IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 3, 2019, accepted June 19, 2019, date of publication July 1, 2019, date of current version July 17, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2925839

Speaker Recognition Based on Long-Term Acoustic Features With Analysis Sparse Representation

TING LIN AND YE ZHANG
School of Electronic Information Engineering, Nanchang University, Nanchang 330031, China

Corresponding author: Ye Zhang (zhangye@ncu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61866024.

ABSTRACT The performance of a speaker recognition system depends highly on which acoustic features are used. Most speaker recognition systems use short-term acoustic features extracted from a single speech frame, and the most popular short-term acoustic features are the Mel-frequency cepstral coefficients (MFCCs). The short-term features are generally static features no dynamic information in the speech signal is included in either cepstral coefficients or an MFCCs frame. Using an analysis sparse representation model, in this paper, we introduce the long-term acoustic (LTA) feature for text-independent speaker recognition, which is a sparse presentation of the static features and dynamic information for the speaker's speech. First, the speech signal is segmented into frames which are overlapping with each other, and then the MFCCs frame features can be extracted to construct some super MFCCs frames by stacking some following frames of the current frame to capture the dynamic information of the speech signal. The super MFCCs frames can be combined into a 2-D MFCCs features map (MFCCsmap). Finally, the speaker model can be built based on the analysis sparse model and the sparse representations of the MFCCsmap are used as the LTA features. A state-of-the-art deep neural network (DNN) is employed as a classifier for speaker recognition. The experimental results illustrate the effectiveness and robustness of the proposed system.

INDEX TERMS Analysis sparse representation, deep neural network, long-term acoustic features, Mel-frequency cepstral coefficients, speaker recognition.

I. INTRODUCTION

Speaker recognition is the process of identifying a person based on the voice of the speaker [1]. Speaker recognition can be considered as a pattern recognition problem in terms of machine learning. In recent years, speaker recognition technology has received extensive attention and can be widely used in various fields such as general business interactions [2], [3], forensics [4], and law enforcement [5].

Usually, the process of speaker recognition involves extracting and identifying unique characteristics of the speech features from a group of speakers. For improving the performance of a speaker recognition system, it is important to select a feature extraction method that optimally combines efficiency and accuracy. Generally, the speech signal is a slowly time-varying or quasi-stationary signal. For stable acoustic characteristics, most existing speaker recognition systems use the speaker short-term acoustic features, such as Mel-frequency cepstral coefficients (MFCCs) [6], [7], linear predictive [8], Gammatone frequency cepstral coefficients (GFCCs) [9], etc, where a speech signal is processed in frames which are overlapping with each other. The length of frame is approximately 20-40 msec, with an overlap of about 30-75%. These short-term acoustic features are usually referred to as static features and no time evolution information is included in these features. However, the dynamic information in speech signal is also different from speaker to speaker. To capture the information about how these acoustic vectors change over time, the traditional method is to compute first and second derivatives of cepstral coefficients [10]. The first order and the second order derivative are called delta coefficients (ΔMFCCs) and delta-delta coefficients (Δ^2MFCCs), respectively. To obtain dynamic information on the speech signal, in [11], a super-vector MFCCs feature

The associate editor coordinating the review of this manuscript and approving it for publication was Huawei Chen.

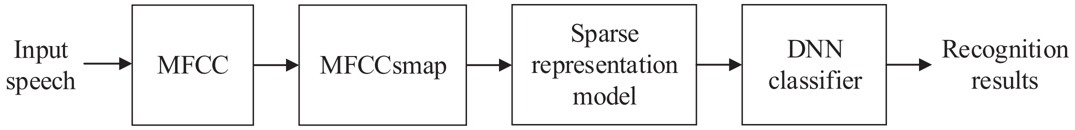


FIGURE 1. The proposed speaker recognition system.

was produced by cascading three neighboring MFCCs frames together, i.e., a center frame, one left context frame, and one right context frame, the difference of magnitude between the context frames and the center frame could be used as complementary features similar to delta features. In [12], the authors constructed the long-term feature analysis (LTFA) features by averaging 4 MFCCs frames, and used the total variability subspace modeling with the speaker LTFA features to realize speaker clustering and recognition. Inspired by these previous studies, in this paper, the super MFCCs frame is constructed by stacking some following frames of the current frame. On the basis of the analysis sparse model, the sparse representations of the super MFCCs frames could be used as long-term acoustic (LTA) features with static and dynamic information of the speech signal.

Recently, the sparse representation of the speaker acoustic features, such as i-vector features [13], the GMM-UBM features [14], tensor features [15], the MFCCs [16] and the Gaussian mixture model mean supervectors [17], was introduced for the speaker recognition with synthesis sparse representation models. In these models, a signal $\mathbf{x} \in R^{M\times 1}$ is represented as a linear combination of a few atoms from an overcomplete dictionary $\mathbf{D} \in R^{M\times Q}\,(Q > M)$, i.e., $\mathbf{x} = \mathbf{Da}$, where $\mathbf{a} \in R^{Q\times 1}$ is the sparse coefficient, i.e., $\|\mathbf{a}\|_0 = L \ll Q$, the ℓ_0 quasi-norm $\|\cdot\|_0$ counts the number of nonzero components in its argument. In general, the dictionary could be obtained using some dictionary learning algorithms, such as the greedy adaptive dictionary algorithm [18], the K-SVD algorithm [19], or by a fixed dictionary, such as the discrete cosine transform (DCT) dictionary [20], the wavelet dictionary [21], etc. However, there is an alternative sparse representation model, i.e., analysis sparse representation (ASR) model. For a signal $\mathbf{x} \in R^{M\times 1}$ and an analysis dictionary or operator $\Omega \in R^{P\times M}$ $(P \geq M)$, this model suggests that $\mathbf{x}$ is approximately sparsifiable with Ω, i.e., $\mathbf{f} = \Omega\mathbf{x}$, where $\mathbf{f} \in R^{P\times 1}$ is the sparse representation of $\mathbf{x}$. More recently, the analysis sparse model has been drawing increasing attention due to its application in image denoising [22], source separation [23], [24], image encryption [25], [26] and image classification [27], [28]. In [27], based on the analysis sparse model, the sparse representations of an image can be learned and used as the features of the image for training a support vector machine to resolve the problems of the image classification. In [28], a nonlinear discriminative cosparse model was proposed to represent the image features and simultaneously a novel discriminative nonlinear analysis operator learning framework was proposed to realize the image classification.

