Hindawi
Journal of Electrical and Computer Engineering
Volume 2017, Article ID 1735698, 9 pages
https://doi.org/10.1155/2017/1735698

Research Article

Speaker Recognition Using Wavelet Packet Entropy, I-Vector, and Cosine Distance Scoring

Lei Lei and She Kun

Laboratory of Cyberspace, School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610054, China

Correspondence should be addressed to She Kun; kun@uestc.edu.cn

Received 16 February 2017; Revised 17 April 2017; Accepted 26 April 2017; Published 14 May 2017

Academic Editor: Lei Zhang

Today, more and more people have benefited from the speaker recognition. However, the accuracy of speaker recognition often drops off rapidly because of the low-quality speech and noise. This paper proposed a new speaker recognition model based on wavelet packet entropy (WPE), i-vector, and cosine distance scoring (CDS). In the proposed model, WPE transforms the speeches into short-term spectrum feature vectors (short vectors) and resists the noise. I-vector is generated from those short vectors and characterizes speech to improve the recognition accuracy. CDS fast compares with the difference between two i-vectors to give out the recognition result. The proposed model is evaluated by TIMIT speech database. The results of the experiments show that the proposed model can obtain good performance in clear and noisy environment and be insensitive to the low-quality speech, but the time cost of the model is high. To reduce the time cost, the parallel computation is used.

1. Introduction

Speaker recognition refers to recognizing the unknown persons from their voices. With the use of speech as a biometric in access system, more and more ordinary persons have benefited from this technology [1]. An example is the automatic speech-based access system. Compared with the conventional password-based system, this system is more suitable for old people whose eyes cannot see clearly and figures are clumsy.

With the development of phone-based service, the speech used for recognition is usually recorded by phone. However, the quality of phone speech is low for recognition because the sampling rate of the phone speech is only 8 KHz. Moreover, the ambient noise and channel noise cannot be completely removed. Therefore, it is necessary to find a speaker recognition model that is not sensitive to those factors such as noise and low-quality speech.

In a speaker recognition model, the speech is firstly transformed into one or many feature vectors that represent unique information for a particular speaker irrespective of the speech content [2]. The most widely used feature vector is the short vector, because it is easy to compute and yield good performance [3]. Usually, the short vector is extracted by Mel frequency cepstral coefficient (MFCC) method [4]. This method can represent the speech spectrum in compacted form, but the extracted short vector represents only the static information of the speech. To represent the dynamic information, the Fused MFCC (FMFCC) method [5] is proposed. This method calculates not only the cepstral coefficients but also the delta derivatives, so the short vector extracted by this method can represent both the static and dynamic information.

Both of the two methods use discrete Fourier transform (DFT) to obtain the frequency spectrum. DFT decomposes the signal into a global frequency domain. If a part of frequency is destroyed by noise, the whole spectrum will be strongly interfered [6]. In other words, the DFT-based extraction methods, such as MFCC and FMFCC, are insensitive to the noise. Wavelet packet transform (WPT) [7] is other type of tool used to obtain the frequency spectrum. Compared with the DFT, WPT decomposes the speech into many small frequency bands that are independent of each other. Because of those independent bands, the ill effect of noise cannot be transmitted over the whole spectrum. In other words, WPT has antinoise ability. Based on WPT, wavelet packet entropy

(WPE) [8] method is proposed to extract the short vector. References [8–11] have shown that the short vector extracted by WPE is insensitive to noise.

I-vector is another type of feature vector. It is a robust way to represent a speech using a single high-dimension vector and it is generated by the short vectors. I-vector considers both of the speaker-dependent and background information, so it usually leads to good accuracy. References [12–14] have used it to enhance the performance of speaker recognition model. Specially, [15] uses the i-vector to improve the discrimination of the low-quality speech. Usually, the i-vector is generated from the short vectors extracted by the MFCC or FMFCC methods, but we employ the WPE to extract those short vectors, because the WPE can resist the ill effect of noise.

Once the speeches are transformed into the feature vectors, a classifier is used to recognize the identity of speaker based on those feature vectors. Gaussian mixture model (GMM) is a conventional classifier. Because it is fast and simple, GMM has been widely used for speaker recognition [4, 16]. However, if the dimension of the feature vector is high, the curse of dimension will destroy this classifier. Unfortunately, i-vector is high-dimensional vector compared with the short vector. Cosine distance scoring (CDS) is another type of classifier used for the speaker recognition [17]. This classifier uses a kernel function to deal with the problem of high-dimension vector, so it is suitable for the i-vector. In this paper, we employ the CDS for speaker classification.

The main work of this paper is to propose a new speaker recognition model by using the wavelet packet entropy (WPE), i-vector, and cosine distance scoring (CDS). WPE is used to extract the short vectors from speeches, because it is robust against the noise. I-vector is generated from those short vectors. It is used to characterize the speeches used for recognition to improve the discrimination of the low-quality speech. CDS is very suitable for high-dimension vector such as i-vector, because it uses a kernel function to deal with the curse of dimension. To improve the discrimination of the i-vector, linear discriminant analysis (LDA) and the covariance normalization (WCNN) are added to the CDS. Our proposed model is evaluated by TIMIT database. The result of the experiments show that the proposed model can deal with the low-quality speech problem and resist the ill effect of noise. However, the time cost of the new model is high, because extracting WPE is time-consuming. This paper calculates the WPE in a parallel way to reduce the time cost.

The rest of this paper is organized as follows. In Section 2, we describe the conventional speaker recognition model. In Section 3, the speaker recognition model based on i-vector is described. We propose a new speaker recognition model in Section 4, and the performance of the proposed model is reported in Section 5. Finally, we give out a conclusion in Section 6.

