symmetry

MDPI

Article

A Neural Beamspace-Domain Filter for Real-Time Multi-Channel Speech Enhancement

Wenzhe Liu [1], Andong Li [1], Xiao Wang [2,3,*], Minmin Yuan [3,4], Yi Chen [2,3], Chengshi Zheng [1,3] and Xiaodong Li [1]

[1] Key Laboratory of Noise and Vibration Research, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China; liuwenzhe@mail.ioa.ac.cn (W.L.); liandong@mail.ioa.ac.cn (A.L.); cszheng@mail.ioa.ac.cn (C.Z.); lxd@mail.ioa.ac.cn (X.L.)
[2] School of Humanities and Management, Southwest Medical University, Luzhou 646099, China; chenyi@swmu.edu.cn
[3] National Environmental Protection Engineering and Technology Center for Road Traffic Noise Control, Beijing 100088, China; mm.yuan@rioh.cn
[4] Research Institute of Highway Ministry of Transport, Beijing 100088, China
* Correspondence: wangxiao@swmu.edu.cn

Abstract: Most deep-learning-based multi-channel speech enhancement methods focus on designing a set of beamforming coefficients, to directly filter the low signal-to-noise ratio signals received by microphones, which hinders the performance of these approaches. To handle these problems, this paper designs a causal neural filter that fully exploits the spectro-temporal-spatial information in the beamspace domain. Specifically, multiple beams are designed to steer towards all directions, using a parameterized super-directive beamformer in the first stage. After that, a deep-learning-based filter is learned by, simultaneously, modeling the spectro-temporal-spatial discriminability of the speech and the interference, so as to extract the desired speech, coarsely, in the second stage. Finally, to further suppress the interference components, especially at low frequencies, a residual estimation module is adopted, to refine the output of the second stage. Experimental results demonstrate that the proposed approach outperforms many state-of-the-art (SOTA) multi-channel methods, on the generated multi-channel speech dataset based on the DNS-Challenge dataset.

Keywords: multi-channel speech enhancement; neural beam filter; deep learning

check for updates

Citation: Liu, W.; Li, A.; Wang, X.; Yuan, M.; Chen Y.; Zheng, C.; Li, X. A Neural Beamspace-Domain Filter for Real-Time Multi-Channel Speech Enhancement. *Symmetry* **2022**, *14*, 1081. https://doi.org/10.3390/sym14061081

Academic Editor: Gianluca Vinti

Received: 23 April 2022
Accepted: 20 May 2022
Published: 24 May 2022

1. Introduction

In the real world, speech is often corrupted by noise and/or reverberation. Speech enhancement aims to extract the clean speech and suppress the noise and reverberation components, which is one of the core problems in audio signal processing. It is reported that multi-channel speech enhancement (MCSE) tends to have superior performance, when compared with monaural speech enhancement, owing to the additional spatial information [1]. Therefore, multi-channel speech enhancement has been widely applied as a preprocessor in video conferencing systems, automatic speech recognition (ASR) systems, and smart TVs. In the past forty years, several beamforming-based [2] and blind-source-separation-based [3] methods have been developed. The deep neural networks (DNNs) are the artificial neural networks (ANNs), with multiple hidden layers between the input and output layers. With the help of their strong nonlinear modeling ability, DNNs have been, widely, used in a variety of audio tasks, such as emotion recognition, ASR, and speech enhancement/separation. Recently, DNNs have facilitated the research in MCSE, yielding notable performance improvements over conventional statistical beamforming techniques [4–11].

Considering the success of DNNs in the single-channel speech enhancement (SCSE) area, a straightforward strategy is to extend the previous SCSE models to extract spatial features, either heuristically or implicitly [4–9]. This paradigm is prone to cause nonlinear speech distortion, such as spectral blackholes in low signal-to-noise (SNR) scenarios, since the advantage of the spatial filter with microphone-array beamforming is not fully exploited to null the directional interference and suppress the ambient noise [10,11]. Another category follows the cascade-style regime. To be specific, in the first stage, an SC-based network was adopted to predict the mask of each acoustic channel in parallel, followed by the steering vector estimation and noise spatial covariance matrix (SCM) calculation. In the second stage, a traditional beamformer, such as minimum variance distortionless response (MVDR) or eigenvalue decomposition (GEV), was adopted for spatial filtering [10,12–14]. These methods have shown their effectiveness in ASR, since ASR can tolerate a latency of hundreds of milliseconds. When the latency should be much lower, such as no more than 20 ms [15] for many practical applications, such as speech communication, hearing aids, and transparency, these methods may degrade their performance, significantly, for these low-latency systems. Moreover, the performance heavily depends on the mask estimation accuracy, which can degrade a lot in complex acoustic scenarios.

As a solution, an intuitive tactic is to enforce the network to directly predict the beamforming weights, which can be done in either the time domain [16,17] or the frequency domain [11,18–20]. Nonetheless, according to the signal theory, the desired beam pattern is required to form its main beam towards the target direction and, meanwhile, form the null towards the interference direction, which tends to be difficult, especially, in low-SNR scenarios, from the optimization perspective. Moreover, slight errors of the estimated weights are able to lead to severe distortions in the beam pattern and, thus, affect the performance of the algorithm.

In this paper, we design a neural filter in the beamspace domain, rather than the spatial domain, for real-time multi-channel speech enhancement. In detail, the multi-channel signals are, first, processed by a set of pre-defined fixed beamformers. A beam set is sampled, uniformly, with various directions in the space. Then, the network is utilized to learn the spectro-temporal-spatial discriminative features of the target speech and noise, which aims to generate the bin-level filtering coefficients to, automatically, weight the beam set. Note, different from the previous neural beamformer-based literature [8,11], where the output weights are applied to multi-channel input signals directly, here the predicted coefficients are to filter the noise component of each pre-generated beam and fuse them. We dub it a neural beamspace-domain filter, to distinguish it from the existing neural beamformer, literally. The rationale of such network design logic is three-fold.

- The target signal can be pre-extracted with the fixed beamformer, and the dominant part should exist within at least one directional beam, serving as the SNR-improved target priori to guide the subsequent beam fusion process. The interference-dominant beam can be obtained, when the beam steers towards the interference direction, providing the interference priori for better distinguishment in a spatial-spectral sense. Besides, the target and interference components may co-exist within each beam, while their distributions are dynamically changed, due to their spectral difference. Therefore, the beam set can be viewed as a reasonable candidate to indicate the spectral and spatial characteristics.
- In addition to the design of beam pattern in the spatial domain, the proposed system can, also, learn the spectral characteristics of the interference components, to cancel residual noise in the spectral domain, completing the enhancement of both the spatial domain and the spectral domain, which can achieve a higher upper limit of performance than the neural spatial network that only performs filtering in the spatial domain.
- From the optimization standpoint, the small error in the beamforming weights may lead to serious distortion of the beam pattern, while the beamspace-domain weights will only leak some undesired components when the error occurs, which has much less direct impact on the performance of the system. Therefore, the beamspace-domain filter is more robust.

As the beam set is only discretely sampled in the space, the information loss tends to arise due to the limited spatial resolution at low frequencies, which causes speech distortion. To this end, a residual branch is designed to refine the fused beam. We have to emphasize that, although the multi-beam concept is used in both [21] and this study, they are very different, as [21] is in essence a parallel single beam enhancement process, while the proposed system can be regarded as the filter and fusion process of the multi-beam. Experiments conducted on the DNS-Challenge corpus [22] show that the proposed neural beam filter outperforms previous state-of-the-art (SOTA) baselines.

