applied sciences

Article

A Unified Speech Enhancement System Based on Neural Beamforming With Parabolic Reflector

Tao Zhang, Yanzhang Geng, Jianhong Sun *, Chen Jiao and Biyun Ding

School of Electronic and Information Engineering, Tianjin University, Tianjin 300072, China;
zhangtao@tju.edu.cn (T.Z.); gregory@tju.edu.cn (Y.G.); Jiaochen@tju.edu.cn (C.J.); 13820693270@163.com (B.D.)
* Correspondence: sunsun@tju.edu.cn

Received: 1 February 2020; Accepted: 18 March 2020; Published: 25 March 2020

Abstract: This paper presents a unified speech enhancement system to remove both background noise and interfering speech in serious noise environments by jointly utilizing the parabolic reflector model and neural beamformer. First, the amplification property of paraboloid is discussed, which significantly improves the Signal-to-Noise Ratio (SNR) of a desired signal. Therefore, an appropriate paraboloid channel is analyzed and designed through the boundary element method. On the other hand, a time-frequency masking approach and a mask-based beamforming approach are discussed and incorporated in an enhancement system. It is worth noticing that signals provided by the paraboloid and the beamformer are exactly complementary. Finally, these signals are employed in a learning-based fusion framework to further improve the system performance in low SNR environments. Experiments demonstrate that our system is effective and robust in five different noisy conditions (speech interfered with factory, pink, destroyer engine, volvo, and babble noise), as well as in different noise levels. Compared with the original noisy speech, significant average objective metrics improvements are about ΔSTOI $= 0.28$, ΔPESQ $= 1.31$, ΔfwSegSNR $= 11.9$.

Keywords: speech enhancement; parabolic reflector; microphone array; deep neural network; beamformer

1. Introduction

Perceived quality and intelligibility of speech signals are degraded by pervasive noise. This presents challenges to many applications, such as speech communication, hearing aids, and speech recognition. For these applications, speech enhancement is crucial to recover signals from the noisy speech. The enhancements offered by multichannel devices are usually greater than those of single-channel devices [1]. Recent studies indicate that it is beneficial to extract a desired speech signal by beamforming in noisy and reverberant environments, especially in high-level background noise [2,3].

Traditional beamforming methods require a priori knowledge of the Direction of Arrival (DoA) or the transfer functions from an acoustic source to microphones [4]. It is a challenging task to estimate the spatial information of a microphone array in adverse acoustic conditions. According to the auditory masking effect, the time-frequency (T-F) masking technique applies a real-valued or binary mask on the signal's spectrum to filter out unwanted components, because the mask reserves speech-dominant T-F units and weakens noise-dominant T-F units [5]. Advanced beamforming operations require an estimate of the cross-power spectral density matrix of the noise. These statistics can be obtained by estimating spectral masks for speech and noise. Then, beamformers with a mask estimation network can also enhance the quality of speech. Networks are first applied in neural beamformers [6,7] to estimate the time-frequency masks and then masks are applied on the signal's spectrum to predict speech and noise statistics. With these statistics, multichannel filter coefficients are computed based on well-studied beamforming designs, such as Minimum Variance Distortion Response (MVDR) beamformers [8], Linearly Constrained Minimum Variance beamformers (LCMV) [9], and Generalized Eigenvalue (GEV) beamformers [10]. Gannot S et al. [11] explored many popular data-dependent spatial filter design criteria and recognized several well-known beamforming criteria as special cases. S. Chakrabarty et al. [12] proposed a Convolutional Neural Network (CNN)-based mask estimation, which was learned from all the channels simultaneously. The results have shown it is beneficial to utilize multi-channel information, while the approach is array-dependent.

Many researchers choose to work on the single channel mask predictor because it can be applied to all kinds of array configurations [13]. Recent studies have mainly focused on how to design an efficient network structure for single channel mask prediction. The prediction can provide a more accurate mask that assigns a proportion of each T-F bin to each of the sources. A Recurrent Neural Network (RNN) with Long Short-Term Memory (LSTM) cells [14] was used to estimate the time-domain spatial filter weights of a filter-and-sum beamformer for each channel separately, which are then convolved with the input signal to obtain the enhanced signal. Then, this work was extended to estimate frequency domain spatial filter weights in [15]. The recording quality of a microphone has an important impact on speech enhancement performance of the system. It is crucial to subtract the channel which captures the signal with less unwanted components as the reference microphone to provide data for mask prediction. Ochiai T et al. [16] proposed an attention mechanism for reference microphone selection, while Lei Sun [17] adopted the data cleaning and augment operations to attain data to yield better performance in following stages. However, these methods mainly focused on a learning-based approach to select desired data from existing channels, which made the system more complicated. Furthermore, there are no significant differences in the different channels of homogenous sensors. Hence, effective signals cannot be acquired. Given the trends above, in this paper, a novel method for speech enhancement that combines acoustic focus and Deep Neural Network (DNN)-based multi-channel beamforming is proposed. The microphone array speech enhancement framework is extended by integrating the speech enhancement component from the parabolic reflector into the mask-based beamformer. The performances of the two different enhancement approaches with two different types of mask application are evaluated. The performances of a traditional Delay and Sum Beamformer (DSB) are evaluated too. The performance of the proposed system is also compared to a Complex Ideal Ratio Mask (CIRM) estimation method using a DNN network, presented in [18]. The performance of the proposed system for different noise types is also analyzed.

The rest of the paper is organized as follows. In Section 2, the design of the proposed system is described, including the signal model, the acoustic focus performance of the paraboloid, as well as masking and beamforming-based enhancement approaches. In Section 3, the experimental setups are presented. In Section 4, results and analyses are reported. Section 5 concludes the paper.

2. Materials and Methods

In this section, the signal model is presented firstly. Then, quantitative analysis of the acoustic focusing characteristics of the paraboloid along with a design of a parabolic reflector is rendered. Following that, two different ways to incorporate the masks in a speech enhancement system are presented. Finally, how to integrate the parabolic reflector (PR) model with multi-microphones beamforming to yield a higher speech enhancement performance is presented.