2. The Conventional Speaker Recognition Model

Conventional speaker recognition model can be divided into two parts such as short vector extraction and speaker classification. The short vector extraction transforms the speech into the short vectors and the speaker classification uses a classifier to give out the recognition result based on the short vectors.

2.1. Short Vector Extraction. Mel frequency cepstral coefficient (MFCC) method is the conventional short vector extraction algorithm. This method firstly decomposes the speech into 20–30 ms speech frames. For each frame, the cepstral coefficient can be calculated as follows [18]:

(1) Take DFT of the frame to obtain the frequency spectrum.

(2) Map the power of the spectrum onto Mel scale using the Mel filter bank.

(3) Calculate the logarithm value of the power spectrum mapped on the Mel scale.

(4) Take DCT of logarithmic power spectrum to obtain the cepstral coefficient.

Usually, the lower 13-14 coefficients are used to form the short vector. Fused MFCC (FMFCC) method is the extension of MFCC. Compared with MFCC, it further calculates the delta derivatives to represent the dynamic information of speech. The derivatives are defined as follows [5]:

$$d_i = \frac{\sum_{p=1}^{2} p\left(cc_{i-p} + cc_{i+p}\right)}{2\sum_{p=1}^{2} p^2},$$
$$dd_i = \frac{\sum_{p=1}^{2} p\left(d_{i-p} + d_{i+p}\right)}{2\sum_{p=1}^{2} p^2}; \quad (1)$$
$$i = 1, 2, 3, \ldots,$$

where cc_i is the ith cepstral coefficient obtained by the MFCC method and p is the offset. d_i is the ith delta coefficient and dd_i is the ith delta-delta coefficient. If the short vector extracted by MFCC is denoted as $[cc_1, cc_2, cc_3, \ldots, cc_I]^T$, then the short vector extracted by FMFCC is denoted as $[cc_1, cc_2, cc_3, \ldots, cc_I;\ d_1, d_2, \ldots, d_I;\ dd_1, dd_2, \ldots, dd_I]^T$.

2.2. Speaker Classification. Gaussian mixture model (GMM) is a conventional classifier. It is defined as

$$p(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m G\left(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\right), \quad (2)$$

where $\mathbf{x}$ is a short vector extracted from an unknown speech. $G(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is the mth Gaussian function in GMM, where $\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m$ are its mean vector and variance matrix, respectively. α_m is the combination weight of the Gaussian function and satisfies $\sum_{m=1}^{M} \alpha_m = 1$. M is the mixture number of the GMM. All of the parameters, such as weights, mean vectors, and variance matrices, are estimated by the famous EM algorithm [19] using the speech samples of a known speaker. In other words, $p(\mathbf{x})$ represents the characteristic of the known speaker's voice, so we use $p(\mathbf{x})$ to recognize the author of

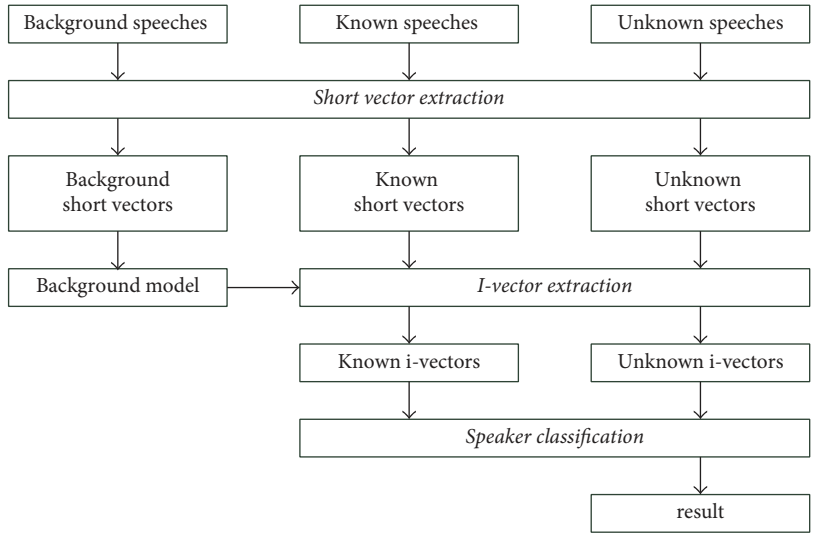

Figure 1: The structure of the speaker recognition using i-vector.

the unknown speeches. Assume that an unknown speech is denoted by $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$, where $\mathbf{y}_i$ represents the ith short vector extracted from $\mathbf{Y}$. Also, assume that the parameters of $p(\mathbf{x})$ are estimated using the speech samples of a known speaker s. The result of recognition is defined as

$$r = \frac{1}{N}\sum_{i=1}^{N} \log\left[p\left(\mathbf{y}_i\right)\right] + \theta, \tag{3}$$

where $\theta > 0$ is the decision threshold and should be adjusted beforehand to obtain the best recognition performance. If $r \leq 0$, then the GMM decides that the author of the unknown speech is not the known speaker s; if the $r > 0$, then the GMM decides that the unknown speech is spoken by the speaker s.

3. The Speaker Recognition Model Using I-Vector

The speaker recognition model using i-vector can be decomposed into three parts such as short vector extraction, i-vector extraction, and speaker classification. Figure 1 shows the structure of the model.

There are three types of speeches used for this model. Background speeches contains thousands of speeches spoken by lots of people, the known speeches are the speech samples of known speakers, and the unknown speeches are spoken by the speaker to be recognized. In the short vector extraction, all of the speeches are transformed into the short vectors by a feature extraction method. In the i-vector extraction, the background short vectors are used to train the background model. The background model is usually represented by a GMM with 2048 mixtures, and all covariance matrices of the GMM are assumed the same for easy computation. Based on the background model, the known and unknown short vectors are used to extract the known and unknown i-vectors, respectively. Note that one i-vector refers to only one speech. In the speaker classification, a classifier is used to match the known i-vector with the unknown i-vector and give out the recognition result.