Our main contributions are summarized as follows:

- We propose a novel multi-channel speech enhancement scheme in the beam-space domain. To the best of our knowledge, this is the first work that shows the effectiveness of the neural beamspace-domain filter for multi-channel speech enhancement.
- We introduce the residual U-Net into the convolutional encoder-decoder architecture, to improve the feature representation capability. A weight estimator module is designed, to predict the time-frequency bin-level filter coefficients, and a residual refinement module is designed to refine the estimated spectrum.
- We validate the superiority of the proposed framework, by comparing it with state-of-the-art algorithms in both the directional interference and diffuse noise scenarios. These evaluation results demonstrate the superiority and potentiality of the proposed method.

The remainder of the paper is organized as follows. We describe the proposed neural beam filter in Section 2. The experimental setting and results are given in Section 3 and Section 4, respectively. Finally, we draw some conclusions in Section 5.

2. Materials and Methods

The aim of this work is to develop a real-time multi-channel speech enhancement system, to extract the clean speech and suppress the noise and reverberant components. The noisy mixtures are recorded by the microphones of an array. The spectra of these signals are used as the inputs of the proposed system. This system comprises three modules, namely fixed beamforming module (FBM), beam filtering module (BFM), and residual refinement module (RRM). The enhanced speech is, then, obtained and transmitted to the telecommunication circuit and/or speech recognition system. The proposed system is presented in Figure 1.

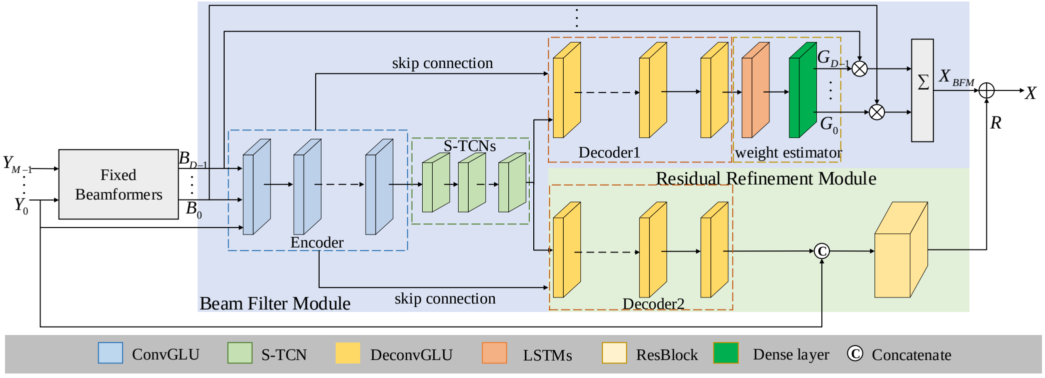


Figure 1. Overview of the proposed framework. Different modules are highlighted with different colors.

2.1. Signal Model

Considering an M-channel microphone array placed in noisy-reverberant environments, the signal received in the m-th microphone can be represented by:

$$y_m(n) = h_m(n) * s(n) + v_m(n), m = 0, 1, \cdots, M-1, \tag{1}$$

where

$Y_m(t,f)$ is the STFT of $y_m(n)$.
$S(t,f)$ is the STFT of $s(n)$.
$V_m(t,f)$ is the STFT of $v_m(n)$.
$t \in \{0,1,\cdots,T-1\}$ refers to the index of frames.
$f \in \{0,1,\cdots,F-1\}$ refers to the index of frequency bins. Considering the symmetry of $Y_m(t,f)$ in frequency, $F = N/2+1$ is chosen throughout this paper.
$X_m^{early}(t,f)$ is the direct-path signal of the speech source and its early reflections.
$X_m^{late}(t,f)$ is the late reverberant speech.

By N-point short-time Fourier transform (STFT), the physical model in the time-frequency domain can be expressed as:

$$Y_m(t,f) = H_m(t,f)S(t,f) + V_m(t,f) = X_m^{early}(t,f) + X_m^{late}(t,f) + V_m(t,f), \tag{2}$$

where

$$X_m^{early}(t,f) = H_m^{early}(t,f)S(t,f), \tag{3}$$

$$X_m^{late} = H_m^{late}(t,f)S(t,f), \tag{4}$$

where

$Y_m(t,f)$ is the STFT of $y_m(n)$.
$S(t,f)$ is the STFT of $s(n)$.
$V_m(t,f)$ is the STFT of $v_m(n)$.
$t \in \{0,1,\cdots,T-1\}$ refers to the index of frames.
$f \in \{0,1,\cdots,F-1\}$ refers to the index of frequency bins. Considering the symmetry of $Y_m(t,f)$ in frequency, $F = N/2+1$ is chosen throughout this paper.
$X_m^{early}(t,f)$ is the direct-path signal of the speech source and its early reflections.
$X_m^{late}(t,f)$ is the late reverberant speech.

In this paper, the aim of the proposed algorithm is to extract the direct-path, plus the early reflected components $X(t,f) = X_{ref}^{early}(t,f)$, from the multi-channel input signals $\mathbf{Y}(t,f) = \{Y_0(t,f),\cdots,Y_{M-1}(t,f)\}$, by the model $\mathcal{F}(\cdot)$, assuming that the 0-th microphone is chosen as the reference microphone, and defining the reflections within the first 100 ms after the direct sound as the early reverberation. From now on, we will omit the subscript (t,f), when no confusion arises. The above process can be formulated as:

$$\hat{X} = \mathcal{F}(\mathbf{Y};\Phi). \tag{5}$$

where

Φ denotes the parameter set of the mapping function $\mathcal{F}(\cdot)$.

After transforming $\hat{X}$ by inverse STFT (iSTFT), the enhanced time-domain signal can be reconstructed by the overlap-add (OLA) method.

2.2. Forward Stream

Figure 1 shows the overall diagram of the proposed architecture, which consists of three components, namely fixed beamforming module (FBM), beam filtering module (BFM), and residual refinement module (RRM).

In FBM, the fixed beamformer is employed to sample the space uniformly and obtain multiple beams steering towards different directions. The beam set denotes $B_d \in \mathbb{C}^{T\times F}$, with $d = 0, \cdots, D-1$, where D denotes the number of resultant multi-beam. The process is, thus, given by:

$$\{B_d\}_{d=0,\cdots,D-1} = \mathcal{F}_{FBM}(\{Y_m\}_{m=0,\cdots,M-1}; \Phi_{FBM}), \tag{6}$$

where

$\mathcal{F}_{FBF}(\cdot)$ is the function of FBM.

Φ_{FBF} denotes the parameter set.

We concatenate the beam set along the channel dimension, serving as the input of BFM, which denotes $\mathbf{B} = Cat(B_0, \cdots, B_{D-1}) \in \mathbb{R}^{2D \times T \times F}$. Here, 2 means that both real and imaginary (RI) parts are considered. As muti-beams can represent both spectral and spatial characteristics, BFM is adopted to learn the spectro-temporal-spatial discriminative information between speech and interference and attempt to assign the filter weights $\widehat{G}_d \in \mathbb{C}^{T \times F}$ for each beam. It is worth noting that as the beam set is discretely sampled in the space, the information loss tends to arise due to the limited spatial resolution. To alleviate this problem, the complex spectrum of the reference channel is, also, incorporated into the input and, meanwhile, similar to [23], the complex residual needs to be estimated with RRM, which aims to compensate for the inherent information loss of the filtered spectrum. This process can be presented as:

$$\widehat{\mathbf{G}} = \mathcal{F}_{BFM}([\{B_d\}_{d=0,\cdots,D-1}, Y_0]; \Phi_{BFM}), \tag{7}$$

$$\widehat{\mathbf{R}} = \mathcal{F}_{RRM}([\{B_d\}_{d=0,\cdots,D-1}, Y_0]; \Phi_{RRM}), \tag{8}$$

where

$\widehat{\mathbf{G}} \in \mathbb{C}^{D \times T \times F}$ is the complex filter estimated by BFM.

