Hindawi Publishing Corporation
Journal of Electrical and Computer Engineering
Volume 2016, Article ID 4908412, 11 pages
http://dx.doi.org/10.1155/2016/4908412

Research Article

Speaker Recognition Using Wavelet Cepstral Coefficient, I-Vector, and Cosine Distance Scoring and Its Application for Forensics

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 9 June 2016; Revised 24 August 2016; Accepted 18 September 2016

Academic Editor: Mariko Nakano-Miyatake

An important application of speaker recognition is forensics. However, the accuracy of speaker recognition in forensic cases often drops off rapidly because of the ill effect of ambient noise, variable channel, different duration of speech data, and so on. Therefore, finding a robust speaker recognition model is very important for forensics. This paper builds a new speaker recognition model based on wavelet cepstral coefficient (WCC), i-vector, and cosine distance scoring (CDS). This model firstly uses the WCC to transform the speech into spectral feature vecors and then uses those spectral feature vectors to train the i-vectors that represent the speeches having different durations. CDS is used to compare the i-vectors to give out the evidence. Moreover, linear discriminant analysis (LDA) and the within-class covariance normalization (WCNN) are added to the CDS algorithm to deal with the channel variability problem. Finally, the likelihood ratio estimates the strength of the evidence. We use the TIMIT database to evaluate the performance of the proposed model. The experimental results show that the proposed model can effectively solve the troubles of forensic scenario, but the time cost of the method is high.

1. Introduction

With the increasing use of computer technology, more and more complex and tedious works can be finished by computer. Automatic speaker recognition (or speaker recognition for short) technique refers to recognizing persons from their voice using the computer software. An important application of speaker recognition is forensics. This technique is usually used for investigation and evidence reporting [1]. In the investigation task, the speaker recognition is used to compare the questioned speech with the speeches of known criminals in police's database to produce a small list of potential suspects. In the evidence reporting task, police has found the suspect and the speaker recognition is used to give out the evidence (is defined as the similarity between the questioned speech and the suspect's speech samples [2]) supporting the fact that the suspect is the author of the criminal speech.

Technologically, the forensic speaker recognition model is very similar to the common speaker recognition model [3]. It firstly uses a feature extractor to transform the digital speech signal into the feature vectors that represent unique information for a particular speaker irrespective of the speech content and then uses a learning algorithm to give out the evidence based on those feature vectors. Finally, it is required to report the likelihood ratio that estimates the strength of the evidence after it gives out the evidence [4]. The forensic context is very challenging for speaker recognition. The ambient noise cannot be controlled, the duration of speech data can vary from a few seconds to several hours, and the speeches are often obtained from different channels such as phone channel and microphone channel. It is important for forensics to find a robust speaker recognition model that is not sensitive to those factors such as noise, duration, and channel.

The speech is usually transformed into the short-term feature vector because this feature vector is simple [5]. In the speaker recognition model used for forensics, Mel-frequency cepstral coefficient (MFCC) method has been widely used to

extract the short-term feature vector [6]. This method calculates only the cepstral coefficients, so the extracted feature vector just represents the static information. Fused MFCC (FMCC) [7] method is an extension of MFCC. This method calculates not only the cepstral coefficients but also the delta derivatives, so the extracted feature vector can represent both the static information and dynamic information. Both of the two methods use discrete Fourier transform (DFT) to obtain the frequency spectrum. Because the DFT is a global signal analysis approach, if a local frequency part of the speech is destroyed by noise, the ill effect of noise will be transmitted to whole feature vector. Wavelet cepstral coefficient (WCC) [8] is another method used to extract the short-term feature vector. This method uses discrete wavelet transform (DWT) to obtain the frequency spectrum. DWT is a local signal analysis approach, so the whole feature vector cannot be interfered strongly by the noise that just destroys the local frequency part. Unfortunately, this method is not used for the forensic speaker recognition. This paper tries to employ this method to extract the short-term feature vector, because it can be robust against the noise.

