Hindawi
Complexity
Volume 2018, Article ID 6153451, 13 pages
https://doi.org/10.1155/2018/6153451

Research Article

Speech Enhancement Control Design Algorithm for Dual-Microphone Systems Using β-NMF in a Complex Environment

Dong-xia Wang, Mao-song Jiang, Fang-lin Niu, Yu-dong Cao, and Cheng-xu Zhou

School of Electronic and Information Engineering, Liaoning University of Technology, Jinzhou, Liaoning 121001, China

Correspondence should be addressed to Dong-xia Wang; dxwang_lg@126.com

Received 15 April 2018; Revised 3 July 2018; Accepted 26 July 2018; Published 9 September 2018

Academic Editor: Junpei Zhong

Single-microphone speech enhancement algorithms by using nonnegative matrix factorization can only utilize the temporal and spectral diversity of the received signal, making the performance of the noise suppression degrade rapidly in a complex environment. Microphone arrays have spatial selection and high signal gain, so it applies to the adverse noise conditions. In this paper, we present a new algorithm for speech enhancement based on two microphones with nonnegative matrix factorization. The interchannel characteristic of each nonnegative matrix factorization basis can be modeled by the adopted method, such as the amplitude ratios and the phase differences between channels. The results of the experiment confirm that the proposed algorithm is superior to other dual-microphone speech enhancement algorithms.

1. Introduction

For the sake of improving the quality and intelligibility of noisy signals, speech enhancement is widely applied in many fields including speech communication, speech coding, and speech recognition. In terms of the number of microphones, speech enhancement methods can be split into two classes: single microphone and microphone array.

In the past, there have been many single-microphone speech enhancement algorithms presented including statistical model method, spectral subtraction, subspace decomposition, and other typical algorithms. These algorithms have a good noise suppression performance under stationary conditions, but at the cost of a priori information loss of clean speech and noise, in which it provides limited performance under a complex environment.

Recently, a new matrix decomposition algorithm called nonnegative matrix factorization (NMF) [1] method has been successfully used to solve a variety of problems in many fields. NMF is a powerful method for machine learning and hidden data discovery; the basic idea of the method is that one nonnegative matrix is decomposed into the product of two nonnegative matrices without making any statistical hypothesis of data. Compared with the traditional matrix decomposition algorithm, it has a strong physical significance, it has small storage, and it is simple and easy to implement. The results show that it has been widely used to effectively solve various problems including pattern clustering and classification tasks [2–5], source separation [6], and speech enhancement [7]. In voice applications, we can obtain a priori information by using train data with NMF instead of the clean signal.

Currently, according to the different methods in machine learning, a single speech enhancement method based on NMF can be categorized into unsupervised learning and supervised learning algorithms [7]. Unsupervised methods are simple and easy to implicate without any prior information on the speech or noise, whose main difficulty is estimating the noise power spectral density (PSD) [8], especially in a complex environment.

For the supervised methods, selecting a proper model needs to consider not only the aspect of the speech and noise signals but also the model parameter estimation using the training samples of those signals. One advantage of these methods is estimating the noise PSD without the need to use other algorithms. Compared with the unsupervised methods

under a complex environment, the studies have been proved that the supervised method is an effective way of obtaining better performance of the enhanced speech signals.

In order to solve the problem of the characteristic of mismatch between training data and testing data, a supervised NMF-based algorithm is proposed in speech enhancement to incorporate with some prior information, including temporal continuity [9] and statistical distribution of the data [10]. More recently, aiming at improving the general subspace constraints, an improved NMF algorithm is proposed by introducing additional terms into the objective function [11]. A framework for decreasing the computational complexity in NMF by using the extreme learning machine (ELM) is designed in [12]. ELM and its variants have been widely applied in different kinds of fields, because of its good scalability and strong generalization performance [13]. With the unceasing development of human-computer interaction recently, higher requirements for speech recognition and computer vision are put forward in a complex environment. In [14–16]; the control scheme for improving the convergence speed is developed to optimize system performance.

In [17], a speech enhancement method for solving the difficult problem of manual selection modes by applying a regularized nonnegative matrix factorization algorithm is presented. In practical application, however, the speech signals have spatial characteristics (spatial diversity of reverberation guidance), which is not present in the single-microphone system. One microphone has good performance in speech enhancement system, however, it only uses both temporal and spectral information of signal and lacks spatial information.

The two-microphone system has attracted much attention for its small size and small amount of calculation, which is in line with the trend of miniaturization of devices. An algorithm for achieving a dual-microphone speech enhancement by using the coherence function is proposed [18]. In [19], the improved method, which incorporates the coherence function and the Kalman filter, is used to obtain enhanced speech signal. These algorithms belong to the unsupervised methods in a sense. Therefore, we propose a novel β-NMF for a dual-microphone speech enhancement. The interchannel characteristic of each NMF basis can be modeled for the method by applying the spatial diversity of speech signals.

The paper is arranged as follows: Section 2 reviews the objective function of standard NMF with β divergence. Section 3 extends it to the dual-microphones system for the NMF basis. Section 4 presents a two-channel speech signal model and details the proposed speech enhancement framework. Section 5 presents simulation results and Section 6 the conclusion.

2. Nonnegative Matrix Factorization with β Divergence

In a single-microphone system, let $\{y(t), t \in \mathrm{R}\}$ be the observed value of one microphone for a specific time duration. By applying the short-time Fourier transform (STFT) to $y(t)$, we can obtain a complex matrix $\mathbf{Y} = [y_{ij}] \in \mathrm{C}^{I\times J}$ ($i \in \{1, 2, \dots, I\}$ denotes the number of frequency bins and $j \in \{1, 2, \dots, J\}$ the number of time frames). Using the standard NMF, the amplitude $\mathbf{Z} = |\mathbf{Y}|$ or equivalently $z_{ij} = |y_{ij}|$ is analyzed in [1]. Finally, the NMF-based algorithm is to find a local optimal decomposition, which is defined as

$$\mathbf{Z} \approx \widehat{\mathbf{Z}} = \mathbf{TV}, \tag{1}$$

where $\mathbf{T} = [t_{ik}] \in \mathrm{R}_+^{I\times K}$ is a basis matrix, $\mathbf{V} = [v_{kj}] \in \mathrm{R}_+^{K\times J}$ is a coefficient matrix, and K is the number of basis vectors.

For the sake of seeking for two nonnegative matrices such that the difference between $\mathbf{Z}$ and the product $\mathbf{TV}$ is minimized, define a measure function D to obtain the optimal decomposition

$$\arg\min_{\mathbf{T},\mathbf{V}} D\left(\mathbf{Z}\middle\|\widehat{\mathbf{Z}}\right), \quad s.t.\ \mathbf{T} \geq 0, \mathbf{V} \geq 0, \tag{2}$$

where $D(\mathbf{Z}\|\widehat{\mathbf{Z}})$ denotes the error divergence function between the observed data $\mathbf{Z}$ and the reconstructed data $\widehat{\mathbf{Z}}$. The different probability models can be derived by (2), and then different types of cost functions are obtained by the maximum likelihood. Selecting an appropriate objective function is the key in formulating the NMF algorithm. Here, the objective function is derived by using a parametric divergence measure, namely, the β divergence [20]

$$D_\beta(\mathbf{Z}|\mathbf{T},\mathbf{V}) = \sum_{ij}\left(z_{ij}\frac{z_{ij}^{\beta-1} - [\mathbf{TV}]_{ij}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{TV}]_{ij}^{\beta-1}\frac{[\mathbf{TV}]_{ij} - z_{ij}}{\beta}\right) \cdot \beta \in \backslash\{0,1\}, \tag{3}$$

where β reflects the reconstruction penalty. The selection of parameter β depends on the statistical distribution characters and requires prior knowledge. When $\beta = 2$, the result is shown as the squared Euclidean distance (ED); when $\beta \to 1$, the result is approximately equal to the Kullback-Leibler (KL) divergence; and when $\beta \to 0$, the result is nearly equal to Itakura-Saito divergence.