$\widehat{\mathbf{R}} \in \mathbb{C}^{D \times T \times F}$ is the complex residual estimated by RRM.

By applying the estimated weights $\{\widehat{G}_d\}_{d=0,\cdots,D-1}$ to filter the beams $\{B_d\}_{d=0,\cdots,D-1}$ and, then, summing them along the channel axis, the fused beam $\widehat{X}_{BFM}$ can be obtained by:

$$\widehat{X}_{BFM} = \sum_d \widehat{G}_d \times B_d, \tag{9}$$

where × denotes the complex-valued multiplication operator. We, then, add the filtered beam and estimated complex residual together, to obtain the final output $\widehat{X}$, i.e.,

$$\widehat{X} = \widehat{X}_{BFM} + \widehat{R}. \tag{10}$$

2.3. Fixed Beamforming Module

In this module, the fixed beamformer is leveraged to transform input multi-channel mixtures into several beams, which steer towards different-looking directions and, uniformly, sample the space. As the fixed beamformer is data-independent, it is robust in adverse environments and has low computational complexity. Moreover, filtering multi-channel mixtures with the fixed beamformer allows our system to be less sensitive to the array geometry. In this paper, we choose the super-directivity (SD) beamformer as the default beamformer, due to its promising performance in high directivity [24]. Note that other fixed beamformers can, also, be adopted, which is out of the scope of the paper. Assuming the target directional angle is θ_d, the weights of the SD beamformer can be calculated as:

$$\mathbf{w}_d(f) = \frac{\mathbf{\Gamma}_{nn}^{-1}(f)\mathbf{v}(\theta_d, f)}{\mathbf{v}^H(\theta_d, f)\mathbf{\Gamma}_{nn}^{-1}(f)\mathbf{v}(\theta_d, f)}, \tag{11}$$

where

$\mathbf{v}(\theta_d, f)$ is the steering vector.

$(\cdot)^H$ is the complex transpose operator.
$\mathbf{\Gamma}_{nn}(f)$ denotes the covariance matrix of a diffuse noise field with the diagonal loading to control the white noise gain.

Note that the diagonal-loading level, often, needs to be chosen carefully, to make a good balance between the white noise gain and the array gain [25]. In this paper, the diagonal loading level is fixed to 1×10^{-5}, and its impact on performance will be studied in the near future. The (i, j)-th element of $\mathbf{\Gamma}_{nn}(f)$ represents the coherence between the signals received by two microphones, with indices i and j in an isotropic diffuse field, which can be formulated as:

$$\mathbf{\Gamma}_{nn}^{(i,j)}(f) = \text{sinc}\left(\frac{2\pi f_s f l_{ij}/N}{c}\right), \tag{12}$$

where

$\text{sinc}(x) = \frac{\sin(x)}{x}$.
l_{ij} is the distance between the i-th and j-th microphones.
c is the speed of sound.
f_s is the sampling rate.

Defining $\mathbf{Y}(t, f) = \{Y_0(t, f), \cdots, Y_{M-1}(t, f)\}$, the output of the d-th SD beamformer can be expressed as:

$$B_d(t, f) = \mathbf{w}_d^H(f)\mathbf{Y}(t, f). \tag{13}$$

2.4. Beam Filter Module

As shown in Figure 1, the beam filter module (BFM) consists of a causal convolutional encoder-decoder (CED) architecture and a weight estimator (WE). For the encoder, it comprises six gated linear units with residual U-Net (GLU-RSU) blocks, to consecutively halve the feature size and extract high-level features, which is described in Section 2.4.2. The decoder is the mirror version of the encoder except that all the convolution operations are replaced by the deconvolutional version (dubbed DeconvGLU). Similar to [26], a stack of squeezed temporal convolutional networks (S-TCNs) is inserted as the bottleneck of CED, to model the temporal correlations among adjacent frames. After that, in the weight estimator, we simulate the filter generation process, where T-F bin-level filter coefficients are assigned for each beam. To be specific, the output embedding tensor of the decoder is, first, normalized by layer normalization (LN), and, then, the LSTM is employed to update the feature frame by frame, with ReLU serving as the intermediate nonlinear activation function. The weights $\widehat{\mathbf{G}}$ are obtained after the output linear layer. Then, these weights are applied to each beam to obtain the target beam.

2.4.1. CED Architecture

Convolutional encoder–decoder architecture is widely used in speech enhancement [27]. It consists of a convolutional encoder, followed by a corresponding decoder. The encoder is a stack of convolutional layers, and the encoder is a stack of deconvolutional layers in the reverse order. The convolution layer uses a filter, namely kernel, to extract the local patterns of the low-level input feature to the high-level embedding. It is widely used in computer vision [28], neural language processing [29], and acoustic signal processing [30,31]. The deconvolution layer is a special convolution layer, which can map low-resolution features to the features with the input feature size. The symmetric CED structure ensures that the output has the same shape as the input, which is, naturally, suitable for the speech enhancement task.

2.4.2. GLU-RSU Block

The GLU-RSU block consists of convolutional gated linear units (ConvGLUs) [32], batch normalization (BN), Parameter ReLU (PReLU), and a residual U-Net (RSU) [33], which is shown in Figure 2.

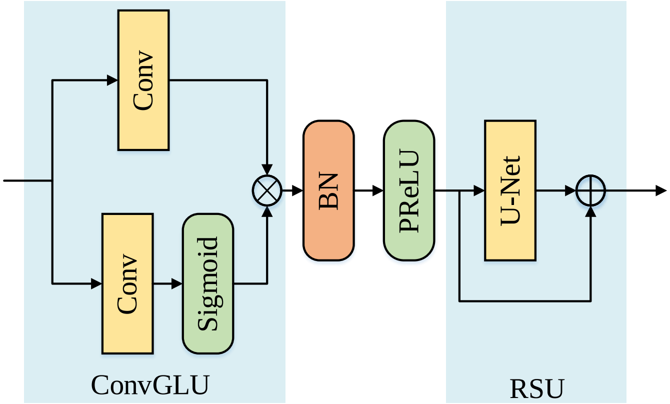


Figure 2. The architecture of GLU-RSU.

Firstly, the input feature $\mathbf{f}_i^{GLURSU}$ is passed by a ConvGLU, which can obtain better modeling capacity than a plain convolutional layer, due to the learnable dynamic feature selection by a gating mechanism, which can be expressed as:

$$\mathbf{f}_o^{GLU} = (\mathbf{f}_i^{GLURSU} * \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{f}_i^{GLURSU} * \mathbf{W}_2 + \mathbf{b}_2), \tag{14}$$

where

$*$ is the convolution operator.
$\odot$ is the Hadamard product operator.
$\mathbf{W}_1$ and $\mathbf{W}_2$ are the weights of these two convolutional layers.
$\mathbf{b}_1$ and $\mathbf{b}_2$ are the bias of these two convolutional layers.
$\sigma(\cdot)$ is the Sigmoid function.

Then, the U-Net in the RSU is used to recalibrate feature distribution, by modeling the spectrum feature in different scales and extracting intra-beam time-frequency discrimination by continuous downsampling. Finally, a residual connection is utilized, to mitigate the gradient-vanishing problem. The above process can be formulated as:

$$\mathbf{f}_o^{GLURSU} = \mathcal{F}_{UNet}(\mathbf{f}_o^{GLU}; \Phi_{UNet}) + \mathbf{f}_o^{GLU}. \tag{15}$$

where

$\mathcal{F}_{UNet}(\cdot)$ is the function of U-Net.
Φ_{UNet} denotes the parameter set.