In the forensic context, the duration of speech can vary from a few seconds to several hours. To compare those speeches having different duration, it is necessary to represent them in a uniform way. I-vector [9] is a newly proposed feature vector and has been employed by the speaker recognition model used for forensics [10]. The main advantage of i-vector is that it can use a signal vector to represent a speech, so it can uniformly represent those speeches having different duration. In its extraction method [11], the speech signal is firstly transformed into short-term feature vectors and then those short-term feature vectors are used to train i-vector.

Based on the feature vector extracted by MFCC or Fused MFCC, Gaussian mixture model (GMM) is the conventional learning algorithm [12]. In [13], it uses the probability distribution estimated during the training phrase to determine whether the suspect is the author of the questioned speech. However, if the dimension of the input vector is very high, the curse of dimension will destroy the algorithm [14]. Cosine distance scoring (CDS) is another type of learning algorithm used for speaker recognition and [1, 15] use it to give out the evidence in the forensic speaker recognition. Because it can deal with the curse of dimension using a cosine kernel, CDS is suitable for i-vector which is high-dimensional vector compared with those short-term feature vectors. Moreover, it does not cost time to estimate the separating hyperplane, so its time cost is low. Two same speeches will become very different, if they are obtained from different channels. This is called channel variability problem. Usually, linear discriminant analysis and the within-class covariance normalization are added to the CDS to deal with this problem [16, 17].

In forensic context, the quality of speech is strongly interfered by noise, variable channel, and different duration. This is very important for forensics to find a speaker recognition model that is unsusceptible to those factors. Based on those above works of the speaker recognition used for forensics, this paper combines the WCC, i-vector approach, and the CDS learning algorithm to build a new speaker recognition model that can be robust against the noise, uniformly represent the speech having different duration, and deal with the channel variability problem. We use TIMIT to evaluate the performance of our system. The experimental results show that the proposed model can solve the trouble of forensic scenario and improve the accuracy of recognition. However, the time cost of the model is high compared with other conventional speaker recognition models.

The rest of the paper is organized as follows. Section 2 briefly describes the conventional speaker recognition model. In Section 3, we describe the i-vector-based forensic speaker recognition model. The proposed model is described in Section 4. In Section 5, we report the result of our experiment. Finally, we give out a conclusion in the last section.

2. Conventional Speaker Recognition Model

2.1. Short-Term Feature Extraction Method. In conventional speaker recognition model, the short-term feature vectors are usually extracted by MFCC and Fused MFCC methods [18]. In MFCC method, a speech signal is firstly divided into 20 ms-long frames with a 10 ms overlap for smoothing the frequency changes over those frames. After the segmentation, the frames whose energy is less than a silence threshold (= 0.0001 in this paper) are discarded as well. For each frame, the cepstral coefficient can be calculated as follows:

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

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

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

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

Usually, only the lower 12–14 cepstral coefficients are used to form the short-term feature vector [19]. This feature vector extracted by MFCC contains only the cepstral coefficients, so it just represents the static information of the frame. Fused MFCC (FMCC) is an extension of MFCC. It is very similar to the MFCC, except that this method calculates not only the cepstral coefficients but also the delta derivatives [20]. Assume that the feature vector extracted by MFCC is denoted as $[cc_1, \ldots, cc_{13}, cc_{14}]$. The delta derivatives are calculated by

$$d_i = \frac{\sum_{p=1}^{2} p\left(cc_{i-p} + cc_{i+p}\right)}{2\sum_{p=1}^{2} p^2},$$
$$dd_i = \frac{\sum_{p=1}^{2} p\left(d_{i-p} + d_{i+p}\right)}{2\sum_{p=1}^{2} p^2}. \tag{1}$$

The feature vector extracted by FMFCC is denoted as $cc_1, \ldots, cc_{13}, cc_{14}, d_1, \ldots, d_{13}, d_{14}, dd_1, \ldots, dd_{13}, dd_{14}$. Compared with the cepstral coefficients vector extracted by MFCC, the feature vector extracted by Fused MFCC can represent the changes over the multiple frames. The detailed algorithm of MFCC and Fused MFCC can be found in [20–22].