2.1. Signal Model

A multi-channel data model with static sources and diffuse noise can be written as follows:

$$y_i(t) = h_{i,j}(t) * s_j(t) + v_i(t), \text{ for } i = 1, 2, \ldots, L,, \;\; j = 1, 2, \ldots, Q, \tag{1}$$

where L denotes the number of microphones, and Q denotes the number of source signals. The notation '$*$' indicates convolution, and t indexes a time sample. $y_i(t)$ denotes the signal at microphone i, and

$s_j(t)$ denotes the j_{th} source signal. $h_{i,j}(t)$ defines the Room Impulse Response (RIR), which models the aspect of sound propagation from source to receiver. An array of L microphones was utilized throughout this work. In the Short Time Fourier Transform (STFT) domain, if the environment can be assumed anechoic, the vector of received signal, $\mathbf{y}(n,k) = [Y(n,k,1),\ldots,Y(n,k,L)]^{\mathrm{T}}$, at time frame n and frequency bin k is given by Equation (2):

$$\mathbf{y}(n,k) = \mathbf{x}_d(n,k) + \mathbf{v}_d(n,k) + \mathbf{v}(n,k). \tag{2}$$

The noise is divided into two components: diffuse noise, denoted by $\mathbf{v}_d(n,k)$, and spatially uncorrelated microphone self-noise, denoted by $\mathbf{v}(n,k)$.

2.2. Analysis of the Paraboloid with Acoustic Focus

2.2.1. Principles

Sten [19] studied the acoustic properties of a paraboloid. Two geometrical characteristics of parabola are illustrated in Figure 1. These are essential for its application as an acoustic reflector.

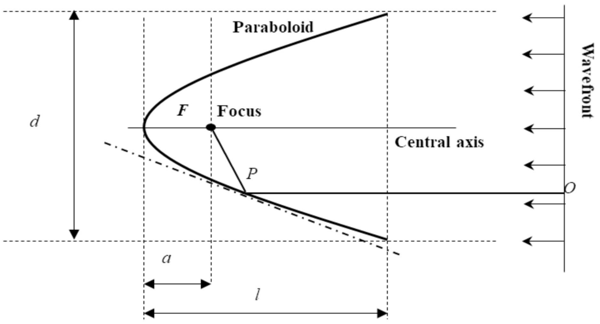


Figure 1. Geometry of parabola and parallel incident waves. *l*

1. The angle between *OP* and the tangent on the point of contact *P* equals the angle between *FP* and the same tangent. In acoustics, any incident wave route in parallel with the central axis will always be focused into the same position, the focus;
2. For a given line at right angles to the axis, the sum of the lengths of *OP* and *PF* is a constant. This means that the sound waves which are parallel to the central axis at the same frequency will have the same phase when reflected to the focus.

and a are the depth and the focal length of paraboloid. Reflector diameter d is a function of ratio l/a. F represents the focus point, P represents the reflection point, and OP is a straight line parallel to the central axis.

As a consequence, the sound pressure is amplified at the focus F. According to [19], the amplification of sounds parallels to the axis is given by Equation (3):

$$F_p = \left\{1 + \left[4\pi\frac{a}{\lambda}\ln(1+\frac{l}{a})\right]^2 + 8\pi\frac{a}{\lambda}\ln(1+\frac{l}{a})\sin 4\pi\frac{a}{\lambda}\right\}^{1/2}, \tag{3}$$

where F_p represents the sound pressure of the amplification factor at the focus, and it is also a pressure factor. λ denotes the wavelength of sound, which equals the ratio of sound speed to sound frequency.

2.2.2. Performance Analysis and Validation

As illustrated in Equation (3), F_p is associated with three factors: a, l, and λ. Figure 2 is given to illustrate the relationship of F_p, l/a, and the sound frequency when the diameter d is assumed as 1 m. The gray plane (F_p = 0 dB) is the base plane, where there is no amplification or attenuation. The depth of paraboloid increases with the increase of the ratio l/a. We can intuitively understand the tendency from Figure 2; that is, the amplification performance will increase, along with an increase in either the frequency or the depth of the paraboloid. In addition, as the depth of the parabola increases, the curve becomes smooth and gradually flattens. Therefore, the magnification performance cannot be improved by deepening the paraboloid.

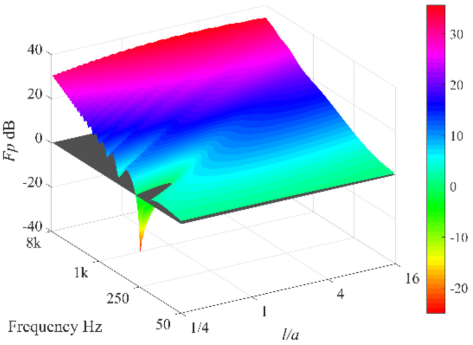


Figure 2. Theoretical amplification at the focus when the diameter is constant.

By means of the indirect boundary element method [20], a series of simulations were implemented in LMS Virtual.Lab software. The results show the convergence of a sound in a parabolic model, which is just the focus with the maximum sound pressure level.

In order to verify the feasibility of the PR-based method, the results obtained from the experiment performed in a real anechoic chamber were compared with computer simulation results. According to the results presented in Figure 3, both the simulation and the experimental results exhibited a similar trend to theoretical rules in Equation (3). It is also noticeable that there were outliers in the measured data curve. This is because the actual paraboloid was made up of plastic rather than a rigid body (theoretically), which made sound waves partially penetrate the paraboloid. So, the ideal focusing could not be achieved.

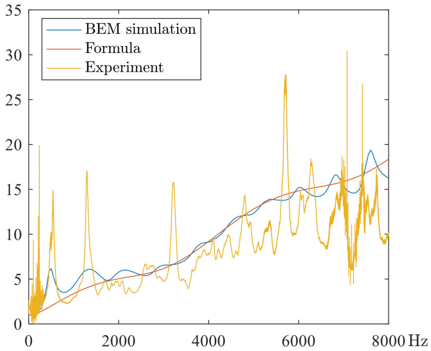


Figure 3. The results of validation.

