Received December 22, 2017, accepted January 16, 2018, date of publication January 23, 2018, date of current version March 13, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2797165

Speech Enhancement Based on NMF Under Electric Vehicle Noise Condition

MINGHE WANG, ERHUA ZHANG, AND ZHENMIN TANG

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Corresponding author: Erhua Zhang (speechstudio@163.com)

This work was supported by the National Science Foundation of China under Grant 61473154.

ABSTRACT Speech-based human–machine interaction (HMI) is essential to electronic navigation, autonomous cars, and intelligent vehicles. The noises generated by the mechanical motion or electric power equipment degrade speech quality and result in HMI failing to work effectively. However, there is relatively little literature available on speech enhancement under electric vehicle noise condition. This paper presents a speech enhancement method based on improved nonnegative matrix factorization (ImNMF). Unlike the traditional nonnegative matrix factorization (NMF) trains its speech dictionary using speech recorded in advance which inevitably contains a little noise component, ImNMF generates the speech dictionary using the spectra of pitch and their harmonics via mathematical model. This purpose is to guarantee the purity of speech dictionary. In addition, in order to alleviate the loss of the information of the noise sample, ImNMF constructs noise dictionary by a combination of the gain adjusted spectrum frames of the noise samples separated online. Compared with traditional NMF, the ImNMF noise atoms are relatively larger. Thus, the representation of speech signal mixed with noise atoms is greatly reduced. Therefore, ImNMF can reduce distortion of reconstructed speech while enhancing the recovered speech quality. Speech enhancement and speaker verification experiments on NUST603 and TIMIT data showed that the proposed ImNMF can effectively enhance speech signal in the noise environment of electric vehicles and further can reduce the equal error rate of the speaker verification system.

INDEX TERMS Electric vehicle noise, speech enhancement, non-negative matrix factorization, speaker verification.

I. INTRODUCTION

Speech based human-machine interaction (HMI) is essential to electronic navigation, autonomous cars, and intelligent vehicles. The noises generated by the mechanical motion or electric power equipment degrade speech signal quality and result in HMI failing to work. The speech enhancement technology [1] (i.e. denoising) not only can improve the signal to noise ratio (SNR) and the auditory perception of recovered speech, but also can effectively enhance the robustness of the speech recognition and speaker verification systems. Therefore, speech enhancement under noisy environments has been a focus of considerable research. However, relatively less research has been conducted on speech enhancement against vehicles noise, especially new energy vehicles, such as electric vehicles. Moreover, with the rapid development of artificial intelligence and information technology, speech recognition and speaker identification (speaker verification) are increasingly used in information security, electronic payments, map navigation, automatic driving, and other applications. Many of them occur in the vehicle environment when people are driving or taking an electric vehicle.

The noise produced by traditional vehicles with oil and gas fuel engines mainly comes from the vibration of engines, frames, gears, and tires. Compared to traditional vehicles, the noise of electric vehicles consists of not only the normal noise generated by the frame, gear, and tire mechanical vibration, but also the electromagnetic noise from high-voltage power supply system, power inverters, motors, and other electrical devices [2]. Therefore, the noise condition becomes more complex in electric vehicles.

Based on the differences of the noise source and interference way, the noise can be divided into convolution noise and additive noise. The noise in a vehicle is additive noise. Thus, the focus of this paper is on speech enhancement under the additive noise condition. To remove additive noise, the simplest methods are spectrum subtraction (SS) and Wiener filtering. The former subtracts an estimated noise spectrum from a noisy speech spectrum. This method was firstly proposed by Boll [3]. The latter based on Wiener filtering was described in [4]. Another important and classic speech enhancement method called minimum mean square error (MMSE) [5], [6] performs non-linear estimation of the short-time spectral amplitude (STSA) of the speech signal to minimize the MSE in the spectral domain.

Recently, signal-subspace-based speech enhancement methods, such as nonnegative matrix factorization (NMF) [7], [8], have attracted more attention. As a basic tool for data representation and analysis, NMF has been successfully applied in the fields of image analysis, text clustering, speech enhancement [8], [9], singer identification [10], speaker recognition [11], and so on. However, while speech enhancement based on NMF improves the speech signal to noise ratio, it inevitably leads to speech signal distortion, and further destroys the integrity of speech spectrum.

The integrity of speech spectrum is the key to determine the quality of the speech signal, which inevitably affects the auditory perception of the utterance. In this paper, the cause of speech distortion during the denoising based on NMF is first analyzed. Then, the dictionary structure of NMF is modified and a novel speech enhancement approach based on improved non-negative matrix factorization (ImNMF) is proposed. In the vehicle interior noise environment, especially in the new energy electric vehicle interior noise condition, speech enhancement evaluation experiment is carried out, and compared with the classical speech enhancement algorithm SS, Wiener, MMSE, NMF, etc.

An evaluation of speech enhancement is carried out under an electric vehicle noise condition. It was compared with classic speech enhancement algorithms, such as SS, Wiener, MMSE, NMF, etc. The speaker verification system is more sensitive to the quality of the speech signal. Thus, to further evaluate the proposed speech enhancement approach, a speaker verification experiment is conducted in electric vehicle noise environments. The results show that the proposed ImNMF can effectively enhance the speech signal in the noisy environment of electric vehicles, while significantly improving the robustness of the speaker verification system.

The rest of this paper is organized as follows: Section 2 introduces the typical model of the noisy speech. Section 3 presents the ImNMF approach and describes as speech enhancement technology. To evaluate the proposed method under electric vehicle noise condition, the experiments on speech enhancement and speaker verification are conducted in Section 4. The conclusion is finally drawn in Section 5.

II. NOISY SPEECH MODEL

In the time domain, an observed speech signal y, which usually degrades, is generated from the clean speech signal x with additive noise n and convolution channel distortions h according to:

$$y = x * h + n, \tag{1}$$

where the $*$ denotes the convolution operator.

At the preprocessing stage, the speech signal is first divided into L frames. Then, the 'hanming' window function and the short-time Fourier transform (STFT) can be applied to every frame of speech signal to generate the spectrum, which is a K-dimensional vector (K is the number of STFT frequency bins). The spectrum of the signals x, h, n, and y obtained by STFT are denoted as $\boldsymbol{X}_l$, $\boldsymbol{H}$, $\boldsymbol{N}_l$ and $\boldsymbol{Y}_l$ respectively. Here, $l \in [1, 2, \ldots, L]$ is the frame index.

In the spectral domain, (1) can be rewritten as:

$$\boldsymbol{Y}_l = \boldsymbol{X}_l \odot \boldsymbol{H} + \boldsymbol{N}_l, \tag{2}$$

where $\odot$ denotes Hadamard product.

Mel-frequency cepstral coefficient (MFCC) is the most widely used acoustic features. To generate MFCC, first a set of Mel scale filters is applied on the spectrum to obtain Mel-filter-bank output. Then the log operator is employed to obtain the log-filter-bank output. Finally, the discrete cosine transform (DCT) is used to generate MFCC in the cepstral domain. The MFCC forms of $\boldsymbol{X}_l$, $\boldsymbol{H}$, $\boldsymbol{N}_l$, and $\boldsymbol{Y}_l$ are denoted as $\widehat{\boldsymbol{X}}_l$, $\widehat{\boldsymbol{H}}$, $\widehat{\boldsymbol{N}}_l$, and $\widehat{\boldsymbol{Y}}_l$ respectively. Obeying above steps, the MFCC of (2) can be given as following [12], [13]:

$$\widehat{\boldsymbol{Y}}_l = \widehat{\boldsymbol{X}}_l + \widehat{\boldsymbol{H}} + \boldsymbol{C} \log\left(1 + \exp\left(\boldsymbol{C}^{-1}\left(\widehat{\boldsymbol{N}}_l - \widehat{\boldsymbol{X}}_l - \widehat{\boldsymbol{H}}\right)\right)\right), \tag{3}$$

where $\boldsymbol{C}$ is the DCT matrix.

From (1) to (3), it can be found that, for channel noises, owing to the effect of the logarithmic operation in the MFCC process, the interaction of the speech and the noise becomes additive and easy to remove by subtraction operator in the Mel-frequency cepstrum domain, while it is convolutive in the time domain or in the spectral domain. The speech enhancement method, cepstral mean subtraction (CMS) is based on this principle. However, for background noise, owing to the effect of the logarithmic operation in the MFCC process, the interaction of the speech and noise in the time or in the spectrum domain is additive and simple, but it becomes nonlinear and very hard to process in the Mel-frequency spectral domain.