4. The Proposed Speaker Recognition Model

The accuracy of recognition system usually drops off rapidly because of the low-quality speech and noise. To deal with the problem, we propose a new speaker recognition model based on wavelet packet entropy (WPE), i-vector, and cosine distance scoring (CDS). In Section 4.1, we describe the WPE method and use it to extract the short vector. Section 4.2 describes how to extract the i-vector using the above short vectors. Finally, the details of CDS are described in Section 4.3.

4.1. Short Vector Extraction. This paper uses WPE to extract the short vector. The WPE is based on the wavelet packet transform (WPT) [20], so the WPT is firstly described. WPT is a local signal processing approach that is used to obtain the frequency spectrum. It decomposes the speech into many local frequency bands at multiple levels and obtains the frequency spectrum based on the bands. For the discrete signal such as digital speech, WPT is usually implemented by the famous Mallat fast algorithm [21]. In the algorithm, WPT is realized by a low-pass filter and a high-pass filter, which are generated by the mother wavelet and the corresponding scale function, respectively. Through the two filters, the speech is iteratively decomposed into a low-frequency and a high-frequency components. We can use a full binary tree to describe the process of WPT. The three structures are shown in Figure 2.

In Figure 2, root is the speech to be analyzed. Each nonroot node represents a component. The left child is the low-frequency component of its parent and the right child is the high-frequency component of its parent. The left branch and the right branch are the low-pass and high-pass filtering

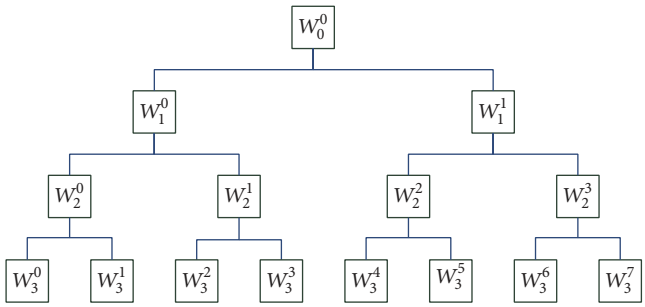

Figure 2: The wavelet packet transform at 2 levels.

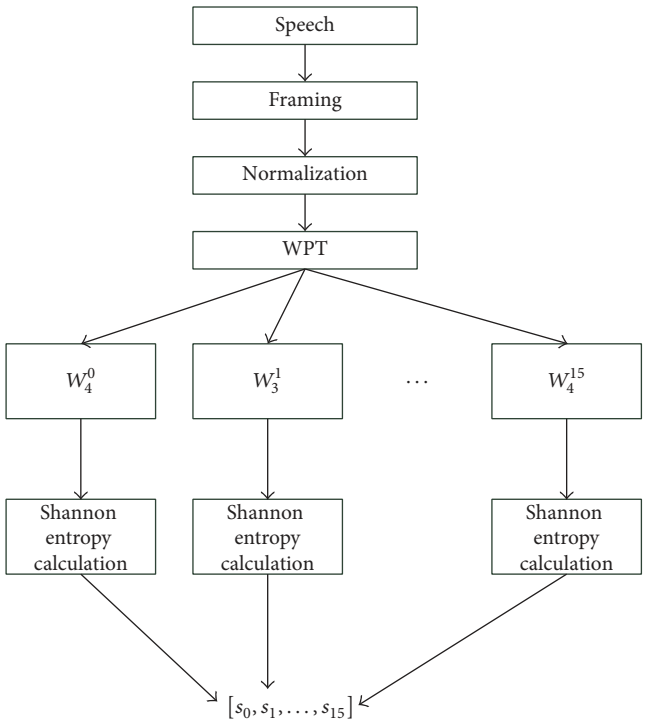

Figure 3: The flow chart of wavelet packet entropy.

processes followed by 2 : 1 downsampling, respectively. The filtering processes are defined as

$$\mathbf{W}_{j+1}^{2k}(n) = \mathbf{W}_j^k * \mathbf{h}(2n),$$
$$\mathbf{W}_{j+1}^{2k+1}(n) = \mathbf{W}_j^k * \mathbf{g}(2n); \quad (4)$$
$$0 \le k \le 2^j,\ 0 \le j \le J,\ 0 \le n \le N_j,$$

where $\mathbf{h}$ and $\mathbf{g}$ are the low-pass and high-pass filter, respectively. N_j is the length of the frequency component at level j. $*$ is the convolution operation. J is the total number of the decomposition levels. Because the WPT satisfies the conservation of energy, each leaf node denotes the spectrum of the frequency bands obtained by WPT. Based on the WPT, the wavelet packet entropy (WPE) method is proposed to extract the short vector and we add a normalization step into the method to reduce the ill effect of the volume in this paper. The flow chart of WPE used in this paper is shown in Figure 3.

Assume that there is a digital speech signal that has finite energy and length. It is firstly decomposed into 20 ms frames, and then each frame is normalized. The normalization process is defined as

$$\overline{f}[m] = \frac{f[m] - \mu}{\sigma}; \quad i = 1, 2, 3, \ldots, M, \quad (5)$$

where f is a signal frame and M is its length. μ is the mean value of the frame and σ is its standard variance. $\overline{f}$ is the normalized frame. After the normalization process, the WPT decomposes the frame at 4 levels using (4). Therefore, we finally obtain 16 frequency bands, and the frequency spectrums in those bands are denoted as $W_4^0, W_4^1, \ldots, W_4^{15}$, respectively. For each spectrum, the Shannon entropy is calculated. The Shannon entropy is denoted as

$$s_k = -\sum_{n=1}^{N} p_{k,n} \log(p_{k,n}); \quad k = 0, 2, 3, \ldots, 7 \quad (6)$$

with

$$p_{k,n} = \frac{\left|W_4^k(n)\right|^2}{e_k},$$
$$e_k = \sum_{n=1}^{N} \left|W_4^k(n)\right|^2, \quad (7)$$

where e_k is the energy of the kth spectrum. $p_{k,n}$ is the energy distribution of the kth spectrum. N is the length of each frequency spectrum. Finally, all of Shannon entropies of all spectrums are calculated and are collected to form a feature vector that is denoted as $[s_0, s_1, \ldots, s_7]^T$.

