhancement task and the speaker separation task are trained 50 and 100 epoch, respectively.

4.3. Baseline

The model proposed herein is compared with different advanced baselines for speech enhancement, including GCRN [36], SADNUNet [37], DC-CRN [21], FasNet+TAC [38], and TPRNN. GCRN and SADNUNet are two advanced single-channel models. DC-CRN is a method based on complex spectral mapping which directly uses the real and the imaginary spectrograms of the multi-channel mixture to estimate those of the clean speech. FasNet+TAC and TPRNN are two advanced time-domain multi-channel models which are designed to be applied to both fixed microphone array and ad hoc microphone array. For the speaker separation task, seven baselines are chosen, namely TasNet, FasNet+TAC, DCRN [39], MC-ConvTasNet [17], TD-GWF-TasNet [40], IC-ConvTasNet [41], and TPRNN. Except for TasNet, other baselines are recently proposed advanced multi-channel speaker

separation systems. All advanced baselines adopt the best configuration mentioned in the original literature and utilize the same dataset mentioned in this study for training, validation and testing.

5. Results

In speech enhancement task, perceptual evaluation of speech quality (PESQ) [42] and short-term objective intelligibility (STOI) [43] are used to assess the enhancement performance. For the speaker separation task, SI-SNR is utilized to evaluate the accuracy of speaker separation.

5.1. Ablation Study

In order to verify the influence of different network components, an ablation of multi-channel noise reverberation speech enhancement task was carried out. Table 1 gives the experimental results, from which the following phenomena can be observed. First of all, increasing the hidden layer size in the GRU can facilitate more accurate spectral estimation. For instance, an average of 0.14, 2.56%, and 0.43 score improvements are observed in terms of PESQ, STOI, and SI-SNR, respectively, from Variant-1 to Variant-2. Secondly, from Variant-2 to Variant-3, significant performance improvements are observed, where Variant-2 only outputs the filter weights of the reference channel, which shows that multi-filter estimation is essential in multi-channel speech enhancement tasks. Thirdly, the performance of the model is further improved by adding the neural beamformer module; the network with neural beamforming achieves a slight improvement in PESQ and STOI, and a significant improvement in SI-SNR, as shown in Variant-3 and Variant-4. One possible reason is that the neural beamformer simulates the traditional MVDR beamformers, which can reduce the distortion of the recovered speech while effectively suppressing the residual noise. This phenomenon is more obvious in the non-causal setting (not reported here). Finally, the non-causal system, i.e., Variant-5, achieved the best performance. This result is very logical, because non-causal systems can take advantage of more information, especially future information, which is crucial for speech-related tasks. In the following experiments, the same parameter configuration as Variant-4 is used unless otherwise specified, as it achieves optimal performance under the causal setting.

Table 1. Ablation study on the simulated 6-mic circular array. "Cau." indicates whether the system is a causal implementation. "MO" and "BF" represent whether the multi-channel masks were output, and neural beamformer, respectively. ↑ represents the increment relative to the mixture. X denotes the hidden dimension in the GRU. The ***bold*** values show the best results.

Variant	Causal	MO	BF	X	Par.(M)	PESQ ↑	STOI ↑	SI-SNR ↑
Variant-1	✓	✗	✗	64	**1.08**	1.22	25.98	9.17
Variant-2	✓	✗	✗	128	1.73	1.36	28.54	9.60
Variant-3	✓	✓	✗	128	1.73	1.51	31.28	11.90
Variant-4	✓	✓	✓	128	1.91	1.53	31.54	12.49
Variant-5	✗	✓	✓	128	2.45	**1.80**	**35.04**	**14.24**