$\mathbf{T}$ and $\mathbf{V}$ are expressed by applying multiplicative iterative updating rules as described in [21]; the update rules are given as

$$\mathbf{T} \leftarrow \mathbf{T} \otimes \frac{\left(\mathbf{Z}/(\mathbf{TV})^{2-\beta}\right)\mathbf{V}^\top}{(\mathbf{TV})^{\beta-1}\mathbf{V}^\top}, \tag{4}$$

$$\mathbf{V} \leftarrow \mathbf{V} \otimes \frac{\mathbf{T}^\top\left(\mathbf{Z}/(\mathbf{TV})^{2-\beta}\right)}{\mathbf{T}^\top(\mathbf{TV})^{\beta-1}}, \tag{5}$$

where the operation $\otimes$ represents an element-wise multiplication, / and the quotient line are performed element-wise division, and the superscript $\top$ is the matrix transpose. As for the initializations of $\mathbf{T}$ and $\mathbf{V}$, positive random numbers are often used.

3. The Dual-Microphone Model for NMF Basis with β Divergence

This section proposes an extension of the standard NMF. Compared with multichannel speech enhancement, dual-channel speech enhancement has advantages in many aspects. Assume that $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ explain the observations of the 1st and 2nd microphones in the time-frequency domain, respectively. In [22], a new interchannel matrix $\mathbf{H}$ is defined, which represents the spatial characteristics between two channels, and they have both common nonnegative matrices $\mathbf{T}$ and $\mathbf{V}$ to model multichannel observations.

3.1. Preprocessing and Modeling. The first is only considering the amplitude observations in the time-frequency domain when we use the standard NMF algorithm for speech enhancement. The observation of the 1st channel is obtained and acted as a reference

$$\mathbf{X}^{(1)} = \mathbf{X}^{(1)} \otimes \left(\frac{\mathbf{X}^{(1)}}{\left| \mathbf{X}^{(1)} \right|} \right)^{*}, \tag{6}$$

where * is the complex conjugate, in order to fully reflect the interchannel characteristic, and then the same is done for the 2nd channel with the expression of

$$\mathbf{X}^{(2)} = \mathbf{X}^{(2)} \otimes \left(\frac{\mathbf{X}^{(1)}}{\left| \mathbf{X}^{(1)} \right|} \right)^{*}. \tag{7}$$

According to the above preprocessing principle, we can find that $\mathbf{X}^{(1)}$ is not only a nonnegative matrix but also a complex matrix. Hence, an accurate modeling for the first channel is designed by using (3), and then an accurate modeling for the second channel is designed by introducing an interchannel matrix $\mathbf{H} = [h_{ik}] \in \mathrm{C}^{I \times K}$, where $\sum_i h_{i:} = 1$ uses random initialization. The interchannel characteristic h_{ik} contains spatial information of the 2nd channel.

3.2. Maximum Likelihood Estimation and Its Cost Function. Using the dual-channel probabilistic model, the likelihood is written as

$$p\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)} | \mathbf{T}, \mathbf{V}, \mathbf{H}\right) \propto p\left(\mathbf{X}^{(1)} | \mathbf{T}, \mathbf{V}\right) p\left(\mathbf{X}^{(2)} | \mathbf{T}, \mathbf{V}, \mathbf{H}\right), \tag{8}$$

where we assume that the data follows the probability distribution. Thus, the maximum negative log-likelihood solution of (8) is represented as

$$\begin{aligned}
&\arg\max_{\mathbf{H}\geq 0, \mathbf{T}\geq 0, \mathbf{V}\geq 0} \log\left(p\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)} | \mathbf{T}, \mathbf{V}, \mathbf{H}\right)\right) \\
&\quad = -\log\left(p\left(\mathbf{X}^{(1)} | \mathbf{T}, \mathbf{V}\right)\right) - \log\left(p\left(\mathbf{X}^{(2)} | \mathbf{T}, \mathbf{V}, \mathbf{H}\right)\right) \\
&\quad = -\sum_{ij} \left(\left[\mathbf{X}^{(1)}\right]_{ij} \frac{\left[\mathbf{X}^{(1)}\right]_{ij}^{\beta-1} - [\mathbf{TV}]_{ij}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{TV}]_{ij}^{\beta-1} \frac{[\mathbf{TV}]_{ij} - \left[\mathbf{X}^{(1)}\right]_{ij}}{\beta} \right) \\
&\qquad - \sum_{ij} \left(\left[\mathbf{X}^{(2)}\right]_{ij} \frac{\left[\mathbf{X}^{(2)}\right]_{ij}^{\beta-1} - [\mathbf{H}\otimes\mathbf{TV}]_{ij}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{H}\otimes\mathbf{TV}]_{ij}^{\beta-1} \frac{[\mathbf{H}\otimes\mathbf{TV}]_{ij} - \left[\mathbf{X}^{(2)}\right]_{ij}}{\beta} \right) \\
&\quad \stackrel{c}{=} D_{\beta}^{(1)}\left(\mathbf{X}^{(1)} \middle\| \mathbf{TV}\right) + D_{\beta}^{(2)}\left(\mathbf{X}^{(2)} \middle\| \mathbf{H}\otimes\mathbf{TV}\right),
\end{aligned} \tag{9}$$

where $=^{c}$ represents equality up to irrelevant constant terms. The former term is explained in Section 2, and now the latter term is given by

$$\begin{aligned}
&D_{\beta}^{(2)}\left(\mathbf{X}^{(2)} \middle\| \mathbf{H}\otimes\mathbf{TV}\right) \\
&\quad = \sum_{ij} \left(\left[\mathbf{X}^{(2)}\right]_{ij} \frac{\left[\mathbf{X}^{(2)}\right]_{ij}^{\beta-1} - [\mathbf{H}\otimes\mathbf{TV}]_{ij}^{\beta-1}}{\beta(\beta-1)} + [\mathbf{H}\otimes\mathbf{TV}]_{ij}^{\beta-1} \frac{[\mathbf{H}\otimes\mathbf{TV}]_{ij} - \left[\mathbf{X}^{(2)}\right]_{ij}}{\beta} \right).
\end{aligned} \tag{10}$$

The gradient is expressed with respect to α of the cost function $\nabla_{\alpha} D$ (The subscript of the cost function of the 2nd term is omitted for convenience, where $\alpha \in \{\mathbf{H}, \mathbf{T}, \mathbf{V}\}$ denotes a variable.) as the difference of two positive terms $\nabla_{\alpha}^{-} D$ and $\nabla_{\alpha}^{+} D$ as

$$\nabla_{\alpha} D = \nabla_{\alpha}^{+} D - \nabla_{\alpha}^{-} D. \tag{11}$$

The solution can be expressed by applying general heuristic multiplicative update rules as

$$\alpha \leftarrow \alpha \otimes \frac{\nabla_{\alpha}^{-} D}{\nabla_{\alpha}^{+} D}. \tag{12}$$