Experimental results verified the effectiveness of the above theories. Based on the corresponding results, the PR system was implemented, and the amplification function of PR was fitted with reference to Equation (3). The ratio was assigned with $l/a = 4$ as a key parameter of the PR system. Other structural parameters of the paraboloid were $a = 40$ mm, $l = 160$ mm, $d = 320$ mm.

2.3. Two Approaches to Utilize Masks

The Ideal Ratio Mask (IRM) is a soft mask and is given as Equation (4):

$$I_{IRM}(n,k) = \frac{\left|X_d(n,k,ref_m)\right|}{\left|Y(n,k,ref_m)\right|}, \tag{4}$$

where ref_m denotes the reference microphone and m denotes the corresponding neural beamformer. $Y(n,k,ref_m)$ denotes the signal recorded at the reference microphone, and $X_d(n,k,ref_m)$ represents the estimated clean speech signal. For a noise mask, its value can be represented as $1{-}I_{IRM}$.

Two different approaches for incorporating a mask to obtain a desired signal are discussed below.

2.3.1. Direct Mask Application

In this approach, the mask can be applied directly to T-F representation of the microphone signal at the reference microphone to calculate the ideal mask. An estimation of the desired signal is given by Equation (5):

$$\hat{X}_d(n,k,ref_m) = I_{(\cdot)}(n,k) \cdot Y(n,k,ref_m), \tag{5}$$

where $I_{(\cdot)}$ represents the mask. Note that the phase of $\hat{X}_d(n,k,ref_m)$ is equal to the phase $Y(n,k,ref_m)$, and the desired signal waveform is obtained by an inverse STFT with the estimated magnitude.

2.3.2. Neural Beamformer

The frequency-domain beamformer is used to reduce computational complexity. $\mathbf{w}(n,k)$ denotes the corresponding beamforming filter coefficients. Using a spatial filtering approach, an estimation of the desired signal is given as a linear combination of the microphone signals $\mathbf{y}(n,k)$, at each $\mathrm{T-F}$ bin, as illustrated in Equation (6):

$$\hat{X}_d(n,k,ref_m) = \mathbf{w}^{\mathrm{H}}(n,k)\mathbf{y}(n,k), \tag{6}$$

where H represents conjugate transpose. In this work, the MVDR criterion was used to compute filter coefficients, and in this criterion, coefficients were found by minimizing the power of noise components at output, given by Equations (7) and (8).

$$\mathbf{w}(n,k) = \underset{\mathbf{w}}{\mathrm{argmin}}\mathbf{w}^{\mathrm{H}}\mathbf{\Phi}_{\mathrm{n}}\mathbf{w}, \tag{7}$$

subject to:

$$\mathbf{w}^{\mathrm{H}}(n,k)\mathbf{a}(n,k) = 1, \tag{8}$$

where $\mathbf{a}(n,k)$ denotes the Relative Transfer Function (RTF) vector.

Considering the individual signal components in Equation (2) to be uncorrelated, the Power Spectral Density (PSD) matrix of microphone signals can be expressed as Equation (9):

$$\mathbf{\Phi}_{\mathbf{y}}(n,k) = \mathrm{E}\left\{\mathrm{y}(n,k)\mathrm{y}^{\mathrm{H}}(n,k)\right\} = \mathbf{\Phi}_{\mathrm{x_d}}(n,k) + \mathbf{\Phi}_{\mathrm{n}}(n,k), \tag{9}$$

where $\mathrm{E}\{\cdot\}$ represents the expectation operator, $\mathbf{\Phi}_{\mathrm{x_d}}(n,k)$ denotes the rank-one PSD matrix of the desired signal, and $\mathbf{\Phi}_{\mathrm{n}}(n,k)$, denotes the PSD matrix of noise components. PSD matrices are robustly estimated using the expectation with respect to time-frequency masks as Equations (10) and (11):

$$\mathbf{\Phi}_{\mathrm{x_d}}(k) = \frac{1}{\sum_n^N I_{(\cdot)}(n,k)} \sum_{n=1}^{N} I_{(\cdot)}(n,k)\mathbf{y}(n,k)\mathbf{y}^{\mathrm{H}}(n,k), \tag{10}$$

$$\mathbf{\Phi}_{\mathrm{n}}(k) = \frac{1}{\sum_n^N 1 - I_{(\cdot)}(n,k)} \sum_{n=1}^{N} (1 - I_{(\cdot)}(n,k))\mathbf{y}(n,k)\mathbf{y}^{\mathrm{H}}(n,k), \tag{11}$$

The closed-form solution to the optimization problem is given by Equation (12) in [21]:

$$\mathbf{w}(n,k) = \frac{\boldsymbol{\Phi}_{\mathrm{n}}^{-1}\mathbf{a}(n,k)}{\mathbf{a}^{\mathrm{H}}(n,k)\boldsymbol{\Phi}_{\mathrm{n}}^{-1}\mathbf{a}(n,k)}. \tag{12}$$

Adopting the optimization formalization [21], the explicit dependence of the above filter on the relative transfer functions can be eliminated, and the following form can be obtained, as illustrated in Equation (13):

$$\mathbf{w} = \frac{(\boldsymbol{\Phi}_{\mathrm{n}}(k))^{-1}\boldsymbol{\Phi}_{\mathrm{x_d}}(k)}{\mathrm{Tr}\Big((\boldsymbol{\Phi}_{\mathrm{n}}(k))^{-1}\boldsymbol{\Phi}_{\mathrm{x_d}}(k)\Big)}\mathbf{u}, \tag{13}$$

where $\mathbf{u}$ is a one-hot vector representing a reference microphone, and $\mathrm{Tr}(\cdot)$ represents matrix trace operation.

The overall flowchart of the proposed speech enhancement framework is illustrated in Figure 4. The circular microphone array captures the noisy speech. The speech is processed by a neural beamformer, which removes most of the interference. So much high frequency information is lost. In this approach, any microphone in the circular array can be defined as the reference microphone to estimate the mask. The physical amplification characteristic of the designed PR model has significant effects in speech enhancement, and some speech distortions are also introduced into the target speech. The signals captured by the microphone at the focus and the output of the neural beamformer1 have complementary information of the desired speech in the frequency domain. By processing the signals from two approaches mentioned earlier, the fusion operation aims to make a trade-off between speech distortion and speech intelligibility. The reflector microphone signal is utilized to estimate the mask, and the direct mask application is used to enhance the desired speech. The Generalized Cross-Correlation (GCC) method [22] is adopted to align the data from different processing methods. The masked data are utilized by MVDR beamformer. During the postfiltering, the output of MVDR beamformer is multiplied by IRM obtained by the mask estimation to get the final output.