Motivated by the recent success of analysis sparse models in image classification, we propose a new speaker recognition system employing the sparse representations of the acoustic features as input to train a DNN classifier. In the proposed system, after the MFCCs frame features are extracted from the speech frames, the super MFCCs frame can be constructed by stacking some following frames of the current frame to capture the static and the dynamic information in speech signal. We combine the super MFCCs frames into a 2-D MFCCs map (MFCCsmap) to learn the analysis dictionary, which could be used as the speaker's model. Then the sparse representations of the MFCCsmap can be obtained with the speaker's model and are utilized as the LTA features to train the DNN classifier for speaker recognition.

The main contributions of this study are twofold. First, we propose a new speaker recognition framework employing a analysis sparse model for measurements combined with an adaptive analysis operator-based prior for the speaker speech. The adaptive analysis sparse model, i.e., the speaker model, is learned with the MFCCsmap of speech signals from training speaker datasets, which saves runtime during recognition. Second, we present the LTA features including the static and dynamic information of the speech signal, which is obtained by using the analysis sparse representations of the MFCCsmap with the speaker model, and the LTA features are used as the input of the DNN classifier.

The remainder of this paper is organized as follows: Section 2 describes the proposed speaker recognition system, and section 3 gives the implementation details and results obtained. Finally section 4 presents the conclusion.

II. THE PROPOSED SPEAKER RECOGNITION SYSTEM

Fig. 1 shows the overview of the system. In the training phase, all training speech signals are divided to a set of time frames with overlapping windows. The MFCCs extracted from each frame are converted to super MFCCs frames to construct the MFCCsmap. Then, the MFCCsmap can be utilized as the training data for building the analysis sparse model. The learning dictionary can be regarded as the speaker model, and the analysis sparse representations of the MFCCsmap can be used as the LTA features of the speaker for training a DNN classifier [29]. In the test phase, the MFCCsmap of the test speech is obtained in the same way as in the training phase, and the long-term acoustic features are generated by the speaker model, and then the LTA features are utilized as the input for the trained DNN classifier to realize the speaker recognition.

A. MFCCSMAP

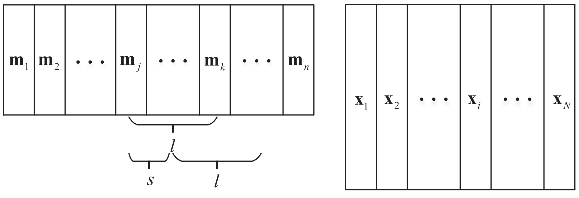
MFCCs features have been used to represent speech signal distribution. MFCCs can be derived through cepstral analysis and warped according to the Mel-scale which is constructed to reflect the frequency sensitivity of the human ear (which is better at low frequencies than at high frequencies). To obtain the short-term MFCCs features, first the higher frequency components of the speech signal are enhanced with a pre-emphasis filter $1 - 0.9z^{-1}$, then the pre-emphasized speech signal is separated into n time frames with a Hamming window, where adjacent time frames overlap by 50%. The frames are transformed from the time domain into the frequency domain by the Fast Fourier Transform (FFT) to obtain the amplitude spectrum. The next processing step is computing the logarithm of the amplitude spectrum, and then the logarithm of the spectrum is frequency warped to transform the spectrum into Mel frequency by using a Mel filter bank. Finally, the log Mel spectrum is converted to the time domain using the discrete cosine transform, and the MFCCs can be obtained retaining a number of leading coefficients. The number of resulting Mel-frequency cepstral coefficients is generally chosen between 12 and 20, and it is set to 12 in this paper.

The super MFCCs frames $\{\mathbf{x}_i\}_{i=1}^{N}$ are built by the MFCCs frames $\{\mathbf{m}_j\}_{j=1}^{n}$ are shown in Fig. 2, where $N < n$. The l MFCCs frames, from the ith MFCCs frame $\mathbf{m}_i$ to the kth MFCCs frame $\mathbf{m}_k$ are shown in Fig. 2(a), are stacked into a frame called the super MFCCs frame, i.e.,

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{m}_i \\ \mathbf{m}_{i+1} \\ \vdots \\ \mathbf{m}_k \end{bmatrix} \tag{1}$$

where $\mathbf{x}_i$ is a super MFCCs frame, and the dimension of the super MFCCs frame is $12l$. For each super MFCCs frame, there are s MFCCs frames which are different from the other super MFCCs frames. Then the N super MFCCs frames can be constructed with n MFCCs frames,

$$N = \left\lfloor \frac{n-l}{s} \right\rfloor \tag{2}$$



(a) MFCCs (b) MFCCsmap

FIGURE 2. The building of MFCCsmap. (a) The block diagram of MFCCs, where $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_j, \ldots, \mathbf{m}_i, \ldots, \mathbf{m}_k, \ldots, \mathbf{m}_n\}$ represent the MFCCs frames. (b) The b lock diagram of MFCCsmap, where $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_N\}$ represent the super MFCCs frames.

where $\lfloor \cdot \rfloor$ means that the fraction $\frac{n-l}{s}$ is rounded down. As shown in Fig. 2(b), the N super MFCCs frames are combined into a $(12l) \times N$ 2-D MFCCsmap $\mathbf{X}$.