2.4.3. Squeezed Temporal Convolutional Network

TCN is used to effectively capture the temporal dependence of speech. Compared with recurrent neural network (RNN), TCN is able to interfere in parallel and achieve better performance, by utilizing 1-D dilated convolutions. S-TCN is a lightweight TCN and consists of several squeezed temporal convolutional modules (S-TCM). From Figure 3, one can see that S-TCM includes the input point convolution, the gated depth-wise dilated convolution (GDD-Conv), and the output point convolution, where the input point convolution and the output point convolution are applied to squeeze and restore the feature dimension, respectively, and GDD-Conv has three differences with the depth-wise dilated convolution in traditional TCM. Firstly, the channel of the dilated causal convolution (DC-Conv) in GDD-Conv is less to effectively represent the information, due to the time-frequency sparseness of the speech spectrum. Moreover, GDD-Conv introduces a gating branch to facilitate information flow in the gradient back-propagation process. The gating branch utilizes the Sigmoid activation function, to map the output of DC-Conv to $(0, 1)$ for changing the feature distribution of the main branch. Note that PReLU and normalization layers are inserted between adjacent convolutional layers, to facilitate the network convergence.

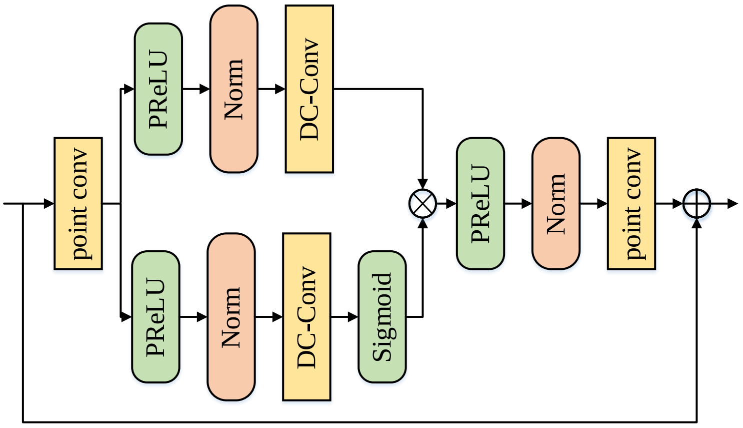


Figure 3. The architecture of the Squeezed Temporal Convolutional Network.

2.5. Residual Refinement Module

Since the SD beamformer tends to amplify the white noise to ensure the array gain at low frequencies, the weighted beam output with BFM, often, contains lots of residual noise components, which need to be further suppressed to improve speech quality. Meanwhile, speech distortion is, often, introduced, due to the mismatch between the main beam steering toward the predefined direction and the true direction of the target speech, which is because the number of the fixed beamformers is limited. To refine the target beam, a residual refinement module (RRM) is proposed, which comprises a decoder module similar to that of BFM and a residual block (ResBlock) containing three residual convolution modules, as shown in Figure 1. The output of S-TCN serves as the input feature of the RRM. After decoding from multiple U^2 blocks, the output tensor $\mathbf{f}^{Res} \in \mathbb{R}^{64\times T\times F}$ is concatenated, with the original complex spectrum of the reference microphone Y_0, and is fed to a point convolution to squeeze the feature dimension to 16. Then, a series of residual convolution modules is applied which comprises a plain convolution layer with a 2×3 kernel and 1×1 stride, BN, PReLU, and a identity shortcut connection. Finally, the complex residual spectrum $\widehat{\mathbf{R}}$ is derived by the output 1×1-conv, to reduce the dimensions to 2, and applied to refine the filtered beam output $\widehat{X}_{BFM}$.

2.6. Loss Function

In this paper, a two-stage training method is used. Firstly, we train the BFM with the magnitude regularized complex spectrum loss, which is defined as:

$$\mathcal{L}_1 = \alpha \sum_{t,f} |X_{ref}^{early}(t,f) - \widehat{X}_{BFM}(t,f)|^2 + (1-\alpha) \sum_{t,f} ||X_{ref}^{early}(t,f)| - |\widehat{X}_{BFM}(t,f)||^2, \tag{16}$$

where α is the regularized factor and is set to 0.5, empirically. The first term and the second term in the loss function are, respectively, the complex spectrum mean squared error (MSE) loss and the magnitude spectrum MSE loss.

Then, we freeze the parameters of BFM when training RRM. The same loss function is utilized:

$$\mathcal{L}_2 = \alpha \sum_{t,f} |X_{ref}^{early}(t,f) - \widehat{X}_{ref}^{early}(t,f)|^2 + (1-\alpha) \sum_{t,f} ||X_{ref}^{early}(t,f)| - |\widehat{X}_{ref}^{early}(t,f)||^2. \tag{17}$$

Note that both the estimations and the targets are adopted with the power compression, to improve the speech enhancement performance. The power-compression process can be expressed as:

$$X^{\beta} = |X|^{\beta}(X/|X|), \tag{18}$$

where the compression factor β is set to 0.5 [34].

2.7. Datasets

In this paper, we conduct two datasets for the performance evaluation in the directional interference situation and the diffused noise situation. The DNS-Challenge corpus (https://github.com/microsoft/DNS-Challenge (accessed on 22 April 2022)) [22] is selected to convolve with multi-channel room impulse responses (RIRs), which represent the transfer function between the sound source and microphones of the array, to generate multi-channel pairs for experiments. To be specific, the clean clips are randomly sampled from the *neutral clean speech* set [22], which includes about 562 h speaking by 11,350 speakers. We split it into two parts without overlap, namely for training and testing. The noise clips in the DNS-Challenge corpus are selected for the directional interference. The utterances in the TIMIT corpus [35] are used to conduct a diffused babble noise field.

For the directional interference situation, around 20,000 types of noise in the DNS-Challenge corpus are selected as the interference source in the training phase, with a duration time of about 55 h [23,26]. For testing, three types of unseen noise are chosen, namely babble, factory1 noises taken from NOISEX92 [36], and cafe noise taken from CHiME3 [37]. We generate the RIRs with the image method [38] using a uniform linear array with 9 microphones, and the distance between two adjacent microphones is around 4 cm. The room size is sampled from $3 \times 3 \times 2.5\ \text{m}^3$ to $10 \times 10 \times 3\ \text{m}^3$, and the reverberation time RT60 ranges from 0.05 s to 0.7 s. The source is randomly located in angle from 0° to 180°, and the distance between the source and the array center ranges from 0.5 m to 3.0 m. The signal-to-noise ratio (SNR) ranges from -6 dB to 6 dB.

For the diffused babble noise situation, we select the utterances from 480 speakers in the TIMIT corpus for training and validation, while the utterances from other speakers are used for generating test diffused noise. In total, 72 different speakers are selected randomly and are assigned to 72 directions $(0°, 5°, \cdots, 355°)$, to simulate diffused babble noise. The SNR ranges from -6 dB to 6 dB with 1 dB interval. The settings of the room size, RT60, and the speaker location are the same as above.

Totally, in each situation, about 80,000 and 4000 multi-channel noisy and reverberant mixtures, respectively, are generated for training and validation. For the testing set, SNR is set to $\{-5, -2, 0, 2, 5\}$ dB, and 150 pairs are generated for each case. Note that the speakers and the room sizes for the test are, also, unseen in both the training and validation sets.