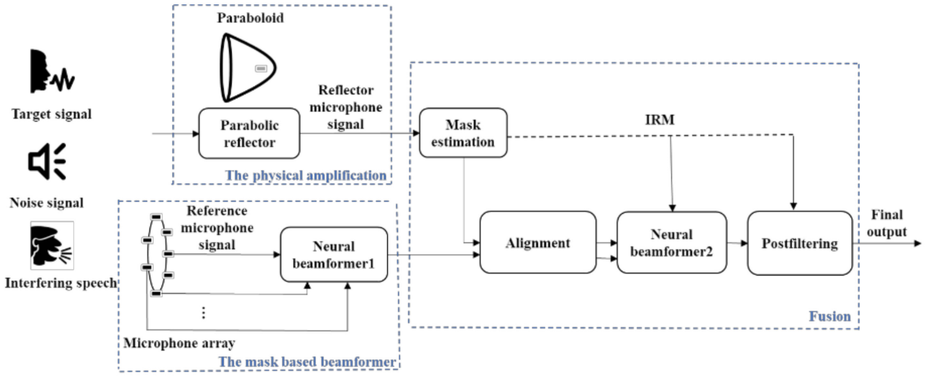


Figure 4. The overall diagram of our proposed system.

In the proposed system, each neural beamformer is set as an MVDR beamformer, based on a single-channel IRM estimation. With the same network structure, two higher quality signals are used as input for the second beamforming. The enhancement performance is improved with a little additional system complexity. It is also achievable to combine different types of neural beamformers to meet our requirements.

3. Experiment and Results

3.1. Experiment Setup

Considering the number of physical channels of the equipment used for data acquisition and the structural parameters of the paraboloid, for all experiments, a uniform circular array was set up, where L = 7 microphones, and the distance between the microphones was 18 cm. The other microphone was at the focus of the parabolic reflector, and the circular array and the paraboloid were combined into a whole, as shown in Figure 5. The input signals, with the sampling frequency of Fs = 16 kHz, were transformed into the STFT domain, which used a Discrete Fourier Transform (DFT) length of 256 and 50% overlap. Signals were divided into 16 ms frames with an 8 ms frame shift in time-domain. McRoomSim [23] was used to generate the Room Impulse Responses (RIRs) required to simulate different acoustic conditions. The room property was set to an anechoic chamber with a sound absorption coefficient of 1, which means that there was no reverberation or other noise in the room. To illustrate the independency of the source spatial position, eight different positions were set at different angles. Table 1 shows the configuration used to generate the dataset. For each position in the training stage, 100 speech signals were convolved with the simulated RIRs corresponding to the specific setup, while for each position in the testing stage 60 speech signals were convolved. The proposed system was evaluated on the IEEE database [24]. Each clean utterance was mixed by adding the speech interference to each isolated noise (babble, volvo, destroyer engine, pink, and factory noise) at different SNRs from –18 dB to 7 dB. The speech interference noise was an utterance of WSJ0 in around 10 s. The other five noises were non-stationary and each signal was around 4 min long. Random cuts from the first half of each noise were mixed with each training utterance to create the training mixtures, and cuts of the second half of that were mixed with each testing utterance to create testing mixtures. Acoustic conditions are shown in Figure 5, and different source positions are shown in Table 1 in detail.

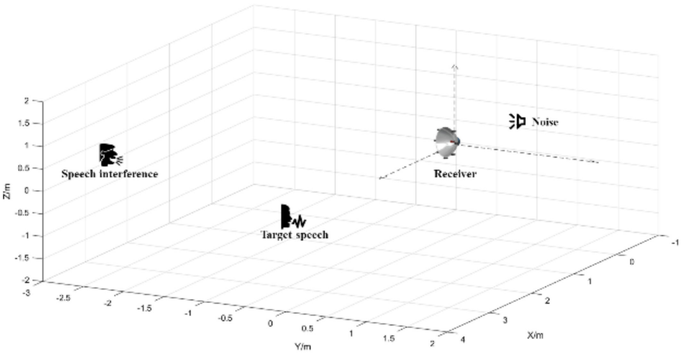


Figure 5. Simulation environment.

Table 1. Different source positions.

	Target Speech	**Noise**	**Speech Interference**
Training Position	[4, 0, 0]	[1.3, 1.52, 2]	[3.6, −3.5, 0.8]
	[3, 0, 0]	[1.8, 2.57, 0.6]	[3.3, −1.58, 2.9]
	[5, 0, 0]	[2.5, 0, 3.66]	[3, −3, 1.96]
	[3, 0, 0]	[1.8, 2.57, 0.6]	[3.3, −1.58, 2.9]
	[4, 0, 0]	[2.3, 2, 1.89]	[3.2, −2.65, 0.38]
Testing Position	[4, 0, 0]	[1.3, −1,52, −2]	[3.2, 2.65, 0.38]
	[4, 0, 0]	[1.8, −2.57, 0.6]	[2.3, 2, 1.89]
	[3, 0, 0]	[1.8, 2.57, 0.6]	[1.3, −1.52, 2]