4.2. I-Vector Extraction. I-vector is a robust feature vector that represents a speech using a single high-dimension vector. Because it considers the background information, i-vector usually improves the accuracy of recognition [22]. Assume that there is a set of speeches. Those speeches are supplied by different speakers and the all speeches are transformed into the short vectors. In the i-vector theory, the speaker- and channel-dependent feature vector is assumed as

$$\overline{\mathbf{m}} = \mathbf{m} + \mathbf{T}\mathbf{w}(\mathbf{U}), \quad (8)$$

where $\overline{\mathbf{m}}$ is the speaker- and channel-dependent feature vector. $\mathbf{m}$ is the background factor. Usually, it is generated by stacking the mean vectors of a background model. Assume that the mean vectors of the background model are denoted by $\boldsymbol{\mu}_1^T, \boldsymbol{\mu}_2^T, \ldots, \boldsymbol{\mu}_N^T$, where each mean vector is a row vector. $\mathbf{m}$ is denoted by $[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_N]^T$. $\mathbf{T}$ is named the total variability matrix and represents a space that contains the speaker- and channel-dependent information. $\mathbf{w}(\mathbf{U})$ is a random vector having standard normal distribution $N(0, 1)$. The i-vector is the expectation of the $\mathbf{w}(\mathbf{U})$. $\mathbf{U}$ is a set of speeches and all of speeches are transformed into the short vectors. Assume that a background model is given, and $\boldsymbol{\Sigma}$ is initialized by covariance matrix of the background model. $\mathbf{T}$ and $\mathbf{w}(\mathbf{U})$ are initialized randomly. $\mathbf{T}$ and $\mathbf{w}(\mathbf{U})$ are estimated by an iteratively process described as follows:

(1) E-step: for each speech in the set $\mathbf{U}$, calculate the parameters of the posterior distribution of $\mathbf{w}(\mathbf{U})$ using the current estimates of $\mathbf{T}$, $\mathbf{\Sigma}$, and $\mathbf{m}$.

(2) M-step: update $\mathbf{T}$ and $\mathbf{\Sigma}$ by a linear regression in which $\mathbf{w}(\mathbf{U})$s play the role of explanatory variables.

(3) Iterate until the expectation of the $\mathbf{w}(\mathbf{U})$ is stable.

The details of the estimation processes of $\mathbf{T}$ and $\mathbf{w}(\mathbf{U})$ are described in [23].

4.3. Speaker Classification. Cosine distance scoring (CDS) is used as the classifier in our proposed model. It uses a kernel function to deal with the curse of dimension, so CDS is very suitable for the i-vector. To describe this classifier easily, we take a two-classification task, for example. Assume that there are two speakers denoted as s_1 and s_2. The two speakers, respectively, speak N_1 and N_2 speeches. All speeches are represented by i-vectors and are denoted by $X^1 = \{\mathbf{x}_1^1, \mathbf{x}_2^1, \ldots, \mathbf{x}_{N_1}^1\}$ and $X^2 = \{\mathbf{x}_1^2, \mathbf{x}_2^2, \ldots, \mathbf{x}_{N_2}^2\}$, where $\mathbf{x}_j^i$ is i-vector representing the jth speech sample of the speaker $\mathrm{s_i}$. We also assume there is an unknown speech represented by i-vector $\mathbf{y}$. The purpose of the classifier is to match the unknown i-vector with the known i-vectors and determine which one speaks the unknown speech. the result of the recognition is defined as

$$D_i(\mathbf{y}) = \frac{1}{N_i}\sum_{n=1}^{N_i} K\left(\mathbf{x}_n^i, \mathbf{y}\right) + \theta; \quad i = 1, 2, \tag{9}$$

where N_i is the total number of speeches supported by the speaker $\mathbf{s_i}$. $\boldsymbol{\theta}$ is the decision threshold. If $D_i(\mathbf{y}) \leq \mathbf{0}$, the unknown speeches are not spoken by the known speaker $\mathbf{s_i}$; if $D_i(\mathbf{y}) > \mathbf{0}$, then author of the unknown speeches is the speaker $\mathbf{s_i}$. $K(\cdot, \cdot)$ is the cosine kernel and is defined as

$$K(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}\mathbf{y}^T}{\sqrt{\mathbf{x}\mathbf{x}^T}\sqrt{\mathbf{y}\mathbf{y}^T}}, \tag{10}$$

where $\mathbf{x}$ is the known i-vector and $\mathbf{y}$ is the unknown i-vector. Usually, the linear discriminant analysis (LDA) and within class covariance normalization (WCCN) are used to implement the discrimination of the i-vector. Therefore, the kernel function is rewritten as

$$K(\mathbf{x}, \mathbf{y}) = \frac{\left(\mathbf{A}^T\mathbf{x}\right)\mathbf{W}^{-1}\left(\mathbf{A}^T\mathbf{y}\right)}{\sqrt{\left(\mathbf{A}^T\mathbf{x}\right)\mathbf{W}^{-1}\left(\mathbf{A}^T\mathbf{x}\right)}\sqrt{\left(\mathbf{A}^T\mathbf{y}\right)\mathbf{W}^{-1}\left(\mathbf{A}^T\mathbf{y}\right)}}, \tag{11}$$

where $\mathbf{A}$ is the LDA projection matrix and $\mathbf{W}$ is WCCN matrix. $\mathbf{A}$ and $\mathbf{W}$ are estimated by using all of the i-vectors and the details of LAD and WCCN are described in [24].