Given additive background noise, the noise model in the time domain or in the spectrum domain is simple and suitable for denoising using methods such as SS and NMF. Since the focus of this paper is speech enhancement under the noise condition of electric vehicles, channel noise $\boldsymbol{H}$ can be ignored. So (2) can be simplified to

$$\boldsymbol{Y}_l = \boldsymbol{X}_l + \boldsymbol{N}_l. \tag{4}$$

In the following sections, we take advantage of (4) to elaborate on the relation between clean and distorted speech. To remove the additive noise efficiently, we consider the speech and noise model in the spectral domain.

III. SPEECH ENHANCEMENT

The major objective of speech enhancement is to recover pure original speech from a noisy speech signal. However, it is difficult to remove noise without distorting speech since the performance of any noise estimation algorithm usually depends on a trade-off between speech distortion and noise reduction. In general, speech enhancement algorithms fall into three categories: filtering techniques, spectral restoration, and model-based methods. Among of them, the model-based speech enhancement methods, such as NMF series based methods, take advantage of the statistical models of both speech and noise to produce estimates of pure speech from noisy observations.

NMF is a recently developed technique for finding linear representations of non-negative data [3], [4]. NMF is a widely used tool for making useful audio representations. NMF factorizes the given non-negative matrix $\boldsymbol{B}$ into two non-negative matrices:

$$\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}, \tag{5}$$

where $\boldsymbol{B} \in R_+^{m_1\times m_2}$, $\boldsymbol{U} \in R_+^{m_1\times r}$, $\boldsymbol{V} \in R_+^{r\times m_2}$. In speech processing, $\boldsymbol{B}$ is usually the spectrogram of the speech signal with spectral vectors stored by column, $\boldsymbol{U}$ is the basis matrix or dictionary, and $\boldsymbol{V}$ is referred to as the NMF coefficient or activation matrix.

Given a clean speech signal $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_L]$, background noise signal $\boldsymbol{N} = [\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_L]$, and observed speech signal $\boldsymbol{Y} = [\boldsymbol{Y}_1, \boldsymbol{Y}_2, , \ldots, \boldsymbol{Y}_L]$, by (4) and (5) one has:

$$\boldsymbol{X} = \boldsymbol{U}^s\boldsymbol{V}^s, \tag{6}$$

$$\boldsymbol{N} = \boldsymbol{U}^n\boldsymbol{V}^n, \tag{7}$$

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}^s & \boldsymbol{U}^n \end{bmatrix}, \tag{8}$$

$$\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}^s \\ \boldsymbol{V}^n \end{bmatrix}, \tag{9}$$

$$\boldsymbol{Y} = \boldsymbol{U}^s\boldsymbol{V}^s + \boldsymbol{U}^n\boldsymbol{V}^n = \begin{bmatrix} \boldsymbol{U}^s & \boldsymbol{U}^n \end{bmatrix}\begin{bmatrix} \boldsymbol{V}^s \\ \boldsymbol{V}^n \end{bmatrix}, \tag{10}$$

In practice the clean signal $\boldsymbol{X}$ and background noise $\boldsymbol{N}$ are unknown, however the corresponding dictionaries $\boldsymbol{U}^s$ and $\boldsymbol{U}^n$ can be designed. In this case, given an observed speech $\boldsymbol{Y}$, by solving (10), one can obtain the estimated coefficient matrices $V^{s'}$ and $\boldsymbol{V}^{n'}$, Then the clean signal $\boldsymbol{X}'$ and $\boldsymbol{N}'$ can be recovered by

$$\boldsymbol{X}' = \boldsymbol{U}^s\boldsymbol{V}^{s'}, \tag{11}$$

$$\boldsymbol{N}' = \boldsymbol{U}^n\boldsymbol{V}^{n'}. \tag{12}$$

Last, the clean speech signal in time domain can be obtained by applying inverse STFT to $\boldsymbol{X}'$.

A. DICTIONARY CONSTRUCTION

The key step of NMF-based speech enhancement is estimation of speech dictionary $\boldsymbol{U}^s$ and noise dictionary $\boldsymbol{U}^n$. So far, the dictionary construction methods mainly include two kinds: the analytic method and the learning (by training) method [14]. The former builds the dictionary via mathematical model, while the later usually trains the dictionary on a large number of speech data recorded in advance.

Lyubimov and Kotov [15] introduced the dictionary construction method by mathematical way in the automatic transcription system to describe the tone and harmonics in polyphony [11] into speech enhancement under noise conditions and presented a NMF with linear constraints (linNMF). Inspired by which we improve the dictionary construction, especially the noise dictionary construction, in order to guarantee the purity of speech dictionary and to alleviate the problem about the loss of the information of the noise sample.

1) THE CONSTRUCTION OF SPEECH DICTIONARY

Most of the traditional NMFs (such as the standard NMF we called NMF in this paper) train the speech dictionary by using speech data recorded in advance which inevitably contains a little noise. The emerging linNMF generates the speech dictionary using the spectrums of the pitches and their harmonics via the mathematical model firstly applied to the automatic transcription system for describing the tone and harmonic in polyphony. We adopt a similar scheme which belongs to analytic methods to guarantee the purity of speech dictionary.

The basic of speech production assumes that in the time domain, the excitation source $e(t)$ and vocal tract filter $z(t)$ are combined into convolution model:

$$x(t) = z(t) * e(t), \tag{13}$$

The excitation signal consists of pitch and its harmonics. Where the excitation signal $e(t)$ itself could be presented by summing up complex sinusoids on frequencies that are multiples of fundamental frequency:

$$e\,(t) = \sum_{k=1}^{P} c_k e^{ik\varpi(t)}, \tag{14}$$

where c_k is the coefficient, and $\varpi\,(t)$ is the fundamental frequency (i.e. pitch). For the τ-frame speech signal, $x(\tau)$, the harmonic function is set as $w(t)$. After STFT, it can be given as:

$$|X_\tau\,(\omega)| = \sum_{k=1}^{P} c_k\,|Z_\tau\,(k\varpi_\tau)|\,|W\,(\omega - k\varpi_\tau)|\,, \tag{15}$$

where $Z(\omega)$ and $W(\omega)$ are STFTs of vocal tract filter $z(t)$ and harmonic $w(t)$.

In MFCC, the power or magnitude spectrum is used, while the phase information is lost. Thus, the speech dictionary can be generated by the linear combination of the pitch and its harmonics after STFT, as following:

Step 1: Computing the fundamental frequencies. The q fundamental frequencies are selected between the range [$f_{\min}$, $f_{\max}$] with interval Δf. In this paper, we set $f_{\min} = 80$, $f_{\max} = 400$, $\Delta f = 10$Hz, thus the corresponding the fundamental frequencies are 80, 90, $\cdots$, 400, which are in total of 33.

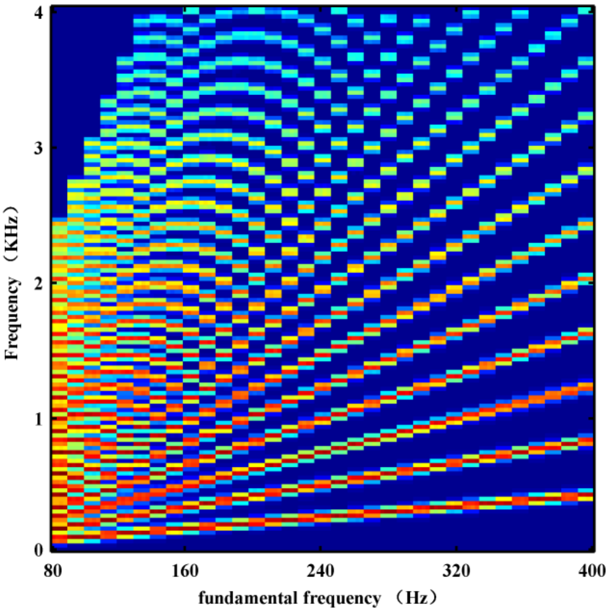


FIGURE 1. The spectral of speech dictionary.

Step 2: Computing the speech dictionary size (total number of speech atoms). For each fundamental frequency we can construct a dictionary containing m atoms. Here we used $m = 4$, and thus the dictionary contains $r_s (= 132)$ atoms.

Step 3: Computing each dictionary atom. For each fundamental frequency, the number of harmonics can be decided by:

$$P = \min\left(\text{fix}\left(f_s / (2f_0)\right), 30\right), \tag{16}$$

where f_s is sampling frequency of speech signal, f_0 is the fundamental frequency, and fix(μ) rounds the element to the nearest integer towards zero. Then the corresponding speech dictionary atoms are:

$$\boldsymbol{\psi}_j^s = [c_1\boldsymbol{\omega}_1, c_2\boldsymbol{\omega}_2, \ldots, c_k\boldsymbol{\omega}_k, \ldots, c_P\boldsymbol{\omega}_P], \tag{17}$$

where $j \in [1, 2, \ldots, r_s]$, $\boldsymbol{\omega}_k \in R_+^K$ is STFT vector of windowed $\cos(2\pi k f_0 t)$, $\boldsymbol{c}_k$ is the gain coefficient.