3.2. Training

Using IRM defined in Equation (4) as the learning target, DNN was designed, where the output could be considered as the probabilities of the existence of speech at each time-frequency bin. The magnitude, the second derivative of the magnitude, and the phase of the received signals for each STFT time frame were directly provided as the input to the system [12]. Amplitude Modulation Spectrogram (AMS), Relative Spectral Transform and Perceptual Linear Prediction (RST-PLP) [25] were also appended to the feature vector. In this paper, Restricted Boltzmann Machine (RBM) [26] based pre-training was used for DNN training. Supervised fine-tuning of the parameters throughout the whole network was performed using the Mean Square Error (MSE) criterion, as illustrated in Equation (14):

$$\rho = \frac{1}{2N}\sum_{n}\sum_{k}\left(I_{IRM}(n,k) - \hat{I}(n,k)\right)^{2}, \tag{14}$$

where $\hat{I}(n,k)$ are the vectors of reference IRM, and N is the total number of frames for the input. The DNN architecture was 2075-1024-1024-1024-129, which denotes that the size was 2075 (415 × 4 + 415, including two left and two right context frames, and one current frame in the input layer), 1024 units for each of the four hidden layers, and 129 for the output layer (DFT length was 256, where 256/2 + 1 is the total number of frequency bins). In this work, the dropout rate was 0.2, and the momentum rate was set to 0.5 for the first five epochs, and afterwards the rate was changed to 0.9 for the remaining 35 epochs. The mini-batch size was set to 512. The sigmoid activation functions were used for all hidden layers and the output layer. Other values were evaluated as well; however, this combination performed best empirically.

In the following experimental evaluations, the method is called PR when processing noisy speech through the physical model. The DSB used for comparison utilized ideal parametric information [27]. Since the estimated IRM from DNN was directly applied to a reference microphone signal, the method was termed as IRM-F. Similarly, with the estimated CIRM, jointly estimating real and imaginary components of STFT [18], the corresponding method is referred to as CIRM-F. When it is used to estimate power spectral density matrices to be used within a MVDR beamformer, the method is referred to as IRM-BF. The method proposed in this work is named PR-IRM-BF.

3.3. Results

The enhanced speech signals from each approach were evaluated in terms of three well-known objective metrics, namely the Perceptual Evaluation of Speech Quality (PESQ), the Short-Time Objective Intelligibility (STOI) score, and the frequency-weighted Segmental SNR (fwSegSNR).

As presented in Table 2, results of the comparison experiments in five mixed noise situations showed that PR and IRM-BF had improved PESQ, STOI, and fwSegSNR performance compared with the original speech in all mixed noise situations. In the presence of noise, a beamformer operation removes speech interference but degrades the continuity of the target speech. PR is able to protect target speech but it is not effective to eliminate high-energy speech interference.

Table 2. The average performance of the system in five mixed noise conditions.

		Destroyer Engine Noise and Speech Interference						
Measure	**SNR**	**Noisy**	**DSB**	**IRM-F**	**CIRM-F**	**IRM-BF**	**PR**	**PR-IRM-BF**
STOI	−18 dB	0.400	0.401	0.445	0.292	0.433	0.575	**0.741**
	−13 dB	0.447	0.499	0.576	0.438	0.595	0.690	**0.827**
	−8 dB	0.541	0.606	0.701	0.602	0.747	0.800	**0.891**
	−3 dB	0.645	0.696	0.783	0.737	0.820	0.884	**0.922**
	2 dB	0.756	0.763	0.861	0.843	0.884	**0.941**	0.938
	7 dB	0.844	0.808	0.916	0.906	0.917	**0.974**	0.945

Table 2. *Cont.*

PESQ	−18 dB	1.164	0.887	1.159	0.806	1.070	1.379	**1.693**
	−13 dB	1.201	1.179	1.272	0.997	1.358	1.601	**2.252**
	−8 dB	1.232	1.517	1.555	1.443	1.852	1.881	**2.690**
	−3 dB	1.401	1.887	1.797	1.771	2.135	2.124	**3.009**
	2 dB	1.718	2.249	2.101	2.212	2.532	2.437	**3.287**
	7 dB	1.964	2.632	2.388	2.565	2.831	2.752	**3.519**
fwSegSNR	−18 dB	−23.880	−17.140	−9.502	−3.233	−2.586	−13.094	**0.644**
	−13 dB	−18.862	−12.072	−5.819	−1.540	−0.926	−8.095	**0.764**
	−8 dB	−13.869	−7.168	−3.177	−0.201	0.584	−3.092	**0.819**
	−3 dB	−8.825	−2.178	−0.098	2.268	1.426	**1.917**	1.249
	2 dB	−3.860	2.688	3.073	4.597	1.697	**6.908**	1.264
	7 dB	1.161	7.455	6.896	6.782	2.432	**11.936**	1.398
				Babble noise and speech interference				
Measure	**SNR**	**Noisy**	**DSB**	**IRM-F**	**CIRM-F**	**IRM-BF**	**PR**	**PR-IRM-BF**
STOI	−18 dB	0.352	0.362	0.389	0.323	0.381	0.510	**0.656**
	−13 dB	0.409	0.447	0.492	0.362	0.484	0.643	**0.779**
	−8 dB	0.490	0.543	0.618	0.511	0.631	0.768	**0.863**
	−3 dB	0.594	0.641	0.726	0.656	0.767	0.867	**0.909**
	2 dB	0.719	0.731	0.829	0.801	0.865	**0.938**	0.936
	7 dB	0.828	0.801	0.896	0.884	0.903	**0.973**	0.946
PESQ	−18 dB	1.013	0.780	1.116	1.044	1.084	1.361	**1.581**
	−13 dB	1.106	0.991	1.168	1.063	1.134	1.651	**2.084**
	−8 dB	1.165	1.296	1.429	1.278	1.513	1.926	**2.546**
	−3 dB	1.412	1.646	1.724	1.693	1.968	2.214	**2.906**
	2 dB	1.747	2.020	2.097	2.071	2.452	2.558	**3.218**
	7 dB	2.052	2.430	2.382	2.474	2.701	2.895	**3.447**
fwSegSNR	−18 dB	−23.749	−20.353	−10.132	−7.863	−3.116	−12.036	**−0.385**
	−13 dB	−18.710	−15.373	−7.122	−3.650	−1.750	−6.971	**−0.253**
	−8 dB	−13.703	−10.374	−4.184	−1.033	−0.353	−1.967	**0.369**
	−3 dB	−8.677	−5.371	−1.356	1.315	0.854	**2.993**	0.982
	2 dB	−3.689	−0.379	2.119	3.639	1.573	**8.035**	1.367
	7 dB	1.315	4.617	5.882	5.974	1.801	**13.036**	1.323
				Factory noise and speech interference				
Measure	**SNR**	**Noisy**	**DSB**	**IRM-F**	**CIRM-F**	**IRM-BF**	**PR**	**PR-IRM-BF**
STOI	−18 dB	0.375	0.382	0.387	0.334	0.382	0.531	**0.658**
	−13 dB	0.423	0.469	0.494	0.375	0.482	0.653	**0.781**
	−8 dB	0.493	0.569	0.624	0.521	0.649	0.771	**0.863**
	−3 dB	0.611	0.662	0.746	0.676	0.790	0.875	**0.912**
	2 dB	0.729	0.731	0.825	0.798	0.855	**0.939**	0.929
	7 dB	0.833	0.773	0.896	0.878	0.899	**0.974**	0.944
PESQ	−18 dB	1.027	0.736	1.106	1.097	1.081	1.333	**1.628**
	−13 dB	1.069	0.998	1.142	1.111	1.083	1.599	**2.077**
	−8 dB	1.228	1.318	1.430	1.373	1.631	1.844	**2.522**
	−3 dB	1.395	1.716	1.770	1.699	2.159	2.173	**2.898**
	2 dB	1.715	2.108	2.037	2.058	2.427	2.508	**3.170**
	7 dB	2.005	2.488	2.346	2.473	2.722	2.850	**3.423**
fwSegSNR	−18 dB	−23.726	−20.386	−10.239	−7.031	−3.282	−12.125	**0.194**
	−13 dB	−18.722	−15.390	−7.514	−1.989	−1.587	−7.123	**0.360**
	−8 dB	−13.706	−10.401	−4.315	−0.456	−0.186	−2.134	**0.502**
	−3 dB	−8.713	−5.389	−1.120	1.576	1.049	**2.879**	0.915
	2 dB	−3.709	−0.407	1.931	3.538	1.473	**7.879**	1.128
	7 dB	1.294	4.582	5.556	5.559	1.898	**12.880**	1.407