5. Experiment and Results

In this section, we report the outcome of our experiments. In Section 5.1, we describe the experimental dataset. In Section 5.2, we carry on an experiment to select the optimal mother wavelet for the WPE algorithm. In Section 5.3, we evaluate the recognition accuracy of our model. In Section 5.4, we evaluate the performance of the proposed model. Finally, the time cost of the model is count in Section 5.5.

5.1. Experimental Dataset. The results of our experiments are performed on the TIMIT speech database [25]. This database contains 630 speakers (192 females and 438 males) who come from 8 different English dialect regions. Each speaker supplies ten speech samples that are sampled at 16 KHz and last 5 seconds. All female speeches are used to obtain background models that represent the common characteristic of the female voice. Also, all male speeches are used to generate another background model characterizing the male voice. 384 speakers (192 females and 192 males) are randomly selected and their speeches are used as the known and unknown speeches. The test results presented in our experiments are collected on a computer with 2.5 GHz Intel Core i5 CPU and 8 GM of memory and the experimental platform is MATLAB R2012b.

5.2. Optimal Mother Wavelet. A good mother wavelet can improve the performance of the WPE algorithm. The performance of a mother wavelet is based on two important elements such as the support size and the number of vanishing moments. If a mother wavelet has large number of vanish moments, the WPE would ignore much of unimportant information; if the mother wavelet has small support size, the WPE would accurately locate important information [26]. Therefore, an optimal mother wavelet should have a large number of vanishing moments and a small support size. In this view, the Daubechies and Symlet wavelets are good wavelets, because they have the largest number of vanishing moments for a given support size. Moreover, those wavelets are orthogonal and are suitable for the Mallat fast algorithm.

In is paper, we use the Energy-to-Shannon Entropy Ratio (ESER) to evaluate those Daubechies and Symlet wavelets to find out the best one. ESER is a way to analyze the performance of mother wavelet and has been employed to select the best mother wavelet in [27]. The ESER is defined as

$$r = \frac{E}{S}, \tag{12}$$

where S is the Shannon entropy of the spectrum obtained by WPT and E was the energy of the spectrum. The high energy means the spectrum obtained by WPT contained much enough information of the speech. The low entropy means that the information in the spectrum is stable. Therefore, the optimal mother wavelet should maximize the energy and meanwhile minimize the entropy.

In this experiment, 8 Daubechies and 8 Symlet wavelets, which are, respectively, denoted as db1–8 and sym1–8, are employed to decompose speeches that are randomly selected from the TIMIT database. We run the experiment 100 times and record the average WSER of those mother wavelets in Table 1.

In Table 1, We find that db4 and sym6 obtain the highest ESER. In other words, the db4 and sysm6 are the best mother wavelets for the speech data. Reference [28] suggests that the sym6 can improve the performance of the speaker recognition model. However, the Symlet wavelets produce the complex coefficients whose imaginary parts are redundant for the real signal such as digital speech, so we abandon the sym6 and choose the db4.

Table 1: The average WSER of the mother wavelets.

Mother wavelet	ESER
db1	888.37
db2	890.32
db3	897.44
db4	**908.49**
db5	901.41
db6	896.53
db7	891.69
db8	890.84
sym1	888.35
sym2	890.36
sym3	894.93
sym4	899.75
sym5	903.82
sym6	**908.59**
sym7	902.44
sym8	898.37

5.3. The Accuracy of Speaker Recognition Model in Clear Environment. This experiment evaluates the accuracy of the speaker recognition model. We randomly select 384 speakers (192 females and 192 males). For each speaker, half of speeches are used as the unknown speech and the other half of speeches are used as the known speeches. For each speaker, the speaker recognition model matches the his/her unknown speeches with all of the known speeches of the 384 speakers and determines who speaks the unknown speeches. If the result is right, the model obtains one score; if the result is wrong, the model gets zero score. Finally, we count the score and calculate the mean accuracy that is defined as

$$\text{accuray} = \frac{\text{score}}{384} \times 100\%. \tag{13}$$

In this experiment, we use four types of speaker recognition models for comparison. The first one is the MFCC-GMM model [4]. This model uses MFCC method to extract 14D short vectors and uses the GMM with 8 mixtures to recognize speaker based on those short vectors. The second one is FMFCC-GMM model [16]. This model is very similar to the MFCC-GMM model, but it uses the FMFCC method to extract the 52D short vectors. The third one is the WPE-GMM model [10]. This model firstly uses WPE to transform the speeches into 16D short vectors and then uses GMM for speaker classification. The last one was the WPE-I-CDS model proposed in this paper. Compared with WPE-GMM model, our model uses the 16D short vectors to generate 400D i-vector and uses CDS to recognize speaker based on the i-vector. We carry on each experiment in this section 25 times to obtain the mean accuracy. The mean accuracy of the above 4 models is shown in Figure 4.