B. SPEAKER MODEL AND LONG-TERM ACOUSTIC FEATURES

Based on the analysis sparse representation model, the speaker model and the LTA features of the speaker's speech can be obtain using the MFCCsmap. For simplicity, the super MFCCs frame of the MFCCsmap is denoted by $\mathbf{x} \in R^{M \times 1}$, $M = 12l$. Generally, the analysis sparse model is representation as

$$\min \|\mathbf{f}\|_0 \quad \text{s.t. } \mathbf{f} = \Omega \mathbf{x} \tag{3}$$

where Ω is the analysis dictionary, $\mathbf{f} = \Omega \mathbf{x}$ is the sparse representation of $\mathbf{x}$. In the proposed speaker recognition system, Ω is learned from a given speaker MFCCsmap, so that Ω describes the feature properties for the speaker and can be used as the adaptive speaker model. The sparse representation of $\mathbf{x}$ can be obtained with Ω and the speaker MFCCsmap, so that $\mathbf{x}$ is dependent on the speaker model and make use of the static and dynamic information of the speaker's speech, and then $\mathbf{x}$ can be used as the LTA feature to make effective speaker-recognition decisions.

To obtain Ω and $\mathbf{x}$ by resolving the problem (3), the ℓ_0 quasi-norm could be replaced by the ℓ_1-norm, i.e.,

$$\min \|\mathbf{f}\|_1 \quad \text{s.t. } \mathbf{f} = \Omega \mathbf{x} \tag{4}$$

Then the constrained optimisation problem (4) can be transformed into an unconstrained optimisation problem with the Lagrange multiplier α

$$\min \|\mathbf{f}\|_1 + \tfrac{\alpha}{2} \|\mathbf{f} - \Omega \mathbf{x}\|_F^2 \tag{5}$$

However, there is a trivial solution for the model (5), that is, $\Omega = 0$, $\mathbf{f} = 0$. To avoid such a solution, a function defined on $\Omega^T \Omega = \mathbf{I}$ has been imposed as a constraint term, which enforces Ω to be a full column rank matrix based on the fact ranks of Ω. This leads to the following new optimization criterion

$$\min \|\mathbf{f}\|_1 + \frac{\alpha}{2} \|\mathbf{f} - \Omega \mathbf{x}\|_F^2 \quad \text{s.t. } \Omega^T \Omega = \mathbf{I} \tag{6}$$

Using a Lagrangian multiplier $\tau > 0$, the optimization problem can be reformulated as

$$\min \|\mathbf{f}\|_1 + \tfrac{\alpha}{2} \|\mathbf{f} - \Omega \mathbf{x}\|_F^2 + \tfrac{\tau}{4} \left\| \Omega^T \Omega - \mathbf{I} \right\|_F^2 \tag{7}$$

Note also that when $\mathbf{f}$ is sparse, minimizing $\|\mathbf{f}\|_1$ can be obtained by the minimization of $\|\mathbf{f} - \Omega \mathbf{x}\|_F^2$, subject to the sparsity constraint. However, both $\mathbf{f}$ and Ω are unknown. To solve the problem, the orthogonality constrained analysis dictionary learning with iterative hard thresholding (OIHT-ADL) algorithm is employed to update the estimation of $\mathbf{f}$ and Ω [30].

1) ESTIMATING f

Given the dictionary Ω, considering the optimization of $\mathbf{f}$ only, the objective function (7) can be modified as

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \|\mathbf{f}\|_1 + \frac{\alpha}{2}\|\mathbf{f} - \Omega\mathbf{x}\|_F^2 \tag{8}$$

The first-order optimality condition of $\mathbf{f}$ implies that

$$\alpha(\mathbf{f} - \Omega\mathbf{x}) + \operatorname{sgn}(\mathbf{f}) = 0 \tag{9}$$

Therefore, we have

$$\hat{f}_p = \begin{cases} (\Omega\mathbf{x})_p - \frac{1}{\alpha}, & (\Omega\mathbf{x})_p > \frac{1}{\alpha}; \\ (\Omega\mathbf{x})_p + \frac{1}{\alpha}, & (\Omega\mathbf{x})_p < -\frac{1}{\alpha}; \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

where $\hat{f}_p$ is the pth vector element of $\mathbf{f}$ with $p = 1, 2, \ldots, P$. It is generally known that the above solution for $\mathbf{f}$ is called soft thresholding. Indeed, the sparsity constraint $\|\mathbf{f}\|_1$ is only approximate for $\|\mathbf{f}\|_0$. In order to promote sparsity and to improve the approximation, the hard thresholding method [31] is used as an alternative, that is, setting the smallest components of the vectors to be zeros while retaining the others:

$$\hat{f}_p = \begin{cases} (\Omega\mathbf{x})_p, & \left|(\Omega\mathbf{x})_p\right| > \varepsilon_p; \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

where ε is the value of the hard thresholding. As such, the solution obtained using the constraint $\|\mathbf{f}\|_1$ will be closer to that using $\|\mathbf{f}\|_0$.

2) UPDATING Ω

For a given $\mathbf{f}$, the cost function (7) of estimating Ω can be rewritten as

$$\begin{aligned}\hat{\Omega} &= \arg\min_{\Omega} L(\Omega) \\ &= \arg\min_{\Omega} \frac{\alpha}{2}\|\mathbf{f} - \Omega\mathbf{x}\|_F^2 + \frac{\tau}{2}\left\|\Omega^T\Omega - \mathbf{I}\right\|_F^2\end{aligned} \tag{12}$$

A simple gradient descent method is used to find the local minimum of the objective function (12), i.e.,

$$\Omega^{t+1} = \Omega^t - \gamma\nabla L(\Omega), \tag{13}$$

where γ is a step size and $\nabla L(\Omega) = -\alpha(\mathbf{f} - \Omega\mathbf{x})\mathbf{x}^T + \tau\left(\Omega\Omega^T - \mathbf{I}\right)\Omega$. Considering that the rows of Ω probably have different scales of norm and even some are possibly zeros, the rows of Ω could be normalized to prevent these situations, if any, by the normalized random vectors, that is,

$$\hat{\omega}_p = \begin{cases} \frac{\hat{\omega}_p}{\left\|\hat{\omega}_p\right\|_2}, & \left\|\hat{\omega}_p\right\|_2 \neq 0; \\ \lambda, & \text{otherwise.} \end{cases} \tag{14}$$

where $\hat{\omega}_p$ is the pth row of $\hat{\Omega}$ and λ is a normalized random vector.