Step 4: At last, the speech dictionary can be constructed by:

$$\boldsymbol{u}_j^s = \boldsymbol{\psi}_j^s \boldsymbol{a}_j, \tag{18}$$

where $\boldsymbol{a}_j \in R^P$ is the vector of harmonic amplitudes, which presented the spectral envelope shape of vowel sound.

For example, when each element of $\boldsymbol{a}_j$ is $1/P$, following above step, the spectral of speech dictionary can be obtained and is shown in Figure 1. Obviously, with varying fundamental frequencies there exist the similar structure between spectrums of vowel and harmonic.

2) THE CONSTRUCTION OF NOISE DICTIONARY

Given a noisy speech signal, the noise samples are obtained by the voice activity detection (VAD) online. These noise samples are connected, and forms the G frames. Then the corresponding G frames noise spectrum can be obtained. One example is shown in Figure. 2.

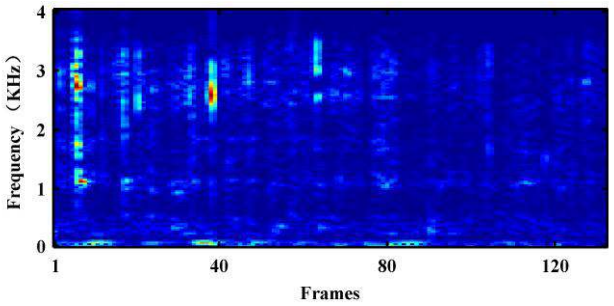


FIGURE 2. The spectral of noise dictionary.

Unlike NMF, linNMF, and denseNMF firstly decompose the noise spectrum samples into 2 matrices: the basis matrix and the coefficient matrix, then directly take the basis matrix as the noise dictionary (NMF) or take the basis matrix with linear constraints as the noise dictionary (linNMF and denseNMF), we take advantage of the noise spectrum samples to construct the noise dictionary without being decomposed by NMF. In the improved scheme, the noise spectrum samples are denoted as $\boldsymbol{w}_g$, $g \in [1, 2, \ldots, G]$ is frame index term, and the corresponding noise dictionary atoms are:

$$\boldsymbol{\psi}_j^n = \left[c_1\boldsymbol{w}_1, c_2\boldsymbol{w}_2, \ldots, c_g\boldsymbol{w}_g, \ldots, c_G\boldsymbol{w}_G\right], \tag{19}$$

where $\boldsymbol{c}_g$ is the gain coefficient. $j \in [r_s + 1, r_s + 2, \ldots, r_s + r_n]$ is frame index term, r_n is the noise dictionary size (total number of noise atoms).

The noise dictionary size r_n is set to 16 in this work. By selecting r_n groups of the different gain coefficients, r_n noise atoms can be obtained. Similarly to speech dictionary, the noise dictionary can be constructed by:

$$\boldsymbol{u}_j^n = \boldsymbol{\psi}_j^n \boldsymbol{a}_j, \tag{20}$$

where $\boldsymbol{\psi}_j^n$ are atoms, $j \in [r_s + 1, r_s + 2, \ldots, r_s + r_n]$.

B. SPEECH RECOVERY WITH lmNMF

Now, to decompose input noisy spectrogram using speech and noise dictionaries, we give the following formula of NMF problem by employing the Kullback-Leibler divergence (KLD),

$$\begin{aligned} &\min\nolimits_{\boldsymbol{V},\boldsymbol{a}_j} \text{KLD}\left(\boldsymbol{Y}||\boldsymbol{UV}\right) + \lambda \left\|\boldsymbol{V}\right\|_1 + \alpha \left\|\boldsymbol{a}_j\right\|_2^2 \\ &\text{s.t. } \boldsymbol{u}_j^s = \boldsymbol{\psi}_j^s \boldsymbol{a}_j, \quad \boldsymbol{u}_j^n = \boldsymbol{\psi}_j^n \boldsymbol{a}_j, \ ||\boldsymbol{a}_j||_1 = 1, \\ &\qquad\qquad\qquad\qquad j = 1, 2, \cdots, r_s + r_n, \end{aligned} \tag{21}$$

where α and λ are balance parameters, KLD is defined by:

$$\text{KLD}\left(\boldsymbol{Y}||\boldsymbol{UV}\right) = \sum_{i,j}\left[\boldsymbol{Y}_{i,j} \log \frac{\boldsymbol{Y}_{i,j}}{(\boldsymbol{UV})_{i,j}} - \boldsymbol{Y}_{i,j} + (\boldsymbol{UV})_{i,j}\right]. \tag{22}$$

It should be noted that although here KLD is employed, other divergences, such as Itakura–Saito Divergence, could

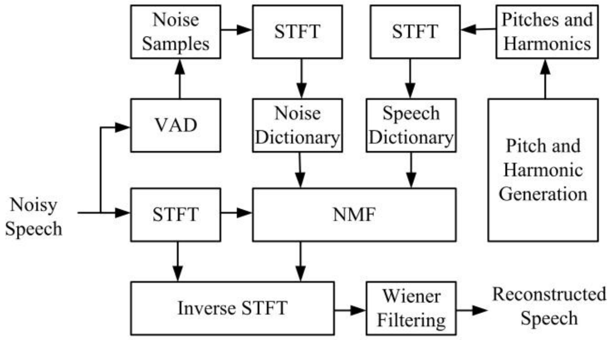


FIGURE 3. The framework of speech enhancement.

be used as well. In (21), the matrix $\boldsymbol{V}$ is assumed to be sparse and measured by L_1 norm. The parameter λ is set different for speech part and noise part, which means the different constraints are used for speech signal and noise signal. λ_s (corresponding to $\boldsymbol{V}^s$) and α ware both set to 0.2, and λ_n (corresponding to $\boldsymbol{V}^n$) is set to 0 in this paper.

Last the optimization (21) can be solved by iteratively updating following two steps:

$$\boldsymbol{a}_j \leftarrow \widetilde{\boldsymbol{a}}_j \cdot \frac{\bar{\mathbf{1}}_j \widetilde{\boldsymbol{a}}_j^T \boldsymbol{\psi}_j^T \tilde{\mathbf{1}} \bar{\boldsymbol{x}}_j^T + \boldsymbol{\psi}_j^T \frac{\mathbf{Y}}{\mathbf{UV}} \bar{\boldsymbol{x}}_j^T + \alpha \bar{\mathbf{1}}_j \widetilde{\boldsymbol{a}}_j^T \widetilde{\boldsymbol{a}}_j}{\boldsymbol{\psi}_j^T \tilde{\mathbf{1}} \bar{\boldsymbol{x}}_j^T + \bar{\mathbf{1}}_j \widetilde{\boldsymbol{a}}_j^T \boldsymbol{\psi}_j^T \frac{\mathbf{Y}}{\mathbf{UV}} \bar{\boldsymbol{x}}_j^T + \alpha \widetilde{\boldsymbol{a}}_j}, \tag{23}$$

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \cdot \frac{\boldsymbol{U}^T \frac{\boldsymbol{Y}}{\boldsymbol{UV}}}{\boldsymbol{U}^T \tilde{\mathbf{1}} + \lambda}, \tag{24}$$

where $\widetilde{\boldsymbol{a}}_j = \boldsymbol{a}_j / ||\boldsymbol{a}_j||_1$, $\bar{\mathbf{1}}_j$ indicates the vector of all-ones of the same size as $\boldsymbol{a}_j$, $\tilde{\mathbf{1}} \in R^{K \times L}$ is the all-ones matrix.

C. THE FRAME OF DENOISING

The framework of ImNMF based denoising approach is shown in Figure 3. The corresponding speech enhancement is as following:

a) Constructing the speech dictionary.

b) Constructing the noise dictionary.

c) Performing STFT and ImNMF for inputing noisy signal.

d) Obtaining the recovery signal by inverse STFT and Wiener filtering.

IV. EXPERIMENT AND ANALYSIS

A. DATA PREPARATION

For performance evaluation, we design and generate an additive noisy corpus based on the TIMIT corpus and partial NUST603 data. The TIMIT corpus contains 6300 recordings of 630 speakers of eight major dialects of American English, each speaker reads 10 phonetically rich sentences. The Mic part of the NUST603 data that recorded with microphone contains 2961 Chinese utterances in total, with durations of 15s-3min, spoken by 423 speakers.