5.2. Speech Enhancement

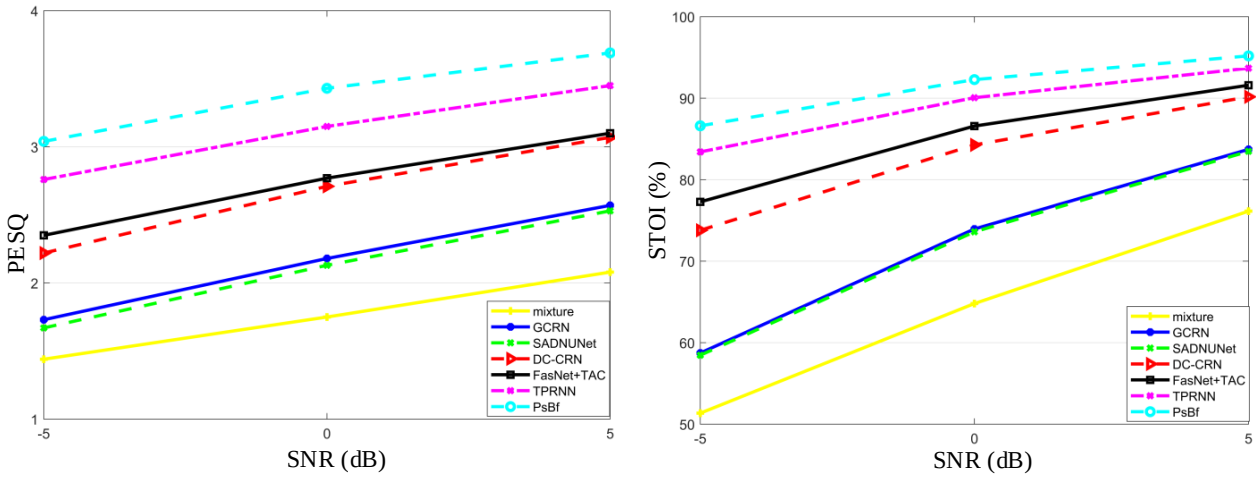
According to the results of the ablation study, Variant-4 is chosen to compare with other baseline methods on speech enhancement tasks. The experiments are performed on the circular arrays first; Table 2 presents STOI and PESQ results at −5 dB, 0 dB, and 5 dB, respectively. First, all models improve PESQ and STOI over the unprocessed mixtures, which demonstrates that the speech enhancement model can obviously improve the perception quality and intelligibility of speech. Second, GCRN and SADNUNet, two advanced single-channel speech enhancement systems, obtain similar metric scores and achieve minor performance improvement over the unprocessed mixture. A possible reason is that single-channel speech enhancement has suffered from the performance bottleneck in

the more challenging noisy reverberant environments. Thirdly, all multi-channel models achieve consistently better enhancement performance than single-channel models, due to the fact that the multi-channel speech enhancement method can improve speech recovery ability by using spatial information provided by the microphone array. Fourthly, the proposed model performes significantly better than all baselines in terms of objective speech intelligibility and quality metrics. For example, compared with FasNet+TAC, a recently proposed time-domain end-to-end multi-channel speech separation method, the proposed PsBf achieves average score improvements of 0.65 and 6.21% according to PESQ and STOI, respectively. Finally, it is very interesting to find that TPRNN, an architecture designed for the ad hoc array, also achieves significant performance improvements in a microphone fixed geometry array configuration. However, it still underperforms the proposed PsBf, which fully demonstrates the superiority of the method proposed.

Table 2. Speech enhancement results on the simulated 6-mic circular array according to PESQ and STOI. "Par." represents the number of training parameters. Best results are shown using ***bold*** values.

Metrics	**Cau.**	**Par.(M)**	**STOI (%)**				**PESQ**			
Test SNR	**-**	**-**	**−5 dB**	**0 dB**	**5 dB**	**Avg.**	**−5 dB**	**0 dB**	**5 dB**	**Avg.**
Mixture	-	-	48.40	55.12	60.51	54.68	1.40	1.57	1.74	1.57
GCRN	✓	9.77	59.48	65.77	69.93	65.06	1.80	2.10	2.30	2.07
SADNUNet	✓	2.63	59.61	65.66	69.81	65.03	1.82	2.04	2.24	2.03
DC-CRN	✓	12.97	70.07	74.66	77.27	74.00	2.16	2.39	2.58	2.38
FasNet+TAC	✓	2.76	69.45	73.86	76.67	73.33	2.09	2.25	2.39	2.24
TPRNN	✗	2.28	81.18	83.82	85.52	83.51	2.76	2.93	3.07	2.92
PsBf	✓	1.91	**84.33**	**86.68**	**88.13**	**86.38**	**2.95**	**3.11**	**3.24**	**3.10**

Furthermore, to investigate the effectiveness of the proposed PsBf in the linear array configuration, the experiments are also conducted on the previously mentioned linear array. From the evaluation results in Figure 3, it can be found that all models obtains similar performance trends as in Table 2. Experimental results demonstrate that method proposed herein is effective for different array configurations.