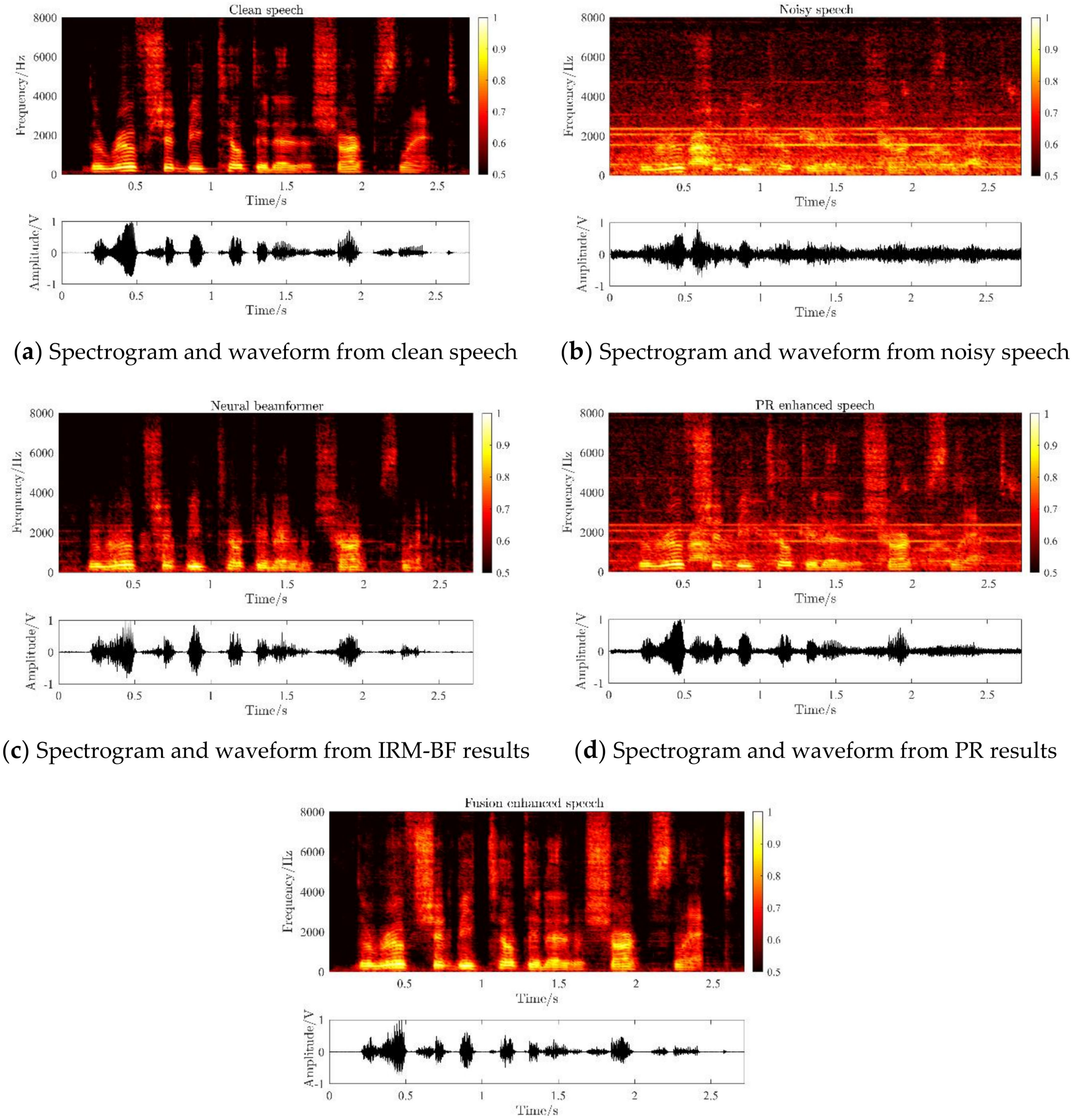
Table 2. *Cont.*

				Volvo noise and speech interference				
Measure	**SNR**	**Noisy**	**DSB**	**IRM-F**	**CIRM-F**	**IRM-BF**	**PR**	**PR-IRM-BF**
STOI	−18 dB	0.361	0.393	0.618	0.479	0.705	0.619	**0.808**
	−13 dB	0.456	0.488	0.722	0.603	0.806	0.744	**0.874**
	−8 dB	0.562	0.589	0.800	0.704	0.870	0.843	**0.910**
	−3 dB	0.677	0.678	0.868	0.805	0.904	0.913	**0.930**
	2 dB	0.784	0.741	0.917	0.838	0.923	**0.957**	0.944
	7 dB	0.868	0.780	0.949	0.923	0.930	**0.981**	0.947
PESQ	−18 dB	0.835	0.738	1.397	1.091	1.776	1.547	**2.196**
	−13 dB	0.993	1.089	1.667	1.394	2.118	1.915	**2.674**
	−8 dB	1.284	1.481	1.965	1.702	2.436	2.215	**3.030**
	−3 dB	1.678	1.872	2.256	2.050	2.721	2.532	**3.276**
	2 dB	1.982	2.265	2.533	3.357	2.877	2.871	**3.486**
	7 dB	2.280	2.665	2.784	2.658	3.091	3.203	**3.636**
fwSegSNR	−18 dB	−23.009	−22.330	−6.675	−2.596	−0.784	−8.496	**0.576**
	−13 dB	−17.892	−17.217	−4.041	−0.757	0.955	−3.413	**0.818**
	−8 dB	−12.890	−12.223	−1.529	1.016	**1.718**	1.588	1.083
	−3 dB	−7.887	−7.227	1.513	3.311	2.687	**6.589**	1.266
	2 dB	−2.885	−2.234	4.541	5.415	2.424	**11.590**	1.319
	7 dB	2.117	2.763	7.748	7.932	2.237	**16.596**	1.376
				Pink noise and speech interference				
Measure	**SNR**	**Noisy**	**DSB**	**IRM-F**	**CIRM-F**	**IRM-BF**	**PR**	**PR-IRM-BF**
STOI	−18 dB	0.390	0.395	0.420	0.320	0.408	0.551	**0.688**
	−13 dB	0.441	0.483	0.541	0.397	0.554	0.669	**0.792**
	−8 dB	0.521	0.580	0.653	0.545	0.691	0.786	**0.867**
	−3 dB	0.627	0.668	0.759	0.698	0.797	0.880	**0.909**
	2 dB	0.736	0.731	0.830	0.811	0.856	**0.941**	0.934
	7 dB	0.839	0.772	0.900	0.886	0.898	**0.975**	0.945
PESQ	−18 dB	1.003	0.746	1.085	1.007	1.072	1.314	**1.738**
	−13 dB	1.028	1.043	1.190	1.040	1.267	1.587	**2.203**
	−8 dB	1.135	1.393	1.488	1.390	1.781	1.839	**2.587**
	−3 dB	1.375	1.783	1.807	1.776	2.200	2.149	**2.916**
	2 dB	1.715	2.177	2.062	2.143	2.422	2.486	**3.185**
	7 dB	1.988	2.554	2.350	2.512	2.775	2.837	**3.469**
fwSegSNR	−18 dB	−23.746	−20.296	−10.597	−4.065	−3.235	−12.889	**0.258**
	−13 dB	−18.743	−15.300	−6.677	−1.282	−1.239	−7.889	**0.265**
	−8 dB	−13.740	−10.305	−3.933	−0.037	0.053	−2.889	**0.388**
	−3 dB	−8.737	−5.308	−0.544	1.862	1.3672	**2.110**	0.7513
	2 dB	−3.749	−0.322	2.200	3.881	1.698	**7.101**	1.039
	7 dB	1.267	4.688	5.628	5.927	1.764	**12.111**	1.243

