DOI: 10.7641/CTA.2017.60600

多重约束非负矩阵分解的非平稳噪声语音增强

邹月娴[†], 刘诗涵, 王迪松

(北京大学 信息工程学院 现代信号与数据处理实验室, 广东 深圳 518055)

摘要: 低信噪比非稳态噪声环境中的语音增强仍是一个开放且具有挑战性的任务. 为了提高传统的基于非负矩阵分解(nonnegative matrix factorization, NMF)的语音增强算法性能, 同时考虑到语音信号的时频稀疏特性和非稳态噪声信号的低秩特性, 本文提出了一种基于多重约束的非负矩阵分解语音增强算法(multi-constraint nonnegative matrix factorization speech enhancement, MC–NMFSE). 在训练阶段, 采用干净语音训练数据集和噪声训练数据集分别构建语音字典和噪声字典. 在语音增强阶段, 在非负矩阵分解目标函数中增加语音分量的稀疏性约束和噪声信号的低秩性约束条件, MC–NMFSE能够更好地从带噪语音中获得语音分量的表示, 从而提高语音增强效果. 通过实验表明, 在大量不同非平稳噪声条件和不同信噪比条件下, 与传统的基于NMF的语音增强方法相比, MC–NMFSE能获得较低的语音失真和更好的非稳态噪声抑制能力.

关键词: 语音增强; 低秩约束; 稀疏约束; 非负矩阵分解; 非稳态噪声

中图分类号: TN912.35 **文献标识码:** A

Enhancing speech corrupted by nonstationary noise using nonnegative matrix factorization with multiple constraints

ZOU Yue-xian[†], LIU Shi-han, WANG Di-song

(Advanced Data & Signal Processing Laboratory, School of Electronic and Computer Engineering,
Peking University, Shenzhen Guangdong 518055, China)

Abstract: The enhancement of speech corrupted by nonstationary noises under low signal-to-noise ratio (SNR) conditions is remaining open and still a very challenging task. To improve the traditional nonnegative matrix factorization (NMF) based speech enhancement, jointly taking the speech sparsity property in time-frequency domain and the low-rank property of nonstationary noise into account, a termed multi-constraint NMF speech enhancement method (MC–NMFSE) is developed. Essentially, in training stage, the speech and noise dictionaries have been constructed by using speech and noise training sets, respectively. In the speech enhancement stage, multi-constraint NMF method is adopted where the data matrix is factorized into two nonnegative sub-matrices with the sparsity and low rank constraints to guarantee the good representation of the speech components from their corrupted version by nonstationary noise. Compared with the traditional NMF speech enhancement method (NMF–SpEnM) and MC–NMFSE, intensive experiments under different nonstationary noise conditions and different signal-to-noise ratios have been carried out to evaluate their performance. Experimental results demonstrate that MC–NMFSE has lower speech distortion and better capability to suppress nonstationary noises.

Key words: speech enhancement; low-rank; sparsity; nonnegative matrix factorization; nonstationary noise

1 Introduction

Enhancing speech from a degraded signal recording is an important task in many speech applications, such as hearing aids, speech recognition, speaker verification/identification, speech emotion classification and so on. Various speech enhancement algorithms, e.g., statistical spectral subtraction (SS)[1–2], the minimum mean square error (MMSE)[3–5] have been proposed to enhance the speech corrupted by stationary or quasi-stationary noise. There are few research outcomes have been reported to deal with speech enhancement with non-stationary noise conditions[6]. Moreover, it is also noted that the traditional speech enhancement methods have limited capability to suppress nonstationary noise, especially when signal-to-noise (SNR) is low. Recently, there are some research developments focus on using deep neural network (DNN)[7] and nonnegative matrix factorization (NMF)[8–10] to enhance the speech quality. Obviously, DNN has been well-known being a deep multiple-layer architecture with a large number of parameters to tune. As a result, a large training data is needed for well train DNN model[11].

Received 11 August 2016; accepted 30 March 2017.

[†]Corresponding author. E-mail: zouyx@pkusz.edu.cn; Tel: +86 755-26032016.

Recommended by Associate Editor YU Zhu-liang.

Supported by National Natural Science Foundation of China (61271309) and Shenzhen Science Research Program (CXZZ20140509093608290).

Compared with DNN-based speech enhancement methods, NMF-based speech enhancement methods (NMF–SpEnM) asks much less training data meanwhile offers a good capacity to represent speech components from the noisy speech[12]. It essentially factorizes the speech data and noise data respectively into two nonnegative sub-matrices, termed as the corresponding dictionary matrix and the encoding matrix in the training phase. And they are then employed to factorize the input noisy speech and determine the enhanced speech[13].

However, after examining the experimental results of the NMF–SpEnM, we found that the performance of the NMF–SpEnM degrades with the decrease of the SNR. The main reason lies on the fact that NMF–SpEnM is developed assuming the subspaces of speech and noise are uncorrelated. When SNR goes down, this assumption is not valid, especially when SNR is lower than 5 dB.

In this study, we strive to improve the performance of NMF–SpEnM under non-stationary noise and low SNR conditions. From signal processing perspective, NMF is a powerful mathematical tool and can be taken to solve real-world problems if there are some application-related domain knowledge. Previous studies reveal that speech has certain sparse property[14] and non-stationary noise shows low-rank property at time-frequency representation[15]. Based on these findings, we are seeking the proper nonnegative matrix factorization method jointly considering the speech sparsity and low rank of non-stationary noise. Specifically, the enforcement of the speech sparsity promotes the effective representation of speech by using few coefficients. Meanwhile, the rank-regularized term enforces the low-rank structure of non-stationary noise. As a result, a novel multi-constrained NMF speech enhancement (MC–NMFSE) algorithm is derived. To evaluate the performance of our proposed MC–NMFSE algorithm, intensive experiments have been carried out. The experimental results also demonstrate the improved speech enhancement performance. The details will be given in Section 3.

The organization of the rest of the paper is as follows. The NMF–SpEnM algorithm and the proposed MC–NMFSE algorithm are presented in Section 2. Section 3 illustrates the experiments and their results, and the conclusion is given in Section 4.