C. DNN CLASSIFIER

For speaker recognition, a deep neural network was applied as the classifier in the paper. The DNN uses four hidden layers, each having R sigmoidal hidden units. The number of units in the input layer equals the dimension of the input feature vector. A softmax layer is used as output layer of the DNN for multicategory classification, the number of output layer units is the total number of speakers to be identified.

The parameters of the hidden layers are initialized through the pre-training based on four restricted Boltzmann machine (RBM) that are stacked, where the hidden layer of the previous RBM serves as the visible layer of the next RBM. The four hidden layers are trained layer-by-layer in a unsupervised greedy manner. The input data of the DNN is the LTA feature $\mathbf{f}$, and the Gaussian-Bernoulli RBM (GB-RBM) is used as the first layer of the DNN to represent the distribution of the $\mathbf{f}$. The probability distribution assigned to the visible-hidden units pair is defined by

$$P\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right) = \frac{e^{-E\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right)}}{Z} \tag{15}$$

where $\mathbf{f}$ is also the state of the visible layer, $\mathbf{h}^1$ denotes the state of hidden layer, and the superscript is the index of hidden layer of the DNN. Z is the normalization term defined by $\sum_{\mathbf{f}, \mathbf{h}^1} e^{-E\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right)}$. $E\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right)$ represents the energy function of the visible layer units and the hidden layer units, i.e.,

$$E\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right) = \sum_p \frac{\left(f_p - o_p^1\right)^2}{2\sigma_p^2} - \sum_r b_r^1 h_r^1 - \sum_{p,r} \frac{f_p}{\sigma_p} h_r^1 w_{p,r}^1 \tag{16}$$

where p, r are the indexes of the visible layer units and hidden layer units, respectively. $\theta^1 = \left(\mathbf{o}^1, \mathbf{b}^1, \mathbf{w}^1\right)$ are the parameters of the first layer. $\mathbf{o}^1$ are biases of the visible layer units, $\mathbf{b}^1$ is the biases of hidden layer, and $\mathbf{w}^1$ are the weights between visible layer units and hidden layer units. σ_p is the standard deviation of the Gaussian noise. The log likelihood of the LTA feature is used as the objective function, i.e.,

$$L\left(\theta^1\right) = \log\prod_p P\left(f_p\right) = \log\frac{\sum_{\mathbf{h}^1} e^{-E\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right)}}{\sum_{\mathbf{f}, \mathbf{h}^1} e^{-E\left(\mathbf{f}, \mathbf{h}^1; \theta^1\right)}} \tag{17}$$

The parameters can be updated by using the stochastic gradient method, i.e.,

$$\theta^1(t+1) = \delta\theta^1(t) - \eta_1\frac{\partial L\left(\theta^1\right)}{\partial\theta^1(t)} \tag{18}$$

where t is the index of the iteration. δ is the momentum factor used to smooth out the weight updates, and η_1 is the learning rate in the training phase. After the first layer has been trained, the states of the binary hidden units $\mathbf{h}^1$ of the GB-RBM are used as the input data for training the next RBM. The rest of the hidden layers of the DNN are pre-trained by the Bernoulli-Bernoulli RBM. For the first BB-RBM example, the energy function of the visible layer units and hidden layer units is defined by

$$E\left(\mathbf{h}^1, \mathbf{h}^2\right) = \sum_r b_r^1 h_r^1 - \sum_r b_r^2 h_r^2 - \sum_{r,r} h_r^1 h_r^2 w_{r,r}^2 \tag{19}$$

where $\mathbf{h}^1$ is the state of visible layer units in the first BB-RBM, and $\mathbf{h}^2$ is the state of hidden layer units.

After pre-training, a randomly initialized softmax layer is used as the output layer of the DNN for multicategory classification, the output of the dth node in the softmax layer is the conditional probability of the current case belonging to the dth speaker. The calculation of the dth node in the output layer can be described as below

$$z_d = \frac{\exp\left(q_d\right)}{\sum_c \exp(q_c)} \tag{20}$$

where c is an index over all output units, and also the index of the speaker, q_d is the input of the dth unit in the softmax layer, and z_d is the output of the dth unit. The cost function C is the cross entropy between the target output $\hat{z}_d$ and the output of the softmax z_d,

$$C = -\sum_d \hat{z}_d \log z_d \tag{21}$$

where $\hat{z}_d$ is the label for the input LTA feature $\mathbf{f}$, taking value of one when the LTA feature belongs to the dth class, otherwise taking value of zero. The gradient descent back-propagation algorithm is carried to fine-tune all parameters for both hidden layers and output layer, i.e.,

$$\theta^{(\rho)}\left(t+1\right) = \theta^{(\rho)}\left(t\right) - \eta_2 \frac{\partial C}{\partial \theta^{(\rho)}\left(t\right)} \tag{22}$$

where η_2 is the learning rate, and ρ is the index of the layer.

III. EXPERIMENTS

We carried out two experiments to evaluate the proposed long-term acoustic features on a speaker recognition task. The evaluation metric is the average classification accuracy rate (ACA)

$$ACA = \frac{number\ of\ correct\ classified\ samples}{number\ of\ total\ testing\ samples} \times 100\% \tag{23}$$

In the first experiment, the effectiveness of the proposed LTA is shown by comparing with the performance of the same recognizer fed with different features, and the recognition performance of the proposed speaker recognition systems is demonstrated by comparing the proposed system with the speaker recognition method in [32]. In the second experiment, the robustness of the LTA features is demonstrated in the presence of white noise.