In Figure 4, we find that MFCC-GMM obtains the lowest accuracy of 88.46%. The result of [4] shows the MFCC-GMM model can obtain accuracy of higher than 90%. This is because we use the GMM with 8 mixtures as the classifier, but [4] uses the GMM with 32 mixtures as the classifier. Large

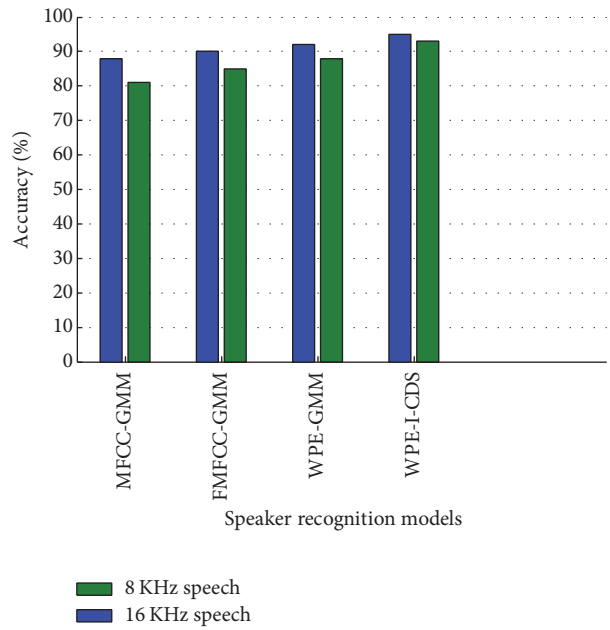

Figure 4: The mean accuracy of 4 models in clean environment.

mixture number can improve the performance of the GMM, but it also causes the very high computational expense. WPE-I-CDS obtain the highest accuracy of 94.36%. This interprets the achievements of i-vector theory. On the other hand, when the 8 KHz speeches (low-quality speeches) are used, all accuracy of speaker recognition models is decreased. The accuracy of MFCC-GMM, FMFCC-GMM, and WPE-GMM decrease by about 6%. Comparatively, the accuracy of WPE-I-CDS decreases by about 1%. This is because the i-vector considers the i-vector to improve the accuracy of the speaker recognition model, and the CDS used the LDA and WCCN to improve the discrimination of the i-vector. Reference [29] also reports that the combination of the i-vector and the CDS can enhance the performance of speaker recognition model used for low-quality speeches such as phone speeches.

5.4. The Accuracy of Speaker Recognition Model in Noisy Environment. It is hard to find a clean speech in the real applications, because the noise in the transmission channel and environment cannot be controlled. In this experiment, we add 30 dB, 20 dB, and 10 dB Gaussian white noise into the speeches to simulate the noisy speeches. All noises are generated by the MATLAB's Gaussian white noise function.

For comparison, this experiment employed three i-vector based models such as MFCC-I-CDS [30], FMFCC-I-CD [31], and our WPE-I-CDS. The two models are very similar to our proposed model, but they use the MFCC and FMFCC to extract the short vectors, respectively. The accuracy of the 3 models in noisy environment is shown in Figure 5.

In Figure 5, the three models obtained high accuracy in clean environment. This also shows that the i-vector can improve the recognition accuracy effectively. However, when we use the noisy speeches to test the 3 models, their accuracies decrease. When 30 dB noise is added to the speeches, the

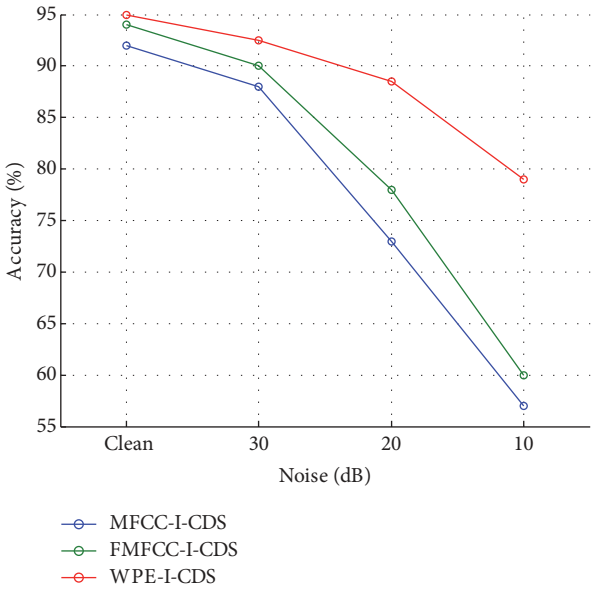

Figure 5: The accuracy of the 3 models in noisy environment.

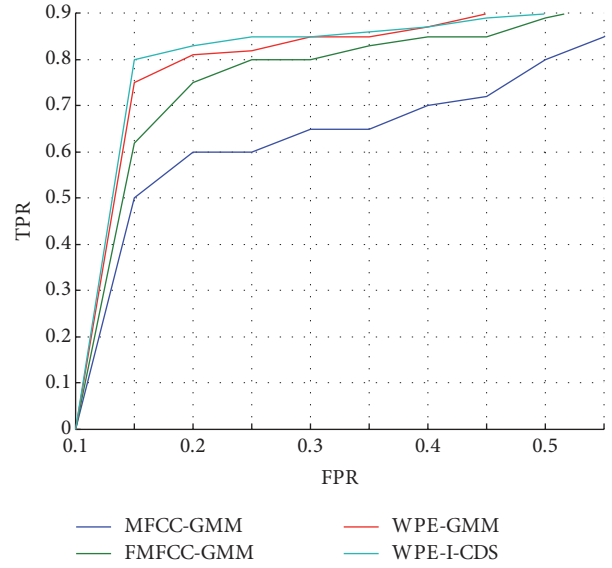

Figure 6: The ROC curves of TPR versus FPR.

accuracy of the three models decreases by about 4%. This shows that all of the models can resist weak noise. However, when we enhance the power of noise, the accuracy of MFCC-I-CDS and FMFCC-I-CDS drops off rapidly. In particular, when the noise increases into 10 dB, the accuracy of the above two models decreases by more than 30%. Comparatively, the WPE-I-CDS's accuracy decreases by less than 12%. Those show that the WPE-I-CDS is robust in noisy environment compared with MFCC-I-CDS and FMFCC-I-CDS. This is because the WPE uses the WPT to obtain the frequency spectrum but MFCC and FMFCC use the DFT to do that. The WPT decomposes the speech into many local frequency bands that can limit the ill effect of noise, but the DFT decomposes the speech into a global frequency domain that is sensitive to the noise.