Two electric vehicles (Beiqi-EV160 and Saichi-X3) are employed to collect the vehicle noises. As shown in Figure 4, the silver electric vehicle on the left is a Beiqi-EV160 charging at the station. A PCM-5560 microphone and a Lenovo T420i laptop are employed for the noise recording.

FIGURE 4. The Beiqi-EV160 (silver electric car on the left) at the charging station.

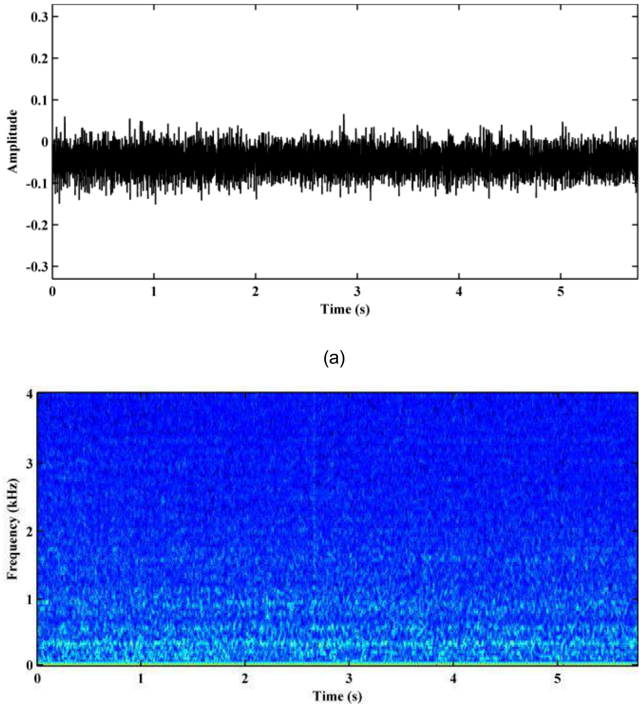
All files of the electric vehicle noise were recorded on the new Sihuan Road in Haidian District, Beijing, China. The vehicle is driven under the real-time traffic conditions. Its speed ranges from 0 to 90 km/h with multiple starting, acceleration and deceleration. During the ride, the air conditioning is on, the horn is occasionally engaged, and there are occasional slight bumps caused by the auto slowdown facilities at the crossroads.

The recording task is completed by 3 people. One is responsible for driving the electric vehicle, another one sitting in the passenger seat puts microphone near the gear shift where usually navigation or intelligent equipment are mounted; and the third one sitting in the back for operating the recording software. The recordings are monaural, 16 bits, and at 16 kHz. For an example, the waveform and spectrum of Beiqi5 noise derived from the EV160 vehicle are shown as Figure 5 (a) and (b), respectively.

For compatibility to the telecommunications system, all the speech, noisy speech, and noise signals are resampled down to 8 kHz. As shown in Figure 5(b), the energy spectrums of the noise in the electric vehicle are mainly assembled at 0-1000 Hz, especially at 0-200 Hz, which is, by and large, the same as that in [16].

It is not easy to estimate the SNR of the in-vehicle speech. The distance between the microphone and the speaker's mouth, the loudness of the individual sound, and the strength of the within vehicle noise are key factors in determining the SNR of in-vehicle speech. The NIST STNR Tools (V2.7) is used to estimate the 20 utterances of several individuals, and the SNR is roughly distributed between 5.75-26.5 dB.

FaNT (filtering and noise adding tool) [17] is an open-source software for speech processing. The software is integrated with the G.712, IRS, MIRS and P.341 filters developed for telecommunications by the international telecommunication union (ITU). Another important function of FaNT is adding noise (according to the designed SNR) to a given original signal. In the following experiments, FaNT is used to mix electric vehicle noise into TIMIT and NUST603 corpuses to generate the desired noisy data. It should be noted that the SNR of the generated noisy speech can be a little lower than the set value because the pre-filtering function of G.712 on the original speech is selected before adding the target noise.



(a)

(b)

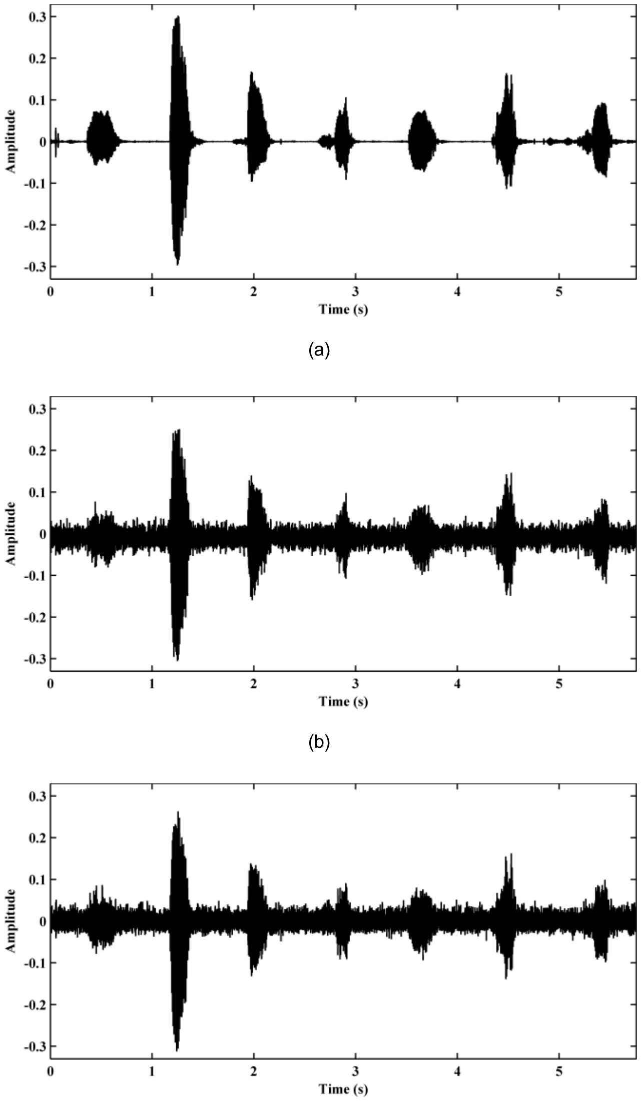
FIGURE 5. The noise of Beiqi-5 for (a) the waveform and (b) the spectrum.

B. DISTORTION COMPARISON

To evaluate the proposed ImNMF for speech enhancement in term of distortion, we compare the recovered speech signals (reconstructed by SS, Wiener, MMSE, NMF, linNMF, denseNMF, and ImNMF) from three aspects of the waveform, the spectrum, and the objective speech quality score, under electric vehicle noise condition.

1) THE WAVEFORMS OF THE RECOVERED SPEECH

For instance, the ‘Beiqi5’ noise and ‘Saichi5’ noise are mixed into the original speech M200-N1 to generate the noisy speech M200-N1-Beiqi5-6dB with ‘Beiqi5’ noise at SNR of 6dB and the noisy speech M200-N1-Saichi5-6dB with ‘Saichi5’ noise at SNR of 6dB, respectively. The waveforms of the original speech signal and the two noisy speech signals are shown in Figure 6. The waveforms of the recovered speech signals via SS, Wiener, MMSE, NMF, linNMF, denseNMF, and ImNMF are shown in Figure 7 and Figure 8, respectively.



(a)

(b)