We use four different databases to investigate the performance of our speaker recognition system, i.e., the TIMIT database [33], VoxForge database [34], THCHS30 database [35], and LibriSpeech database [36]. Ten speakers were selected from the TIMIT database, i.e., seven male speakers and three female speakers, with 10 English speech sentences of each speaker. We downloaded 12 speakers, consisting of 8 males and 4 females, from the online VoxForge website, with 8 English voice samples per speaker. Ten more speakers, 3 male and 7 female, were selected from the THCHS30 database, which is an open Chinese speech database, and each speaker had five voice samples. Ten speakers, 5 male and 5 female, were selected from the LibriSpeech database which is an English speech database with 8 voice samples per speakers. The sampling rate of all speech materials was 16 kHz, and the sample size was 16 bits. The speakers' material were randomly selected from these four database, and the voice samples of each speaker composed 8s of speech for the training dataset, and 2s for the test dataset. The languages used for the training and testing are same, so the system is language dependent. But, the texts of the speech used in the training and testing are different, so the system is text independent.

A. THE PERFORMANCE OF THE PROPOSED SPEAKER RECOGNITION SYSTEM

This experiment was designed to test the recognition performance of the proposed speaker recognition systems. In the training phase, the training speech for each speaker was firstly pre-emphasized, and then segmented into 799 frames with a 20ms Hanning window size and a 10ms step size. The FFT of each frame was converted from a power spectrum to the mel scale by 24 triangular mel-filters, and then 12-dimensional MFCCs were computed by applying log compression firstly and the DCT transform. With setting $l = 6, s = 1$, 793 super MFCCs frames could be constructed from the 799 MFCCs frames to build the MFCCsmap $\mathbf{X}$, which dimension was 72×793. To get the LTA features $\mathbf{f}_c$ of the training speech, the speaker model Ω_c, i.e., the dictionary, was set as 792×72, and then was randomly initialized. The Ω_c was firstly learned from the training MFCCsmap $\mathbf{X}_c$ by the OIHT-ADL algorithm, where the parameters were selected as follows: $\alpha = 1 \times 10^{-3}$, $\tau = 1$, $\gamma = 1 \times 10^{-3}$ and the iteration number was 30. Then, we multiplied the Ω_c by the $\mathbf{X}_c$, and the $\mathbf{f}_c$ were obtained from the product by using the hard thresholding method, where the 50 smallest values in $\Omega \mathbf{x}_p$ were set as the value of the hard thresholding ε_p. The obtained $\mathbf{f}_c$ were used to train the corresponding DNN classifier. During pre-training, 5 iterations were used to pre-train each hidden layer of the DNN with a learning rate of 0.00025 and a momentum term of 0.9. For fine-tuning, 300 iterations were used and the learning rate was set to 0.01. These parameters of the DNN classifier were selected by some experiences. In the test phase, the MFCCs were first extracted from the test speech signal. The test MFCCsmap of all speech $\hat{\mathbf{X}}$ was built, then the product of the $\hat{\mathbf{X}}$ and the learned speaker model Ω_c was calculated, the $\hat{\mathbf{f}}_c$ was obtained from the product by using the hard thresholding method. Finally, the $\hat{\mathbf{f}}_c$ was recognized by the corresponding DNN classifier, where the probability value of the softmax layer was calculated. The ACA of this experiment was gained by comparing the output of all DNNs. Each experiment did ten times under the same conditions and its ACAs were recorded.

The average value of ten ACAs was calculated and presented in Fig. 3, from which we can observe that the proposed system performed about 90% correct speaker identification.

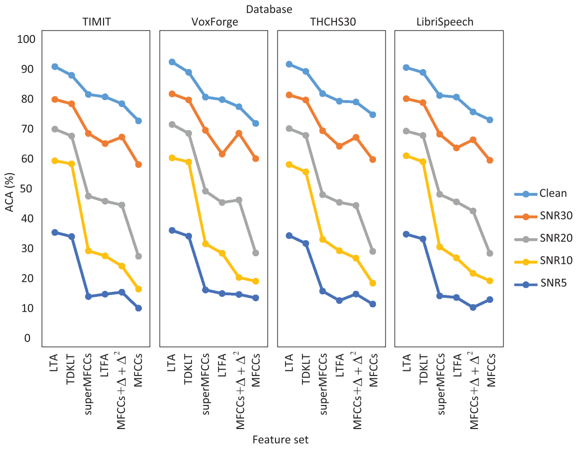


FIGURE 3. ACA (%) by the different speaker recognition system for four different databases.

We compared the recognition performance of the proposed long-term features with that of a static MFCCs feature extraction method and the three MFCCs-based methods of dynamic feature extraction. Both the training and test speech are converted to the following four features:

- MFCCs features: The speech signal was first pre-emphasized, and then segmented into 799 frames with a 20ms Hanning window size and a 10ms step size. The FFT of each frame was converted from power spectrum into mel scale by 24 triangular mel-filters, and then 12-dimensional MFCCs were computed by applying log compression firstly and the DCT transform. The MFCCs were the static only features and shared by the static part of all features.
- MFCCs $+ \Delta + \Delta^2$ features: Firstly, the static MFCCs was computed with the same method as above. Then, the 12 dimensional ΔMFCCs frame was computed by one MFCCs frame and its neighbor 4 MFCCs frames. The 12 dimensional Δ^2MFCCs frame was computed by one ΔMFCCs frame and its neighbor 4 ΔMFCCs frames. The 36-dimensional MFCCs $+ \Delta + \Delta^2$ vector was built by a MFCCs frame, a ΔMFCCs frame, and a Δ^2MFCCs frame.
- The super MFCCs features: This baseline followed the idea proposed by Yu *et al.* [11] with slight modification. The 12-dimensional MFCCs were first extracted from the speech of a speaker, and then the 72-dimensional super MFCCs frames were built by the method in Section II-B with the parameter $l = 6$.
- LTFA featrues: The LTFA features were first proposed in [12], which 12-dimensional MFCCs were extracted, and the average of four successive MFCCs frames was calculated as a 12-dimensional LTFA feature vector.