2 Multi-constraint nonnegative matrix factorization speech enhancement approach

To make the presentation clear, the basic principle of NMF, speech enhancement based on NMF (NMF–SpEnM) and our proposed multi-constraint NMF based speech enhancement (MC–NMFSE) algorithm will be presented in details, then we discuss the complexity of MC–NMFSE algorithm.

2.1 Nonnegative matrix factorization (NMF)

In this subsection, the principle of NMF will be given. Essentially, NMF is a matrix factorization technique which factorizes one nonnegative input matrix $\boldsymbol{V}(\boldsymbol{V} \in \mathbb{R}^{m\times n})$ into two matrices $\boldsymbol{W}(\boldsymbol{W} \in \mathbb{R}^{m\times r})$ and $\boldsymbol{H}(\boldsymbol{H} \in \mathbb{R}^{r\times n})$ with nonnegativity constraints, which can be denoted as follows:

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H},\ \boldsymbol{W}, \boldsymbol{H} \geqslant 0, \tag{1}$$

where the matrix $\boldsymbol{W}$ is termed as a dictionary matrix or a basis matrix, while the matrix $\boldsymbol{H}$ is termed as the weighting matrix. $\boldsymbol{r}$ is the rank of factorization, which is chosen to be smaller than m and n. For basis matrix $\boldsymbol{W}$, each column represents a basis vector. For the weighting matrix $\boldsymbol{H}$, each row represents their weight in each column of the input matrix $\boldsymbol{V}$. Alternatively, in terms of column-wise approximation, we can get

$$\boldsymbol{v}_i \approx \boldsymbol{W}\boldsymbol{h}_i, \tag{2}$$

where $\boldsymbol{v}_i$ is the ith column of $\boldsymbol{v}$, $\boldsymbol{h}_i$ is the ith column of $\boldsymbol{H}$. From (2), we can see that each input vector $\boldsymbol{v}_i$ is a linearly representation by the basis matrix and its corresponding weight coefficients.

Mathematically, NMF performs the decomposition by minimizing the following cost function:

$$\begin{aligned}&\min D(\boldsymbol{V}, \boldsymbol{W}\boldsymbol{H}),\\ &\text{s.t. } \boldsymbol{W} > 0,\ \boldsymbol{H} > 0,\end{aligned} \tag{3}$$

where $D(\cdot)$ is a defined as a distance matric which measures the distance between two nonnegative matrices $\boldsymbol{V}$ and $\boldsymbol{W}\boldsymbol{H}$. An iterative approach can be used to obtain the optimal solution of (3). Besides, the initialization is performed using positive random initial conditions for matrices $\boldsymbol{W}$ and $\boldsymbol{H}$. Moreover, the convergence of the process has also been proved. It is easy to see that, from (3), choosing different distance metric $D(\cdot)$ will lead to different matrix factorization, different approximation of matrix $\boldsymbol{V}$, dictionary $\boldsymbol{W}$ and coding matrix $\boldsymbol{H}$. There are several commonly used $D(\cdot)$ functions, such as L_1, L_2, earth mover's distance (EMD), and Kullback-Leibler divergence (KLD). Some research outcomes have shown that the KLD is proved to give a better performance in speech applications compared with other distance metrics[12]. Hence, in this study, we only consider using KLD as the NMF distance matric which has the following expression:

$$\begin{aligned}&D(\boldsymbol{V}||\boldsymbol{W}\boldsymbol{H}) =\\ &\sum_{ij}(\boldsymbol{V}_{ij}\log\frac{\boldsymbol{V}_{ij}}{(\boldsymbol{W}\boldsymbol{H})_{ij}} - \boldsymbol{V}_{ij} + (\boldsymbol{W}\boldsymbol{H})_{ij}).\end{aligned} \tag{4}$$

It is noted that KLD is lower bounded by zero, and vanishes if and only if $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H}$. Minimizing KLD cost function in (4), the multiplicative update rule has been adopted since it gives good compromise between convergence speed and the implementation of Kullback-

Leibler divergence[13]:

$$\boldsymbol{W}_{ia} \leftarrow \boldsymbol{W}_{ia} \frac{\sum\limits_{\mu} \boldsymbol{H}_{a\mu} \boldsymbol{V}_{i\mu} / (\boldsymbol{W}\boldsymbol{H})_{i\mu}}{\sum\limits_{v} \boldsymbol{H}_{av}}, \tag{5}$$

$$\boldsymbol{H}_{a\mu} \leftarrow \boldsymbol{H}_{a\mu} \frac{\sum\limits_{i} \boldsymbol{W}_{ia} \boldsymbol{V}_{i\mu} / (\boldsymbol{W}\boldsymbol{H})_{i\mu}}{\sum\limits_{k} \boldsymbol{W}_{ka}}. \tag{6}$$

2.2 NMF based speech enhancement

Speech enhancement is intended to recover the clean speech $s(t)$ from its corrupted version $x(t)$. Considering additive noise, $x(t)$ is given by

$$x(t) = s(t) + n(t). \tag{7}$$

Taking short-time fourier transform (STFT) on Eq.(7), we obtain the corresponding data model in the time-frequency domain as

$$\boldsymbol{V} \approx \boldsymbol{V}_{\rm s} + \boldsymbol{V}_{\rm n}, \tag{8}$$

where $\boldsymbol{V}_{\rm s}$, $\boldsymbol{V}_{\rm n}$, and $\boldsymbol{V}$ are the spectra magnitude matrix of clean speech, noise, and noisy speech, respectively. It is noted that for the non-stationary noise, $\boldsymbol{V}_{\rm n}$ has a low rank[15]. Obviously, $\boldsymbol{V}_{\rm s}$, $\boldsymbol{V}_{\rm n}$, and $\boldsymbol{V}$ are all nonnegative matrices and the NMF technique can be employed to represent these matrices. Then factorization $\boldsymbol{V}_{\rm s}$, $\boldsymbol{V}_{\rm n}$, and $\boldsymbol{V}$ by (4), we have

$$\boldsymbol{V} = \boldsymbol{W}\boldsymbol{H},\ \boldsymbol{V}_{\rm s} = \boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s},\ \boldsymbol{V}_{\rm n} = \boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n},$$

respectively As a result, (8) can be expressed as follows:

$$\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H} = \boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s} + \boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n} = [\boldsymbol{W}_{\rm s}\ \ \boldsymbol{W}_{\rm n}] \begin{bmatrix} \boldsymbol{H}_{\rm s} \\ \boldsymbol{H}_{\rm n} \end{bmatrix}. \tag{9}$$