From the above analysis, the two isolated systems (PR, IRM-BF) have their own shortcomings when addressing serious noises. In extremely low SNR environments, the proposed system significantly outperformed the compared methods by utilizing the complementarity of PR and IRM-BF and made a trade-off to get much better results over individual enhancements. In the case of 7 dB, the proposed system caused performance degradation in terms of fwSegSNR and STOI, possibly due to signal distortions. Specifically, when the original noisy speech was of relatively high quality, the PR model failed to show its superiority in improving the SNR of the desired signal, while the fusion operations introduced extra noise. For IRM estimation, the improvement of fwSegSNR and STOI achieved by each application of the mask was much higher than that of DSB beamformers. The CIRM-F method led to larger improvements in terms of fwSegSNR but suffered from lower PESQ and STOI improvement.

4. Discussion

To illustrate the effectiveness of our speech enhancement system more clearly, an utterance corrupted by mixed noise (destroyer engine + speech interference) at −2 dB from test data and enhanced by our proposed system is presented, as shown in Figure 6.



(**a**) Spectrogram and waveform from clean speech

(**b**) Spectrogram and waveform from noisy speech

(**c**) Spectrogram and waveform from IRM-BF results

(**d**) Spectrogram and waveform from PR results

(**e**) Spectrogram and waveform from the proposed method

Figure 6. (**a**–**d**) are the spectrograms and waveforms from clean speech, noisy speech, PR model, and multi-channel beamforming preprocessing, respectively. The spectrogram of our final output is listed in (**e**). (**a**) Spectrogram and waveform from clean speech; (**b**) Spectrogram and waveform from noisy speech; (**c**) Spectrogram and waveform from IRM-BF results; (**d**) Spectrogram and waveform from PR results; (**e**) Spectrogram and waveform from the proposed method.

Compared with the spectrogram of PR enhanced speech, speech processed by neural beamformer removed most of the interference parts, while losing a lot of high-frequency information. It also verified the description that there are some similarities, differences, and supplementary parts in these twofold signals. Although the PR model also introduced some speech distortions to the target speech, the spectrogram indicated that the PR model has significant performance in data cleaning.

The processed speech makes a trade-off between speech distortion and speech intelligibility by fusing operation, yielding better enhancement performance. As seen in Figure 6e, the power and the strength of noise are largely suppressed.