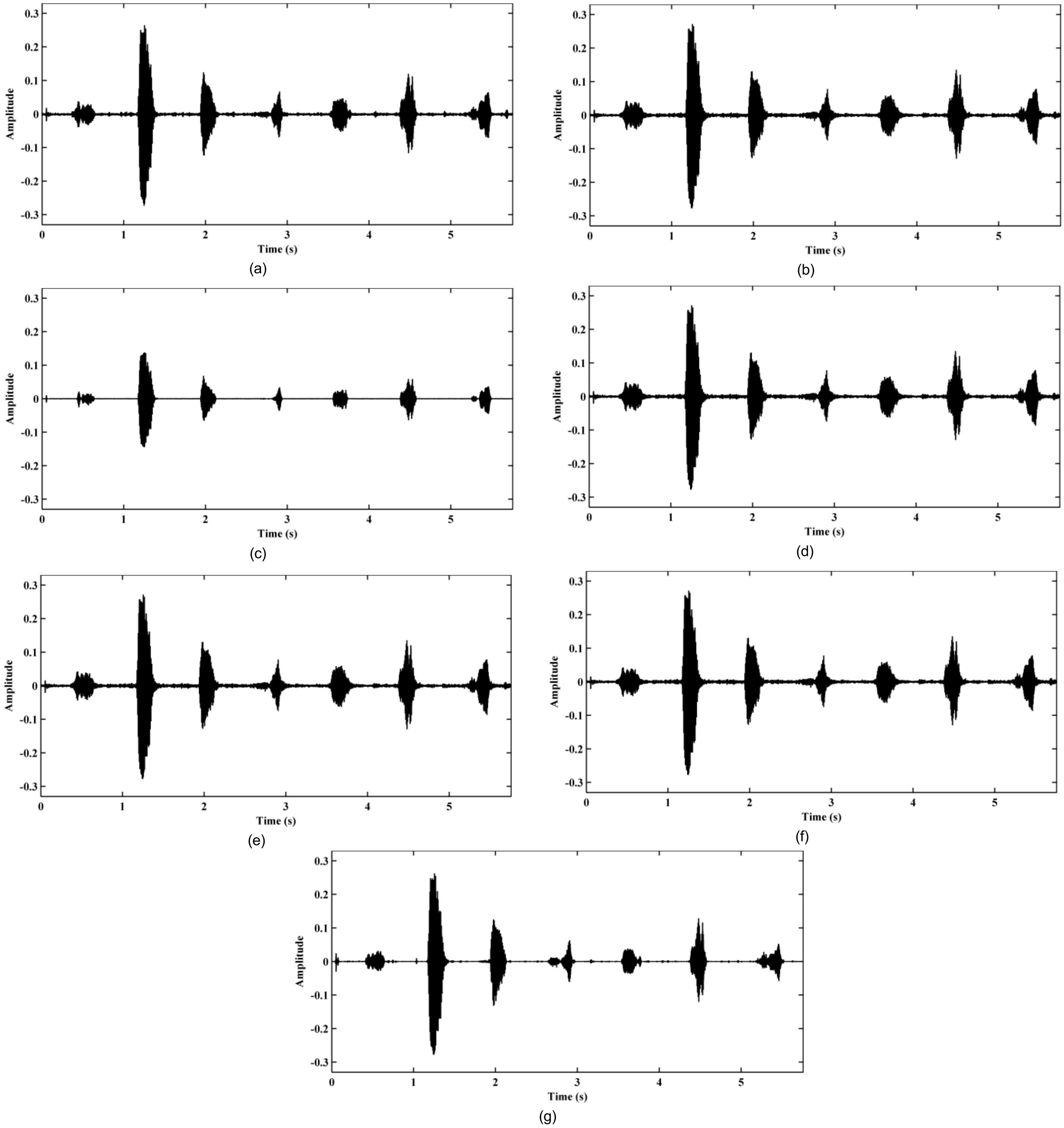
(c)

FIGURE 6. The waveforms of (a) the original speech M200-N1, (b) the noisy speech M200-N1-Beiqi5-6dB, and (c) the noisy speech M200-N1-Saichi5-6dB.

From Figure 7(a), 7(b), 8(a), and 8(b), we can find that the speech signals recovered by the classical SS and MMSE speech enhancement approaches contain a little residual noise because of the inaccuracy of noise estimation. As shown in figures 7(c) and 8(c), the algorithm of Wiener achieves high SNR of the recovered speech, and it also causes a great distortion while removing noise. Figure 7(d), 7(f), 8(d), and 8(f) show that the traditional method NMF and the emerging method denseNMF less eliminate the noise against Beiqi5 leading to much noise remained, while over eliminate the noise against Saichi5 leading to serious speech distortion. As shown in Figure 7(e) and 8(e), there is much residual noise in the enhanced speech against both of the two electric vehicle noises based on linNMF. The waveforms of the recovered speech signals via the proposed ImNMF are showed in Figure 7(g) and 8(g), from which we find that against either Beiqi5 or Saichi5 noise, ImNMF can effectively enhance the noisy speech as well. Among of the competitive denoising methods, ImNMF performs the best trade off between the reduction of noise and the speech distortion.

2) THE SPECTRUMS OF THE RECOVERED SPEECH

The spectrum of the original speech signal is shown in Figure 9. The spectrums of the recovered speech signals via SS, Wiener, MMSE, NMF, linNMF, denseNMF, and ImNMF are shown in Figure 10 and 11, respectively. The formants



(a)

(b)

(c)

(d)

(e)

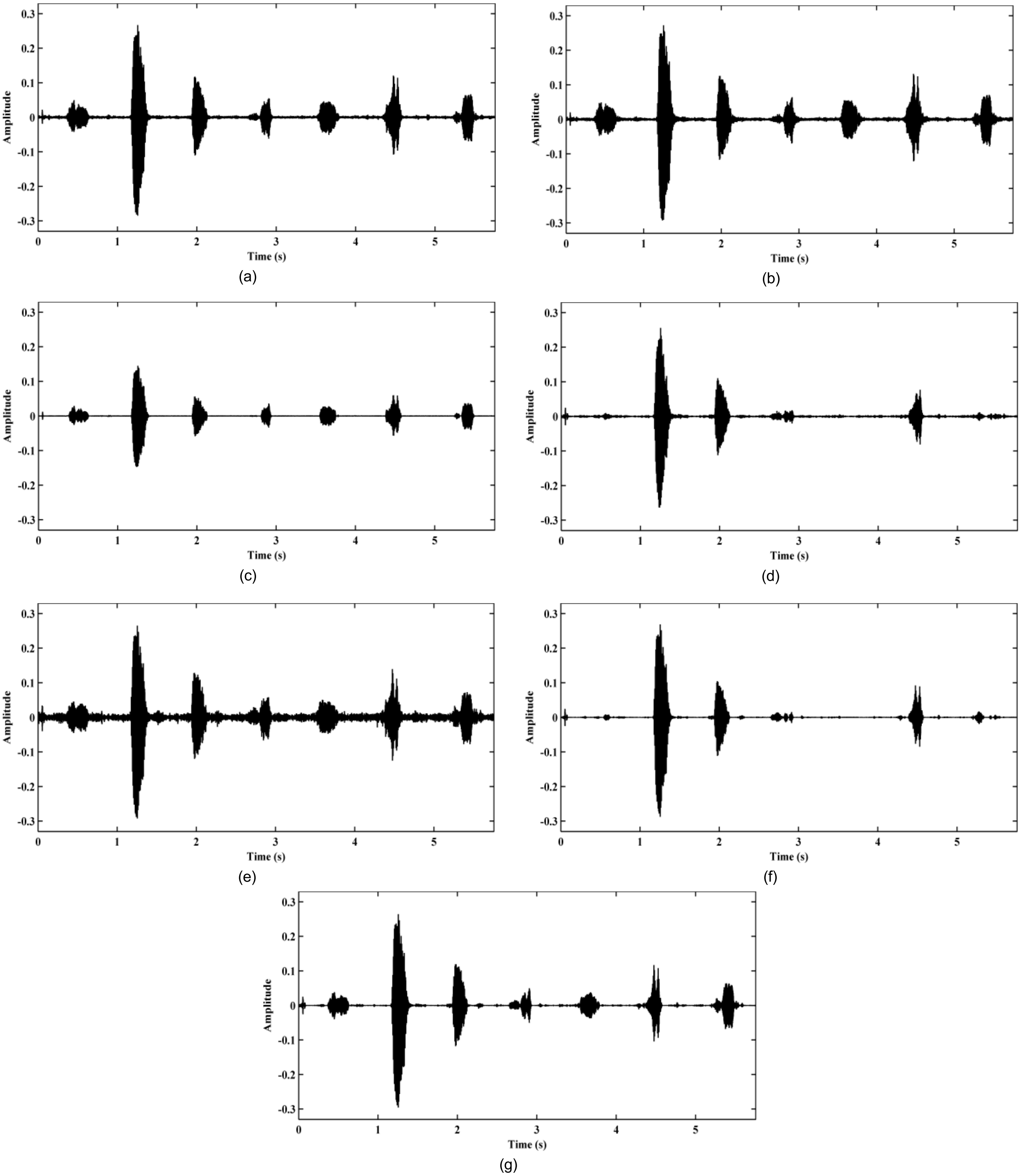
(f)

(g)

FIGURE 7. The waveforms of recovered speech from noisy speech M200-N1-Beiqi5-6dB via (a) SS, (b) MMSE, (c) Wiener, (d) NMF, (e) linNMF, (f) denseNMF, and (g) ImNMF.

corresponding to vowels of speech are clearly present in these figures.