According to (9), it is clear that factorizing the input data matrix $\boldsymbol{V}$ by NMF yields both $\boldsymbol{W}$ and $\boldsymbol{H}$. From the last term of (9), we can see that if $\boldsymbol{W}_{\rm s}$ and $\boldsymbol{W}_{\rm n}$ are determined in the training stage for properly representing information of speech and noise, then in the speech enhancement stage, with the input data matrix $\boldsymbol{V}$ and constructed basis matrix $\boldsymbol{W} = [\boldsymbol{W}_{\rm s}\ \ \boldsymbol{W}_{\rm n}]$, the weighting matrix $\boldsymbol{H}$ can be determined by NMF to properly represent the weighting coefficients. As a result, $\boldsymbol{H}_{\rm s}$ and $\boldsymbol{H}_{\rm n}$. can be obtained from $\boldsymbol{H}$. Therefore, the enhanced speech component $\boldsymbol{V}_{\rm s}$ in (8) is reconstructed by

$$\hat{\mathbf{V}}_{\rm s} = \boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}. \tag{10}$$

The block diagram of the NMF based speech enhancement (NMF–SpEnM) algorithm is shown in Fig.1. Clearly, NMF–SpEnM has two stages. In the training stage, $\boldsymbol{W}_{\rm s}$ and $\boldsymbol{W}_{\rm n}$ are trained separately with training speech dataset and training noise dataset, respectively with the same NMF procedure. In the speech enhancement stage, the input data matrix $\boldsymbol{V}$ is factorized by NMF with trained $\boldsymbol{W} = [\boldsymbol{W}_{\rm s}\ \ \boldsymbol{W}_{\rm n}]$ to generate the coefficient matrix $\boldsymbol{H}$. As shown in (9), the coefficient matrices $\boldsymbol{H}_{\rm s}$ and $\boldsymbol{H}_{\rm n}$ can be computed from the generated $\boldsymbol{H}$.

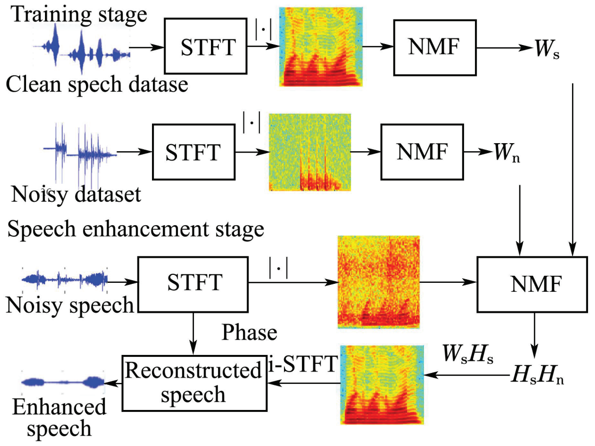


Fig. 1 Block diagram of the proposed MC–NMF speech enhancement method

Researchers have observed that the enhanced speech by (10) may suffer from the speech distortion. In order to improve the intelligibility of the enhanced speech, an indirectly speech enhancement method has been proposed, where an ideal ratio mask (IRM) is typically generated according to the following formulation[16]

$$\boldsymbol{M} = \frac{\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}}{\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s} + \boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n}}. \tag{11}$$

In Eq.(11), $\boldsymbol{W}_{\rm s}$ and $\boldsymbol{W}_{\rm n}$ have been obtained in the training stage. $\boldsymbol{H}_{\rm s}$ and $\boldsymbol{H}_{\rm n}$ are computed in the speech enhancement stage. It is noted that the estimated mask $\boldsymbol{M}$ indicates a ratio of speech component to the received signal at each time-frequency point (τ, f). Analyzing (11) and (8) gives following observations: 1) when no additive noise at (τ, f), $\boldsymbol{M}(\tau, f)$ equals one; 2) when the noise dominates (τ, f), that is $\boldsymbol{V}_{\rm n}$ is much larger than $\boldsymbol{V}_{\rm s}$, or $\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n}$ is much larger than $\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}$, then $\boldsymbol{M}(\tau, f)$ is much smaller than one or approaches zero. As a result, an enhanced spectrogram $\boldsymbol{V}_{\rm enhanced}$ can be computed as

$$\boldsymbol{V}_{\rm enhanced} = \boldsymbol{V} \otimes \boldsymbol{M}. \tag{12}$$

As discussed above, with (12), the signal at speech-dominant (τ, f) is remained almost unchanged since $\boldsymbol{M}(\tau, f)$ equals or approximates to one. However, the signal at noise-dominant (τ, f) is suppressed since $\boldsymbol{M}(\tau, f)$ is approaching zero with the decrease of the SNR. At last, the inverse FFT (i-FFT) is employed to reconstruct the enhanced speech signal in time domain using $\boldsymbol{V}_{\rm enhanced}$ and its phase computed from noisy speech.

2.3 Proposed multi-constraint NMF speech enhancement method

In our previous work[17], the time correlation of speech signal is used to train an expressive speech dictionary using NMF technique. It is encouraged to see the improved speech enhancement performance. In this subsection, we propose a novel multi-constraint NM-

F based speech enhancement (MC–NMFSE) algorithm which considered the characteristics of both speech and noise to guarantee the effectiveness representation of the speech components corrupted by nonstationary noise.

