JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 35, 291-305 (2019)
DOI: 10.6688/JISE.201903_35(2).0003

Robust Speaker Verification Against Additive Noise*

MING-HE WANG, ER-HUA ZHANG+ AND ZHEN-MIN TANG
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, 210094 P.R. China
E-mail: sdwmh@163.com; speechstudio@163.com; tzm.cs@mail.njust.edu.cn

Recent studies on speaker verification show total variability space (TVS) based approaches followed by Gaussian probabilistic linear discriminant analysis (GPLDA) are effective in dealing with convolutional noises (such as channel noise), even with additive noises. However, issues arise owing to the various types of noise that are unseen and non-stationary in real-world applications. To remove these noises, we introduce robust principal component analysis (RPCA) into a TVS modeled speaker verification system, called the RPCA-TVS. In which the noise spectrum is considered as the low-rank component and the speech spectrum as the sparse component in the short-time Fourier transform domain. The aim of this paper is to improve the robustness of speaker verification under additive noisy environments, particularly for non-stationary and unseen noises. Experimental results demonstrate that the proposed RPCA-TVS performs better than the competing methods at various signal to noise ratio levels. In particular, the RPCA-TVS reduces the equal error rate (EER) by 4.7% on the whole, compared with the multi-condition system, under the six additive noise conditions at the SNR of 5, 10, and 25 dB.

Keywords: robust speaker verification, additive noise, total variability space, robust principal component analysis, Gaussian probabilistic linear discriminant analysis

1. INTRODUCTION

Among the various biometric identification technologies, speaker recognition (include speaker identification and speaker verification) [1-4] is one of the most promising technologies. Speaker verification is the process of automatically recognizing who is speaking based on the individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, and human-machine Interaction (HMI) [5, 6]. Current state-of-the-art speaker recognition systems can achieve encouraging performances on clean data recorded under a quiet environment. However speech signals in real-world scenarios are often distorted by noises and these noises are often unknown and non-stationary. An ideal technique should offer good performances, even in unseen noise conditions (*i.e.* the noises used at the testing stage are unseen at the training stage) and not be limited to several known noise types [7]. Consequently, the robustness of speaker recognition has been one of the most important challenges.

In recent decades, many noise-robust speaker recognition methods have been proposed and applied in some commercial credit systems. These methods are generally categorized into four cases: speech enhancement, feature compensation, robust model-

Received September 17, 2017; revised December 29, 2017; accepted February 2, 2018.
Communicated by Berlin Chen.
+ Corresponding author.
* This research was supported by the National Science Foundation of China (Grant No. 61473154).

ing, and score compensation [8]. This paper mainly focuses on the speech enhancement for speaker verification. The major objective of speech enhancement is to recover pure original speech from a noisy speech signal. However, removing noise without distorting speech is a challenging issue since the performance of any noise estimation algorithm usually depends on a trade-off between speech distortion and noise reduction [9]. 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 spec trum. This method was firstly proposed by Boll in [10]. The latter based on Wiener filtering was described in [11]. Another important and classic speech enhancement method called minimum mean square error (MMSE) [12, 13] performs non-linear estimation of the short-time spectral amplitude (STSA) of the speech signal to minimize the mean square error in the spectral domain. As one of the signal-subspace-based speech enhancement methods, nonnegative matrix factorization (NMF) [14, 15], have attracted more attention. As a basic tool for data representation and analysis, NMF has been successfully applied in the speech enhancement and speaker recognition. Although speech enhancement can improve the intelligibility of a noisy speech signal [16], not all speech enhancements can always positively affect the speaker verification accuracy under various noise conditions, since the hidden speaker factor may be distorted during speech enhancement.

Recently, the sparse and low-rank representation (LRR) have been successfully used in exploring the multiple subspace structures of data [17]. As another benchmarking approach based on the sparseness and low-rank of data, Huang [18] introduces the robust principal component analysis (RPCA) [19], initially used for face recognition, for separating the singing voices from the accompanying music, taking advantage of the low-rank (*i.e.*, the repetition of the music rhythm) and the sparsity of the speech signal in the spectral domain. Many types of noise have repeating structures similar to music, thus the RPCA has the potential to recover clean speech from noisy speech, under various noise conditions.

On another hand, most researches on robust speaker verification have focused on the problem of compensating the mismatch between the training and test speech segments, caused by the transmission channel. Many state-of-the-art systems including the Gaussian mixture model-universal background model (GMM-UBM) [1, 3], joint factor analysis (JFA) [20], and I-vector-based techniques (*i.e.* TVS) [21-24] have been proposed for dealing with individual challenges such as inadequate utterances for training or channel noise distortion. A TVS-modeled identity vector (I-vector) followed by Gaussian probabilistic linear discriminant analysis (GPLDA) [25-27] has demonstrated high performances and is popular in text-independent speaker verification systems. It enhances the speaker verification accuracy under additive noisy environments as well as under channel noise conditions. However, their optimal subspaces for discriminating the speakers are noise-level dependent and the I-vectors shift owing to noise variability, causing the noise contaminated I-vectors to form clusters in the I-vector space [18, 20].