The derivative of the cost function of the 2nd term in (10) with respect to $\mathbf{H}$, $\mathbf{T}$, and $\mathbf{V}$ are shown as

$$\begin{aligned}
\frac{\partial D}{\partial \mathbf{T}} &= -\frac{\mathbf{X}^{(2)}}{(\mathbf{H}\otimes\mathbf{TV})^{2-\beta}} \mathbf{V}^{\top}\mathbf{H}^{\top} 1_{I\times K} + (\mathbf{H}\otimes\mathbf{TV})^{\beta-1} \mathbf{V}^{\top}\mathbf{H}^{\top} 1_{I\times K}, \\
\frac{\partial D}{\partial \mathbf{V}} &= -\mathbf{T}^{\top} \otimes \mathbf{H}^{\top} \frac{\mathbf{X}^{(2)}}{(\mathbf{H}\otimes\mathbf{TV})^{2-\beta}} + \mathbf{T}^{\top} \otimes \mathbf{H}^{\top} (\mathbf{H}\otimes\mathbf{TV})^{\beta-1}, \\
\frac{\partial D}{\partial \mathbf{H}} &= -\frac{\mathbf{X}^{(2)}}{(\mathbf{H}\otimes\mathbf{TV})^{2-\beta}} \mathbf{V}^{\top}\mathbf{T}^{\top} 1_{I\times K} + (\mathbf{H}\otimes\mathbf{TV})^{\beta-1} \mathbf{V}^{\top}\mathbf{T}^{\top} 1_{I\times K}.
\end{aligned} \tag{13}$$

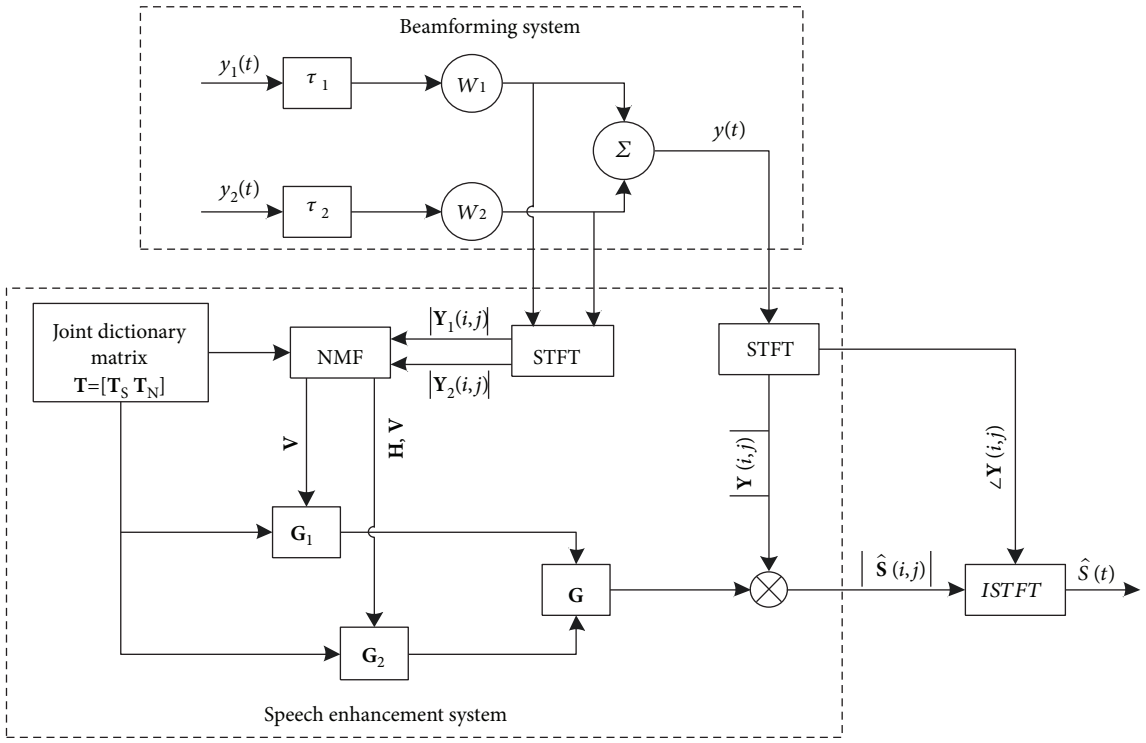

Figure 1: The block diagram of the proposed algorithm.

This leads to the following updating rules by using the cost function of (9), and then the complex matrices and nonnegative matrices $\mathbf{T}$ and $\mathbf{V}$ are estimated by using the update rule of [21]; we can obtain the gradient of the cost function which is rewritten as

$$\mathbf{T} \leftarrow \mathbf{T} \otimes \frac{\left|\left(\mathbf{X}^{(1)}/(\mathbf{TV})^{2-\beta}\right)\mathbf{V}^{\top} + \left(\mathbf{X}^{(2)}/(\mathbf{H}\otimes\mathbf{TV})^{2-\beta}\right)(\mathbf{HV})^{\top}1_{I\times K}\right|}{(\mathbf{TV})^{\beta-1}\mathbf{V}^{\top} + \left|(\mathbf{H}\otimes\mathbf{TV})^{\beta-1}(\mathbf{HV})^{\top}1_{I\times K}\right|}, \tag{14}$$

$$\mathbf{V} \leftarrow \mathbf{V} \otimes \frac{\left|\mathbf{T}^{\top}\left(\mathbf{X}^{(1)}/(\mathbf{TV})^{2-\beta}\right) + (\mathbf{H}\otimes\mathbf{T})^{\top}\left(\mathbf{X}^{(2)}/(\mathbf{H}\otimes\mathbf{TV})^{2-\beta}\right)\right|}{\mathbf{T}^{\top}(\mathbf{TV})^{\beta-1} + \left|(\mathbf{H}\otimes\mathbf{T})^{\top}(\mathbf{H}\otimes\mathbf{TV})^{\beta-1}\right|}, \tag{15}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\left(\mathbf{X}^{(2)}/(\mathbf{H}\otimes\mathbf{TV})^{2-\beta}\right)(\mathbf{TV})^{\top}1_{I\times K}}{(\mathbf{H}\otimes\mathbf{TV})^{\beta-1}(\mathbf{TV})^{\top}1_{I\times K}}, \tag{16}$$

where $1_{I\times K}$ is a $I \times K$ matrix of ones. As is shown by Formulas (14), (15), and (16) derived above, it can reduce to single-channel counterparts (4) and (5) if only one microphone is used, and the interchannel matrix $\mathbf{H}$ is a unit matrix.

4. Proposed NMF-Based Speech Enhancement Algorithm

Assuming that dual microphones are set up in a complex environment, and the noise and target speech signals are spatially separated. Let $s(t)$ be the target speech, and then the noisy speech signal of the mth microphone $y_m(t)$ can be defined with the expression of

$$y_m(t) = a_m(t) * s(t) + n_m(t) \quad (m = 1, 2), \tag{17}$$

where $*$ is the operator of conjunction, m is the microphone index, t is the sample index, and $a_m(t)$ and $n_m(t)$ represent room reverberation and noise, corresponding to the mth microphone, respectively. The block diagram of the proposed algorithm is described in Figure 1, which mainly includes two parts: the training stage and the enhancement stage.