Research shows that NMF tends to return a sparse and part-based representation of speech spectrogram[18]. However, sparsity in NMF occurs as a byproduct due to nonnegativity constraints, rather than being designed objectively. As a result of that, the sparsity is not actually taken full use of. In our study, considering the sparsity of speech spectra magnitude matrix $\boldsymbol{V}_{\rm s}(\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s})$ and the low-rank of the non-stationary noise spectra magnitude matrix $\boldsymbol{V}_{\rm n}(\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n})$, the NMF factorization model can be written as:

$$\begin{aligned}&\min D(\boldsymbol{V},\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}+\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n}),\\&\text{s.t. } \|\boldsymbol{H}_{\rm s}\|_0<k_1,\ \|\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n}\|_*<k_2,\end{aligned}\tag{13}$$

where $D(\cdot)$ represents KL divergence shown in equation (4), $\|\cdot\|_0$ is l_0 norm and $\|\cdot\|_*$ refers to the nuclear norm of the matrix, which is the summation of its singular values, which is a proxy for minimizing the rank of $\boldsymbol{V}_{\rm n}$[15]. k_1 and k_2 are constant parameters to control the degree of sparsity and rank. With the sparsity and low-rank constraints, the model (13) is able to estimate the speech and noise components more accurately. It is clear that the optimal solution in (13) is NP-hard task. Alternatively, the desired Hs can be efficiently computed by minimizing the l_1 norm instead of l_0 norm. Then, using augmented Langrangian technique, (13) can be reformulated as

$$\begin{aligned}&\min\ D(\boldsymbol{V},\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}+\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n})+\\&\quad\lambda_{\rm s}\|\boldsymbol{H}_{\rm s}\|_1+\lambda_{\rm n}\|\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n}\|_*,\end{aligned}\tag{14}$$

where $\lambda_{\rm s}$ and $\lambda_{\rm n}$ are termed as the speech sparsity regularization parameter and the low-rank of noise regularization parameter, respectively. Research in [19] shows that the sum of the Frobenius norms of the nonnegative matrix $\boldsymbol{W}$ and $\boldsymbol{H}$ gives upper bound on the nuclear norm of their product as

$$\|\boldsymbol{W}\boldsymbol{H}\|_*\leqslant\frac{1}{2}\|\boldsymbol{W}\|_{\rm F}^2+\frac{1}{2}\|\boldsymbol{H}\|_{\rm F}^2.\tag{15}$$

Therefore, the cost function shown in (14) can be rewritten as

$$\begin{aligned}&\min\ D(\boldsymbol{V},\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}+\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n})+\\&\quad\lambda_{\rm s}\|\boldsymbol{H}_{\rm s}\|_1+\frac{\lambda_{\rm n}}{2}\|\boldsymbol{H}_{\rm n}\|_{\rm F}^2,\end{aligned}\tag{16}$$

where the $\|\boldsymbol{W}_{\rm n}\|_{\rm F}$ is omitted since it has been pretrained and fixed, and the parameter $\lambda_{\rm s}$ and $\lambda_{\rm n}$ are set following[20], shown as:

$$\lambda_{\rm s}=\sqrt{2N}\sigma,\ \lambda_{\rm n}=\sqrt{2}\sigma,\tag{17}$$

where N represents the number of frames in noisy spectrogram, σ represents mean square error of the noisy spectrogram matrix. Such a setting guarantees that if the noisy speech data $\boldsymbol{V}$ consists of n frames of zero-mean white noise of variance σ^2, then both $\boldsymbol{W}_{\rm s}\boldsymbol{H}_{\rm s}$ and $\boldsymbol{W}_{\rm n}\boldsymbol{H}_{\rm n}$ are zero[19]. Another advantage of this setting is that the regularization parameters are set as data dependent instead of an empirical value. As seen in Eq. (9), $\boldsymbol{H}=[\boldsymbol{H}_{\rm s}^{\rm T}\ \ \boldsymbol{H}_{\rm n}^{\rm T}]^{\rm T}$, assuming that the dimension of $\boldsymbol{H}_{\rm s}$ and $\boldsymbol{H}_{\rm n}$ are $r_{\rm s}\times n$ and $r_{\rm n}\times n$ respectively, then $r=r_{\rm s}+r_{\rm n}$, where r is the number of rows of $\boldsymbol{H}$. Similar to [21], through gradient descent method[13], the following update rules are a good solution of problem (16).

Theorem 1 The cost function in Eq.(16) is nonincreasing under the update rules

$$\boldsymbol{H}_{a\mu}\leftarrow\boldsymbol{H}_{a\mu}\frac{\sum\limits_i[\boldsymbol{W}_{ia}\boldsymbol{V}_{i\mu}/(\boldsymbol{W}\boldsymbol{H})_{i\mu}]}{\sum\limits_k\boldsymbol{W}_{ka}+\lambda_{\rm s}},\ 1\leqslant a\leqslant r_{\rm s},\tag{18}$$

$$\begin{aligned}\boldsymbol{H}_{a\mu}\leftarrow&\ \frac{-\sum\limits_k\boldsymbol{W}_{ka}}{2\lambda_{\rm n}}+\\&\frac{\sqrt{(\sum\limits_k\boldsymbol{W}_{ka})^2+4\lambda_{\rm n}\boldsymbol{H}_{a\mu}\sum\limits_i\boldsymbol{W}_{ia}\boldsymbol{V}_{i\mu}/\boldsymbol{W}\boldsymbol{H}_{i\mu}}}{2\lambda_{\rm n}},\\&a\geqslant r_{\rm s}.\end{aligned}\tag{19}$$

The divergence is invariant under these updates if and only if W and H are at a stationary point of the divergence. To prove Theorem 1, we firstly introduce the auxiliary function and one lemma that has been proved in [13].