Obviously, a high signal to noise ratio (SNR) can be advantageous for the robustness of speaker verification systems. However, efforts are rarely made at combining the TVS model with speech enhancements for denoising and improving the SNR of the speech signals in advance. This paper aims to design a robust speaker verification system using the TVS model and RPCA. The RPCA, primarily applied in image recovery, aims

to recover underlying clean data and remove the noise data from corrupted data. Inspired by the success of the singing voice separation using the RPCA in [18], we employ the RPCA for separating speech and noise for speaker verification. The objective of this work is to investigate the promotion of the robustness of speaker verification under additive noisy environments, particularly under non-stationary and unseen noise conditions.

The frame of our speaker verification system is shown as Fig. 1. Section 2 focuses on the "RPCA based Mel-frequency cepstral coefficient (MFCC) extracting" module, where we discuss the methods for advantageously utilizing the RPCA to recover clean speech signals from noisy speech. Section 3 designs the "I-vector extracting" module, where we construct a new robust speaker verification system using the TVS model. The experimental results are given in Section 4 and the conclusions in Section 5.

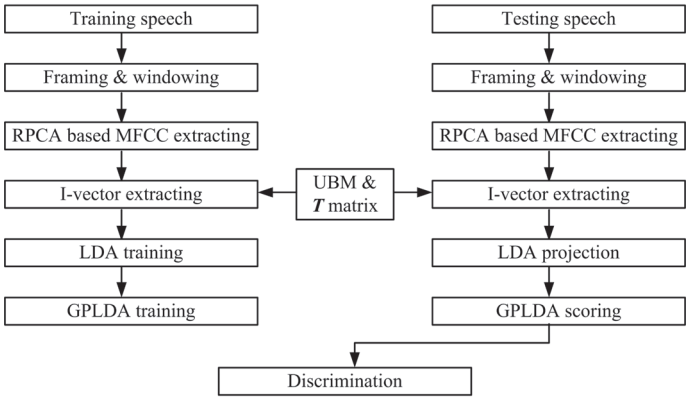

Fig. 1. The speaker verification frame with the RPCA-TVS.

2. RPCA BASED MFCC EXTRACTING

This section first outlines the RPCA, and then illustrates how applying it to speech enhancement. At last, we illustrate how to extract feature by embedding an RPCA-based speech denoising algorithm into the MFCC extraction.

2.1 RPCA

Candès *et al.* [19] proposed RPCA for recovering low-ranked matrices distorted by noise, when the noise matrix is sufficiently sparse. The RPCA aims at decomposing a data matrix $\boldsymbol{S}$ into $\boldsymbol{D}+\boldsymbol{E}$, where $\boldsymbol{S}$ is a matrix corrupted by errors, $\boldsymbol{D}$ is a low-rank matrix, and $\boldsymbol{E}$ is a sparse matrix. When the rank of $\boldsymbol{D}$ is not too large and $\boldsymbol{E}$ is enough sparse, we can determine the solution by optimizing the following problem:

$$\min_{\boldsymbol{D},\boldsymbol{E}} \operatorname{rank}(\boldsymbol{D}) + \lambda\|\boldsymbol{E}\|_0 \qquad \text{s.t. } \boldsymbol{S} = \boldsymbol{D} + \boldsymbol{E}, \tag{1}$$

where $\boldsymbol{S}$, $\boldsymbol{D}$, and $\boldsymbol{E} \in \mathbb{R}^{J\times K}$. The rank$(\cdot)$ is rank function; $\|\cdot\|_0$ denotes the L_0-norm (number of non-zero matrix entries); $\lambda > 0$ is a trade-off parameter. The principal component pursuit approach suggests solving the convex optimization problem:

$$\min_{\boldsymbol{D},\boldsymbol{E}} \|\boldsymbol{D}\|_* + \lambda\|\boldsymbol{E}\|_1 \quad \text{s.t. } \boldsymbol{S} = \boldsymbol{D} + \boldsymbol{E}, \tag{2}$$

where $\|\cdot\|_*$ denotes the nuclear norm (sum of the singular values) and $\|\cdot\|_1$ denotes the L_1-norm (sum of the absolute values of the matrix entries).

2.2 RPCA Based Denoising

In the time domain, for a clean speech x distorted by both the convolutional noise h and the additive noise n, the noisy speech signal y can be formulated [1, 28] by:

$$y = x * h + n, \tag{3}$$

where $*$ denotes the convolution operator. In this paper we mainly focus on additive noise rather than the convolutional noise, and thus h can be discarded in this paper. Then, in the short-time Fourier transform (SFT) domain, Eq. (3) becomes Eq. (4):

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

where $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{N}$ denote the spectrum matrices of clean speech, noisy speech, and the noise, respectively.