5. Conclusions

The parabolic reflector, a physical amplification, was proposed. It focuses the target speech considerably from noisy speech and provides a heterogeneous channel. The channel captures relatively clean data to estimate masks. The source-to-array distance is so long so that it is difficult to collect effective signals. The physical amplification model reduces the system complexity and provides favorable data. Moreover, by utilizing complementary information between the parabolic reflector and the microphone array, the proposed signal fusion system achieves better performance for noise and speech interference mixed conditions, especially in low SNR environments. In the future, we will extend the current work in several ways, such as upgrading a neural beamformer module to track more phase information from microphone arrays. Our most important future work is to acquire RIR by measuring and utilizing the framework in far-field multi-talker microphone array speech enhancement.

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

Funding: This research was funded by the Zhejiang Lab's International Talent Fund for Young Professionals.

Acknowledgments: The authors want to express their acknowledgements to the Zhejiang Lab's International Talent Fund for Young Professionals.

Conflicts of Interest: The authors declare no conflict of interest and the funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Vincent, E.; Gribonval, R.; Plumbley, M. Oracle estimators for the benchmarking of source separation algorithms. *Signal Process.* **2007**, *87*, 1933–1950. [CrossRef]
2. Stenzel, S.; Freudenberger, J.; Schmidt, G.; Schmidt, G. A Minimum variance beamformer for spatially distributed microphones using a soft reference selection. In Proceedings of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA), Villers-les-Nancy, France, 12–14 May 2014; pp. 127–131.
3. Cohen, I.; Benesty, J.; Gannot, S. (Eds.) *Speech Processing in Modern Communication: Challenges and Perspectives*; Springer: New York, NY, USA, 2010.
4. Hawkes, M.; Nehorai, A.; Arye, N. Acoustic vector-sensor beamforming and Capon direction estimation. *IEEE Trans. Signal Process.* **1998**, *46*, 2291–2304. [CrossRef]
5. Wang, Y.; Narayanan, A.; Wang, D.L. On Training Targets for Supervised Speech Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2014**, *22*, 1849–1858. [CrossRef] [PubMed]
6. Nakatani, T.; Ito, N.; Higuchi, T.; Araki, S.; Kinoshita, K. Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 286–290.
7. Heymann, J.; Drude, L.; Boeddeker, C.; Hanebrink, P.; Haeb-Umbach, R. Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 5325–5329.
8. Souden, M.; Benesty, J.; Affes, S. On Optimal Frequency-Domain Multichannel Linear Filtering for Noise Reduction. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 260–276. [CrossRef]
9. 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]

10. Warsitz, E.; Haeb-Umbach, R. Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition. *IEEE Trans. Audio Speech Lang. Process.* **2007**, *15*, 1529–1539. [CrossRef]
11. Gannot, S.; Vincent, E.; Markovich-Golan, S.; Ozerov, A. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 692–730. [CrossRef]
12. Chakrabarty, S.; Wang, D.; Habets, E.A.P. Time-Frequency Masking Based Online Speech Enhancement with Multi-Channel Data Using Convolutional Neural Networks. In Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 476–480.
13. Xiao, X.; Zhao, S.; Jones, U.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.
14. Li, B.; Sainath, T.N.; Weiss, R.J.; Wilson, K.W.; Bacchiani, M. Neural Network Adaptive Beamforming for Robust Multichannel Speech Recognition. In Proceedings of the Interspeech Conference, San Francisco, CA, USA, 8–12 September 2016; pp. 1976–1980.
15. Sainath, T.N.; Weiss, R.J.; Wilson, K.W.; Li, B.; Narayanan, A.; Variani, E.; Bacchiani, M.; Shafran, I.; Senior, A.; Chin, K.W.; et al. Multichannel Signal Processing With Deep Neural Networks for Automatic Speech Recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 965–979. [CrossRef]
16. Ochiai, T.; Watanabe, S.; Hori, T.; Hershey, J.R.; Xiao, X. Unified Architecture for Multichannel End-to-End Speech Recognition With Neural Beamforming. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 1274–1288. [CrossRef]
17. Sun, L.; Du, J.; Gao, T.; Fang, Y.; Ma, F.; Lee, C.-H. A Speaker-Dependent Approach to Separation of Far-Field Multi-Talker Microphone Array Speech for Front-End Processing in the CHiME-5 Challenge. *IEEE J. Sel. Top. Signal Process.* **2019**, *13*, 827–840. [CrossRef]
18. Williamson, D.S.; Wang, Y.; Wang, D. Complex ratio masking for joint enhancement of magnitude and phase. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Institute of Electrical and Electronics Engineers (IEEE), Shanghai, China, 20–25 March 2016; pp. 5220–5224.
19. Wahlstrom, S. The Parabolic Reflector as an Acoustical Amplifier. *Audio Eng. Soc.* **1985**, *33*, 418–429.
20. Inci, E.O.; Coox, L.; Atak, O.; Deckers, E.; Desmet, W. Applications of an isogeometric indirect boundary element method and the importance of accurate geometrical representation in acoustic problems. *Eng. Anal. Bound. Elem.* **2020**, *110*, 124–136. [CrossRef]
21. Cox, H.; Zeskind, R.; Owen, M. Robust adaptive beamforming. *IEEE Trans. Acoust. Speech Signal Process.* **1987**, *35*, 1365–1376. [CrossRef]
22. Zhang, Q.; Zhang, L. An improved delay algorithm based on generalized cross correlation. In Proceedings of the IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 3–5 October 2017; pp. 395–399.
23. Wabnitz, A.; Epain, N.; Jin, C.; van Schaik, A. Room Acoustics Simulation for Multichannel Microphone Arrays. In Proceedings of the International Symposium on Room Acoustics, Melbourne, Australia, 29–31 August 2010; pp. 1–6.
24. Rosenthal, S. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.* **1969**, *17*, 225–246.
25. Wang, Y.; Han, K.; Wang, D.L. Exploring Monaural Features for Classification-Based Speech Segregation. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *21*, 270–279. [CrossRef]
26. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507. [CrossRef] [PubMed]
27. Brandstein, M.S.; Ward, D.B. (Eds.) *Microphone Arrays: Signal Processing Techniques and Applications*; Springer: Berlin, Germany, 2001.