4.1. Training Stage. By applying STFT, (17) can be represented in the frequency domain

$$\mathbf{Y}_m(i,j) = \mathbf{A}_m(i,j)\mathbf{S}(i,j) + \mathbf{N}_m(i,j). \tag{18}$$

At the stage of training, we chose the magnitude spectra of the clean speech and noise from the database as the data matrix for the β-NMF processing to produce the basis matrices $\mathbf{T}_S$ and $\mathbf{T}_N$, by using multiplicative iterative updating rules given in (4) and (5) to the corresponding training data, separately. The basis matrices are saved as a joint dictionary matrix, namely, $\mathbf{T} = [\mathbf{T}_S\ \mathbf{T}_N]$, and as a priori information of the enhancement stage.

4.2. Enhancement Stage. The proposed enhancement stage consists of three parts, firstly beamforming, secondly signal gain estimation, and finally speech signal reconstruction, which are explained in the next section.

4.2.1. Beamforming. Beamforming is one of the most popular algorithms which are the basis of microphone array speech enhancement. In general, the most common fixed beamformers are the delay-and-sum and superdirective beamformers. In the paper, we can use the delay-and-sum as

$$y(t) = \sum_{i=1}^{m} w_i y_i(t - \tau_i), \tag{19}$$

where w_i represents weight and τ_i denotes the time delay compensation obtained by estimation.

4.2.2. Signal Gain Estimation. Firstly, two noisy speeches $y_1(t)$ and $y_2(t)$ are used as input signals of this stage after delay compensation, and then we obtain the magnitude spectra of noise by applying STFT, namely, $|\mathbf{Y}_1|$ and $|\mathbf{Y}_2|$. Next, they are factorized via the extension of NMF with the fixed joint dictionary matrix $\mathbf{T} = [\mathbf{T}_S\, \mathbf{T}_N]$, which is just derived from the training stage via using the update rules given in (15) and (16). Accordingly, the magnitude spectra can be approximately decomposed into an interchannel matrix $\mathbf{H} = [\mathbf{H}_S\, \mathbf{H}_N]$ and a coefficient matrix$\mathbf{V} = [\mathbf{V}_S\, \mathbf{V}_N]$.

(1) Based on the above results, we can obtain the 1st channel (as reference channel) gain function G_1 which is defined with the expression of

$$\mathbf{G}_1 = (\mathbf{T}_S \mathbf{V}_S) \cdot / (\mathbf{T}\mathbf{V}). \tag{20}$$

(2) By using the interchannel matrix $\mathbf{H}$, we can also obtain the 2nd channel gain function $\mathbf{G}_2$ which is represented as

$$\mathbf{G}_2 = (\mathbf{H}_S \otimes \mathbf{T}_S \mathbf{V}_S) \cdot / (\mathbf{H} \otimes \mathbf{T}\mathbf{V}). \tag{21}$$

(3) The final gain function $\mathbf{G}$ can be obtained for this work by Formulas (20) and (21). Furthermore, the gain estimation is achieved by

$$\mathbf{G} = \mathbf{G}_1 \otimes \mathbf{G}_2, \tag{22}$$

where / is the element-wise division.

4.2.3. NMF-Based Signal Reconstruction. This stage is similar to a Wiener filtering process; the gain function $\mathbf{G}$ is obtained by using (22) and acts as a Wiener filter. First, we obtain the magnitude spectra of $y(t)$ by using STFT, namely, $|\mathbf{Y}|$, and then the magnitude spectra of the enhanced speech $|\widehat{\mathbf{S}}|$ is approximately represented by

$$\left|\widehat{\mathbf{S}}\right| = \mathbf{G} \otimes |\mathbf{Y}|. \tag{23}$$

Therein, the enhanced speech waveform $\hat{s}(t)$ is estimated by using the inverse STFT.

5. Experimental Results

In this section, we perform an experiment to evaluate the performance of these methods with respect to quality and intelligibility. We compare the proposed method with the speech enhancement algorithm coherence based in [18] and the standard NMF in terms of performance. The performance of the proposed method is evaluated using a perceptual evaluation of speech quality (PESQ) [23], source-to-distortion (SDR) [24], and segmental SNR (SSNR) which are used as the objective measures, where a higher value indicates a better result.

5.1. Experimental Setup. The selection of the clean speech and the noise is the TIMIT database [25] and the NOISEX database [26], where using downsampling we can adjust the sampling rate of all signals to 8 kHz. In this study, the training for the clean speech contains 20 sentences (60 seconds) pronounced by 10 males and 10 females. Each of the test speech signals for the speech enhancement work is one sentence. We select two background noises in the paper: the Hfchannel and Factory1 noises. Besides, training data and test data in the experiment are disjoint. For the proposed framework, the window function, the applied frame size, and the frame shift are Hamming window, 512 samples and 128 samples, respectively. According to the standard decision of $K \leq IJ/(I + J)$, assuming the clean speech and noise basis vectors, K is set to 30, respectively, and let the maximum iteration number be equal to 50. The two microphones with a 4 cm spacing distance picked up noisy speech signals which were generated by convolving the target and noise sources with a set of HRTFs measured inside a mildly reverberant room (T60 $\approx$ 220 ms) with dimensions $4.3 \times 3.8 \times 2.3\,\mathrm{m}^3$ (length $\times$ width $\times$ height), by adding the noise to the clean testing speech to generate the noisy signals at four signal-to-noise ratios (SNRs): −10, −5, 0, and 5 dB. The distance between the target source and the midpoint of the two microphones is set to 1.2 m. The direction of arrival (DOA) was chosen, respectively, according to $\theta \propto \{0^\circ, 30^\circ, 60^\circ, 90^\circ, 120^\circ, 150^\circ, 180^\circ\}$. The squared Euclidean distance $\beta = 2$ is used for simplicity.

Figure 2 shows the results of the PESQ, SDR, and SSNR metric with the variation of the θ values while input SNR is set to 0 dB under the Factory1 noise condition. As can be seen, DOA of the target source has little influence on the PESQ metric for these methods, but a great effect on the other metrics. In the following experiments we ultimately chose $\theta = 60^\circ$ for consistency and simplicity. Figure 2 also indicates that the proposed method can suppress not only the background noise level effectively but also comparability when the angle of the source is set to 60°.

5.2. Speech Quality and Intelligibility Evaluation. To investigate the achievable gain estimation performance, we chose two background noises in a complex environment: the