Comparing the spectrums in the subfigures of Figure 10, we can find that, there are different degrees of energy loss in the formant structures of the seven enhanced speech signals against Beiqi5 noise. Among of them, in detail the spectrum of the recovered speech via ImNMF is the most integrated and closest to the original signal, while that via Wiener is the most negative one. The spectrums of the recovered speech signals from the noisy speech with Saichi5 noise are showed as Figure 11. Compared to the spectrums of original speech, there is much distortion (i.e. the formant loss) in the recovered speech via Wiener, NMF, linNMF, and denseNMF, and there is a slight distortion in the recovered speech via SS and MMSE. Generally, the recovered speech via ImNMF is the closest to original signal and with the least distortion.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

FIGURE 8. **The waveforms of recovered speech from noisy speech M200-N1-Saichi5-6dB via (a) SS, (b) MMSE, (c) Wiener, (d) NMF, (e) linNMF, (f) denseNMF, and (g) ImNMF.**

3) THE OBJECTIVE QUALITY EVALUATION

There are many ways to evaluate the recovered speech quality. The following, we select five quantitative and objective scoring methods to evaluate the quality of the reconstructed speech after denoising. The first criterion is the signal-to-distortion ratio (SDR) of the enhanced speech proposed by BSS-EVAL [18] to show the impact on noise separation and suppression of the algorithms. The second metric is the perceptual evaluation of speech quality (PESQ) score, which measures the subjective speech quality [19]. The third one is the short-time objective intelligibility (STOI) score described in [20], which is expected to have a monotonic relation with

TABLE 1. The quality scores of the noisy speech and the recovered speech.

Quality	Noisy	SS	Wiener	MMSE	NMF	Lin-NMF	Dense NMF	ImNMF (proposed)
SDR	3.3	4.4	3.8	4.3	4.2	4.5	**4.8**	4.7
STOI	0.65	0.78	0.71	0.77	0.76	0.77	0.78	**0.83**
PESQ	2.3	3.1	2.5	**3.3**	2.9	2.9	3.1	**3.3**
OVRL	2.7	5.6	3.4	3.5	5.2	5.3	5.4	**5.9**
SIG	2.7	3.2	2.8	3.3	3.0	3.1	3.1	**3.6**

TABLE 2. The structure of the noisy corpuses for training and testing.

Data Set	Number of males	Number of females	Number of utterances per speaker	Number of noise types	SNR	Total number of utterances
Noisy TIMIT (training)	326	136	10	6	8,15,25,original	87780
Noisy TIMIT (testing)	112	56	10	2	8,15,25,original	11760
Noisy NUST603 (training)	140	142	7	6	8,15,25,original	37506
Noisy NUST603 (testing)	70	71	7	6	8,15,25,original	18753

TABLE 3. The schemes of the training and testing.

Data Set	GMM & UBM Training	***T*** Training	LDA, GPLDA Training	Testing
Noisy TIMIT (training set)	x	x	x	
Noisy TIMIT (testing set)				x
Noisy NUST603 (training set)	x	x		
Noisy NUST603 (testing set)	x	x		

TABLE 4. The EER and MINDCF-08 on TIMIT Corpus distorted by EV noise at the SNR of 8 dB (%).

Test Item	Noise	SS	Wiener	MMSE	NMF	linNMF	Dense-NMF	ImNMF (proposed)
EER	Beiqi5	16.7	32.2	12.5	16.6	13.1	13.5	**11.7**
EER	Saichi5	16.5	32.4	12.0	18.4	13.7	14.3	**11.3**
minDCF-08	Beiqi5	7.7	9.5	**6.4**	7.6	7.0	7.2	**6.4**
minDCF-08	Saichi5	7.6	9.4	6.5	7.4	6.7	7.0	**6.1**

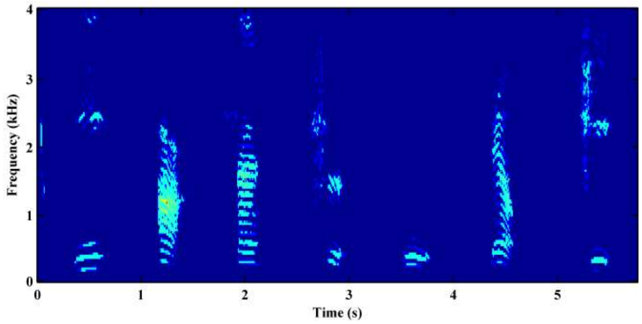
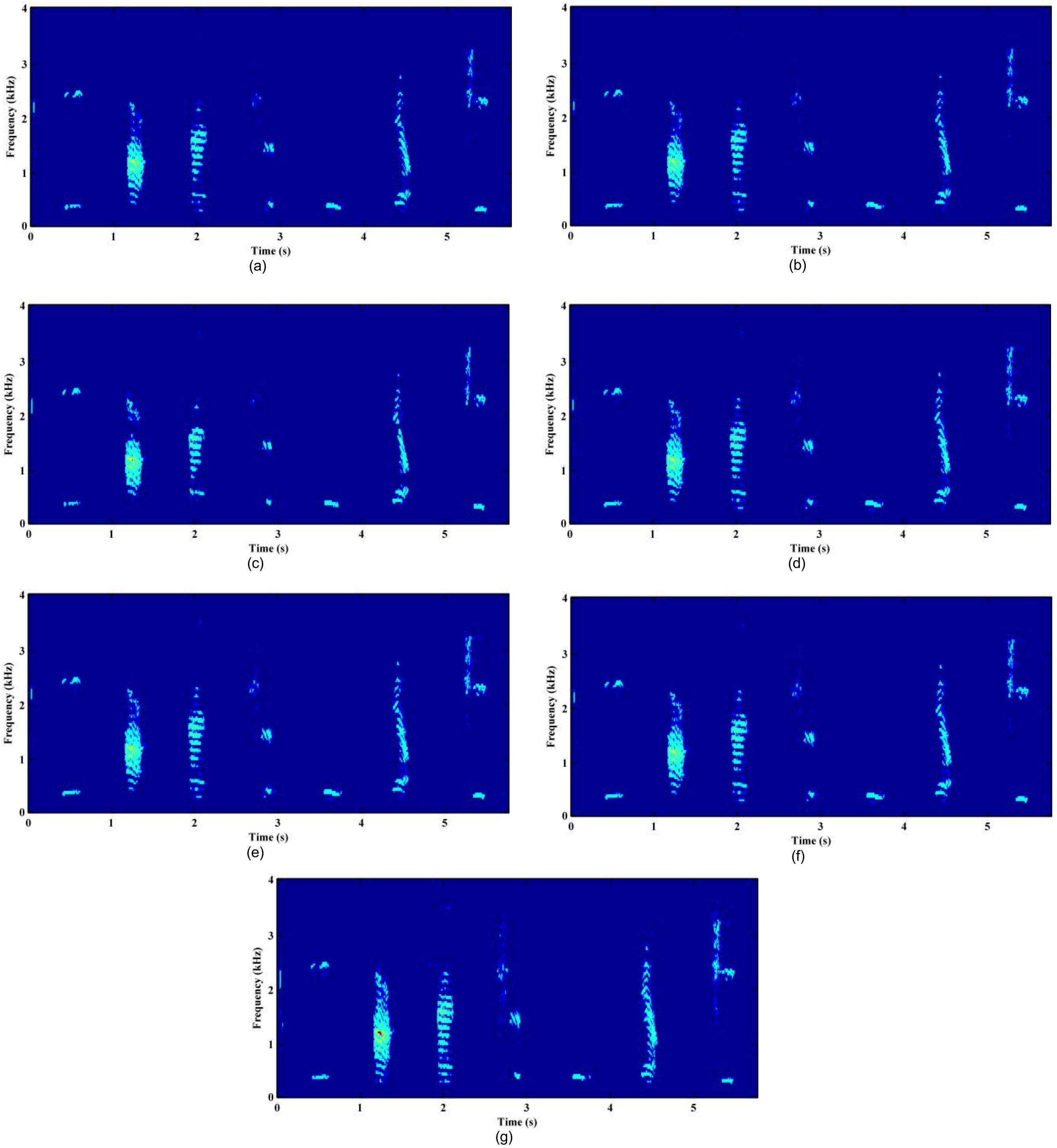


FIGURE 9. The spectrum of the original speech M200-N1.

the subjective speech intelligibility. The overall effect of the enhanced speech using the scale of the mean opinion score (OVRL) [19], [21] is the fourth measure. The last criterion is the SIG described in [21], which employs a five-point scale of signal distortion. For all the above metrics, a larger score indicates better performance.

50 utterances are randomly chosen from TIMIT and NUST603 corpuses, and then are mixed with the two electric vehicle noises at a SNR of 6 dB to generate a total of 100 noisy utterances for evaluating the denoising performance. The average score of the all noisy utterances is considered as the speech quality score. The speech quality scores of the noisy speech and the recovered speech via SS, Wiener, MMSE, NMF, linNMF, denseNMF, and ImNMF are shown in Table 1. From which it can be find that, excepting SDR, the proposed ImNMF is slightly lower than denseNMF to be suboptimal, and for all other terms, achieves better performance than the competing methods.