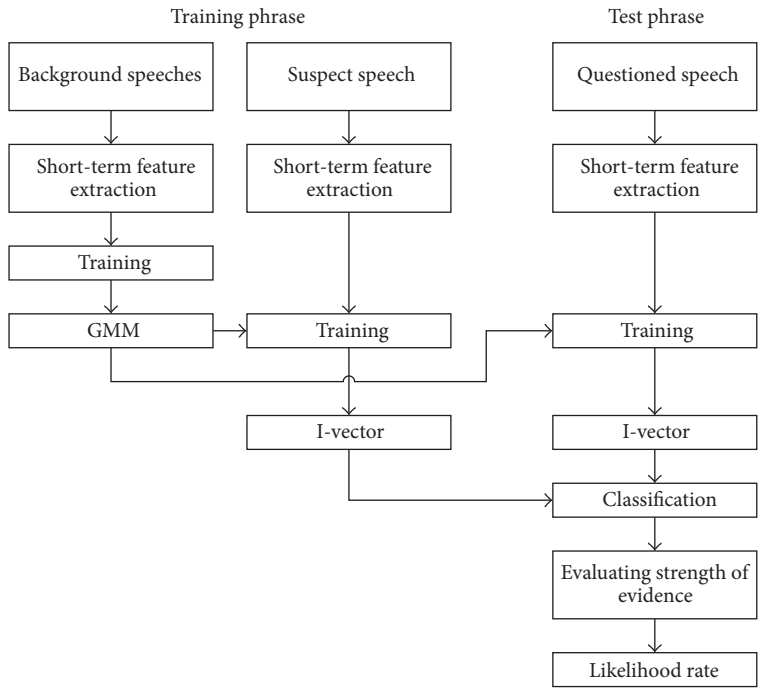

Figure 1: Forensic speaker recognition system.

2.2. Speaker Classification. Gaussian mixture model (GMM) is the conventional learning algorithm used for speaker classification in the conventional speaker model. This model tries to estimate the probability distribution governing the known speech data of a particular speaker and then uses the probability distribution to test an unknown speech and determine whether the unknown speech is spoken by the speaker. It is linear combination of Gaussian functions. The experiment in [23] shows that, with increasing of the number of Gaussian functions which GMM contains, the recognition accuracy of GMM will become very high. However, the large number of Gaussian functions also reduces excessive computational complexity.

Neural network (NN) is another type of learning algorithm used for speaker classification. This algorithm simulates the human brain to learn the knowledge of the known speech data of a particular speaker by iteratively adjusting the weights that connect between two neurons in the adjacent two layers and uses the knowledge to determine whether the speaker was the author of an unknown speech. Probabilistic neural network (PNN) is a special case of NN, where the sigmoid activation function is replaced by an exponential function. It uses the NN structure to directly implement the Bayesian decision and does not cost time to estimate the probability distribution compared with GMM, so the time cost of this algorithm is very low.

3. I-Vector-Based Speaker Recognition Model Used for Forensics

The i-vector-based speaker model used for forensics is shown in Figure 1.

A forensic speaker recognition model can be decomposed into training phrase and test phrase. The first two columns denote the training phrase and the last column denotes the test phrase. In Figure 1, background speeches usually contain thousands of speeches spoken by a huge number of people [5]. Suspect speech and questioned speech are collected from the suspect and the criminal scene. All the tree types of speeches have different length and are full of noise. Furthermore, they are usually obtained from different channels.