(**a**) (**b**)

Figure 3. Speech enhancement results on the simulated 6-mic linear array in terms of PESQ and STOI. (**a**) PESQ; (**b**) STOI.

5.3. Speaker Separation

In this subsection, the effectiveness of the proposed PsBf on the noisy reverberant speaker separation task is verified; the PsBf is also trained on the previously mentioned multi-channel two speaker separation dataset. Table 3 shows the results of different speaker separation methods on this dataset. From Table 3, it is observed that as the overlap ratio between two speakers increases, the separation performances of all systems consistently decrease, which shows that the overlap ratio has a significant influence on speaker separation performance. However, the PsBf still obtains acceptable separation performance. In particular, under the extreme condition wherein the overlap ratio between two speakers exceeds 75%, the proposed system improves Si-SNR from 8.50 dB to 10.74 dB compared to TD-GWF-TasNet. At the same time, the greater the average angle of the speaker with respect to the microphone center, the better the separation performance. For instance, from $<15^\circ$ to $>90^\circ$, all models have significant performance improvements, regardless of single-channel or multi-channel models (e.g., TasNet-filter: 8.03 vs. 8.94, TD-GWF-TasNet: 10.70 vs. 13.60 and PsBf: 13.83 vs. 15.30). This is very reasonable, as a large angle can facilitate spatial differentiation ability between different sources. Furthermore, the proposed method outperforms all baselines by a large margin under all conditions, i.e., the speaker angle and overlap ratio, which demonstrates that the proposed PsBf can effectually recover speech with minimal distortion. For instance, compared with FasNet-TAC, PsBf obtains 3.46 dB average improvement in terms of SI-SDR. Finally, although PsBf obtains the best separation performance, it has fewer parameters than other baseline models except for IC-ConvTasNet, which verifies the higher parameter effectiveness of the model.

Table 3. Experiment results on the simulated 6-mic circular array. The angle indicates the average angle of the speaker with respect to the microphone center. The percentages denote the overlap ratios between two speakers. SI-SNR is reported on a decibel scale. The best results are shown using ***bold*** values.

Metrics	Par. (M)	Speaker Angle				Overlap Ratio				Avg.
		$<15^\circ$	15–45°	45–90°	$>90^\circ$	<25%	25–50%	50–75%	>75%	
TasNet-filter	2.9	8.03	8.35	8.71	8.94	13.2	9.6	6.85	4.39	8.51
FasNet+TAC	2.9	8.63	10.65	12.21	13.04	15.2	12.0	9.65	7.59	11.11
DCRN	18.67	9.23	9.61	9.84	10.13	14.34	10.74	7.80	5.76	9.66
MC-ConvTasNet	5.09	8.47	8.89	9.31	10.06	13.03	9.90	7.76	6.02	9.18
TD-GWF-TasNet	2.6	10.70	11.90	12.90	13.60	16.30	13.30	11.00	8.50	12.30
IC-ConvTasNet	1.74	10.27	11.68	12.45	12.54	16.44	13.19	10.67	6.65	11.74
TPRNN	2.28	11.16	13.24	14.40	15.16	17.32	14.27	12.08	10.24	13.48
PsBf	**2.27**	**13.83**	**14.51**	**14.63**	**15.30**	**18.52**	**15.55**	**13.46**	**10.74**	**14.57**

6. Conclusions

A pre-separation and all-neural beamformer framework for microphone array speech separation is proposed in this paper. It is composed of two modules, namely the pre-separation module and the all-neural beamforming module. The pre-separation module is used to obtain pre-separated speech and interference, and the all-neural beamforming module utilizes GRU and FC to calculate frame-level beamforming weights. Experiments on multi-channel speech enhancement and speaker separation tasks are performed, and the evaluation results show that the proposed method significantly outperforms other baseline models. Furthermore, this approach can be used for symmetrical stereo speech. Future

work will focus on making sure that PsBf can perform denoising and dereverberation at the same time, and trying to optimize PsBf with ASR to improve robustness jointly.