Definition 1 $G(h,h')$ is an auxiliary function for $F(h)$ if the conditions hold

$$G(h,h')\geqslant F(h),\ G(h,h)=F(h).\tag{20}$$

Lemma 1 If G is an auxiliary function, then F is nonincreasing under the update

$$h^{t+1}=\arg\min_h G(h,h^t).\tag{21}$$

As discussed in [13], by iterating the update in Eq.(21), a sequence of estimates that converge to a local minimum $h_{\min}=\arg\min_h F(h)$ can be obtained:

$$\begin{aligned}&F(h_{\min})\leqslant\cdots\leqslant F(h^{t+1})\leqslant\\&F(h^t)\leqslant\cdots\leqslant F(h^1)\leqslant F(h^0).\end{aligned}\tag{22}$$

Then we have the following Lemma that can be easily proved as follows[13]:

Lemma 2 Define

$$\begin{aligned}G(h,h^t)=&\sum_i(v_i\log v_i-v_i)+\sum_{ia}W_{ia}h_a-\\&\sum_{ia}v_i\frac{W_{ia}h_a^t}{\sum\limits_b W_{ib}h_b^t}(\log W_{ia}h_a)+\\&\sum_{ia}v_i\frac{W_{ia}h_a^t}{\sum\limits_b W_{ib}h_b^t}(\frac{W_{ia}h_a^t}{\sum\limits_b W_{ib}h_b^t})+\\&\lambda_{\rm s}\sum_{1\leqslant a\leqslant r_{\rm s}}h_a+\frac{\lambda_{\rm n}}{2}\sum_{r_{\rm s}\leqslant a\leqslant r}h_a^2.\end{aligned}\tag{23}$$

This is an auxiliary function for

$$F(h) = \sum_i [v_i \log(\frac{v_i}{\sum_a W_{ia}h_a}) - v_i] + \sum_a W_{ia}h_a + \lambda_{\rm s} \sum_{1\leqslant a\leqslant r_{\rm s}} h_a + \frac{\lambda_{\rm n}}{2} \sum_{r_{\rm s}\leqslant a\leqslant r} h_a^2. \quad (24)$$

At last, we give the proof of Theorem 1 as follows:

Proof The minimum of $G(h, h^t)$ with respect to h is determined by setting the gradient to zero, when $1 \leqslant a \leqslant r_{\rm s}$, we have

$$\frac{\partial G(h,h^t)}{\partial h_a} = -\sum_i v_i \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t}\frac{1}{h_a} + \sum_i W_{ia} + \lambda_{\rm s} = 0. \quad (25)$$

Thus the update rule of Eq.(21) takes the form

$$h_a^{t+1} = \frac{h_a^t}{\sum_b W_{kb} + \lambda_{\rm s}} \sum_i \frac{v_i W_{ia}}{\sum_b W_{ib}h_b^t}. \quad (26)$$

When $r_{\rm s} \leqslant a \leqslant r$, we have

$$\frac{\partial G(h,h^t)}{\partial h_a} = -\sum_i v_i \frac{W_{ia}h_a^t}{\sum_b W_{ib}h_b^t}\frac{1}{h_a} + \sum_i W_{ia} + \lambda_{\rm n}h_a = 0. \quad (27)$$

Considering the nonnegative of h_a, the update rule of Eq.(21) takes the form

$$h_a^{t+1} = \frac{-\sum_k W_{ka}}{2\lambda_{\rm n}} + \frac{\sqrt{(\sum_k W_{ka})^2 + 4\lambda_{\rm n} \sum_i v_i W_{ia} h_a^t / \sum_b W_{ib} h_b^t}}{2\lambda_{\rm n}}. \quad (28)$$

Since G is an auxiliary function, F in Eq.(24) is nonincreasing under these updates. Rewritten in matrix form, Eqs.(26) and (28) are equivalent to the update rules in Eqs.(18) and (19). With the solution of Eq.(16) obtained via Eqs.(18) and (19), the noisy speech can be denoised, and our proposed algorithm can be divided into the training stage and enhancement stage.

In the training stage: 1) Convert the training clean speech data and noise data into time-frequency (TF) domain by STFT. Take the magnitude spectra of the speech frames and noise frames to form the input data matrix $\boldsymbol{V}$; 2) Compute $\boldsymbol{W}_{\rm s}$ and $\boldsymbol{W}_{\rm n}$ by NMF using cost function shown in Eq.(3) and update equations in (5) and (6).

In the enhancement stage: 1) Convert the noisy speech data into TF domain by STFT, keep phase components unchanged and take the magnitude spectra of the noisy speech frames to form the input data matrix $\boldsymbol{V}$; 2) Construct $\boldsymbol{W}$ by using the trained dictionaries ($\boldsymbol{W} = [\boldsymbol{W}_{\rm s} \ \ \boldsymbol{W}_{\rm n}]$); 3) Compute $\boldsymbol{H}$ by MC–NMF cost function shown in Eq.(16) and update equation in (18); 4) Separate $\boldsymbol{H}$ to get $\boldsymbol{H}_{\rm s}$ and $\boldsymbol{H}_{\rm n}$; 5) Compute the mask (IRM) $\boldsymbol{M}$ by Eq.(11); 6) Compute the enhanced speech from the noisy spectrogram by Eq.(12).

2.4 Complexity analysis of MC–NMFSE algorithm

Assuming that the noisy speech signal is transformed into κ frames, and the length of STFT is χ. By using the fast fourier transform (FFT), the time complexity of calculation for each frame can be denoted as $O(\chi \log \chi)$, thus the time complexity of FFT calculation for κ frames is $\mathrm{O}(\kappa\chi\log\chi)$. Besides, assuming that the number of atoms in speech and noise dictionary are k_1 and k_2 respectively, and the number of updates of (18) and (19) is τ. Then the time complexity for solving problem (16) is $\mathrm{O}(\tau k_1 \kappa\chi) + \mathrm{O}(\tau k_2 \kappa\chi)$, which can be denoted as $\mathrm{O}(\kappa\chi)$ since τ, k_1 and k_2 are constants. Once the enhanced spectrogram is obtained, $\boldsymbol{V}_{\rm enhanced}$ should be transformed into time domain. Similar to the FFT, the time complexity of inverse FFT is $\mathrm{O}(\kappa\chi\log\chi)$. Therefore, the time complexity of MC–NMFSE algorithm is

$$\mathrm{O}(\kappa\chi\log\chi) + \mathrm{O}(\kappa\chi) + \mathrm{O}(\kappa\chi\log\chi) = \mathrm{O}(\kappa\chi\log\chi). \quad (29)$$

3 Experiments

Several experiments are conducted in this subsection to evaluate the performance of the proposed MC–NMFSE algorithm.