3. Experiments

3.1. Baselines

In this paper, three multi-channel speech enhancement systems, namely MC-Conv-TasNet [39], FaSNet-TAC [16], and MIMO-UNet [20], are chosen as the comparative systems. MC-Conv-TasNet is the multiple-input version of Conv-TasNet, which is one of the most effective time-domain speech enhancement and separation models. FaSNet-TAC is an end-to-end filter-and-sum style time-domain multi-channel speech enhancement system, which can achieve better performance than mask-based beamformers. MIMO-UNet is a frequency-domain neural beamformer, which is the winner of the INTERSPEECH Far-field Multi-Channel Speech Enhancement Challenge for Video Conferencing [40]. Note that all the models are set with causal configuration, that is, no future frames are involved in the calculation and inference of the current frame.

3.2. Experiment Setup

3.2.1. Training Detail

All the utterances are sampled at 16 kHz, and a 32 ms Hann window is utilized, with 50% overlap between adjacent frames. Accordingly, 512-point FFT is utilized, leading to 257-D spectral features. Adam optimizer is applied, with the initial learning rate set to 5×10^{-4}. If validation loss does not decrease for two consecutive epochs, the learning rate will be halved. All models are trained for 60 epochs. The system is completely built with Python 3.6 and PyTorch 1.6.0. We carry out the training procedures on a workstation with Montage(R) Jintide(R) C2460 1 CPU and one TESLA V100 PCIe GPU. More detailed information of the hardware is listed in Table 1.

Table 1. The parameters of the hardware.

Parameters	Description
Architecture	x86-64
CPU op-mode(s)	32-bit, 64-bit
CPU(s)	96
On-line CPU(s) list	0-95
Thread(s) per core	2
Vendor ID	GenuineIntel
CPU family	6
Model	85
Model name	Montage(R) Jintide(R) C2460 1
Stepping	4
CPU MHz	1001.030
CPU max MHz	2101.0000
CPU min MHz	1000.0000
BogoMIPS	4202.12
Memory device(s)	32 GB × 16
GPU	TESLA V100 PCIe 32 GB

3.2.2. Network Detail

Assuming the input feature of the BFM is $\mathbf{B} \in \mathbb{R}^{2D \times T \times F}$, in the encoder and decoder parts, the kernel size and stride of the 2D convolution layers are 2×3 and 1×2, except for the first layer, which are 2×5 and 1×2, respectively. For six GLU-RSU blocks in the encoder, the number of encoding layers in these U-Net blocks Q is set to $\{4, 3, 2, 2, 1, 0\}$, and the number of decoding layers of these U-Net blocks is $\{0, 1, 2, 2, 3, 4\}$. The kernel size of all the convolutional layers in these U-Net blocks is 1×3, and the stride is 1×2. The number of channels remains 64, by default. Three S-TCNs are adopted, each of which consists of 6 S-TCMs, with kernel size and dilation rate being 5 and $\{1, 2, 4, 8, 16, 32\}$, respectively.

After the CED, an embedding is generated and its size is $64 \times T \times F$. In the weight estimator module, two uni-directional LSTM layers, with 64 hidden nodes and two full-connected layers with 64 nodes and $2D$ nodes, are employed to predict time-frequency bin-level weights. In the ResBlock, the kernel size and stride are 2×3 and 1×1, respectively.

4. Results and Discussion

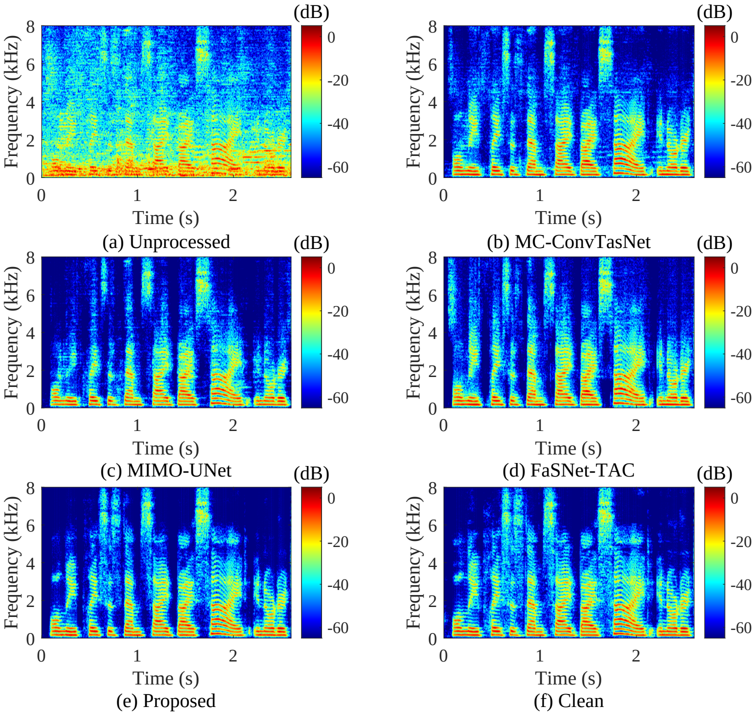
We choose perceptual evaluation speech quality (PESQ) [41] and extended short-time objective intelligibility (ESTOI) [42] as objective metrics, to compare the performance of different models. The PESQ score is used to evaluate speech quality of the enhanced utterance, which is obtained from the clean speech and the enhanced speech. Its value ranges from -0.5 to 4.5. The higher the PESQ score is, the better the speech perceptual quality. The ESTOI score is chosen to evaluate speech intelligibility. The higher the ESTOI score is, the better the speech intelligibility.

4.1. Results Comparison in the Directional Interference Case

The objective results of different SE systems are shown in Tables 2–4. For comparison, the number of beams D is set to 10, which means that the sampling resolution is 20°. We evaluate these systems in terms of PESQ and ESTOI.

From these tables, several observations can be made. First, compared with SCSE-extension-based multi-channel speech enhancement approaches, such as MC-ConvTasNet, end-to-end neural spatial filters, such as FaSNet-TAC and MIMO-UNet, yield notable performance improvements, consistently, thanks to linear filtering of the multi-channel signals, which can reduce speech distortion. For example, compared with MC-ConvTasNet, FaSNet-TAC achieves 0.21 and 2.83% improvements, in terms of PESQ and ESTOI under the cafe noise, and MIMO-UNet gets 0.20 PESQ improvement and 4.96% ESTOI improvement. Second, the proposed system outperforms neural beamforming-based approaches by a large margin, in all cases. For example, compared with FaSNet-TAC, our system achieves 0.61 and 13.63% improvements, in terms of PESQ and ESTOI for cafe noise, respectively. Moreover, our model outperforms MIMO-UNet by 0.62 and 11.50% in PESQ and ESTOI, respectively. This demonstrates the superiority of filtering the beams over the best neural spatial filters, based on frequency and time domains. This is because designing weights for beams is easier to optimize than approximating the desired beam pattern. Moreover, speech- and noise-dominant beams help the network learn their discriminative features. Finally, noise-dominant beams enable the noise characteristic to better cancel the residual noise in speech-dominant beams.

Figure 4 shows the spectrograms of the speech, corrupted by the cafe interference and its processed utterances. One can find that the proposed method has better noise suppression and less speech distortion, when compared with baselines. In particular, the proposed algorithm recovers harmonic structures well and obtains less speech distortion in the high-frequency (around 4 kHz–8 kHz) low-SNR bins, since the fixed beamformer, which has good spatial resolution at high frequencies, is able to better suppress high-frequency interference.



(a) Unprocessed

(b) MC-ConvTasNet

(c) MIMO-UNet

(d) FaSNet-TAC

(e) Proposed

(f) Clean

Figure 4. Spectrogram processed by different methods: Spectrogram of (**a**) a noisy reverberant mixture and enhanced signals processed by (**b**) MC-ConvTasNet, (**c**) MIMO-UNet, (**d**) FaSNet-TAC, and (**e**) the proposed system. (**f**) Spectrogram of clean speech.