Most noises in the real-world manifest as repeating or quasi-repeating structures similar to the rhythm of music. On the other hand, the human voice spectrum is sparse or quasi-sparse in the SFT domain. Therefore, we assume noise as a low-rank matrix and clean speech as a sparse matrix for recovering the clean speech signal from noisy speech. Now, given noisy speech $\boldsymbol{Y}$, RPCA can be employed for recovering the clean speech signal $\boldsymbol{X}$ under additive noise $\boldsymbol{N}$ condition. Then, Eq. (4) can be rewritten as:

$$(\boldsymbol{X}, \boldsymbol{N}) = \arg\min_{\hat{\boldsymbol{X}},\hat{\boldsymbol{N}}} \lambda\|\hat{\boldsymbol{X}}\|_1 + \|\hat{\boldsymbol{N}}\|_* \quad \text{s.t. } \boldsymbol{Y} = \hat{\boldsymbol{X}} + \hat{\boldsymbol{N}} \tag{5}$$

where $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{N}}$ denote the spectrum matrices of the recovered speech and the estimated noise, respectively. There exist many algorithms for RPCA, such as inexact augmented Lagrange multiplier algorithms [29], gradient descent approach [30], and sub-gradient based algorithm [30]. The gradient descent approach is adopted in this paper.

An important difference between this work and face recognition is that in face recognition systems the clean face images are regarded as low-rank component and the noises are regarded as sparse component, while in this work, the original voice signals are regarded as the sparse component and the noises as the low-rank component.

Given the F100_Mic_n1_Factory2_8dB utterance y with the "Factory2" noise at an SNR of 8 dB, we calculate its spectrum matrix $\boldsymbol{Y}$ in the SFT domain, and then solve the model (5). The parameter λ can be tuned to obtain satisfactory recovery results. The results are depicted in Figs. 2 when $\lambda = 0.018$. It can be observed from Fig. 2 (a) that the "Factory2" noise is randomly appears at both the silent and vocal segments. However, the noise matrix has low rank due to its repetition on the time axis. As shown in Fig. 2 (b), the low rank matrix $\hat{\boldsymbol{N}}$ mainly contains noise but also has some speech signals. From Fig. 2 (c), we can observe there are distinct format structures in the sparse matrix $\hat{\boldsymbol{X}}$ indicating vocal activity; it also contains a little remaining noise. Therefore, the RPCA can effectively separate the voice component from noisy speech and significantly enhance the quality of the speech signal under additive noise conditions.

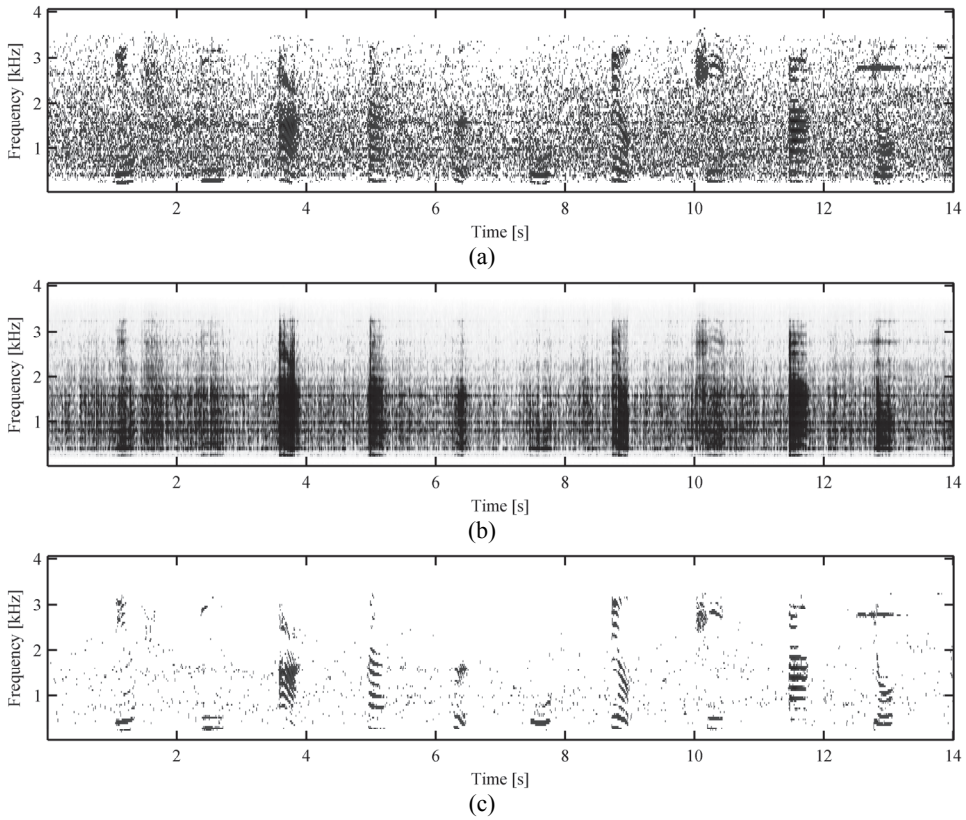

Fig. 2. The RPCA results for the F100_Mic_n1_factory2_8dB from NUST603 contaminated with noise at an SNR of 8 dB. Where, (a) the spectrum of $\boldsymbol{Y}$ denoting noisy speech with a "factory2" noise, (b) the low-ranked matrix $\hat{\boldsymbol{N}}$ denoting the separated noise spectrum, and (c) the sparse matrix $\hat{\boldsymbol{X}}$ denoting the recovered clean speech.