3.1 Dataset and parameter setting

In order to evaluate the performance of the proposed MC–NMFSE algorithm in different language conditions, TIMIT[22] database the most widely used English database in speech enhancement and the CCTV news database of Mandarin are used. Three types of nonstationary noise from NOISEX–92, namely, machinegun, subway, destroyerops, are taken as noise sources. Noisy signals are obtained by mixing a sentence with one type of noise at -5 dB, -3 dB and 0 dB, respectively. In the training phase, 530 utterances from 630 speakers are randomly chosen, which gives 30 minutes training speech. The training speech is down-sampled to 8 kHz with the frame length of 256 samples (32 ms) and a frame shift of 128 samples. Then it is transformed to 513 dimensions spectra magnitude by STFT to form the training data set for NMF–type algorithms, which is used to train the speech dictionary matrix (SDM) Ws. The number of the SDM atoms is set to 40 by empirical value. For each type of noise, a specific noise dictionary $\mathbf{W}_{\rm n}$ is also trained with NMF using 15 minutes noise signal. And the number of noise dictionary matrix (NDM) atoms is set to 20 by empirical value. For speech enhancement stage, the testing noisy speech dataset is constructed in the same way as to construct the training dataset. Signal-to-noise ratio (SNR), log-spectral-distance (LSD), and perceptual evaluation of speech quality score (PESQ)[23] which are the commonly used measurements for speech enhancement, are taken to evaluate the performance of the proposed MC–NMFSE algorithm as compared with the MMSE[24], NMFSpEnM and online–BNMF[25] algorithms.

3.2 Experimental results and analysis

Experiment 1 In this experiment, we aim to evaluate the performance of the proposed MC–NMFSE algorithm under different language and noise conditions.

First of all, the training and testing speech dataset are formed by TIMIT database. Three different type of noises are considered. The SNR, PESQ and LSD performance of the algorithms under different noise and SNR conditions are shown in Tables 1–3.

Table 1 The SNR, PESQ and LSD results of MMSE, NMF–SpEnM, online–BNMF and MC–NMFSE at different SNRs of machine-gun noise conditions

Method	Measurements		
	SNR	PESQ	LSD
MMSE(-5 dB)	-4.1303	1.3618	1.8554
MMSE(-3 dB)	-2.3688	1.5219	1.7522
MMSE(0 dB)	0.0462	1.6943	1.6188
NMF–SpEnM(-5 dB)	4.6065	1.9788	1.6314
NMF–SpEnM(-3 dB)	5.9006	2.2656	1.4639
NMF–SpEnM(0 dB)	7.7815	2.4609	1.3350
Online-BNMF(-5 dB)	2.0970	1.1307	1.6309
Online-BNMF(-3 dB)	2.1562	1.2764	1.5411
Online-BNMF(0 dB)	2.2746	1.5426	1.4397
MC–NMFSE(-5 dB)	**6.7771**	**2.2341**	**1.5826**
MC–NMFSE(-3 dB)	**7.1472**	**2.5247**	**1.3830**
MC–NMFSE(0 dB)	**7.8180**	**2.6296**	**1.3118**

Table 2 The SNR, PESQ and LSD results of MMSE, NMF–SpEnM, online–BNMF and MC–NMFSE at different SNRs of subway noise conditions

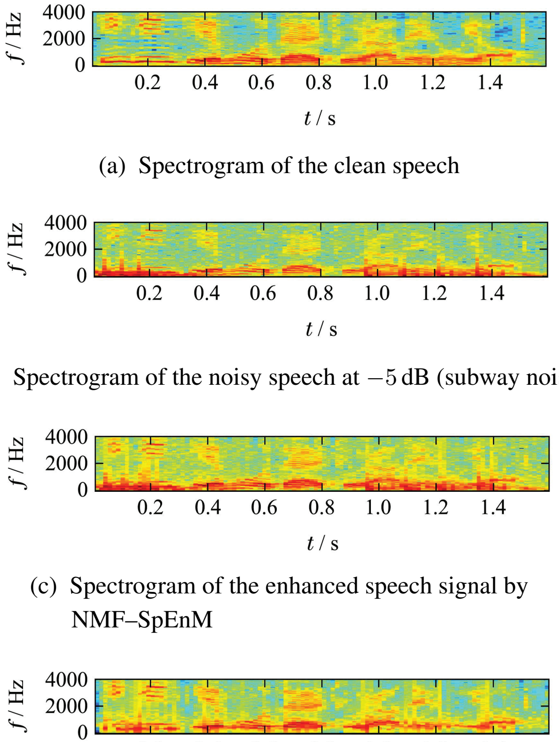
Method	Measurements		
	SNR	PESQ	LSD
MMSE(-5 dB)	-3.2397	0.7467	2.4764
MMSE(-3 dB)	-1.7117	0.8970	2.4134
MMSE(0 dB)	0.4731	1.1312	2.3134
NMF–SpEnM(-5 dB)	-0.9765	1.5027	2.3472
NMF–SpEnM(-3 dB)	0.8073	1.6243	2.2668
NMF–SpEnM(0 dB)	2.4859	1.8157	2.1200
Online-BNMF(-5 dB)	1.0499	1.1410	2.2446
Online-BNMF(-3 dB)	1.8764	1.3728	2.0088
Online-BNMF(0 dB)	2.0915	1.6346	**1.8813**
MC–NMFSE(-5 dB)	**1.0588**	**1.6548**	**2.1085**
MC–NMFSE(-3 dB)	**2.4859**	**1.7841**	**2.0235**
MC–NMFSE(0 dB)	**4.3519**	**1.9693**	1.8879

Table 3 The SNR, PESQ and LSD results of MMSE, NMF–SpEnM, online–BNMF and MC–NMFSE at different SNRs of destroyerops noise conditions

Method	Measurements		
	SNR	PESQ	LSD
MMSE(-5 dB)	1.1235	1.0349	**1.8706**
MMSE(-3 dB)	1.9539	1.2028	**1.8634**
MMSE(0 dB)	3.1093	1.4505	1.8559
NMF–SpEnM(-5 dB)	1.1352	1.7846	2.1092
NMF–SpEnM(-3 dB)	2.7632	1.9100	2.0138
NMF–SpEnM(0 dB)	5.0005	2.0953	**1.8541**
Online-BNMF(-5 dB)	0.9290	1.0377	2.2403
Online-BNMF(-3 dB)	1.1797	1.1792	2.2278
Online-BNMF(0 dB)	1.7514	1.4026	1.8804
MC–NMFSE(-5 dB)	**1.1820**	**1.7860**	2.1309
MC–NMFSE(-3 dB)	**2.8081**	**1.9109**	2.0334
MC–NMFSE(0 dB)	**5.0587**	**2.0981**	1.8706