Firstly, the model transforms the three types of speeches into short-term feature vectors. The short-term feature vectors extracted from the background speeches are used to train the background model. This model represents the speaker- and channel-independent information and is implemented by a GMM. Once the GMM model is created, the short-term feature vectors extracted from the suspect speech and the questioned speech are used to train i-vector. One i-vector is trained using only the short-term feature vectors extracted from a speech. After the suspect speech and questioned speech are transformed into i-vector, a learning algorithm is used to compare those i-vectors to give out the evidence in forensics and then report the strength of the evidence as a likelihood ratio.

4. The Proposed Model

In the forensic context, the duration of speech can vary from few seconds to several hours and the recording condition is full of noise. Moreover, the speeches are usually obtained from different channels. To deal with those problems, this paper proposed a new speaker recognition model by employing WCC, i-vector, and CDS. The WCC is used to extract

the short-term feature vectors that are used to train i-vector. I-vector is used to represent the speeches whose durations are very different in a uniformly way for comparing those speeches easily. CDS is used to give out the evidence. Section 4.1 describes the WCC, Section 4.2 describes i-vector, and Section 4.3 describes the CDS. Finally, the likelihood-ratio algorithm is described in Section 4.4.

4.1. Short-Term Feature Extraction. This paper uses the wavelet cepstral coefficient (WCC) to extract the short-term feature vectors, because it is able to effectively limit the ill effect of noise using discrete wavelet transform (DWT).

Wavelet transform is a type of signal processing tool that is used to obtain the frequency spectrum. A standard wavelet transform is defined by

$$Wf(n,m) = \frac{1}{\sqrt{m}} \int_{-\infty}^{+\infty} S(t) \cdot \psi\left(\frac{t-n}{m}\right) dt, \quad (2)$$

where $Wf(n,m)$ is a continue signal frame which has finite energy. $\psi(\cdot)$ is the mother wavelet and $Wf(n,m)$ represents the nth wavelet coefficient at level m. For analyzing the discrete digital signal, the discrete wavelet transform (DWT) is proposed. The DWT is usually implemented by the famous Mallet algorithm [24]. In the algorithm, the DWT is realized through a pair of low-pass and a high-pass wavelet filters that are reconstructed from a selected mother wavelet and its corresponding scaling function. Through those filters, the signal is decomposed into a low-frequency part and a high-frequency part. The low-frequency part can be further decomposed at the next decomposition level to obtain higher low-frequency resolution. The low- and high-passed filtering processes are implemented by

$$\begin{aligned} A_{m+1} &= A_m * h[2n], \\ D_{m+1} &= A_m * g[2n], \\ A_0[n] &= S[n], \\ & n = 1, 2, 3, \ldots, N, \end{aligned} \quad (3)$$

where N is the length of the analyzed signal. g and h represent the low-pass and high-pass conjugate mirror filters, respectively. $*$ is the convolution operation. Compared with DFT used in MFCC or Fused MFCC, DWT can decompose the signal into many small local frequency domains and obtain the local frequency spectrum. In other words, if one of frequency parts of signal is destroyed by noise, whole frequency spectrums will not be interfered strongly. This means that the frequency spectrum obtained by wavelet is robust against noise.

Recently, researchers have widely used a new type of feature extractor named wavelet cepstral coefficient method for short-term feature extraction [8]. The flow chart of WCC extraction algorithm used in this paper is shown in Figure 2.

In Figure 2, the speech is firstly decomposed into 20 ms-long short-term frames with a 10 ms overlap. After the segmentation, the silence frames are discarded using an energy threshold (= 0.0001). After the silence frame removing, we add a normalization method to the WCC approach to remove the ill effect of sound volume. The normalization method [25] is given by

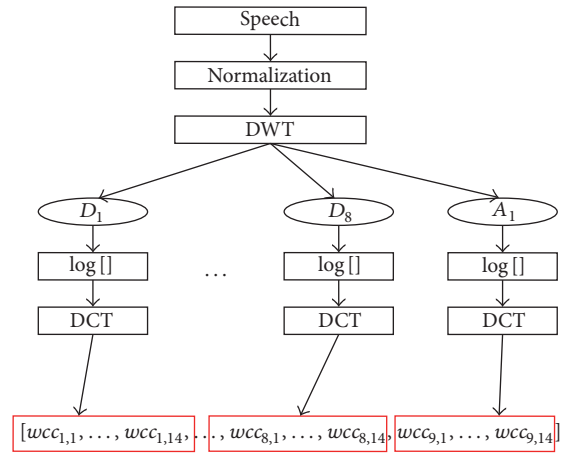

Figure 2: The flow chart of WCC extraction algorithm.