The noise atoms of NMF, linNMF, and denseNMF are generated by NMF training which decomposes the noise samples into tiny fragments and results in losing noise information more or less. In the process of noisy speech



(a)

(b)

(c)

(d)

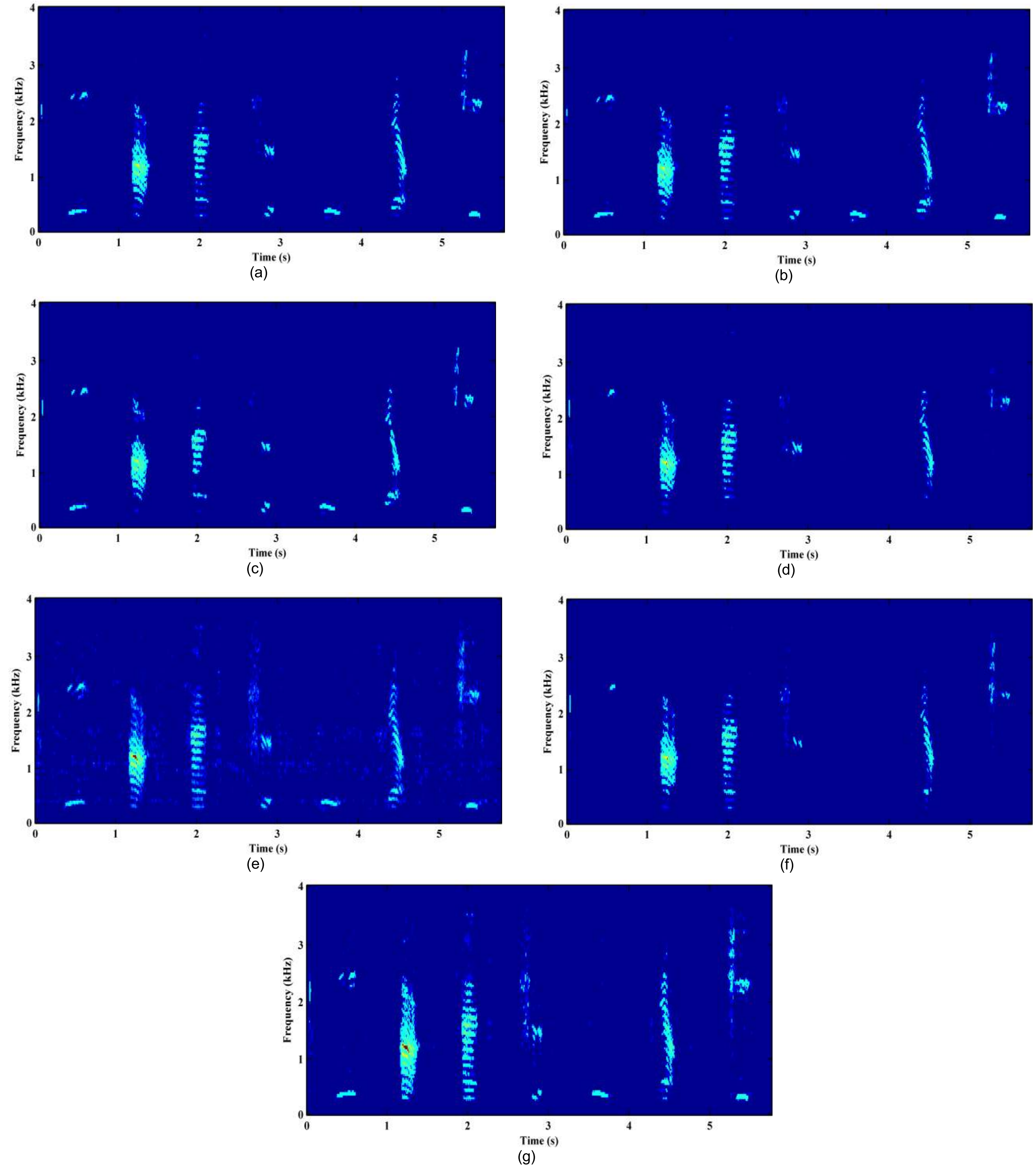
(e)

(f)

(g)

FIGURE 10. **The spectrums of recovered speech from noisy speech M200-N1-Beiqi5-6dB via (a) SS, (b) MMSE, (c) Wiener, (d) NMF, (e) linNMF, (f) denseNMF, and (g) ImNMF.**

decomposition, the particle of noise atoms is too small, which leads the noise atoms to be easily mixed into the speech signal representation. Thus, part of the speech signal incorrectly represented by the noise atom is discarded as noise in the reconstruction, finally. In order to prevent this drawback, ImNMF generates the speech dictionary using the mathematical model and constructs noise dictionary by means of linear combination of the spectrum frames of the noise samples. Compared with traditional NMF, linNMF, and denseNMF, the noise atoms of ImNMF are relatively larger and preserve more information of the noise samples. Thus, representation of speech signal mixed with noise



(a)

(b)

(c)

(d)

(e)

(f)

(g)

FIGURE 11. The spectrums of recovered speech from noisy speech M200-N1-Saichi5-6dB via (a) SS, (b) MMSE, (c) Wiener, (d) NMF, (e) linNMF, (f) denseNMF, and (g) ImNMF.

atoms is greatly reduced, which further resulting in less distortion.

C. SPEAKER VERIFICATION

The speaker verification system is more sensitive to the quality of the speech signal. Therefore, the speaker verification can indirectly evaluate the performance of the speech enhancement under electric vehicle noise condition. The noisy utterances are generated by adding noises on TIMIT and the 'Mic' part of NUST603. As listed in Table 2, these noisy utterances are divided into training set and testing set, which are used for model training and performance testing, respectively. The SNRs are set to 8, 15, 25 dB, the 'original' notes the original speech signal without adding noise. In consideration of the unseen noise condition (the noise for the testing is unseen in the training) [22], the babble, factory1, volvo, pink, light-rain, and music-box noises are mixed into the training sets, while the beiqi5 and saichi5 noises are added to the test sets, respectively.

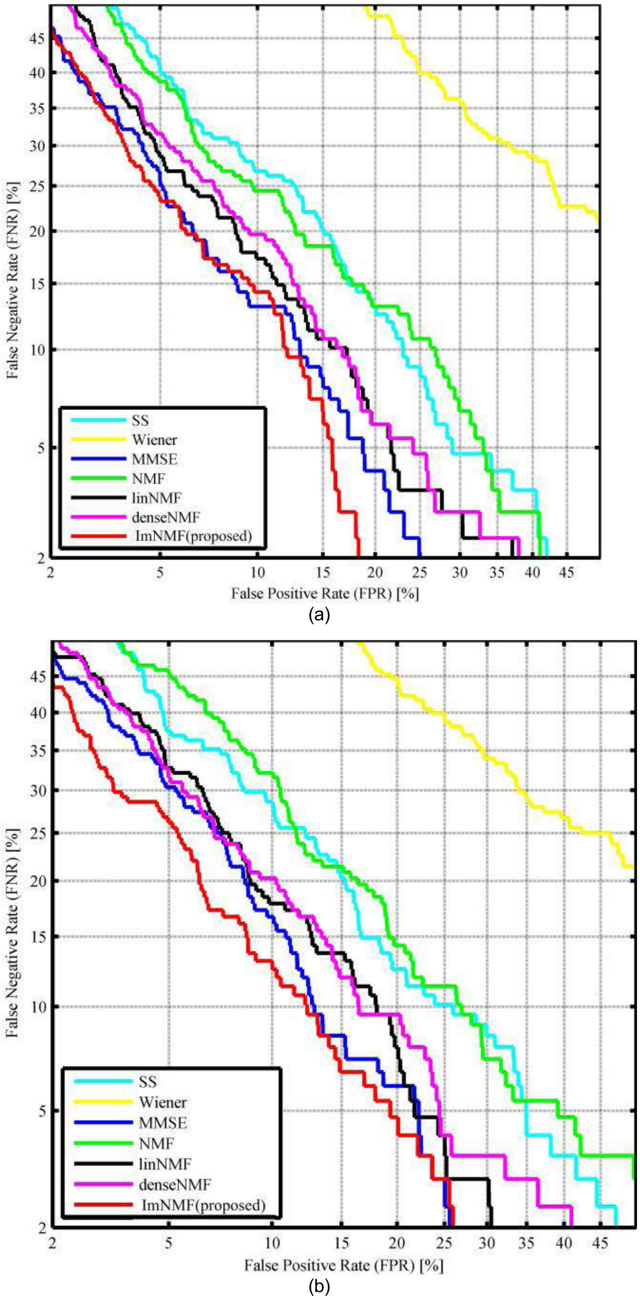
At the Training Stage: Firstly, according the schemes listed in Table 3, the Gaussian mixture model (GMM) [23]–[25], the universal background model (UBM) [26], [27], and the total subspace (T) [23], [25] are trained in advance. Then the 400-dimensional i-vector (identify vector) [23]–[28] features are extracted for training sets. Finally, the linear discriminant analysis (LDA) model [23]–[25] for dimensionality reduction and the Gaussian probabilistic linear discriminant analysis (GPLDA) [23]–[25], [28] model are trained, using the i-vector features of the training sets.