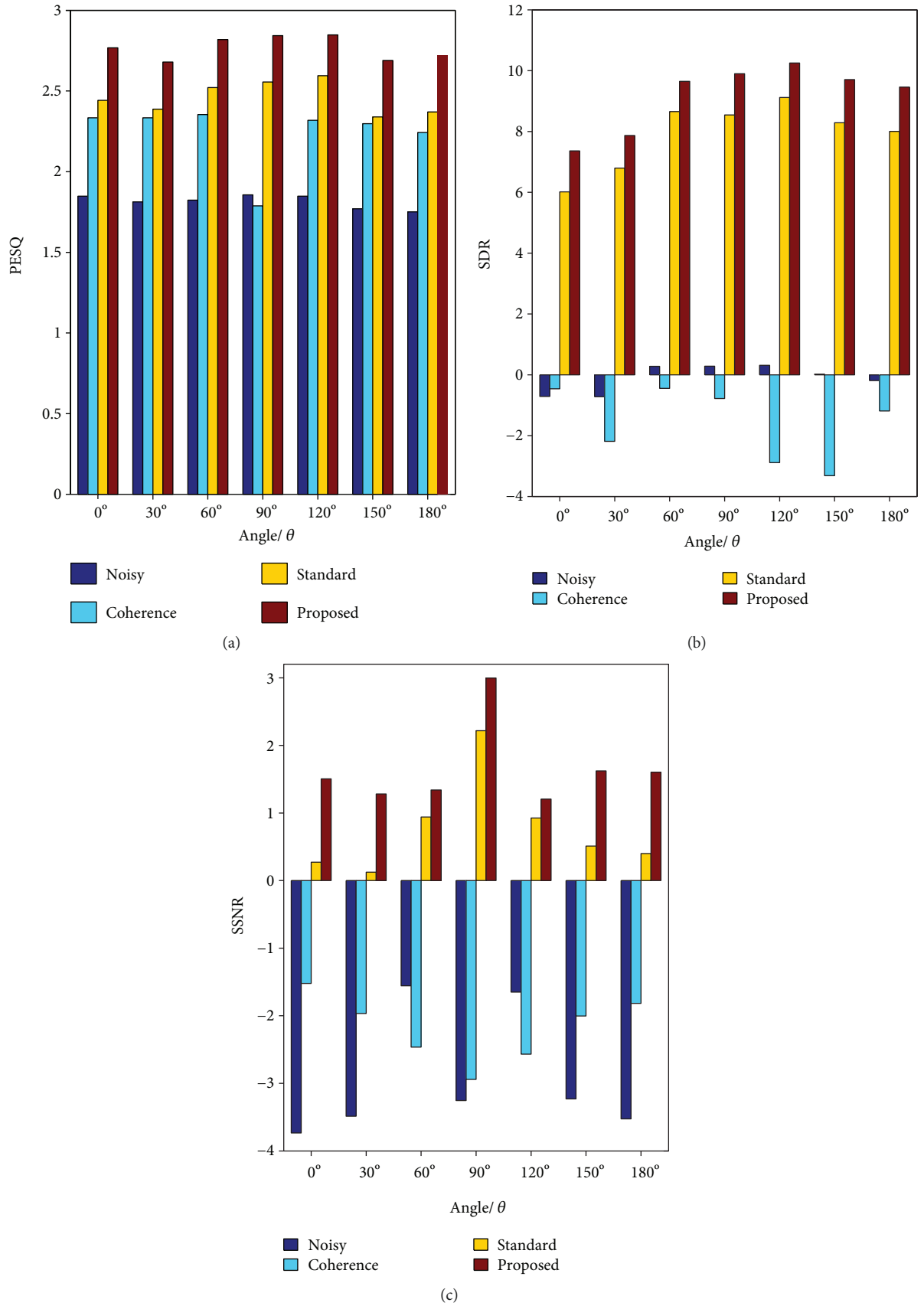

Figure 2: PESQ, SDR, and SSNR values of the enhanced speech from Factory1 noise at 0 dB input SNR.

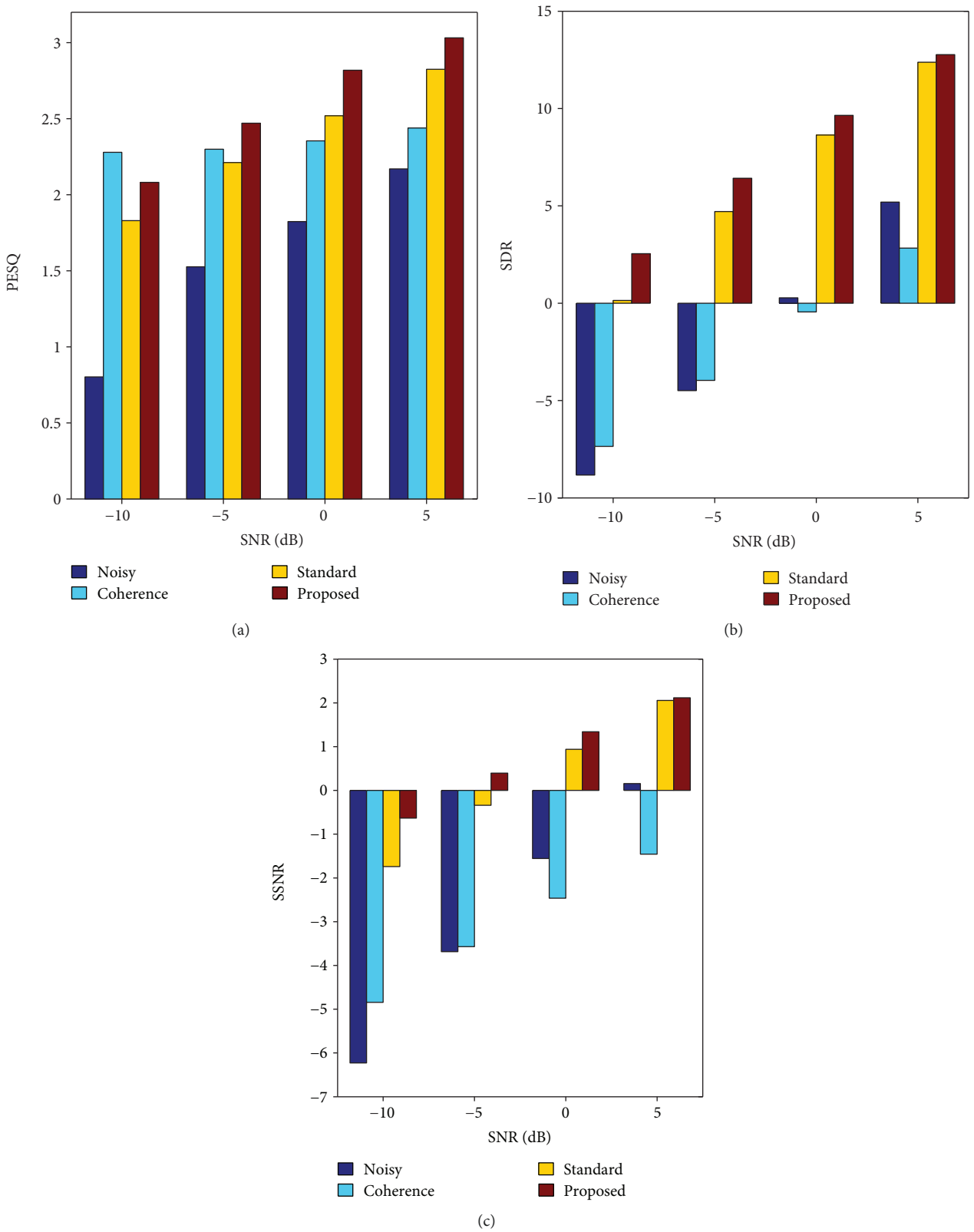

Figure 3: PESQ, SDR, and SSNR scores in Factory1 noise scenarios.