5.5. The Performance of the Speaker Recognition Model. Usually, the speaker recognition model is used in the access control system. Therefore, a good speaker recognition model should have ability to accept the login of the correct people and meanwhile to reject the access of the imposter, as a gatekeeper does. In this experiment, we use the receiver operating characteristic (ROC) curve to evaluate the ability of our model. The ROC curve shows the true positive rate (TPR) as a function of the false positive rate (FPR) for different values of the decision threshold and has been employed in [2].

In this experiment, we randomly select 384 speakers (192 males and 192 females) to calculate the ROC curve. Half of those speakers are used as the correct people and another half of the speakers are used as the imposters. We firstly use the speeches of the correct people to test the speaker recognition model to calculate the TPR, and then we use the speeches of the imposters to attack the speaker recognition model to calculate the FPR. The 4 models, such as MFCC-GMM, FMFCC-GMM, WPE-GMM, and our WPE-I-CDS, are used for comparison. To plot the ROC curve, we adjusted the decision thresholds to obtain different ROC points. The ROC curves of those 4 models were shown in Figure 6.

Low FPR shows that the speaker recognition model can effectively resist the attack coming from the imposters, and high TPR shows that the speaker recognition model can accurately accept the correct speakers' login. In other words, a speaker recognition model can be useful if its TPR is high for a low FPR. In Figure 6, when FPR is higher than 0.45, all models obtain the high TPR, but WPE-I-CDS obtain higher TPR than other 3 models for a given FPR that is less than 4.5. This shows that the WPE-I-CDS can more effectively achieve the access control task than other models.

5.6. Time Cost. This section tests the time cost of the fast MFCC-GMM, the conventional MFCC-I-CDS, and our WPE-I-CDS. We used 200 5-second-long speeches to test each model and calculated the average time cost. The result of this experiment was shown in Table 2.

In Table 2, MFCC-GMM does not employ the i-vector for speech representation, so it does not cost time to extract the i-vector. Comparatively, the WPE-I-CDS should cost time to extract the i-vector. The WPE-I-CDS cost the most time to extract the short vector compared with the MFCC-GMM. This is because the WPT used by WPE is more complex than the DFT used by the MFCC. On the other hand, the parameters of GMM should be estimated beforehand, as MFCC-GMM cost time to train the classifier. CDS needs not cost time to estimate the parameters, but it should cost time to estimate the matrices of the LDA and WCNN in the training classifier step. In all, the i-vector can improve the recognition accuracy at cost of increasing the time consumption and calculating the WPE costs too much time compared with

Table 2: The time cost of the different speaker recognition models.

Speaker recognition model	Feature extraction (s/speech)		Speaker classification (s/speech)	
	Short vector extraction	I-vector extraction	Training classifier	Recognition
MFCC-GMM	0.46	—	1.92	0.64
MFCC-I-CDS	0.45	2.37	1.81	0.73
WPE-I-CDS	2.81	2.24	1.82	0.75
WPE-I-CDS (parallel computation)	0.78	2.23	1.82	0.74

calculating the MFCC. Therefore, it is very important to find way to reduce the time cost of the WPE.

Parallel computation is an effective way to reduce the time cost, because the loops in the linear computation can be finished at once using a parallel algorithm. For example, a signal, whose length is N, is decomposed by WPT at M levels. In the conventional linear algorithm of WPT, we have to run a filtering process whose time complexity was $O(\log N)$ $M \times N$ times for each decomposition level, so the total time cost of WPT is $O(MN \log N)$. If we used N independent computational cores to implant the WPT using a parallel algorithm, the time complexity of WPT can reduce to $O(M \log N)$. This paper uses 16 independent computational cores to implement the WPE parallel algorithm, and the last line of Table 2 shows that the time cost of WPE is reduced very much.

6. Conclusions

With the development of the computer technique, the speaker recognition has been widely used for speech-based access system. In the real environment, the quality of the speech may be low and noise in the transformation channel cannot be controlled. Therefore, it is necessary to find a speaker recognition model that is not sensitive to those factors such as noise and low-quality speech.

This paper proposes a new speaker recognition model by employing wavelet packet entropy (WPE), i-vector, and CDS, and we name the model WPE-I-CDS. WPE used a local analysis tool named WPT rather than the DFT to decompose the signal. Because WPT decomposes the signal into many independent frequency bands that limit the ill effect of noise, the WPE is robust in the noisy environment. I-vector is a type of robust feature vector. Because it considers the background information, i-vector can improve the accuracy of recognition. CDS uses a kernel function to deal with the curse of dimension, so it is suitable for the high-dimension feature vector such as i-vector. The result of the experiments in this paper shows that the proposed speaker recognition models can improve the performance of recognition compared with the conventional models such as MFCC-GMM, FMFCC-GMM, and WPE-GMM in clean environment. Moreover, the WPE-I-CDS obtains higher accuracy than other i-vector-based models such as MFCC-I-CDS and FMFCC-I-CDS in noisy environment. However, the time cost of the proposed model is very higher. To reduce the time cost, we employ the parallel algorithm to implement the WPE and i-vector extraction methods.

In the future, we will combine audio and visual feature to improve the performance of the speaker recognition system.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors also thank Professor Kun She for their assistance in preparation of the manuscript. This paper is supported by the Project of Sichuan Provincial Science Plan (M112016GZ0073) and National Nature Foundation (Grant no. 61672136).