2.3 Feature Extracting With MFCC

At a conventional MFCC extraction stage, the log operation renders the convolutional noise in the spectral domain additive and simple in the cepstral domains; however, it renders additive noise in the spectral domain more complex in the cepstral domain. Unlike most traditional noise robust speaker verification approaches for compensating the noise impact after the MFCC or I-vector extraction [3, 23], which do not deal with the additive noise directly, we embed an RPCA-based denoising algorithm into the MFCC extraction phase, as illustrated in Fig. 3.

For a given noisy speech, the spectrum is first generated by preprocessing (such as framing and windowing) and SFT. Then, the RPCA decomposes the noisy speech spectrum into two matrices: the low-rank and sparse matrices. The former denotes the noise component and should be discarded, whereas, the later denotes recovered speech and is used as the input to the Mel-filter group, followed by the logarithm. Finally, we can obtain the MFCC by discrete cosine transformation (DCT). The study of singer identi-

fication [28] demonstrates that singing voice refinement can improve the SNR in contrast to the singing voice separation, but with respect to singer identification, the separated singing voice, *i.e.*, the unrefined singing voice, is more appropriate than the refined singing voice. Therefore, we directly use the sparse portion obtained by the RPCA without refinement.

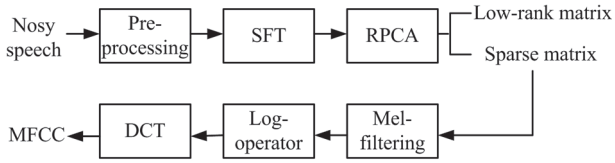

Fig. 3. The flowchart of the MFCC extraction via RPCA.

3. TVS MODEL

This section first explains how to combine the TVS model with RPCA based speech enhancements. For completion purpose, we also give the illustration of related modules in Fig. 1.

3.1 TVS Model

In recent years, a feature extractor called I-vectors, which uses a simple factor analysis in the TVS, has been proposed in [21-24]. The total factor is a hidden variable, defined by its posterior distribution, conditioned to the Baum-Welch statistics for a given utterance. Unlike the JFA that separately models the inter-speaker and channel variability, the TVS directly models both the speaker and channel information in a single low dimensional subspace, enhancing the speaker verification robustness to a certain degree, under additive noise environments as well as under channel noise environments. It is difficult to comprehend intuitively in theory because of the log operation, while extracting the MFCC.

Generally, a Gaussian mixture model-based UBM is estimated using maximum a posteriori (MAP) adaptation on a universal database, in advance. Given an utterance, the TVS-UBM super-vector is defined as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{x} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{M}$ is the mean of a given utterance, $\boldsymbol{m}$ represents the mean of all the speaker utterances and both are estimated by the UBM; $\boldsymbol{T}$ is a rectangular matrix of low rank and $\boldsymbol{x}$ is the total factor, which is a variable vector with a standard normal distribution, $N(0, 1)$; ε is the residual, generally caused by other unconcerned variabilities. For a given utterance, assuming that the zero order and the first order Baum-Welch statistics have already been obtained, the I-vector feature can be extracted as described in [21].

The TVS model has demonstrated high performances under channel noise conditions and can enhance speaker verification accuracy under additive noisy conditions.

However, it is noise level dependent. Efforts have rarely been made for utilizing the TVS model combined with the speech enhancements, in advance. Hence, we design a robust speaker verification system using the TVS model and the speech enhancement technologies described in section 2. We employ the RPCA to recover pure speech from noisy speech and build an RPCA-based speaker verification system on the TVS model.

3.2 Score Calculation

The GPLDA-based I-vector system [25-27] scoring is calculated using a batch likelihood ratio, which is computationally expensive but more effective than the traditional cosine similarity scoring. It is proved that the GPLDA is effective in both channel noise and additive noise conditions. Given two I-vectors ($\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$), the batch likelihood ratio can be calculated as described in [27]:

$$g = \ln \frac{P(\boldsymbol{Y}_1, \boldsymbol{Y}_2 \mid H_0)}{P(\boldsymbol{Y}_1 \mid H_1) P(\boldsymbol{Y}_2 \mid H_1)}, \tag{7}$$

where g is the score of the batch likelihood ratio; H_1: the speakers are the same and H_0: the speakers are different.

4. EXPERIMENTAL AND ANALYSIS

This section verifies the accuracy and reliability of the proposed scheme through simulation and comparison of the performance with several well-known schemes.

4.1 Experimental Setup

For evaluating the performance, we design and generate an additive noisy corpus based on the TIMIT corpus and the Mic part (recorded by a microphone channel) of NUST603 data [24] by filtering and noise adding tool (NaFT) [31] with 12 noise samples. As described in Table 1, the TIMIT contains a total of 6300 English sentences with duration of about 3 s, spoken by 630 speakers, 438 males and 192 females. The Mic part of the NUST603 data contains 2961 Chinese utterances in total, with durations of 20 s-3 min, spoken by 423 speakers, 210 males and 213 females. The 12 noise samples are: "babble," "factory1," "volvo," "pink," "music-box," "light-rain," "air-conditioning," "factory2," "keyboard-typing," "happy-birds," "white," and "wind" noises. Some of them are derived from the NOISEX-92 noise database and the others are free-download from the website (http://www.freesound.org).

Table 1. The TIMIT corpus and NUST603 data.