Table 2. Objective result comparisons among different causal MCSE models, in terms of PESQ and ESTOI for babble noise. **BOLD** indicates the best score in each case.

Metrics	**PESQ**						**ESTOI (%)**					
	−5 dB	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**	**−5 dB**	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**
Unprocessed	1.38	1.54	1.58	1.70	1.91	1.62	25.81	35.86	40.42	47.51	56.77	41.27
MC-ConvTasNet	1.93	2.18	2.20	2.33	2.41	2.21	56.80	64.42	64.60	68.93	71.99	65.35
MIMO-UNet	1.93	2.18	2.27	2.41	2.57	2.27	54.99	63.73	66.81	71.77	75.74	66.61
FaSNet-TAC	2.03	2.31	2.34	2.47	2.66	2.36	54.64	64.46	66.20	71.03	75.73	66.41
Proposed	**2.52**	**2.87**	**2.96**	**3.10**	**3.30**	**2.95**	**67.98**	**77.03**	**79.68**	**82.79**	**87.36**	**78.97**

Table 3. Objective result comparisons among different causal MCSE models, in terms of PESQ and ESTOI for factory1 noise. **BOLD** indicates the best score in each case.

Metrics	**PESQ**						**ESTOI (%)**					
	−5 dB	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**	**−5 dB**	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**
Unprocessed	1.27	1.35	1.43	1.54	1.76	1.47	24.35	32.89	39.00	44.20	54.32	38.95
MC-ConvTasNet	1.98	2.09	2.24	2.34	2.41	2.21	55.61	59.62	63.49	67.29	68.36	62.87
MIMO-UNet	2.11	2.35	2.51	2.56	2.70	2.45	57.06	65.17	69.45	71.84	75.65	67.83
FaSNet-TAC	2.11	2.23	2.40	2.48	2.63	2.37	55.10	60.87	66.30	69.29	73.32	64.98
Proposed	**2.59**	**2.78**	**2.97**	**3.11**	**3.26**	**2.94**	**67.20**	**73.58**	**78.46**	**81.47**	**85.44**	**77.23**

Table 4. Objective result comparisons among different causal MCSE models, in terms of PESQ and ESTOI for cafe noise. **BOLD** indicates the best score in each case.

Metrics	**PESQ**						**ESTOI (%)**					
	−5 dB	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**	**−5 dB**	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**
Unprocessed	1.38	1.54	1.68	1.78	1.94	1.66	29.91	38.43	45.49	51.73	60.65	45.24
MC-ConvTasNet	2.00	2.13	2.24	2.31	2.41	2.22	56.52	62.84	64.07	67.38	70.94	64.35
MIMO-UNet	2.09	2.34	2.47	2.53	2.69	2.42	58.21	66.97	70.40	73.38	77.59	69.31
FaSNet-TAC	2.14	2.32	2.45	2.57	2.69	2.43	56.97	64.63	66.79	71.56	75.97	67.18
Proposed-10beams	**2.68**	**2.88**	**3.07**	**3.22**	**3.38**	**3.05**	**71.73**	**77.23**	**81.68**	**84.84**	**88.58**	**80.81**

4.2. Results Comparison in the Diffused Babble Noise Case

The evaluation results of different multi-channel speech enhancement models in the diffused babble noise scenario are shown in Table 5.

Table 5. Objective result comparisons among different causal MCSE models, in terms of PESQ and ESTOI for diffused babble noise. **BOLD** indicates the best score in each case.

Metrics	**PESQ**						**ESTOI (%)**					
	−5 dB	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**	**−5 dB**	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**
Unprocessed	1.36	1.49	1.54	1.62	1.78	1.56	26.54	35.27	39.86	45.49	53.87	40.21
MC-ConvTasNet	1.89	2.16	2.28	2.38	2.55	2.25	51.67	59.84	64.02	67.37	72.53	63.08
MIMO-UNet	2.29	2.55	2.66	2.76	2.92	2.63	60.44	68.67	72.29	75.50	79.58	71.30
FaSNet-TAC	2.14	2.35	2.44	2.52	2.65	2.42	56.45	63.35	66.61	69.24	73.73	65.88
Proposed-10beams (w/o U-Net)	2.64	2.89	2.98	3.08	3.24	2.97	69.83	76.83	80.07	82.36	85.97	79.01
Proposed-10beams	**2.73**	**2.97**	**3.07**	**3.17**	**3.32**	**3.05**	**71.93**	**78.43**	**81.69**	**83.76**	**87.10**	**80.58**

It can be seen that the trend of the model performance is similar to that of the directional interference scenario. The neural spatial filters, such as FaSNet-TAC and MIMO-UNet, are, consistently, superior to MC-ConvTasNet. The proposed algorithm significantly outperformed all baseline systems in the PESQ and ESTOI metrics of each SNR. For example, going from FaSNet-TAC to the proposed system, average 0.63 and 14.70% improvements are achieved, in terms of PESQ and ESTOI, respectively. Moreover, it improves the MIMO-UNet baseline by 0.42 PESQ, and 9.28% ESTOI, on average. This demonstrates the superiority of the proposed system, over the best neural spatial filters in the diffused babble noise case.

4.3. Ablation Analysis

We also validate the role of FBM, BFM, and RRM. Table 6 shows the average results of three directional interferences in each case.

Table 6. Objective results of ablation experiments.

Metrics	**PESQ**						**ESTOI (%)**					
	−5 dB	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**	**−5 dB**	**−2 dB**	**0 dB**	**2 dB**	**5 dB**	**Avg.**
Unprocessed	1.34	1.48	1.56	1.67	1.87	1.58	26.69	35.73	41.64	47.81	57.25	41.82
Proposed-10beams (w/o RSU)	2.35	2.64	2.79	2.95	3.14	2.77	63.34	71.86	76.29	80.33	85.14	75.39
Proposed-10beams (w/o RRM)	2.48	2.70	2.85	3.01	3.20	2.85	65.73	73.64	77.94	81.46	85.79	76.91
Proposed-7beams	2.57	2.82	2.97	3.12	3.29	2.95	68.45	75.49	79.52	82.72	86.85	78.61
Proposed-10beams	2.60	2.84	3.00	3.14	3.31	2.98	68.97	75.95	79.94	83.03	87.13	79.00
Proposed-19beams	2.61	2.87	3.01	3.15	3.33	2.99	69.69	76.52	80.17	83.24	87.44	79.41

To analyze the effectiveness of FBM, we set another two candidates of D of 7 (30°) and 19 (10°), where 7 (30°) means $D = 7$, and each main beam width is about 30°; 19 (10°), analogously. It can be seen that the performance of the beam neural filter gradually improves with the increase in D, which reveals the importance of FBM. However, the relative performance improvement decreases as the spatial sampling interval becomes progressively smaller, although there is still a mismatch between beam pointing and source direction, which indicates that the proposed model is robust to direction mismatch, whereas the spatial filter is more sensitive to direction estimation error.

To show the effectiveness of BFM, we visualize the norm of estimated complex weights in Figure 5. The input signals are mixed by a speech radiating from 85° and a Factory1 noise source from 45°. We can find that greater weights are assigned to beams steering toward the surroundings of the target direction, while beams steering to other directions, including those steering toward the interference direction, are given little weights during speaking, while all weights are small in non-speech segments.