From Tables 1–3, it is clear to see that the SNR and PESQ of the proposed MC–NMFSE algorithm perform best under three nonstationary noise conditions compared with other algorithms, which demonstrates the powerful denoising ability of MC–NMFSE. As for the LSD results, our proposed MC–NMFSE outperforms other algorithms under machine-gun noise conditions with different SNR levels and subway noise conditions with SNR to be -5 dB and -3 dB as shown in Tables 1–3, but its LSD performance is inferior to online-BNMF under subway noise conditions when S-NR is 0 dB, since the online-BNMF introduces the least distortion in the enhanced speech signal while performing moderate noise reduction[25]. Besides, under destroyerops noise conditions, MMSE and NMF–SpEnM algorithm give the best LSD results when SNRs are low (-5 dB and -3 dB) and SNR is 0 dB respectively. These results are reasonable because the property of destroyerops is closer to stationary noise compared with another two noises. And MMSE is effective for suppressing stationary noise, but it brought the decrease of speech quality. From the discussions above, we can conclude that the proposed MC–NMFSE algorithm outperforms in nonstationary noise conditions compared with that of NMF–SpEnM, online-BNMF and MMSE algorithm. But the performance of the proposed MC–NMFSE algorithm is comparable in stationary noise condition compared with that of NMF–SpEnM algorithm and is inferior to that of the MMSE algorithm.

Moreover, in order to evaluate the performance of the proposed MC–NMFSE algorithm in different language conditions, CCTV news database is used. All experimental settings are keep the same except replacing the TIMIT training dataset by CCTV news database. The spectrograms of the enhanced speech by different methods are illustrated in Fig.2. to visualize the performance of the MC–NMFSE algorithm. We can observe that, compared to the NMF–SpEnM, the proposed MC–NMFSE algorithm discards more noise components in low frequency bands. All these experiments validate the speech enhancement capability of the proposed MC–NMFSE algorithm under low–SNR and nonstationary noise conditions.



(a) Spectrogram of the clean speech

(b) Spectrogram of the noisy speech at −5 dB (subway noise)

(c) Spectrogram of the enhanced speech signal by NMF–SpEnM

(d) Spectrogram of the enhanced speech signal by MC–NMFSE

Fig. 2 Illustration of speech spectrograms (with CCTV news database)

Experiment 2 This experiment is carried out to evaluate the impact of the atoms of SDM and NDM on the PESQ performance of MC–NMFSE algorithm. It is noted that the number of dictionary atoms (r) is an important parameter for NMF-based speech enhancement methods. The experimental settings are the same as those in Experiment 1 except that we vary r from 20 to 100 in SDM and vary r from 15 to 50 in NDM. The results are shown in Fig.3. It can be seen that for SDM, when $r = 30$, the PESQ reaches highest value under machine-gun and subway noise conditions. For NDM, when $r = 15$ for machine-gun noise and $r = 30$ for subway noise, the PESQ reaches its highest value. Besides, when $r < 30$, the PESQ is not sensitive to the choice of r. Considering the tradeoff between the computational complexity and the speech enhancement performance, we choose to set r as 30 in SDM and set r as 15 in NDM in our experiments. In Fig.3, The SNR is set to −5 dB. The up figure shows PESQ versus atom number of SDM, and the down one shows PESQ versus atom number of NDM.

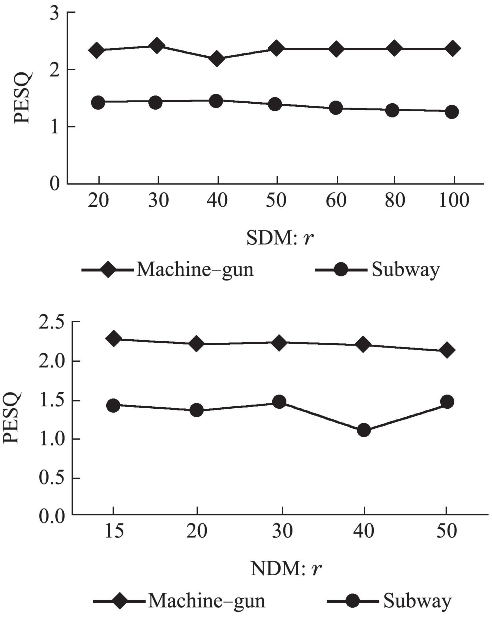


Fig. 3 PESQ performance of the proposed MC–NMFSE algorithm versus number of atoms

Experiment 3 This experiment aims at evaluating the impact of number of training data frames on the SNR performance of the MC–NMFSE algorithm. It is noted that our proposed MC–NMFSE algorithm is a learning based algorithm. Its performance may vary with the number of training data frames used. The experimental settings are the same as those in Experiment 1 except that we vary the number of training data frames from 30 thousands to 120 thousands. The experimental results are shown in Table 4.

Table 4 The SNR performance of the MCNMF–SE algorithm under different number of training data frame

Metrics (SNR)	Number of frames (thousands)			
	30	60	90	120
Output-SNR (−5 dB)	7.0578	7.1376	7.0757	**7.1421**
PESQ (−5 dB)	2.4728	2.4709	**2.4939**	2.4741
LSD (−5 dB)	1.4461	1.4466	1.4496	**1.4379**
Output-SNR (−3 dB)	7.4440	**7.5400**	7.4144	7.5044
PESQ (−5 dB)	2.5548	2.5530	**2.5733**	2.5542
LSD (−3 dB)	1.3945	1.3957	1.4000	**1.3882**
Output-SNR (0 dB)	7.9061	**8.0242**	7.8262	7.9482
PESQ (0 dB)	2.6639	2.6639	**2.6812**	2.6603
LSD (0 dB)	1.3244	1.3270	1.3341	**1.3221**

From Table 4 we can see that the number of training data frames do impact the performance of our MC–NMFSE algorithm. Specifically, when input SNR is very low (such as −5 dB), more training data benefits the speech enhancement performance. For example, when the training data frame number s is 120 thousands, the SNR and LSD reach their best values. With the increase of the input SNR, such as SNR= 0 dB, SNR, PESQ and LSD reach their highest values at $s = 60, 90$ and 120 thousands, respectively. In average, these results indicate that longer training data may lead to better noise suppress while keep lower speech distortion.