Corpus	Chanel	Males	Females	Utterances per Speaker	Total Utterances
TIMIT (training set)	Mic	326	136	10	4620
TIMIT (testing set)	Mic	112	56	10	1680
NUST603 (training set)	Mic	140	142	7	1974
NUST603 (testing set)	Mic	70	71	7	987

This paper focuses on the robustness of speaker verification under additive noisy environments, particularly, under non-stationary and unseen noise conditions. The noisy data sets are generated by NaFT tool, adding noise at the SNR of 0, 5, 8, 10, 15, and 25 dB ('original' denotes the original pure speech without adding noise), as described in Table 2. Their SNR would be a slightly lower than those of the setup because the input signal is preprocessed by the G.721 filter according to our configuration during adding a noise. There are four challenges in our experiment: First, some non-stationary noises (such as "wind," "keyboard-typing," and "happy-birds" noise) are selected for the testing task. Second, to highlighting the unseen noise condition, all the noise samples are divided into two groups: one group consisting of the first 6 noises is used at the training stage, while another group containing the rest 6 noises is used at the testing stage. Third, the noise levers in training and testing are independent each other because that the noisy utterances with noise levers at SNR of 0, 8, 15 dB, and the original utterances of the training sets are selected for training, while the noisy utterances of the testing set with noise levers at SNR of 5, 10, 25 dB are selected for testing. The original utterances of the testing set are used for speaker enrolling only. Finally, the noise data for testing include a "happy-birds" noise, in the spectrum of which, there are formant structures similar to human audio signals.

Table 2. The design of the noisy TIMIT and noisy NUST603.

Distorted Data Set	Noise Types	SNR Vary	Total Utterances
Noisy TIMIT (training set)	The first 6	0,8,15,orignal	87780
Noisy TIMIT (testing set)	The rest 6	5,10,25,orignal	31920
Noisy NUST603 (training set)	The first 6	0,8,15,orignal	37506
Noisy NUST603 (testing set)	The first 6	0,8,15,orignal	18753

The waveforms of the selected 6 noise samples for testing are shown in Fig. 4. The group of noise samples selected for testing includes either indoor or outdoor noise, stationary or non-stationary noise, artificial or natural noise. Therefore, the noise samples for testing in the evaluation experiments are broadly representative and the evaluation condition is close to practical applications.

For focusing on the speaker verification robustness for short utterances, we mainly evaluate the system performance on the TIMIT corpus with noise. However, both the TIMIT and NUST603 data, after the adding noises, are used for training the UBM and TVS models. In our experiment, the training schemes for the UBM, TVS ($\boldsymbol{T}$ space), LDA, and GPLDA are illustrated in Table 3. The proposed approach is compared against the baseline, the multi-condition approach [32, 33] and five other methods based on popular speech enhancement technologies combined with TVS. The speech enhancement methods include: SS, Wiener, MMSE, NMF, and LLR. Since this paper pays more attention to the speech enhancement impact on the TVS modeled speaker verification, we employ LRR to recover the clean speech for denoising as a competing method, which is different from that in [17] using LRR to construct the over-complete dictionary that is composed of the I-vectors of target training samples and non-target background training samples.

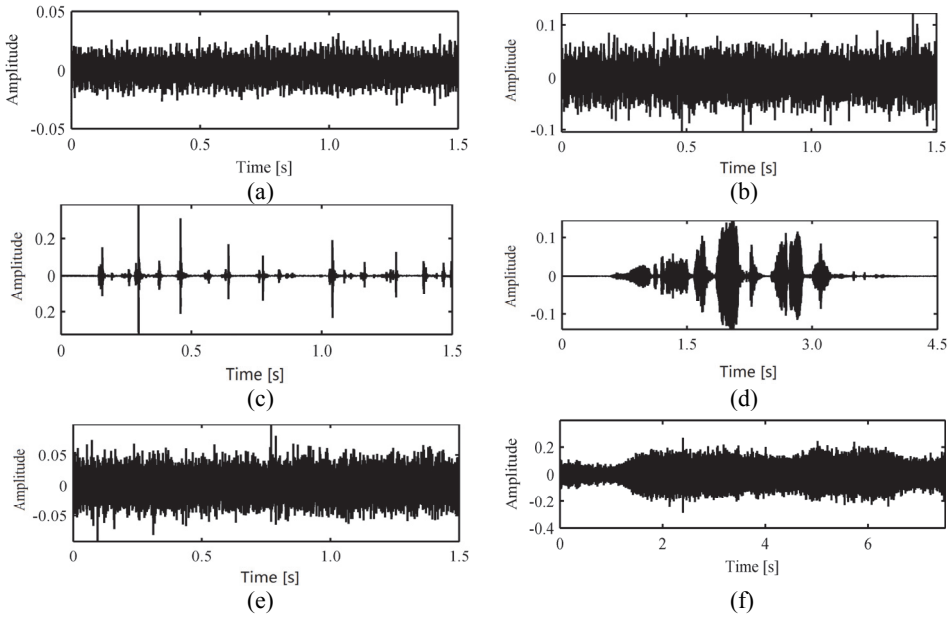

Fig. 4. The waveforms of the six noise samples for testing: (a) "air-conditioning," (b) "factory2," (c) "keyboard-typing," (d) "happy-birds," (e) "white," and (f) "wind" noises.