$$\overline{f}(n) = \frac{f(n) - \mu}{\sigma}, \quad n = 1, 2, 3, \ldots, N, \quad (4)$$

where $f(n)$ is a short-term frame that has finite energy and length. μ and σ are the mean and standard variance of the frame, respectively. N is the length of frame and $\overline{f}(n)$ is the normalized frame. The result of normalization is shown in Figure 3.

After the normalization, DWT is used to obtain the local frequency spectrum. This paper decomposes the speech into 8 levels by DWT, and therefore we obtain 9 local frequency parts such as 8 high-frequency parts denoted by $D_1, D_2, \ldots, D_8$ and one low-frequency part denoted by A_1. For each frequency part, log[] and DCT are also used to calculate the 14 cepstral coefficients. WCC is very similar to the MFCC, but the difference is that the cepstral coefficient in MFCC is calculated on a global frequency domain, but the cepstral coefficient in WCC is calculated on many local frequency domains obtained by DWT.

4.2. Training I-Vector. After the speech is transformed to short-term feature vectors, we can use those vectors to train i-vector. A background model should be trained at first. The background model represents the speaker- and channel-independent information and is implemented by a GMM. This GMM is trained by a huge set of short-term feature vectors extracted from the background speech set that contains thousands of speeches spoken by large number of speakers. For forensics, the background speech set may contain all speeches of all known criminals in police database, and two gender-dependent GMMs that generalize the characteristic of gender-dependent voice are trained using female's speeches and male's speeches, respectively [1]. Once the GMMs are trained, the suspect i-vector and questioned i-vector are trained using a particular suspect speech and questioned speech, respectively.