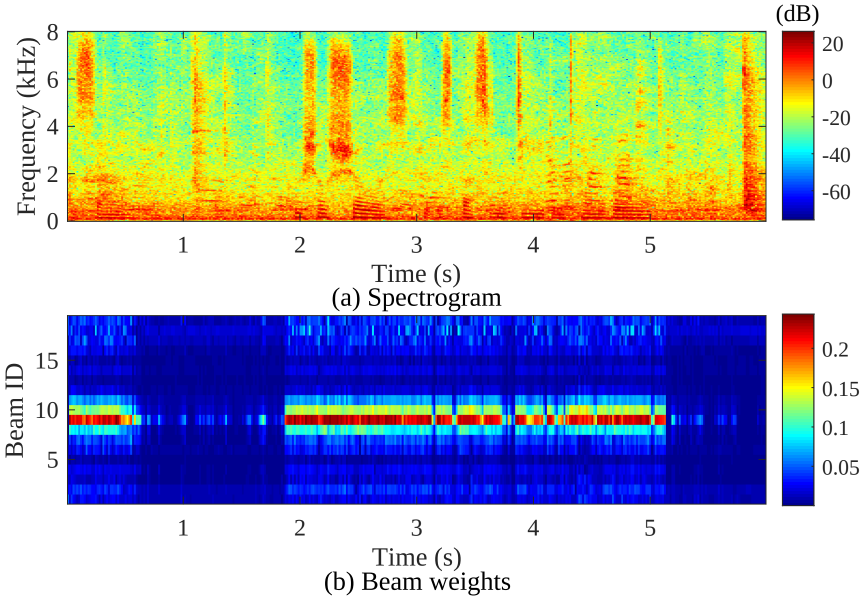


Figure 5. Visualization results of the filter weights, estimated by the proposed system: (**a**) the spectrogram of the signal received by reference microphone, (**b**) visualization results of the norm of the complex weights, estimated by the neural network.

Besides, the proposed system with RRM achieves PESQ improvements of 0.13 and ESTOI improvements of 2.09%. Comparing the visualization results of the model with and without RRM from Figure 6, one can see that the residual noise components are further suppressed at low frequencies, and some missing speech components are recovered, which confirms the effectiveness of RRM in the proposed system.

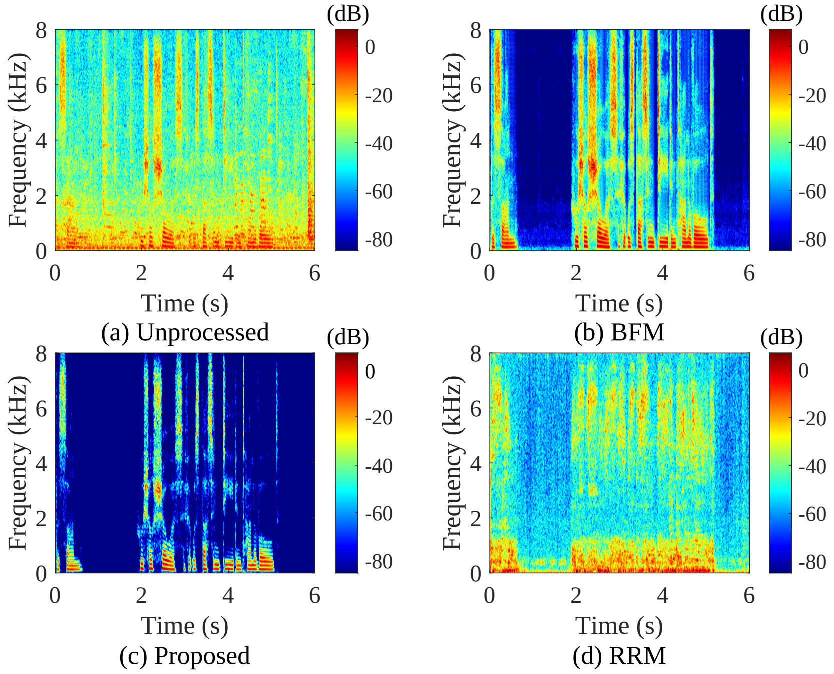


Figure 6. Visualization results of (**a**) unprocessed (SIR = 0 dB), (**b**) BFM, (**c**) proposed-19beams, (**d**) RRM.

Finally, we can find that using RSU, followed by the (De)ConvGLU, can achieve significant performance improvements compared to using (De)ConvGLU only, and achieve 0.21 and 3.53% PESQ and ESTOI average improvements in the cafe interference scenario, demonstrating that U-Net can extract stronger discriminating feature characterizatio,n by modeling multi-beam information at different scales.

5. Conclusions

Speech signals are often distorted by background noise and reverberation in daily listening environments. Such distortions severely degrade speech intelligibility and quality for human hearing, as well as make automatic speech recognition more difficult. In this paper, we propose a causal neural beamspace-domain filter for real-time multi-channel speech enhancement, to recover clean speech from the noisy mixtures received by the microphone array. It comprises three components, namely FBM, BFM, and RRM. Firstly, FBM is adopted, to separate the sources from different directions. Then, BFM maps filter weights, by jointly learning the spectro-temporal-spatial discriminability of speech and interference. Finally, RRM is adopted, to refine the weighting beam output.

From the experimental results, we have the following conclusions:

- The proposed system achieves better speech quality and intelligibility over previous SOTA approaches in the directional interference case.
- In the diffused babble noise scenario, our method, also, achieves better performance than previous systems.
- From the spectrograms of BFM and RRM, one can see that RRM is helpful to refine the missing components of the output of BFM.
- From the ablation study, RSU is able to learn stronger discriminating features to improve the performance.

Video conferencing plays a very crucial part in our daily social interactions, due to the COVID-19 virus. This proposed method can be used to suppress noise and reverberation during a video conference, to improve speech quality and intelligibility. Moreover, it also can be applied to human–machine interaction systems and mobile communication devices.

Future work could concentrate on designing an MCSE system for amultiple speakers scenario, based on the proposed method. Moroever, a more effective feature extraction module of BFM can be explored.

Author Contributions: Conceptualization, W.L. and A.L.; methodology, software, and validation, W.L. and X.W.; writing–original draft preparation, W.L. and A.L.; conceptualization, writing–review, and editing, M.Y., C.Z., and Y.C.; supervision, C.Z. and X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Environmental Protection Engineering and Technology Center for Road Traffic Noise Control, with grant number "F20221018" (create a database about the degree of annoyance of mixed traffic noise among different people during different time periods).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: We wish to confirm that there are no known conflicts of interest associated with this publication, and there was no significant financial support for this work that could have influenced its outcome. We confirm that the manuscript has been read and approved by all named authors, and that there are no other persons, who satisfied the criteria for authorship, that are not listed. We, further, confirm that the order of authors listed in the manuscript was approved by all of us.