These recognition systems used also the DNN with four hidden layers to classify the different speakers. The number of the four DNNs input layer unit was equal to the dimension of the input feature vectors, and the number of the hidden layers units were same, i.e., {1584, 1584, 1584, 1584}. The output layers of the four DNNs were the softmax layer, the number of units depends on the number of speakers. To get the best performance of these system, 5 iterations were used for pre-training each RBM with a learning rate of 0.00025 and a momentum term of 0.9, and in the fine-tuning, the mean square error was used as the cost function and a learning rate of 0.01 for the 500 iterations.

We also compared the proposed speaker recognition with the recent state of the art techniques, i.e., the truncated discrete Karhunen-Loeve transform (TDKLT) features [32]. In [32], the speech signal was then divided into overlapping frames of 25ms, with frame shift of 10ms. Each frame was cleaned up by a noise reduction block based on the Wiener filter. Further signal enhancements were then performed by a SNR-dependent waveform processing phase. Then, the 12-dimensional TDKLT features were computed with the truncated discrete Karhunen-Loeve transform. Finally, the TDKLT features were used for training a truncated Bayesian classification. The results of experiment were shown in Fig. 3, and the performance of the proposed system was better than the method in [32].

The results of the five speaker recognition systems were shown in Fig. 3. Clearly, the proposed LTA features can achieve as high an accuracy rate as 92.39%. The best ACA of the speaker recognition system based the TDKLT was the 89.32%. The proposed method is more effective for speaker recognition task. In comparison with the static MFCCs features, the performances of the others features added the dynamic information have been improved. The MFCCs $+ \Delta + \Delta^2$ features did not achieve good recognition results in this experiment, and the average ACA across the four databases was 78.01% during the test. The super MFCCs features can yield better results than that of the LTFA features, and the average result of the super MFCCs features was 81.57%, and the LTFA features was 80.42%. Experimental results indicate that the super MFCCs features can offer a higher discriminability than the other two dynamic features. By the analysis sparse representation, the average recognition rate of the proposed LTA features across the four databases was increased to 91.37%.

B. THE ROBUSTNESS OF THE LTA FEATURES IN WHITE NOISE CORRUPTION

For evaluating the robustness of this speaker recognition system, we will now examine the recognition accuracy of the proposed features under white noise conditions. In this experiment, the training speech was still the 8s clean speech signal, while the test speech was the noisy speech signal which was generated by adding Gaussian white noise to the 2s clean speech with a specified signal-to-noise ratio (SNR). The noise series were generated by the Gaussian white noise

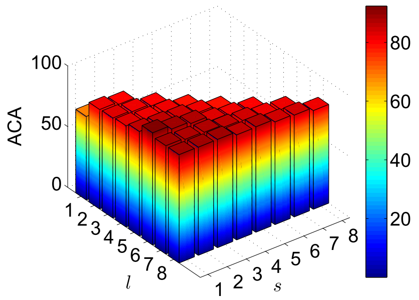


FIGURE 4. ACA(%) of the speaker recognition system as a function of *l* and *s* on MFCCsmap.

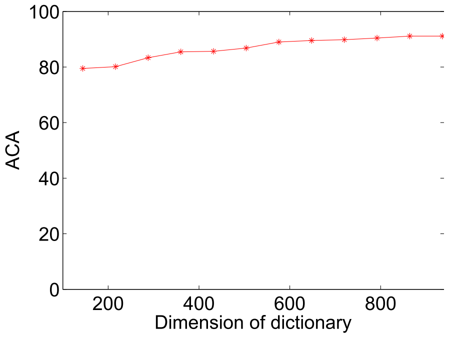
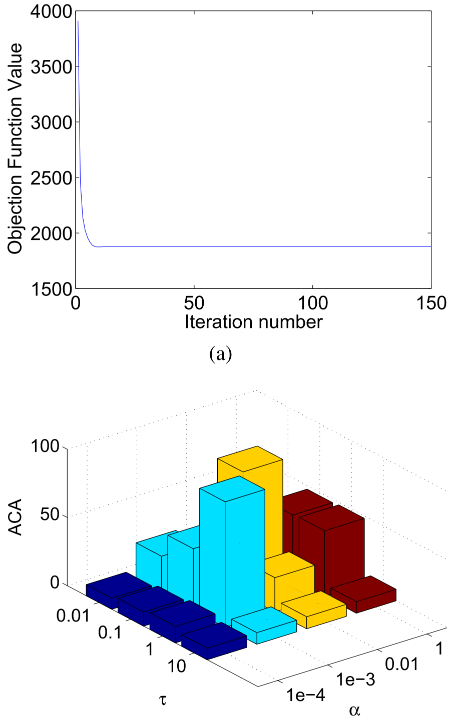


FIGURE 5. ACA(%) of speaker recognition based on the ASR with different dictionary dimensions.



(a)

(b)

FIGURE 6. (a) Convergence rate of ASR for solving its proposed objective function with $\alpha = 1 \times 10^{-3}$ and $\tau = 1$; (b) ACA(%) of the speaker recognition on different databases with different parameters for ASR.

function of Matlab, and the signal noise ratio (SNR) was set at values of 5, 10, 20 and 30 dB. The results of the experiment were shown in Fig. 3.

As shown in Fig. 3, the ACAs of using LTA features had some reduction with the decrease of the SNR, while the accuracy of speaker recognition systems using the MFCCs features or the others dynamic features were diminished rapidly by the white noise. As shown in Fig.3, the speaker recognition system with LTA features performed better than the MFCCs features and the others features. At 30 dB, the average ACA of using the LTA feature across the four databases was 81.05%. In the same condition, the ACA of the MFCCs feature was 60.17%, and the best performance of the other dynamic features was the super MFCCs features, with an average ACA of 69.50% across the four databases. From what had been mentioned above, it was clear that the robustness of LTA features was better than the super MFCCs features. It is probably because that the analysis sparse model has robustness.