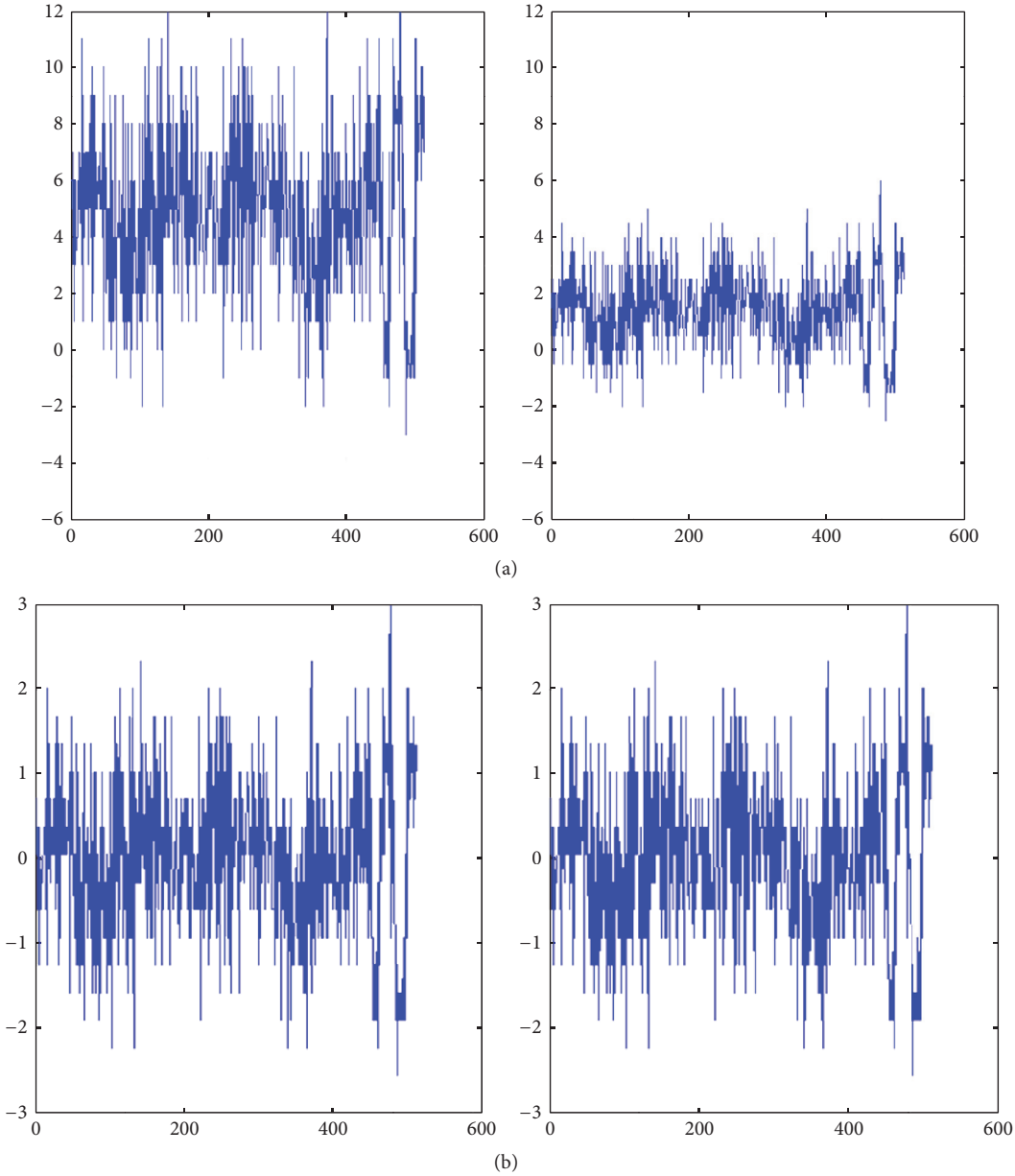

Figure 3: (a) Two speech signals that have different sound volume before normalization. (b) Two speech signals that have different sound volume after normalization.

Given a speech, i-vector approach assumes that the ideal speaker- and channel-dependent feature vector used to represent the speech can be modeled as

$$M = m + Zx, \tag{5}$$

where M is the ideal feature vector. m is a vector that consists of all the mean vectors of the GMM. If the mean vectors of GMM are denoted by $\mu_1^T, \mu_2^T, \ldots, \mu_I^T$, where I is the number of the mean vectors and each μ is a row vector, then m is denoted by $[\mu_1, \mu_2, \ldots, \mu_I]^T$. Z is a low rank matrix and is called the total variability matrix. x is i-vector and obeys standard normal distribution. For a given speaker and channel, m and Z are changeless, so the ideal feature vector is dependent on the value of x. In other words, x can represent the speech sample. Based on the assumption shown in (5), the i-vector training algorithm is used to iteratively estimate x and Z. x is estimated using a speech, and Z is estimated using all speeches generated from a speaker. The details of the training algorithm are described in [26].

4.3. Evidence Reporting. Cosine distance scoring (CDS) is a famous learning algorithm. It uses a cosine kernel to directly compare two input feature vectors and give out the degree of similarity between the two feature vectors. The cosine kernel is defined as

$$E = K(x_1, x_2) = \frac{x_1 x_2}{\sqrt{x_1 x_1^T}\sqrt{x_2 x_2^T}}, \tag{6}$$

where x_1, x_2 are the two feature vectors. In this paper, they are two i-vectors that represent a suspect's speech sample and a questioned sample, respectively. $K(x_1, x_2)$ is the degree of similarity between the suspect's speech and the questioned speech, so it is the evidence reported by the forensic speaker recognition model.