Author Contributions: W.X. and X.X. designed and programmed the proposed approach and wrote the paper. X.Z. and G.L. participated in algorighm design. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the required data are cited in the manuscript.

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

References

1. 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. [CrossRef] [PubMed]
2. Xiang, X.; Zhang, X.; Chen, H. A Convolutional Network With Multi-Scale and Attention Mechanisms for End-to-End Single-Channel Speech Enhancement. *IEEE Signal Process. Lett.* **2021**, *28*, 1455–1459. [CrossRef]
3. Pandey, A.; Xu, B.; Kumar, A.; Donley, J.; Calamia, P.; Wang, D. TPARN: Triple-path Attentive Recurrent Network for Time-domain Multichannel Speech Enhancement. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6497–6501. [CrossRef]
4. Wang, Y.; Narayanan, A.; Wang, D. On Training Targets for Supervised Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef] [PubMed]
5. Li, A.; Liu, W.; Zheng, C.; Fan, C.; Li, X. Two Heads are Better Than One: A Two-Stage Complex Spectral Mapping Approach for Monaural Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1829–1843. [CrossRef]
6. Liu, Y.; Wang, D. Divide and Conquer: A Deep CASA Approach to Talker-Independent Monaural Speaker Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 2092–2102. [CrossRef]
7. Liu, Y.; Wang, D. Causal Deep CASA for Monaural Talker-Independent Speaker Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2109–2118. [CrossRef]
8. Li, A.; Liu, W.; Zheng, C.; Li, X. Embedding and Beamforming: All-neural Causal Beamformer for Multichannel Speech Enhancement. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6487–6491. [CrossRef]
9. Affes, S.; Grenier, Y. A signal subspace tracking algorithm for microphone array processing of speech. *IEEE Trans. Speech Audio Process.* **1997**, *5*, 425–437. [CrossRef]
10. Gannot, S.; Burshtein, D.; Weinstein, E. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Trans. Signal Process.* **2001**, *49*, 1614–1626. [CrossRef]
11. Erdogan, H.; Hershey, J.; Watanabe, S.; Mandel, M.I.; Roux, J.L. Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks. In Proceedings of the Interspeech 2016, San Francisco, CA, USA, 8–12 September 2016; pp. 1981–1985. [CrossRef]
12. 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), New Orleans, LA, USA, 5–9 March 2017; pp. 3246–3250. [CrossRef]
13. Qian, K.; Zhang, Y.; Chang, S.; Yang, X.; Florencio, D.; Hasegawa-Johnson, M. Deep Learning Based Speech Beamforming. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5389–5393. [CrossRef]
14. Wang, Z.; Wang, P.; Wang, D. Complex Spectral Mapping for Single- and Multi-Channel Speech Enhancement and Robust ASR. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1778–1787. [CrossRef]
15. Gu, R.; Zhang, S.; Chen, L.; Xu, Y.; Yu, M.; Su, D.; Zou, Y.; Yu, D. Enhancing End-to-End Multi-Channel Speech Separation Via Spatial Feature Learning. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 7319–7323. [CrossRef]
16. Wang, Z.; Wang, D. Combining Spectral and Spatial Features for Deep Learning Based Blind Speaker Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 457–468. [CrossRef]
17. Zhang, J.; Zorilă, C.; Doddipatla, R.; Barker, J. On End-to-end Multi-channel Time Domain Speech Separation in Reverberant Environments. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6389–6393. [CrossRef]
18. Tan, K.; Xu, Y.; Zhang, S.X.; Yu, M.; Yu, D. Audio-Visual Speech Separation and Dereverberation With a Two-Stage Multimodal Network. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 542–553. [CrossRef]