C. SELECTION OF THE PARAMETERS

In the proposed speaker recognition system, the parameters of constructing the MFCCsmap and the OIHT-ADL algorithm should be selected. Firstly, the parameters of constructing the MFCCsmap were selected. The ACA of the proposed speaker recognition was used to show their performance. To build the MFCCsmap of the speaker, the l MFCCs frames were stacked to construct a super MFCCs frame, and each super MFCCs frame had s MFCCs frame(s) which were different from each other. The influences of the parameters l and s were examined, where the parameter l was selected from the range of $1 \leq l \leq 8$, and the range of s was less than or equal to l, but never less than 1. Thus, the range of the super MFCCs frame's dimension was from 12 to 96, and all the super MFCCs frames were combined into the MFCCsmap. The speaker model Ω was built by the method described in section II-B, where the dimension of the dictionary was ten times the super MFCCs frame' dimension, and the parameters were set with $\alpha = 1 \times 10^{-3}$, $\tau = 1$, and $\gamma = 1 \times 10^{-3}$. The experimental results on the VoxForge database were shown in the Fig. 4, from which we could infer that increasing l improves the recognition accuracy of the system, but when l was more than 6, the rate does not increase. In addition, increasing s degrades the recognition accuracy of system. For all databases, $l = 6$ and $s = 1$ were set for constructing MFCCsmap.

Next, the parameters of the OIHT-ADL algorithm were selected. The convergence curves of the objective function in (7) and the ACA of the proposed speaker recognition system were used to show their performance, and the parameters of the algorithm were set empirically by experimental tests. The dimension of dictionary Ω was first considered in the experiment. We tested the dimension of dictionary with different values in the ranges of $144 \leq P \leq 936$. The results were shown in Fig. 5. With the increasing the dimension of dictionary, the performance of the classification became better. The dictionary with a bigger size may lead to a sparser representation of the signal but may have a higher computational cost. For obtaining stable classification performance, the size of dictionary Ω was set as 792×72. Then, the convergence rate of ASR was calculated on the objective function values, and the parameters of the objective function were selected for the algorithm. We tested the parameters γ with different values in the ranges of $1 \times 10^{-3} \leq \gamma \leq 100$, and set $\gamma = 1 \times 10^{-3}$. The parameters were tested with different values in the ranges of $1 \times 10^{-4} \leq \alpha \leq 1$ and $1 \times 10^{-2} \leq \tau \leq 10$; the results were shown in Fig. 6. As shown in Fig. 6(a), the objective function values of the algorithms with $\alpha = 1 \times 10^{-3}$ and $\tau = 1$ rapidly decreased and became stable after about 30 iterations on the VoxForge database. Then, the sensitivities of the parameters α and τ were studied and shown in Fig. 6(b) with the ACA on the VoxForge database. According to the results shown in Fig. 6, $\alpha = 1 \times 10^{-3}$, $\tau = 1$, were set for the VoxForge database. The experiments were repeated for the other three databases, and it was found that the parameters could be applied to the other databases.

IV. CONCLUSION

In this paper, the speaker model is built based on the analysis sparse representation, and LTA features were extracted from the MFCCsmap of the speaker's speech with the speaker model. Both static and dynamic information of the speech signal could be included in the LTA feature. The four types of the features (MFCCs features, the MFCCs $+ \Delta + \Delta^2$ features, the super-vector MFCCs features, and the LTFA features) with DNN classifier, and the TDKL method were used as baseline methods for experimental comparison. The proposed LTA features were found to be robust and outperformed all baseline conditions. In this paper, some small speaker databases were used to test the speaker recognition system. Larger speaker database can be covered by increasing the complexity of DNN, increasing the number of hidden layers or the node number of hidden layers. However, along with the increase of the DNN's complexity, the requirement of computational resources will greatly increase. Due to the restriction of the available computational resources, the experimental verification of using large database has not been completed. This problem should be addressed by finding an even more effective DNN classifier in the future.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful comments and suggestions.

REFERENCES

[1] J. P. Campbell, Jr., "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[2] H. Anwer, S. Anjum, and N. A. Saqib, "Robust speaker recognition for e-commerce system," in *Proc. Int. Conf. Radar, Antenna, Microw., Electron. Telecommun.*, Oct. 2015, pp. 92–97.

[3] E. Martinson and W. Lawson, "Learning speaker recognition models through human-robot interaction," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2011, pp. 3915–3920.

[4] F. Guapo, P. Correia, D. Meuwly, and D. van der Vloed, "Empirical validation of likelihood ratio methods—A case study in forensic speaker recognition," in *Proc. 4th Int. Conf. Biometrics Forensics (IWBF)*, Mar. 2016, pp. 1–5.

[5] K. Khelif, Y. Mombrun, G. Backfried, F. Sahito, L. Scarpato, P. Motlicek, S. Madikeri, D. Kelly, G. Hazzani, and E. Chatzigavriil, "Towards a breakthrough speaker identification approach for law enforcement agencies: SIIP," in *Proc. Eur. Intell. Secur. Inform. Conf.*, Sep. 2017, pp. 32–39.

[6] A. Awais, S. Kun, Y. Yu, S. Hayat, A. Ahmed, and T. Tu, "Speaker recognition using mel frequency cepstral coefficient and locality sensitive hashing," in *Proc. Int. Conf. Artif. Intell. Big Data*, May 2018, pp. 271–276.

[7] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 52–55, Jan. 2006.

[8] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Process. Lett.*, vol. 17, no. 6, pp. 599–602, Jun. 2010.

[9] X. Zhao, Y. Shao, and D. Wang, "CASA-based robust speaker identification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1608–1616, Jul. 2012.

[10] S. Memon, I. A. Jokhio, S. H. Arisar, M. Lech, and N. Maddage, "Delta-MFCC features and information theoretic expectation maximization based text-independent speaker verification system," *IETE J. Res.*, vol. 58, no. 1, pp. 5–12, 2012.