At the Testing Stage: The recovered utterances via SS, Wiener, MMSE, NMF, linNMF, denseNMF, and ImNMF are put into speaker verification system to be tested, respectively. Firstly, the 400-dimensional i-vector features of each utterance for testing are extracted. Then, the dimensionality of them is reduced via LDA, and the scores are calculated based on GPLDA. Finally, the speaker verification decisions are made and the equal error rate (EER) [23] and minimum decision cost function of 2008 (minDCF-08) [23] are calculated based on the scores at SNR $= 8$ dB. The results are listed in Table 4, for the two metrics, a small value indicates better performance. The detection error tradeoff (DET) curves [23] (as a means of representing performance on detection tasks that involve a tradeoff of error types) of speaker verification with various denoising methods are shown as Figure 12. For the DET curves, which is closer to the coordinate origin indicates the better performance.

As listed in Table 4, ImNMF achieves the best performance among the seven competing methods in terms of the EER and minDCF-08. MMSE is a classic speech enhancement method and gains the suboptimal performance (be next only to ImNMF). The same conclusion can be drawn from Figure 12, that is, ImNMF outperforms all other six methods in speaker verification system under the noise condition of electric vehicles.

It also can be fond from Table 4 and Figure 12 that the methods such as linNMF, denseNMF, and ImNMF, which construct speech dictionary by an analytic way, have a distinct advantage in speech representation. Unlike the noise dictionaries of NMF, linNMF, and denseNMF are trained by a standard NMF on noise data, the noise dictionary of ImNMF is a linear combinations of the noise spectra samples with gain coefficients, but without being decomposed by NMF, which makes the noise atom larger and less losing the information of the noise data. Thus, ImNMF not only improves the noise dictionary's capability to represent the noise signal, but also reduces the noise dictionary's probability to represent the speech signal.

From the above discussion, we can come to conclude that ImNMF improves the SNR and reduces the distortion as well in speech enhancement. It achieves the best trade off between the noise reduction and the speech distortion.



(a)

(b)

FIGURE 12. The DET curves of speaker verification with denoising via SS, MMSE, Wiener, NMF, linNMF, denseNMF, and ImNMF on TIMIT Corpus Distorted by (a) Beiqi5 noise and (b) Saichi5 noise at SNR of 8 dB.

V. CONCLUSION

This paper has proposed a speech enhancement method ImNMF based on the improved nonnegative matrix factorization and applied it to the speaker verification system. In ImNMF, the atoms of its speech dictionary are constructed via a mathematical model to guarantee the speech dictionary purity. The noise dictionary consists of the linear combination of the noise spectrum samples separated online, which preserves the information of the noise samples as much as possible. The speech enhancement experiments have been conducted on the TIMIT and NUST603 data with the electric vehicle noises. The results demonstrated that the proposed ImNMF can effectively enhance the noisy speech and can improve the robustness of the speaker verification system under electric vehicle noise conditions.

ACKNOWLEDGMENT

The statements made herein are solely the responsibility of the authors.

REFERENCES

[1] J. Yang, B. Xia, Y. Shang, W. Huang, and C. Mi, "Improved battery parameter estimation method considering operating scenarios for HEV/EV applications," *Energies*, vol. 10, no. 1, p. 5, 2016.

[2] L. Zhai, X. Zhang, N. Bondarenko, D. Loken, T. P. Van Doren, and D. G. Beetner, "Mitigation emission strategy based on resonances from a power inverter system in electric vehicles," *Energies*, vol. 9, no. 6, p. 49, 2016.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[4] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

[5] B. M. Mahmmod, A. R. Ramli, S. H. Abdulhussain, S. A. R. Al-Haddad, and W. A. Jassim, "Low-distortion MMSE speech enhancement estimator based on Laplacian prior," *IEEE Access*, vol. 5, pp. 9866–9881, 2017.

[6] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

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

[9] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech enhancement under low SNR conditions via noise estimation using sparse and low-rank NMF with Kullback–Leibler divergence," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 7, pp. 1233–1242, Jul. 2015.

[10] Y. Hu and G. Liu, "Separation of singing voice using nonnegative matrix partial co-factorization for singer identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 643–653, Apr. 2015.

[11] N. Lyubimov, M. Nastasenko, M. Kotov, and D. Doroshin, "Exploiting non-negative matrix factorization with linear constraints in noise-robust speaker identification," in *Proc. 16th SPECOM*, Novi Sad, Serbia, 2014, pp. 200–208.

[12] T. Yoshioka and T. Nakatani, "Noise model transfer: Novel approach to robustness against nonstationary noise," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 10, pp. 2182–2192, Oct. 2013.

[13] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[14] Q. S. Lian, B. S. Shi, and S. Z. Chen, "Research advances on dictionary learning models, algorithms and applications," *Acta Automat. Sinica*, vol. 41, no. 2, pp. 240–260, 2015.

[15] N. Lyubimov and M. Kotov, "Non-negative matrix factorization with linear constraints for single-channel speech enhancement," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, Lyon, France, 2013, pp. 446–450.

[16] J. Hou, Y. Liu, C. Zhang, and S. Huang, "An in-car Chinese noise corpus for speech recognition," in *Proc. Int. Conf. Asian Lang. Process.*, Nov. 2011, pp. 228–231.

[17] H.-G. Hirsch. *FaNT: Filtering and Noise Adding Tool*. Accessed: 2005. [Online]. Available: http://dnt.kr.hs-niederrhein.de/download/fant_manual.pdf

[18] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[19] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, May 2001, pp. 749–752.

[20] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[21] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. 9th Int. Conf. Spoken Lang. Process. (INTERSPEECH)*, Pittsburgh, PA, USA, Sep. 2006, pp. 229–238.

[22] M. Sun, X. Zhang, and T. F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 93–104, Jan. 2016.

[23] Y. Jiang and Z. M. Tang, "Research on the speaker identification based on short utterance," *Acta Electron. Sinica*, vol. 39, no. 4, pp. 953–957, 2011.

[24] A. Kanagasundaram, D. Dean, S. Sridharan, M. McLaren, and R. Vogt, "I-vector based speaker recognition using advanced channel compensation techniques," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 121–140 2014.

[25] Y.-F. Xu, H. Yang, R.-H. Zhou, and Y.-H. Yan, "Gaussian PLDA for speaker verification and joint estimation," *Acta Automat. Sinica*, vol. 40, no. 6, pp. 1068–1074, 2014.

[26] A. R. Avila, M. Sarria-Paja, F. J. Fraga, D. O'Shaughnessy, and T. H. Falk, "Improving the performance of far-field speaker verification using multi-condition training: The case of GMM-UBM and i-vector systems," in *Proc. 15th Conf. Int. Speech Commun. Assoc. (ISCA)*, Singapore, 2014, pp. 1096–1100.

[27] T. Stafylakis, P. Kenny, M. J. Alam, and M. Kockmann, "Speaker and channel factors in text-dependent speaker recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 65–78, Jan. 2016.

[28] M.-W. Mak, X. Pang, and J.-T. Chien, "Mixture of PLDA for noise robust i-vector speaker verification," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 1, pp. 130–142, Jan. 2016.

MINGHE WANG received the M.S. degree from Nanjing Tech University in 2009. He is currently pursuing the Ph.D. degree with the Nanjing University of Science and Technology. His main research interests include signal processing, speech enhancement, and speaker verification.

ERHUA ZHANG received the B.S., M.S., and Ph.D. degrees from the China University of Geosciences in 1988, 1991, and 2000, respectively. He is currently an Associate Professor with the Nanjing University of Science and Technology. His main research interests include signal processing, speaker recognition, and 3-D data visualization.

ZHENMIN TANG received the B.S. degree from Harbin Engineering University in 1982, the M.S. degree from the East China Institute of Technology in 1988, and the Ph.D. degree from the Nanjing University of Science and Technology in 2002. He is currently a Professor with the Nanjing University of Science and Technology and a CCF Senior Member. His main research interests include speech recognition, image processing, and intelligent robots.