19. Wang, Z.; Wang, P.; Wang, D. Multi-microphone Complex Spectral Mapping for Utterance-wise and Continuous Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2001–2014. [CrossRef] [PubMed]
20. Tan, K.; Zhang, X.; Wang, D. Deep Learning Based Real-Time Speech Enhancement for Dual-Microphone Mobile Phones. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1853–1863. [CrossRef] [PubMed]
21. Tan, K.; Wang, Z.; Wang, D. Neural Spectrospatial Filtering. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 605–621. [CrossRef]
22. Zhang, Z.; Xu, Y.; Yu, M.; Zhang, S.X.; Chen, L.; Yu, D. ADL-MVDR: All Deep Learning MVDR Beamformer for Target Speech Separation. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 6089–6093. [CrossRef]
23. Xu, Y.; Zhang, Z.; Yu, M.; Zhang, S.X.; Yu, D. Generalized Spatio-Temporal RNN Beamformer for Target Speech Separation. In Proceedings of the Interspeech 2021, Brno, Czech Republic, 30 August–3 September 2021; pp. 3076–3080. [CrossRef]
24. Luo, Y.; Han, C.; Mesgarani, N.; Ceolini, E.; Liu, S.C. FaSNet: Low-Latency Adaptive Beamforming for Multi-Microphone Audio Processing. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 260–267. [CrossRef]
25. Xiang, X.; Zhang, X.; Xie, W. Distributed Microphones Speech Separation by Learning Spatial Information With Recurrent Neural Network. *IEEE Signal Process. Lett.* **2022**, *29*, 1541–1545. [CrossRef]
26. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1026–1034. [CrossRef]
28. Ballas, N.; Yao, L.; Pal, C.; Courville, A. Delving Deeper into Convolutional Networks for Learning Video Representations. *arXiv* **2015**, arXiv:1511.06432.
29. Pandey, A.; Wang, D. Dense CNN With Self-Attention for Time-Domain Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1270–1279. [CrossRef]
30. Kolbæk, M.; Yu, D.; Tan, Z.H.; Jensen, J. Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1901–1913. [CrossRef]
31. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210. [CrossRef]
32. Reddy, C.K.A.; Gopal, V.; Cutler, R.; Beyrami, E.; CHENG, R.; Dubey, H.; Matusevych, S.; Aichner, R.; Aazami, A.; Braun, S.; et al. The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results. In Proceedings of the Interspeech 2020, Shanghai, China, 25–29 October 2020; pp. 2492–2496. [CrossRef]
33. Varga, A.; Steeneken, H.J.M. Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. *Speech Commun.* **1993**, *12*, 247–251. [CrossRef]
34. Luo, Y.; Han, C.; Mesgarani, N. Group Communication With Context Codec for Lightweight Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 1752–1761. [CrossRef]
35. Kingma, D.P.; Ba, J.L. Adam: A method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015.
36. Tan, K.; Wang, D. Learning Complex Spectral Mapping With Gated Convolutional Recurrent Networks for Monaural Speech Enhancement. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 380–390. [CrossRef] [PubMed]
37. Xiang, X.; Zhang, X.; Chen, H. A Nested U-Net With Self-Attention and Dense Connectivity for Monaural Speech Enhancement. *IEEE Signal Process. Lett.* **2022**, *29*, 105–109. [CrossRef]
38. Luo, Y.; Chen, Z.; Mesgarani, N.; Yoshioka, T. End-to-end Microphone Permutation and Number Invariant Multi-channel Speech Separation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6394–6398. [CrossRef]
39. Wang, Z.; Wang, D. Multi-Microphone Complex Spectral Mapping for Speech Dereverberation. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 486–490. [CrossRef]
40. Luo, Y. A Time-domain Generalized Wiener Filter for Multi-channel Speech Separation. *arXiv* **2021**, arXiv:2112.03533.
41. Lee, D.; Kim, S.; Choi, J.W. Inter-channel Conv-TasNet for Multichannel Speech Enhancement. *arXiv* **2021**, arXiv:2111.04312.
42. Rix, A.W.; Beerends, J.G.; Hollier, M.P.; Hekstra, A.P. Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs. In Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752. [CrossRef]
43. Taal, C.H.; Hendriks, R.C.; Heusdens, R.; Jensen, J. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *19*, 2125–2136. [CrossRef]