[11] H. Yu, Z. Ma, M. Li, and J. Guo, "Histogram transform model using MFCC features for text-independent speaker identification," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 500–504.

[12] C.-L. Huang, C. Hori, H. Kashioka, and B. Ma, "Speaker clustering using vector representation with long-term feature for lecture speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 3532–3536.

[13] Y.-H. Chin, J.-C. Wang, C.-L. Huang, K.-Y. Wang, and C.-H. Wu, "Speaker identification using discriminative features and sparse representation," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 8, pp. 1979–1987, Aug. 2017.

[14] J. M. K. Kua, J. Epps, and E. Ambikairajah, "*i*-vector with sparse representation classification for speaker verification," *Speech Commun.*, vol. 55, no. 5, pp. 707–720, 2013.

[15] Q. Wu and L. Zhang, "Auditory sparse representation for robust speaker recognition based on tensor structure," *EURASIP J. Audio, Speech, Music Process.*, vol. 2008, no. 1, Nov. 2008, Art. no. 578612.

[16] M. Hasheminejad and H. Farsi, "Frame level sparse representation classification for speaker verification," *Multimedia Tools Appl.*, vol. 76, no. 20, pp. 21211–21224, Oct. 2017.

[17] B. C. Haris and R. Sinha, "Sparse representation of total variability smoothed GMM mean supervectors for speaker verification," in *Proc. Int. Conf. Signal Process. Commun.*, Jul. 2012, pp. 1–5.

[18] M. G. Jafari and M. D. Plumbley, "Speech denoising based on a greedy adaptive dictionary algorithm," in *Proc. 17th Eur. Signal Process. Conf.*, Aug. 2009, pp. 1423–1426.

[19] B. C. Haris and R. Sinha, "Speaker verification using sparse representation over KSVD learned dictionary," in *Proc. Nat. Conf. Commun.*, Feb. 2012, pp. 1–5.

[20] Y. Ohta and T. Aida, "Sparse representation approach to inverse halftoning in terms of DCT dictionary," in *Proc. 14th Int. Conf. Control, Automat. Syst.*, Oct. 2014, pp. 1377–1380.

[21] S. Ayas and M. Ekinci, "Single image super resolution based on sparse representation using discrete wavelet transform," *Multimedia Tools Appl.*, vol. 77, no. 13, pp. 16685–16698, Jul. 2018.

[22] Z. Wu, H. Gao, Y. Chen, and H. Kang, "A new image sparse reconstruction method for mixed Gaussian–Poisson noise with multiple constraints," in *Computer Vision*. Singapore: Springer, 2017, pp. 345–356.

[23] Z. He, S. Xie, and Y. Fu, "Sparse representation and blind source separation of ill-posed mixtures," *Sci. Chin. F, Inf. Sci.*, vol. 49, no. 5, pp. 639–652, Oct. 2006.

[24] W. Fang, H. Wang, B. Xu, and Y. Zhang, "Blind source separation using analysis sparse constraint," *Electron. Lett.*, vol. 52, no. 13, pp. 1112–1114, 2016.

[25] Y. Zhang, B. Xu, and N. Zhou, "A novel image compression–encryption hybrid algorithm based on the analysis sparse representation," *Opt. Commun.*, vol. 392, pp. 223–233, Jun. 2017.

[26] R. Thanki and S. Borra, "Fragile watermarking for copyright authentication and tamper detection of medical images using compressive sensing (CS) based encryption and contourlet domain processing," *Multimedia Tools Appl.*, vol. 78, no. 10, pp. 13905–13924, Oct. 2018.

[27] S. Shekhar, V. M. Patel, and R. Chellappa, "Analysis sparse coding models for image-based classification," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2015, pp. 5207–5211.

[28] Z. Wen, B. Hou, and L. Jiao, "Discriminative nonlinear analysis operator learning: When cosparse model meets image classification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3449–3462, Jul. 2017.

[29] Z. Qawaqneh, A. A. Mallouh, and B. D. Barkana, "Deep neural network framework and transformed MFCCs for speaker's age and gender classification," *Knowl.-Based Syst.*, vol. 115, pp. 5–14, Jan. 2017.

[30] Y. Zhang, T. Yu, and W. Wang, "An analysis dictionary learning algorithm under a noisy data model with orthogonality constraint," *Sci. World J.*, vol. 2014, no. 11, Jul. 2014, Art. no. 852978, 2014.

[31] M. Elad, M. A. T. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proc. IEEE*, vol. 98, no. 6, pp. 972–982, Jun. 2010.

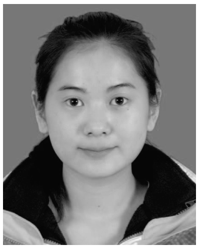
[32] G. Biagetti, P. Crippa, L. Falaschetti, S. Orcioni, and C. Turchetti, "An investigation on the accuracy of truncated DKLT representation for speaker identification with short sequences of speech frames," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4235–4249, Dec. 2017.

[33] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue. (1993). *Timit Acoustic-Phonetic Continuous Speech Corpus*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1

[34] K. Maclean. (2018). *Voxforge*. [Online]. Available: http://www.voxforge.org/

[35] D. Wang and X. Zhang, "THCHS-30: A free Chinese speech corpus," 2015, *arXiv:1512.01882*. [Online]. Available: https://arxiv.org/abs/1512.01882

[36] P. Vassil, C. Guoguo, P. Daniel, and K. Sanjeev. (2015). *Librispeech: An ASR Corpus Based on Public Domain Audio Books*. [Online]. Available: http://www.openslr.org



TING LIN received the B.E. degree from Liaocheng University, China, in 2016. She is currently pursuing the master's degree with Nanchang University, China. Her current research interest includes speaker recognition.

YE ZHANG received the Ph.D. degree in information and communication engineering from Shanghai University, Shanghai, China, in 2009. He is currently a Professor with Nanchang University, Nanchang, China. His current research interests include blind signal processing, speech and image signal processing, pattern recognition, and machine learning.