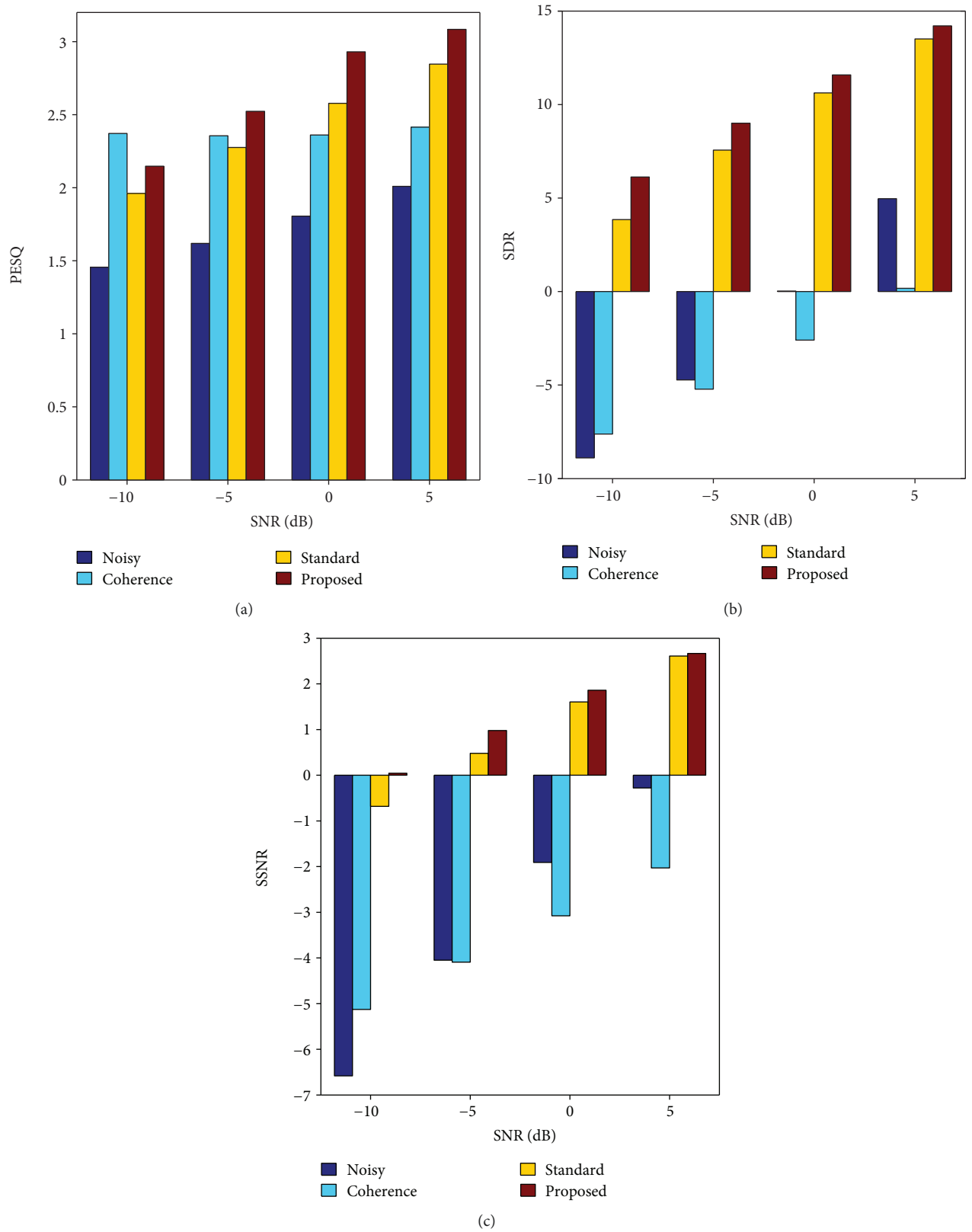

Figure 4: PESQ, SDR, and SSNR scores in Hfchannel noise scenarios.

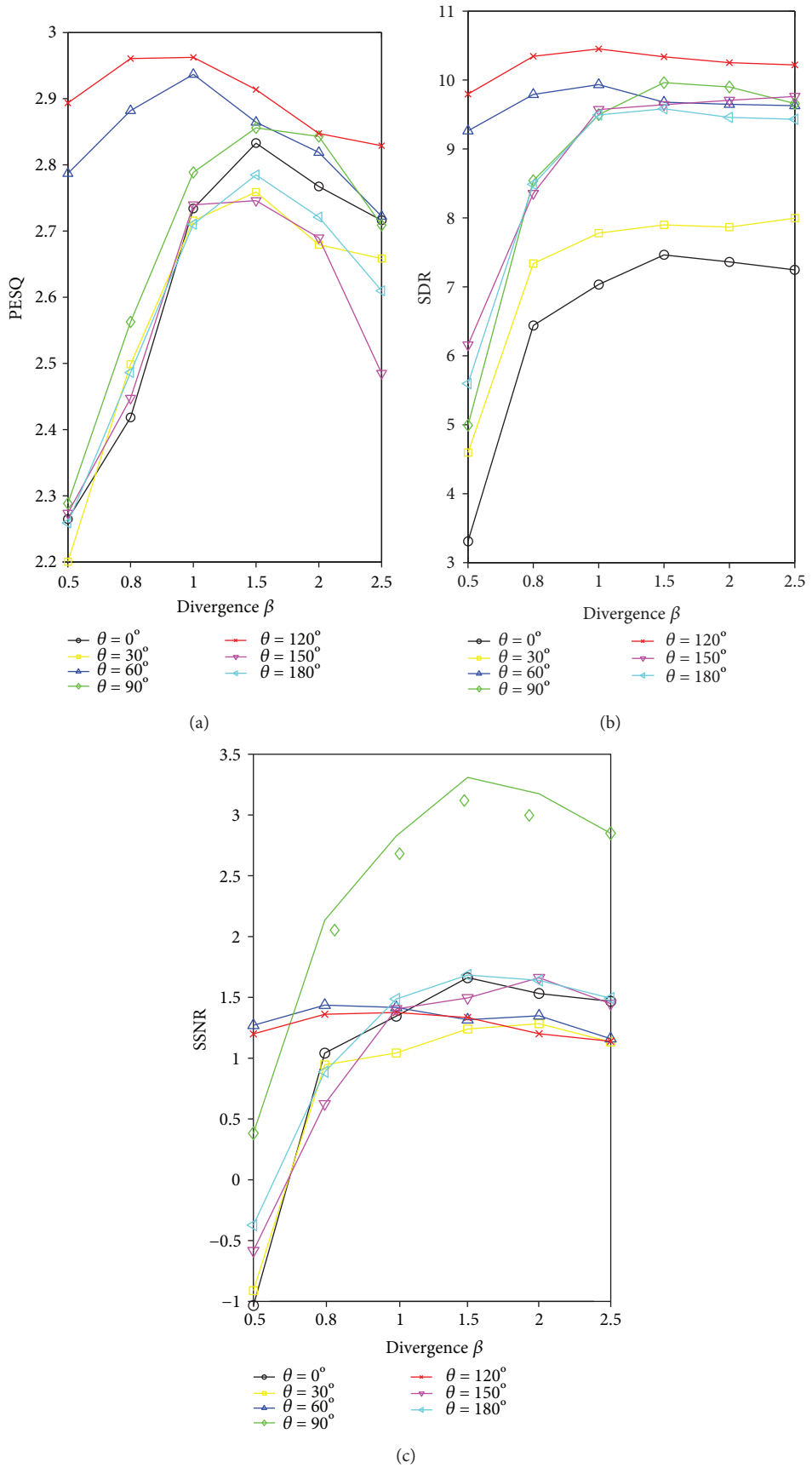

Figure 5: PESQ, SDR, and SSNR values of the enhanced speech from Factory1 noise at 0 dB input SNR.