4 Conclusions

In this paper, a novel multi-constraint NMF based speech enhancement (MC–NMFSE) against the low-SNR nonstationary noise is proposed. Specifically, sparsity property of speech and low rank property of nonstationary noise are employed to constraint the factorization, then corresponding solution is obtained. The results of the experiments with mixtures containing various noise types show that the proposed MC–NMFSE algorithm outperforms the conventional NMF algorithm both with TIMIT database and CCTV News database. Besides, the proposed MC–NMFSE algorithm outperforms MMSE algorithm in terms of SNR and PESQ under nonstationary noise in low SNR conditions, but it is slightly inferior to MMSE algorithm under destroyerops noise condition since the property of destroyerops is closer to stationary noise.

References:

[1] KAMATH S, LOIZOU P. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise [C] //*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Orlando, USA: IEEE, 2002, 4: 4164.

[2] PALIWAL K, WÓJCICKI K, SCHWERIN B. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain [J]. *Speech Communication*, 2010, 52(5): 450 – 475.

[3] EPHRAIM Y, MALAH D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator [J]. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1985, 33(2): 443 – 445.

[4] LOIZOU P. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum [J]. *IEEE Transactions on Speech and Audio Processing*, 2005, 13(5): 857 – 869.

[5] GERKMANN T, HENDRIKS R C. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(4): 1383 – 1393.

[6] SUN M, LI Y, GEMMEKE J F, et al. Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback－Leibler divergence [J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2015, 23(7): 1233 – 1242.

[7] XU Y, DU J, DAI L R, et al. An experimental study on speech enhancement based on deep neural networks [J]. *IEEE Signal Processing Letters*, 2014, 21(1): 65 – 68.

[8] RAJ B, VIRTANEN T, CHAUDHURI S, et al. Non-negative matrix factorization based compensation of music for automatic speech recognition [C] //*Interspeech*. Makuhari, Japan: IEEE, 2010: 717 – 720.

[9] JODER C, WENINGER F, EYBEN F, et al. Real-time speech separation by semi-supervised nonnegative matrix factorization [C] //*International Conference on Latent Variable Analysis and Signal Separation*. Berlin: Springer, 2012: 322 – 329.

[10] KWON K, SHIN J W, SONOWAT S, et al. Speech enhancement combining statistical models and NMF with update of speech and noise bases [C] //*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Florence, Italy: IEEE, 2014: 7053 – 7057.

[11] XU Y, DU J, DAI L R, et al. A regression approach to speech enhancement based on deep neural networks [J]. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 2015, 23(1): 7 – 19.

[12] WILSON K W, RAJ B, SMARAGDIS P, et al. Speech denoising using nonnegative matrix factorization with priors [C] //*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, USA: IEEE, 2008: 4029 – 4032.

[13] LEE D D, SEUNG H S. Algorithms for non-negative matrix factorization [C] //*Advances in Neural Information Processing Systems*. Vancouver, Canada: IEEE, 2001: 556 – 562.

[14] SIGG C D, DIKK T, BUHMANN J M. Speech enhancement with sparse coding in learned dictionaries [C] //*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, Texas, USA: IEEE, 2010: 4758 – 4761.

[15] CHEN Z, ELLIS D P W. Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition [C] //*IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz, USA: IEEE, 2013: 1 – 4.

[16] WENINGER F, LE R J, HERSHEY J R, et al. Discriminative NM-F and its application to single-channel source separation [C] //*Interspeech*. Singapore: IEEE, 2014: 865 – 869.

[17] LIU S H, ZOU Y X, NING H K. Nonnegative matrix factorization based noise robust speaker verification [C] //*IEEE China Summit and International Conference on Signal and Information Processing*. Chengdu, China: IEEE, 2015: 35 – 39.

[18] PEHARZ R, PERNKOPF F. Sparse nonnegative matrix factorization with 0-constraints [J]. *Neurocomputing*, 2012, 80(1): 38 – 46.

[19] SPRECHAMANN P, BRONSTEIN A, BRONSTEIN M, et al. Learnable low rank sparse models for speech denoising [C] //*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada: IEEE, 2013: 136 – 140.

[20] CANDèS E J, LI X, MA Y, et al. Robust principal component analysis [J]. *Journal of the ACM (JACM)*, 2011, 58(3): 11.

[21] FéVOTTE C, IDIER J. Algorithms for nonnegative matrix factorization with the β-divergence [J]. *Neural Computation*, 2011, 23(9): 2421 – 2456.

[22] GAROFOLO J S. *Getting started with the DARPA TIMIT CD–ROM: An acoustic phonetic continuous speech database* [R]. Gaithersburgh: National Institute of Standards and Technology (NIST), 1988: 107.

[23] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs [C] //*Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*. Salt Lake, USA: IEEE, 2001, 2: 749 – 752.

[24] COHEN I. Speech enhancement using a noncausal a priori SNR estimator [J]. *IEEE Signal Processing Letters*, 2004, 11(9): 725 – 728.

[25] MOHAMMADIHA N, SMARAGDIS P, LEIJON A. Supervised and unsupervised speech enhancement using nonnegative matrix factorization [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, 21(10): 2140 – 2151.

作者简介:

邹月娴 (1964–), 女, 博士, 教授, 博士生导师, 目前研究方向为语者DOA估计、语音增强、声纹识别和情感识别, E-mail: zouyx@pkusz.edu.cn;

刘诗涵 (1991–), 女, 硕士研究生, 目前研究方向为单通道语音增强, E-mail: 490050401@qq.com;

王迪松 (1993–), 男, 硕士研究生, 目前研究方向为语音增强和DOA估计, E-mail: 1276749811@qq.com.