In the multi-condition system, all the clean and noisy utterances are pooled together for LDA and GPLDA training, and the standard features of the MFCC are used at both the training and testing stages. In the proposed method, we embed the RPCA-based denoising algorithm into the MFCC extraction phase and use the generated new features in the entire system. In the other five speech enhancement based competing systems, the LDA and GPLDA are trained using the recovered speech after denoising via their corresponding speech enhancement technologies. Using the training data drawn from the TIMIT and NUST603 corpuses, a gender-independent UBM of 512-mixture Gaussians in a 39-dimensional MFCC feature space with the first and second derivatives appended is built by adapting the UBM with the MAP algorithm. The TVS model is estimated simultaneously using the same training data set; then, the 400-dimensional I-vectors are extracted. The original utterances (clean speech) and the noisy utterances of the test data set are used for enrolling and testing, respectively. The equal error rate (EER) and the minimum decision cost function of 2008 (minDCF-08) are used as the performance metrics.

Table 3. The schemes of training and testing.

Data set	UBM training	***T*** training	LDA & GPLDA training	Testing
TIMIT (training set)	×	×	×	
TIMIT (testing set)				×
NUST603 (training set)	×	×		
NUST603 (testing set)	×	×		

4.2 Parameter Analysis

We explore various values of λ for testing different tradeoffs between the rank of noise and the sparsity of voice, in order to adjust λ slightly for obtaining the best possible result. The detection error tradeoff (DET) curves of the proposed RPCA-TVS at different λ are shown in Fig. 5. It can be observed that the best performance occurs at $\lambda = 0.018$, whereas, the singing-voice separation system based on the RPCA obtains the best result at $\lambda = \frac{1}{\sqrt{\max(J,K)}}$ as a good rule of thumb [18, 19]. This is because the short utterances with a fixed J (the number of frequencies in the SFT) and a small K (the number of frames that is approximately 300) in our experiment, result in a λ (approximately 0.06) that is too large. For simplicity, we set $\lambda = 0.018$ instead of $\lambda = 0.025$ which present in our previous work [24], and obtain satisfactory results in terms of the EER.

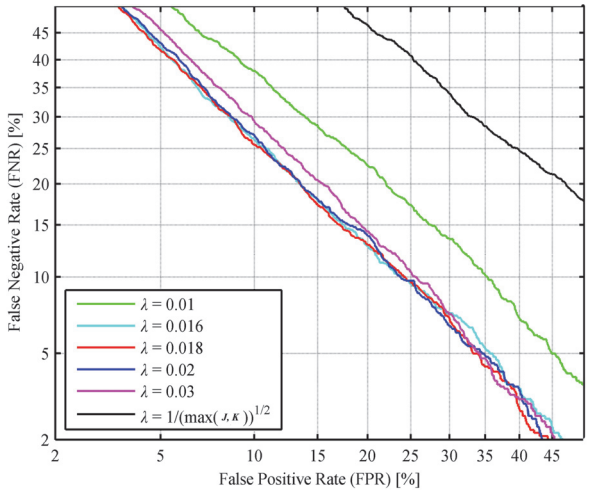

Fig. 5. The DET curves of the RPCA-TVS at various λ.

4.3 EER Evaluation

The EERs of the multi-condition, SS, Wiener, NMF, MMSE, LLR, and the proposed RPCA-TVS based systems are presented in Table 4. It can be observed that, in most cases, the RPCA-TVS works better than the other competing methods. The Wiener method shows signs of performance degradation instead of performance improvement. Thus, we arrive at a conclusion similar to the singer identification in [28] that refinement does not assist speaker verification because the speaker's acoustic character is probably destroyed during refinements such as Wiener filtering.

Finally, as shown in Table 4 and Fig. 2, by taking advantage of the repetitive structures of noise and the sparsity of clean speech, the RPCA-TVS approach works effectively under additive noise conditions, even when the noise is non-stationary and unseen. It is worth mentioning that the RPCA-TVS is the best and the only one that can outperform the baseline under the 'happy-birds' noise condition.

Table 4. The EER on the Noisy TIMIT corpus (%).

Noise	SNR	Multi-Condition	SS	Wiener	NMF	MMSE	LRR	RPCA-TVS
Air-conditioning	5dB	35.2	26.9	32.2	**19.1**	24.4	35.2	28.1
	10dB	29.2	20.0	26.3	**15.8**	15.9	29.2	19.4
	25dB	12.8	13.3	23.2	13.2	11.8	12.8	**11.5**
Factory2	5dB	27.1	23.3	35.2	23.4	20.2	24.7	**19.8**
	10dB	17.9	16.1	27.4	17.9	15.2	17.9	**14.2**
	25dB	9.9	11.4	23.3	21.1	10.3	9.9	**9.5**
Keyboard-typing	5dB	26.1	22.9	33.9	20.4	21.6	27.1	**19.6**
	10dB	20.3	19.1	27.6	18.4	17.3	23.3	**16.0**
	25dB	13.2	13.8	24.5	15.2	12.3	13.2	**11.1**
Happy-birds	5dB	22.6	21.8	32.2	23.0	20.1	22.7	**19.7**
	10dB	16.9	18.7	29.7	19.4	17.7	17.9	**16.0**
	25dB	11.2	12.0	20.2	16.7	11.4	11.7	**10.3**
White	5dB	26.8	29.0	30.3	31.6	29.8	27.8	**22.0**
	10dB	19.1	23.3	26.0	20.1	19.7	19.4	**16.8**
	25dB	11.2	11.9	24.4	15.4	**9.2**	9.5	10.0
Wind	5dB	30.8	28.8	37.0	30.9	**25.1**	30.4	25.9
	10dB	23.4	22.7	31.6	24.6	21.4	24.7	**19.8**
	25dB	11.4	11.0	23.7	16.7	11.1	11.4	**9.5**