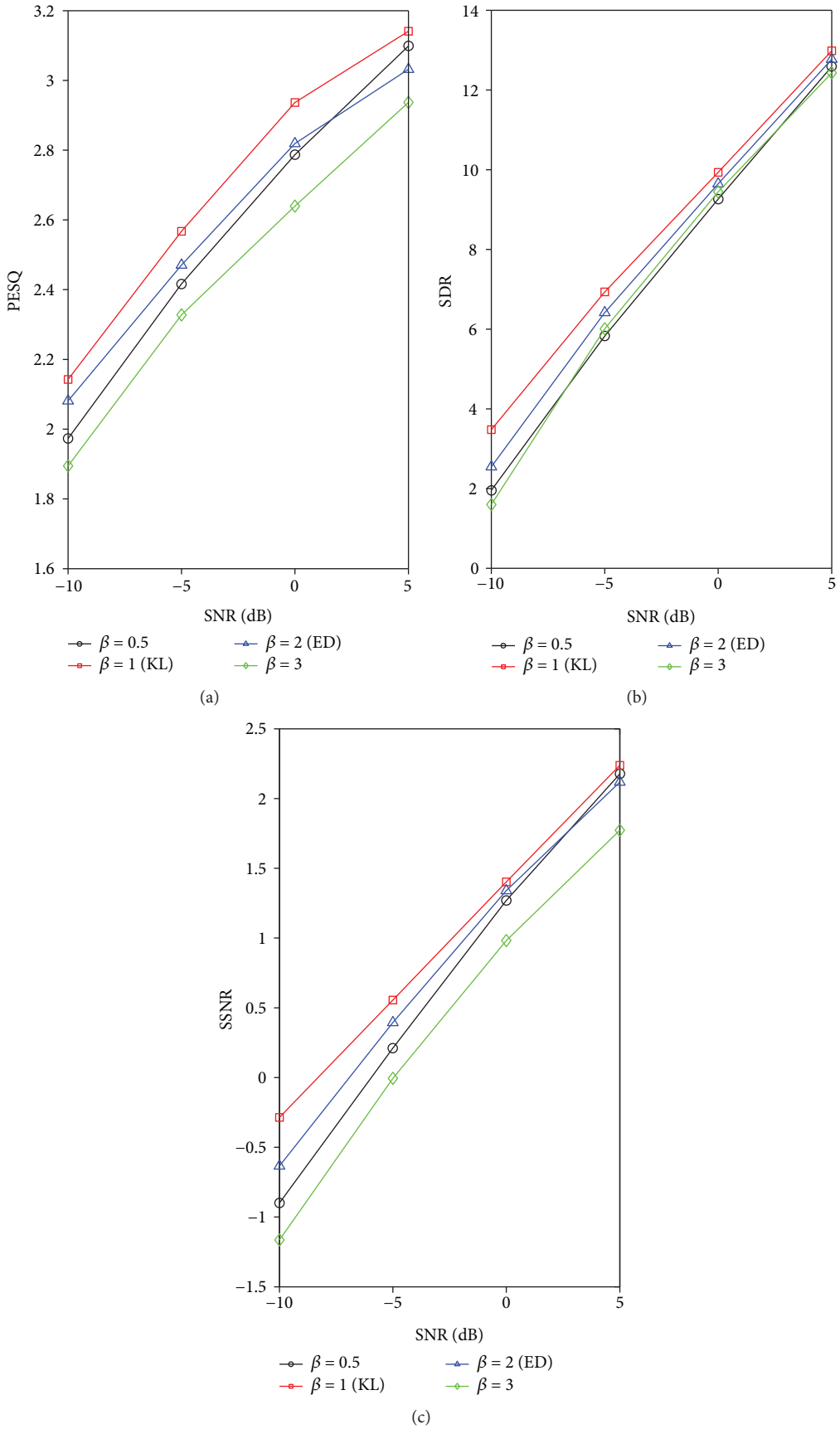

Figure 6: PESQ, SDR, and SSNR scores for the different divergence values from Factory1 noise at different SNR levels.

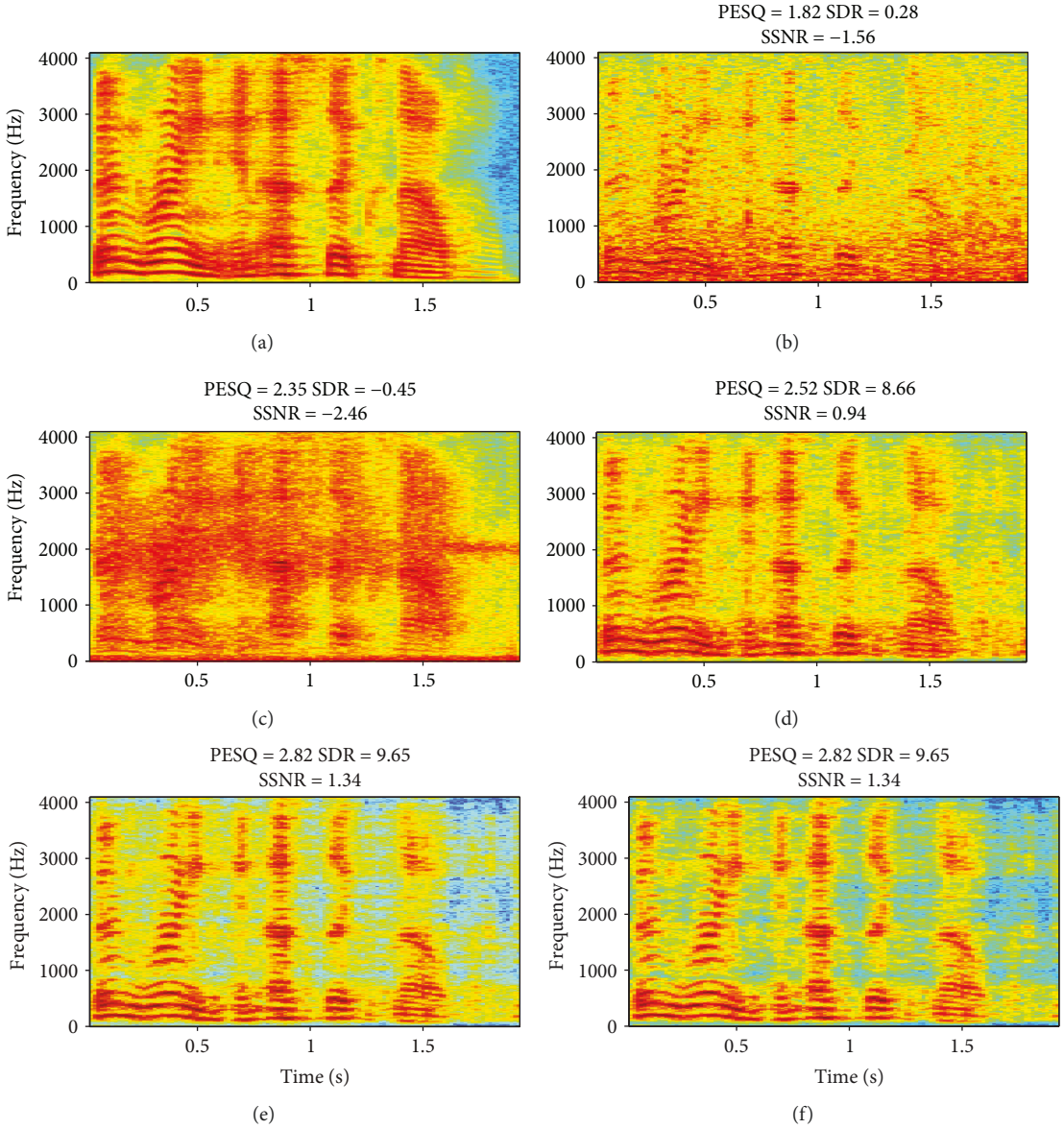

Figure 7: The spectrogram of (a) the clean speech signal, (b) the noisy signal obtained via the delay-and-sum, (c) the noisy speech enhanced by the coherence-based method, (d) the noisy speech enhanced by the standard NMF, (e) the noisy speech enhanced by the proposed method ($\beta = 2$), and (f) the noisy speech enhanced by the proposed method ($\beta = 1$).