Two same speech samples will become very different, if they are obtained from different channels. This is called channel variability problem. In the forensic context, the speech data are usually obtained from different channels. For example, in the criminal case where a victim receives a threatening call, the questioned speech is a phone recording, but police usually record the speeches of suspect by microphone. In other words, the question speech is obtained from the phone channel, but the suspect's speeches are obtained from the microphone channel. To deal with the channel variability problem, LDA and WCCN are added to the CDS. The cosine kernel is denoted as

$$E = K(x_1, x_2) = \frac{(A^T x_1) W^{-1} (A^T x_2)}{\sqrt{(A^T x_1) W^{-1} (A^T x_1)} \sqrt{(A^T x_2) W^{-1} (A^T x_2)}}, \quad (7)$$

where A is the LDA projection matrix and W is WCCN matrix. The details of LDA and WCCN are described in [27].

4.4. Evaluating the Strength of Evidence. For evidence reporting, the speaker recognition model should report the strength of evidence as a likelihood ratio (LR). To calculate the LR, two competing hypotheses H_0 and H_1 are given. H_0 assumes that the suspected speaker is the author of the questioned speech and H_1 assumes that the suspected speaker is not the author of the questioned speech. Based on the two hypotheses, the LR [28] is defined as

$$\mathrm{LR} = \frac{P(E \mid H_0)}{P(E \mid H_1)}, \quad (8)$$

where E is the evidence calculated by (7). Firstly, we estimate the probability distribution $P(E \mid H_0)$ and $P(E \mid H_1)$. In conventional aural speaker recognition, the degree of similarity between the questioned sample and the suspect's sample is estimated by a seven-level verbal scale shown in Table 1 [28].

To simulate this, we transform E into 7-level scale. The transform function is defined as

$$T(E) = \begin{cases} 1 & -1 \le E < -0.7 \\ \vdots & \vdots \\ 4 & -0.1 \le E < 0.1 \\ \vdots & \vdots \\ 7 & 0.7 \le E \le 1. \end{cases} \quad (9)$$

Assume that there are s ($\ge$ 2) known criminals in the police database and each one speaks m ($\ge$ 2) speeches. We iteratively select two speeches spoken by the same speaker and calculate $T(E)$. If $T(E)$ happens n_1 times, then $P(T(E) = e \mid H_0)$ is calculated as

$$P(T(E) = e \mid H_0) = \frac{n_1}{C_s^1 C_m^2} = \frac{2n_1}{sm(m-1)}, \quad (10)$$

where C_m^n denotes the number of n-combinations in a set of m elements. On the other hand, we iteratively select two speech samples spoken by different speakers and calculate $T(E)$. If $T(E) = e$ happens n_2 times, then $P(T(E) = e \mid H_1)$ is calculated as

$$P(T(E) = e \mid H_1) = \frac{n_2}{C_s^2 C_m^1 C_m^1} = \frac{2n_2}{sm^2(m-1)}. \quad (11)$$

We calculate $P(T(E) = e \mid H_0)$ and $P(T(E) = e \mid H_1)$ for each $T(E)$ to obtain the distribution of E for H_0 and H_1. Assume that there is an evidence E and $T(E) = e_1$. If we want to report its strength, we firstly search the distributions and find the value of $P(e_1 \mid H_0)$ and $P(e_1 \mid H_1)$ and then calculate the LR.

Table 1: The verbal scales used in aural speaker recognition.

Level	Verbal equivalent
1	I am certain that the two speakers are not the same
2	I am almost certain that the two speaker are not the same
3	It is possible that the two speakers are not the same
4	I am unable to decide
5	It is possible that the two speakers are the same
6	I am almost certain that the two speakers are the same
7	I am certain that the two speakers are the same

5. Results

In this section we report the outcome of our experiments. In Section 5.1, we describe the experimental dataset and procedure. In Section 5.2, we carry out an experiment to select the optimal mother wavelet for WCC method. In Section 5.3, we evaluate the performance of the proposed speaker recognition model used for investigation. In Section 5.4, we evaluate the performance of the proposed model used for evidence reporting. In Section 5.5, the time cost of the proposed model is counted.