References

1. Wang, D.; Chen, J. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2018**, *26*, 1702–1726. [CrossRef] [PubMed]
2. Benesty, J.; Makino, S.; Chen, J. *Speech Enhancement*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005; pp. 199–228.
3. Makino, S.; Lee, T.W.; Sawada, H. *Blind Speech Separation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007; pp. 3–45.
4. Tawara, N.; Kobayashi, T.; Ogawa, T. Multi-channel speech enhancement using time-domain convolutional denoising autoencoder. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 86–90.
5. Liu, C.; Fu, S.; Li, Y.; Huang, J.; Wang, H.; Tsao, Y. Multichannel speech enhancement by raw waveform-mapping using fully convolutional networks. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2020**, *28*, 1888–1900. [CrossRef]
6. Tan, K.; Xu, Y.; Zhang, S.; Yu, M.; Yu, D.; Tsao, Y. Audio-visual speech separation and dereverberation with a two-stage multimodal network. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 542–553. [CrossRef]
7. Wu, J.; Chen, Z.; Li, J.; Yoshioka, T.; Tan, Z.; Lin, E.; Luo, Y.; Xie, L. An end-to-end architecture of online multi-channel speech separation. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 81–85.
8. Gu, R.; Chen, L.; Zhang, S.; Zheng, J.; Xu, Y.; Yu, M.; Su, D.; Zou, Y; Yu, D. Neural spatial filter: Target speaker speech separation assisted with directional information. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 4290–4294.
9. Fu, Y.; Wu, J.; Hu, Y.; Xing, M.; Xie, L. Desnet: A multi-channel network for simultaneous speech dereverberation, enhancement and separation. In Proceedings of the IEEE Spoken Language Technology Workshop, Shenzhen, China, 19–22 January 2021; pp. 857–864.
10. Xu, Y.; Yu, M.; Zhang, S.; Chen, L.; Weng, C.; Liu, J.; Yu, D. Neural Spatio-Temporal Beamformer for Target Speech Separation. In Proceedings of the Interspeech, Shanghai, China, 25–29 October 2020; pp. 56–60.
11. Zhang, Z.; Xu, Y.; Yu, M.; Zhang, S.; Chen, L.; Yu, D. Adl-mvdr: All deep learning mvdr beamformer for target speech separation. In Proceedings of the ICASSP, Toronto, ON, Canada, 6–11 June 2021; pp. 6089–6093.
12. Heymann, J.; Drude, L.; Haeb-Umbach, R. Neural network based spectral mask estimation for acoustic beamforming. In Proceedings of the ICASSP, Shanghai, China, 20–25 March 2016; pp. 196–200.
13. Zhang, X.; Wang, Z.; Wang, D. A speech enhancement algorithm by iterating single- and multi-microphone processing and its application to robust asr. In Proceedings of the ICASSP, New Orleans, LA, USA, 5–9 March 2017; pp. 276–280.
14. Gu, R.; Zhang, S.; Zou, Y; Yu, D. Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain. *IEEE Signal Proc. Let.* **2021**, *28*, 1370–1374. [CrossRef]
15. Zheng, C.; Liu, W.; Li, A.; Ke, Y.; Li, X. Low-latency monaural speech enhancement with deep filter-bank equalizer. *J. Acoust. Soc. Am.* **2022**, *151*, 3291—3304. [CrossRef]
16. Luo, Y.; Han, C.; Mesgarani, N.; Ceolini, E.; Liu, S. Fasnet: Low-latency adaptive beamforming for multi-microphone audio processing. In Proceedings of the ASRU, Sentosa, Singapore, 14–18 December 2019; pp. 260–267.
17. 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, Virtual, 4–8 May 2020; pp. 6394–6398.
18. Xiao, X.; Watanabe, S.; Erdogan, H.; Lu, L.; Hershey, J.; Seltzer, M.L.; Chen, G.; Zhang, Y.; Mandel, M.; Yu, D. Deep beamforming networks for multi-channel speech recognition. In Proceedings of the ICASSP, Shanghai, China, 20–25 March 2016; pp. 5745–5749.
19. Xu, Y.; Zhang, Z.; Yu, M.; Zhang, S.; Yu, D. Generalized spatio-temporal rnn beamformer for target speech separation. *arXiv* **2021**, arXiv:2101.01280.
20. Ren, X.; Zhang, X.; Chen, L.; Zheng, X.; Zhang, X.; Guo, L.; Yu, B. A causal u-net based neural beamforming network for real-time multi-channel speech enhancement. In Proceedings of the Interspeech, Brno, Czechia, 30 August 2021; pp. 1832–1836.
21. Chen, J.; Li, J.; Xiao, X.; Yoshioka, T.; Wang, H.; Wang, Z.; Gong, Y. Fasnet: Cracking the cocktail party problem by multi-beam deep attractor network. In Proceedings of the ASRU, Okinawa, Japan, 16–20 December 2017; pp. 437–444.
22. Reddy, C.; Dubey, H.; Gopal, V.; Cutler, R.; Braun, S.; Gamper, H.; Aichner, R.; Srinivasan, S. Icassp 2021 deep noise suppression challenge. In Proceedings of the ICASSP, Toronto, ON, Canada, 6–11 June 2021; pp. 6623–6627.
23. Li, A.; Zheng, C.; Zhang, L.; Li, X. Glance and gaze: A collaborative learning framework for single-channel speech enhancement. *arXiv* **2021**, arXiv:2106.11789.
24. Parsons, A.T. Maximum directivity proof for three-dimensional arrays. *J. Acoust. Soc. Am.* **1987**, *82*, 179–182. [CrossRef]
25. Pan, C.; Chen, J.; Benesty, J. Reduced-Order Robust Superdirective Beamforming With Uniform Linear Microphone Arrays. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2016**, *24*, 1548–1559. [CrossRef]
26. 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. Proc.* **2021**, *29*, 1829–1843. [CrossRef]
27. Tan, K.; Wang, D. A convolutional recurrent neural network for real-time speech enhancement. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 3229–3233.
28. Long, J.; Shelhamer, E; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the CVPR, Boston, MA, USA, 8–10 June 2015; pp. 3431–3440.
29. Kalchbrenner, N.; Grefenstette, E.; Blunsom, P. A convolutional neural network for modelling sentences. In Proceedings of the ACL, Balimore, MD, USA, 22–27 June 2014; pp. 655–665.

30. Ciaburro, G.; Iannace, G. Improving smart cities safety using sound events detection based on deep neural network algorithms. *Informatics* **2020**, *7*, 23. [CrossRef]
31. Ciaburro, G. Sound Event Detection in Underground Parking Garage Using Convolutional Neural Network. *Big Data Cogn. Comput.* **2020**, *4*, 20. [CrossRef]
32. Tan, K.; Wang, D. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2019**, *28*, 380–390. [CrossRef] [PubMed]
33. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* **2020**, *106*, 107404. [CrossRef]
34. Liu, W.; Li, A.; Zheng, C.; Li, X. A separation and interaction framework for causal multi-channel speech enhancement. *Digital Signal Process.* **2022**, *126*, 103519. [CrossRef]
35. Zue, V.; Seneff, S.; Glass, J. Speech database development at mit: Timit and beyond, *Speech Commun.* **1990**, *9*, 351–356. [CrossRef]
36. Varga, A.; Steeneken, H. 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]
37. Barker, J.; Marxer, R.; Vincent, E.; Watanabe, S. The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines. In Proceedings of the ASRU, Scottsdale, AZ, USA, 13–17 December 2015; pp. 504–511.
38. Allen, J.; Berkley, D. Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **1979**, *65*, 943–950. [CrossRef]
39. 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, Virtual, 4–8 May 2020; pp. 6389–6393.
40. Rao, W.; Fu, Y.; Hu, Y.; Xu, X.; Jv, Y.; Han, J.; Shang, S.; Jiang, Z.; Xie, L.; Wang, Y.; et al. Interspeech 2021 conferencingspeech challenge: Towards far-field multi-channel speech enhancement for video conferencing. *arXiv* **2021**, arXiv:2104.00960.
41. Rix, A.; Beerends, J.; Hollier, M.; Hekstra, A. Perceptual evaluation of speech quality (pesq)—A new method for speech quality assessment of telephone networks and codecs. In Proceedings of the ICASSP, Salt Palace Convention Center, Salt Lake City, UT, USA, 7–11 May 2001; pp. 749–752.
42. Jensen, J.; Taal, C. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Trans. Audio Speech Lang. Proc.* **2016**, *24*, 2009–2022. [CrossRef]