Hfchannel and Factory1 noises. Figures 3 and 4 give some results between the noisy signal and the enhanced signal with the different methods where parameter β is set to 2.

From Figures 3 and 4, we can find that the proposed method leads to higher PESQ, SDR, and SSNR scores than the coherence-based method [18] and the standard NMF algorithm [7] in almost all cases, which reveals that this algorithm could also prominently improve both the quality and intelligibility of speech signals. The analysis of PESQ scores shows that the method in [18] has good stability and scarcely affected by SNR, but with lower performance corresponding with other metrics. The latter tends to much distortion. It can be also seen that the advantage of these algorithms becomes less evident with SNR increased. Compared with the coherence-based method, the proposed methods based on NMF still attain improvement in objective measures.

Figure 5 shows the results of the PESQ, SDR, and SSNR metric for the change in the incidence of angle under different β parameter conditions. From Figure 5, we can see that the change in the incidence angle has a significant influence on the performance of the proposed method. Based on the observation of SSNR values, for $\theta = 90^{\circ}$, the proposed method has better scores, but at the expense of speech quality and intelligibility. For $\theta = 120^{\circ}$, we can get an optimum solution of the angle of incidence. Besides, by comparing analysis of the PESQ and SDR values with different β parameters, it is found from simulating results that parameter β has great influence on speech quality more than speech intelligibility

under the same angle conditions. For $\beta = 1$, it not only can guarantee the accuracy of the proposed method but also can suppress the background noise level effectively without introducing much distortion.

The simulation experiment shows the performance of the proposed algorithm with the different divergence and noises in Figure 6. We can find that PESQ, SDR, and SSNR scores become better and better when the SNR increases under the same conditions. For the same SNR conditions, an optimum solution with $\beta \to 1$ can be obtained where divergence tends to the KL divergence. In fact, this observation can be interpreted that the proposed method based on the KL divergence can improve speech quality and intelligibility better than other parameter properties. Besides, this result indicates that under the same conditions the proposed method has obvious improvement of PESQ, SDR, and SSNR scores, especially at low SNR. Hence, the proposed method can provide aural quality and noisy speech intelligibility.

5.3. Signal Spectrogram. By comparing the color depth of speech spectrograms, we can obtain the structural characteristics of residual noise and speech distortion. The spectrograms of the different signals are presented in Figure 7. It reflects that the performance of this method is better than that of those methods. Comparing to them for Factory1 noise while input SNR is set to 0 dB, it is easy to see that the proposed method based on NMF exhibits lower speech distortion and residues than the traditional coherence-based method and the standard NMF method do in the restored spectrogram.

Besides, the β parameter influences the SDR scores. In the paper, the method based on the KL divergence is shown to be superior to the squared Euclidean distance in speech enhancement capability. Finally, the proposed speech enhancement framework based on KL-NMF provides the significant improvement in both quality and intelligibility justified by the higher evaluation scores.

6. Conclusions

We propose a dual-microphone speech enhancement framework based on β-NMF in the paper. This method extends single-microphone speech enhancement based on NMF by introducing the interchannel matrix to the cost function. It can express the interchannel characteristic of each NMF basis very well by applying a priori information. The results of the experiment express that the presented method is effective in nonstationary and low SNR conditions.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Scientific Public Welfare Research of Liaoning Province (20170056), National Natural Science Foundation of China (no. 60901063), the Program for Liaoning Innovative Research Team in University under Grant LT2016006, and the Program for Distinguished Professor of Liaoning Province.

References

[1] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[2] W. S. Chen, Y. Zhao, B. Pan, and B. Chen, "Supervised kernel nonnegative matrix factorization for face recognition," *Neurocomputing*, vol. 205, pp. 165–181, 2016.

[3] M. Han and B. Liu, "Ensemble of extreme learning machine for remote sensing image classification," *Neurocomputing*, vol. 149, pp. 65–70, 2015.

[4] M. Babaee, S. Tsoukalas, G. Rigoll, and M. Datcu, "Immersive visualization of visual data using nonnegative matrix factorization," *Neurocomputing*, vol. 173, pp. 245–255, 2016.

[5] Z. Xia, X. Feng, J. Peng, J. Wu, and J. Fan, "A regularized optimization framework for tag completion and image retrieval," *Neurocomputing*, vol. 147, pp. 500–508, 2015.

[6] A. Lefèvre, F. Glineur, and P. A. Absil, "A convex formulation for informed source separation in the single channel setting," *Neurocomputing*, vol. 141, no. 141, pp. 26–36, 2014.

[7] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[8] C. M. Nelke, N. Chatlani, C. Beaugeant, and P. Vary, "Single microphone wind noise PSD estimation using signal centroids," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7063–7067, Florence, Italy, May 2014.

[9] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17–20, Prague, Czech Republic, May 2011.

[10] H. Chung, E. Plourde, and B. Champagne, "Regularized NMF based speech enhancement with spectral componenets modeled by Gaussian mixtures," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Reims, France, September 2014.

[11] Y. Liu, Y. Liao, L. Tang, F. Tang, and W. Liu, "General subspace constrained non-negative matrix factorization for data representation," *Neurocomputing*, vol. 173, pp. 224–232, 2016.

[12] Q. He, X. Jin, C. Du, F. Zhuang, and Z. Shi, "Clustering in extreme learning machine feature space," *Neurocomputing*, vol. 128, pp. 88–95, 2014.

[13] C. Yang, K. Huang, H. Cheng, Y. Li, and C.-Y. Su, "Haptic identification by ELM-controlled uncertain manipulator," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 8, pp. 2398–2409, 2017.

[14] C. Yang, Y. Jiang, W. He, J. Na, Z. Li, and B. Xu, "Adaptive parameter estimation and control design for robot

manipulators with finite-time convergence," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 10, pp. 8112–8123, 2018.

[15] L. Ma and D. P. Li, "Adaptive neural networks control using Barrier Lyapunov functions for DC motor system with time-varying state constraints," *Complexity*, vol. 2018, Article ID 5082401, 9 pages, 2018.

[16] S. M. Lu, D. P. Li, and Y. J. Liu, "Adaptive neural network control for uncertain time-varying state constrained robotics systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, no. 99, pp. 1–8, 2017.

[17] K. K. Sunnydayal and S. Cruces, "An iterative posterior NMF method for speech enhancement in the presence of additive Gaussian noise," *Neurocomputing*, vol. 230, pp. 312–315, 2017.

[18] N. Yousefian and P. C. Loizou, "A dual-microphone speech enhancement algorithm based on the coherence function," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 599–609, 2012.

[19] W. Nabi, N. Aloui, and A. Cherif, "Speech enhancement in dual-microphone mobile phones using Kalman filter," *Applied Acoustics*, vol. 109, pp. 1–4, 2016.

[20] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2010.

[21] P. D. O'Grady and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, vol. 72, no. 1-3, pp. 88–101, 2008.

[22] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Formulations and algorithms for multichannel complex NMF," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 229–232, Prague, Czech Republic, May 2011.

[23] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] C. Lopes and F. Perdigao, *Phone Recognition on TIMIT Database*, Speech Technologies, 2011.

[26] A. Varga and H. J. M. Steeneken, "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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.