4.4 DET Curves and minDCF-08

In the last experiment, we place all the utterances used for testing in section 4.3 together, to generate a test data set for evaluating the general performances of the seven speaker verification systems based on different technologies described above. Their DET curves are shown in Fig. 6, and the corresponding EER and minDCF-08 are listed in Table 5. As Fig. 6 showing, the RPCA-TVS has an excellent performance and achieves the best performance in terms of the EER and minDCF-08 compared to the baseline (multi-condition) and the other approaches based on speech enhancements such as the SS, Wiener, NMF, LRR, and MMSE. From Table 5, we can determine that the RPCA-TVS achieves the best EER and minDCF-08 among the seven systems based on their respective technologies.

It can be assumed that many noises exist in a low-rank subspace because of their repetitive structure in the spectrum domain. On the other hand, the human voices can be regarded as relatively sparse. Using the APG algorithm for solving the convex minimization problem, the RPCA can effectively decompose noisy speech into a noise component and a human voice component, thereby, improving the speaker verification accuracy significantly.

Table 5. The EERs and minDCFs on the whole with all the testing utterances pooled together (%).

Index	Multi-Condition	SS	Wiener	NMF	MMSE	LRR	RPCA-TVS
EER	24.5	22.2	30.7	22.8	20.7	24.2	**19.8**
minDCF	8.7	8.4	9.7	8.4	8.1	8.3	**8.0**

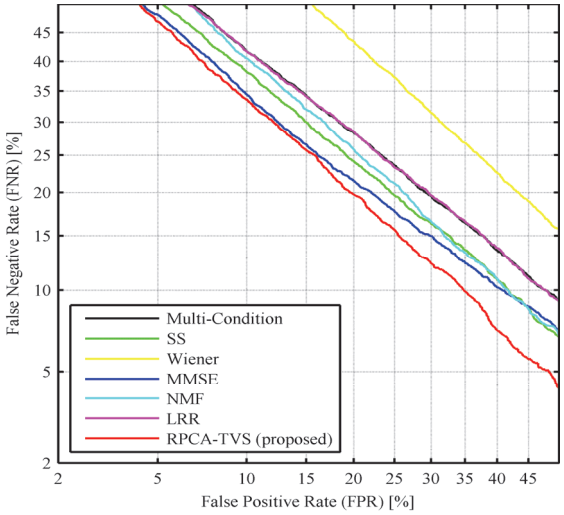

Fig. 6. The DET curves of the competing systems with all the testing utterances pooled together.

From the above discussion, it is clear that the RPCA-TVS approach achieves better performances in comparison with the baseline and the competing approaches, under complex noise conditions at various SNR levels, even under non-stationary and unseen noise conditions. It should be worth noted that the proposed denoising technology based on the RPCA can work effectively without prior training and without the need for noise segments separated online, compared with the previous denoising approaches.

5. CONCLUSIONS

In this paper, we have proposed a robust speaker verification approach, termed RPCA-TVS, which employs the RPCA in TVS-modeled speaker verification systems. To evaluate the performance, we generated an additive noisy corpus based on TIMIT and NUST603 data, using the NaFT tool, with 12 types of noise samples. We selected the first 6 of the noise samples for training and the remaining for testing, independently. Extensive experiments were conducted on the generated corpus and the results demonstrate that the RPCA-TVS achieves encouraging performances under additive noisy conditions at various SNR levels, even under non-stationary and unseen noise conditions. The proposed RPCA-TVS achieves better performances than the competing approaches and reduces the EER by 4.7%, on the whole, than the multi-condition system, under the additive noise conditions at the SNR of 5, 10, and 25 dB.