References

[1] H. Perez, J. Martinez, and I. Espinosa, "Using acoustic paraliguistic information to assess the interaction quality in speech-based system for elder users," *International Jounrnal of Human-Computer Studies*, vol. 98, pp. 1–13, 2017.

[2] N. Almaadeed, A. Aggoun, and A. Amira, "Speaker identification using multimodal neural networks and wavelet analysis," *IET Biometrics*, vol. 4, no. 1, pp. 18–28, 2015.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[5] K. S. Ahmad, A. S. Thosar, J. H. Nirmal, and V. S. Pande, "A unique approach in text-independent speaker recognition using mfcc feature sets and probabilistic neural network," in *Proceedings of the 8th International Conference on Advances in Pattern Recognition (ICAPR '15)*, Kolkata, India, 2015.

[6] X.-Y. Zhang, J. Bai, and W.-Z. Liang, "The speech recognition system based on bark wavelet MFCC," in *Proceedings of the 8th International Conference on Signal Processing (ICSP '06)*, pp. 16–20, Beijing, China, 2006.

[7] A. Biswas, P. K. Sahu, and M. Chandra, "Admissible wavelet packet features based on human inner ear frequency response for Hindi consonant recongnition," *Computer & Communication Technology*, vol. 40, pp. 1111–1122, 2014.

[8] K. Daqrouq and T. A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers," *Applied Soft Computing*, vol. 27, pp. 231–239, 2015.

[9] D. Avci, "An expert system for speaker identification using adaptive wavelet sure entropy," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6295–6300, 2009.

[10] K. Daqrouq, "Wavelet entropy and neural network for text-independent speaker identification," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 5, pp. 796–802, 2011.

[11] M. K. Elyaberani, S. H. Mahmoodian, and G. Sheikhi, "Wavelet packet entropy in speaker-identification emotional state detection from speech signal," *Journal of Intelligent Procedures in Electrical Technology*, vol. 20, pp. 67–74, 2015.

[12] M. Li, A. Tsiartas, M. Van Segbroeck, and S. S. Narayanan, "Speaker verification using simplified and supervised i-vector modeling," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 7199–7203, Vancouver, Canada, 2013.

[13] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 4047–4051, Florence, Italy, 2014.

[14] A. Kanagasundaram, D. Dean, and S. Sridharan, "I-vector based speaker recognition using advanced channel compensation techniques," *Computer Speech & Labguage*, vol. 28, pp. 121–140, 2014.

[15] M. H. Bahari, R. Saeidi, H. Van Hamme, and D. Van Leeuwen, "Accent recognition using i-vector, gaussian mean supervector and gaussian posterior probability supervector for spontaneous telephone speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7344–7348, IEEE, Vancouver, Canada, 2013.

[16] B. Saha and K. Kamarauslas, "Evaluation of effectiveness of different methods in speaker recognition," *Elektronika ir Elektrochnika*, vol. 98, pp. 67–70, 2015.

[17] K. K. George, C. S. Kumar, K. I. Ramachandran, and A. Panda, "Cosine distance features for robust speaker verification," in *Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH '15)*, pp. 234–238, Dresden, Germany, 2015.

[18] B. G. Nagaraja and H. S. Jayanna, "Efficient window for monolingual and crosslingual speaker identification using MFCC," in *Proceedings of the International Conference on Advanced Computing and Communication Systems (ICACCS '13)*, pp. 1–4, Coimbatore, India, 2013.

[19] M. Medeiros, G. Araújo, H. Macedo, M. Chella, and L. Matos, "Multi-kernel approach to parallelization of EM algorithm for GMM training," in *Proceedings of the 3rd Brazilian Conference on Intelligent Systems (BRACIS '14)*, pp. 158–165, Sao Paulo, Brazil, 2014.

[20] H. R. Tohidypour, S. A. Seyyedsalehi, and H. Behbood, "Comparison between wavelet packet transform, Bark Wavelet & MFCC for robust speech recognition tasks," in *Proceedings of the 2nd International Conference on Industrial Mechatronics and Automation (ICIMA '10)*, pp. 329–332, Wuhan, China, 2010.

[21] X.-F. Lv and P. Tao, "Mallat algorithm of wavelet for time-varying system parametric identification," in *Proceedings of the 25th Chinese Control and Decision Conference (CCDC '13)*, pp. 1554–1556, Guiyang, China, 2013.

[22] D. Marthnez, O. Pichot, and L. Burget, "Language recognition in I-vector space," in *Proceedongs of the Interspeech*, pp. 861–864, Florence, Italy, 2011.

[23] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[24] M. McLaren and D. Van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," in *Proceedings of the 36th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '11)*, pp. 5460–5463, Prague, Czech Republic, 2011.

[25] A. Biswas, P. K. Sahu, A. Bhowmick, and M. Chandra, "Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition," *International Journal of Speech Technology*, vol. 17, no. 4, pp. 389–399, 2014.

[26] S. G. Mallat, *A Wavelet Tour of Signal Processing*, ELsevier, Amsterdam, Netherlands, 2012.

[27] Q. Yang and J. Wang, "Multi-level wavelet shannon entropy-based method for single-sensor fault location," *Entropy*, vol. 17, no. 10, pp. 7101–7117, 2015.

[28] T. Ganchev, M. Siafarikas, I. Mporas, and T. Stoyanova, "Wavelet basis selection for enhanced speech parametrization in speaker verification," *International Journal of Speech Technology*, vol. 17, no. 1, pp. 27–36, 2014.

[29] M. Seboussaoui, P. Kenny, and N. Dehak, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," *Odyssey*, vol. 6, pp. 1–6, 2011.

[30] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH '09)*, pp. 1559–1562, Brighton, United Kingdom, 2009.

[31] M. I. Mandasari, M. McLaren, and D. Van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH '11)*, pp. 21–24, Florence, Italy, 2011.