REFERENCES

1. J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 22, 2014, pp. 745-777.

2. W. T. Hong, "Minimum classification error training of hidden conditional random fields for speech and speaker recognition," *Journal of Information Science and Engineering*, Vol. 29, 2013, pp. 729-742.
3. J. Gonzalez-Rodriguez, "Evaluating automatic speaker recognition systems: An overview of the NIST speaker recognition evaluations (1996-2014)," *Loquens*, Vol. 1, 2014, pp. 1-15.
4. W. H. Tsai and S. J. Liao, "Speaker identification in overlapping speech," *Journal of Information Science and Engineering*, Vol. 26, 2010, pp. 1891-1903.
5. C. H. Yang, J. C. Wang, J. F. Wang, H. P. Lee, C. H. Wu, and K. H. Chang, "Multiband subspace tracking speech enhancement for in-car human computer speech interaction," *Journal of Information Science and Engineering*, Vol. 22, 2008, pp. 1093-1107.
6. W. Gao and Q. Cao, "Frequency warping for speaker adaptation in hmm-based speech synthesis," *Journal of Information Science and Engineering*, Vol. 30, 2014, pp. 1149-1166.
7. M. Sun, X. Zhang, H. V. Hamme, and T. F. Zheng, "Unseen noise estimation using separable deep auto encoder for speech enhancement," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 24, 2015, pp. 93-104.
8. Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2012, pp. 4253-4256.
9. S. Lee and G. Lee, "Noise estimation and suppression using nonlinear function with a priori speech absence probability in speech enhancement," *Journal of Sensors*, Vol. 6, 2016, pp. 1-7.
10. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics*, *Speech and Signal Processing*, Vol. 27, 1979, pp. 113-120.
11. J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, Vol. 67, 2005, pp. 1586-1604.
12. 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. 1, 2017, pp. 9866-9881.
13. H. Y. Chang, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, 2005, pp. 475-486.
14. D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, Vol. 401, 1999, pp. 788-791.
15. 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, 2013, pp. 2140-2151.
16. F. Chen, "Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation," *Biomedical Signal Processing and Control*, Vol. 24, 2016, pp. 109-113.
17. T. T. Dat, Y. K. Jin, H. G. Kim, and K. R. Lee, "Robust speaker verification using low-rank recovery under total variability space," in *Proceedings of International Con-*

ference on It Convergence and Security, 2015, pp. 1-4.
18. P. S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2012, pp. 57-60.
19. E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, Vol. 58, 2011, pp. 1-37.
20. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio Speech and Language Processing*, Vol. 19, 2011, pp. 788-798.
21. A. Kanagasundaram, D. Dean, S. Sridharan, and M. Mclaren, "I-vector based speaker recognition using advanced channel compensation techniques," *Computer Speech and Language*, Vol. 28, 2014, pp. 121-140.
22. W. Li, T. Fu, and J. Zhu. "An improved i-vector extraction algorithm for speaker verification," *Eurasip Journal on Audio Speech and Music Processing*, Vol. 1, 2015, pp. 1-9.
23. W. B. Kheder, D. Matrouf, J. F. Bonastre, M. Ajili, and P. M. Bousquet, "Additive noise compensation in the i-vector space for speaker recognition," in *Proceedings of IEEE International Conference on Acoustics*, *Speech and Signal Processing*, 2015, pp. 4190-4194.
24. M. Wang, E. Zhang, and Z. Tang, "Robust principal component analysis based speaker verification under additive noise conditions," in *Proceedings of the 7th Chinese Conference on Pattern Recognition*, Part II, 2016, pp. 598-606.
25. Y. Jiang, K. A. Lee, and L. B. Wang, "PLDA in the I-SUPERVECTOR space for text-independent speaker verification," *EURASIP Journal on Audio*, *Speech*, *and Music Processing*, Vol. 2014, 2014, pp. 1-13.
26. N. Li and M. W. Mak, "SNR-invariant PLDA modeling in nonparametric subspace for robust speaker verification," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 23, 2015, pp. 1648-1659.
27. M. W. Mak, X. Pang, and J. T. Chien, "Mixture of PLDA for noise robust i-vector speaker verification," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 24, 2016, pp. 130-142.
28. Y. Hu and G. Liu, "Separation of singing voice using nonnegative matrix partial cofactorization for singer identification," *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 23, 2015, pp. 643-653.
29. M. Chen, Z. Lin, Y. Ma, and L. Wu, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," *Eprint Arxiv*, Vol. 9, 2010, pp. 1-20.
30. X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust PCA via gradient descent," in *Proceedings of the 29th International Conference on Neural Information Processing Systems*, 2016, pp. 4152-4160.
31. H. Guenter Hirsch, "FaNT-filtering and noise adding tool," http://dnt.kr.hsnr.de/download/fant_manual.pdf, 2005, pp. 1-4.
32. 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 *Proceedings of the 15th Conference of the International Speech Communication Association*, 2014, pp. 1096-1100.
33. B. W. Mekonnen and B. D. Dufera, "Noise robust speaker verification using GMM-UBM multi-condition training," in *Proceedings of IEEE International Conference on Green Innovation for African Renaissance*, 2015, pp. 1-5.

Ming-He Wang (王明合) received the M.S. degree from Nanjing Tech University in 2009. Now he is a doctoral candidate in Nanjing University of Science and Technology. His main research interests include signal processing, speech enhancement, and speaker verification.

Er-Hua Zhang (張二華) received his B.S., M.S., and Ph.D. degrees from China University of Geosciences in 1988, 1991 and 2000, respectively. Now he is an Associate Professor in Nanjing University of Science and Technology. His main research interests include signal processing, speaker recognition, and 3-dimensional data visualization.

Zhen-Min Tang (唐振民) received his B.S. degree from Harbin Engineering University in 1982, received his M.S. degree in East China Institute of Technology in 1988, and received his Ph.D. degree in Nanjing University of Science and Technology in 2002. Now he is a Professor in Nanjing University of Science and Technology and a CCF senior member. His main research interests include speech recognition, image processing, and intelligent robot.