5.1. Experimental Dataset. The results of our experiments were performed on TIMIT speech database [29]. This database contained 630 speakers (192 females and 438 males) who came from 8 different dialect regions. Each speaker supplied ten 5-second-long speeches that were sampled at 16 KHz. In forensic context, the speeches had different length and were full of noise. Moreover, the questioned speech and suspect's speeches are usually obtained from different channels. For each speaker, 3 speeches were downsampled to 8 KHz and other 7 speeches were still sampled at 16 KHz. This simulated the speech data obtained from different channels. Moreover, the 3 speeches sampled at 8 KHz were combined

in a 15-second-long speech, and other 7 speeches still lasted 5 seconds. This simulates the speech data having different duration. Three types of noises such as 10 dB, 20 dB, and 30 dB were added to those speeches to simulate the speeches that were full of noise in Section 5.3. All of female speeches and all of male speeches were used to train two gender-dependent background models, respectively. The test results presented in our experiments were collected on a computer with a 2.5 GHz Intel Core i5 processor and 8 GM of memory. The experimental platform was MATLAB R2012b.

5.2. Mother Wavelet. The mother wavelet was a key issue for DWT, and good mother wavelet could improve the performance of the wavelet-based spectral speech feature such as the WCC. The goal of this experiment was to find the optimal mother wavelet for WCC. The number of the vanishing movements and the size of support were two important elements for a mother wavelet. In the theory of mother wavelet [30], if the mother wavelet had enough vanishing movements, the DWT would ignore much of unimportant information; if the mother wavelet had small enough support, the wavelet coefficients obtained by DWT would sparely and accurately represent the important information of a signal. Therefore, an optimal mother wavelet should have large number of vanishing movements and meanwhile have small support. However, we had to take tradeoff between the number of vanishing movements and the size of support, because they should satisfy the following equation:

$$L \geq 2p - 1, \tag{12}$$

where L is the size of support and p is the number of vanishing movements. In this view, Daubechies wavelets [31] were the optimal wavelets, because they had the smallest support for given vanishing movements. Moreover, these wavelets had orthogonal conjugate mirror filters which were suitable for the Mallat fast DWT algorithm.

In this experiment, we employ the normalized partial energy (NPE) [32] to evaluate the performance of Daubechies wavelets. NPE was used to quantify how well a particular transform, such as DWT, performed in capturing the important information of a signal. Assume that there was a wavelet coefficient series denoted by $\{c_1, c_2, \ldots, c_N\}$, where N is the total number of the wavelet coefficients. Form the squared magnitudes $|c_t|^2$ and order them such that

$$\left|c_{(1)}\right|^2 \geq \left|c_{(2)}\right|^2 \geq \cdots \geq \left|c_{(N)}\right|^2. \tag{13}$$

The NPE was defined by

$$\mathrm{NPE}(n) = \frac{\sum_{u=1}^{n} \left|c_{(u)}\right|^2}{\sum_{u=1}^{N} \left|c_{(u)}\right|^2}, \quad n = 1, 2, 3, \ldots, N. \tag{14}$$

We can see that NPE(n) varied from 0 to 1 for all n. If the NPE would be close to 1 for small n, the DWT was able to capture the key information. In other words, the mother wavelet used in the DWT was optimal. The Daubechies wavelet was denoted by db$N$, where N is its number of vanishing movements. This experiment employed db 1–8 to decompose 200 speeches that were randomly selected from our dataset. For each mother wavelet, we calculated 200 NPEs and count the average value of those NPEs. Table 2 shows the average NPE of those mother wavelets when n was equal to 5, 10, and 20, respectively.

Table 2: The NPEs of the DWT using different mother wavelets.

Mother wavelets	$n = 5$	$n = 10$	$n = 20$
db1	0.51	0.73	0.91
db2	0.66	0.75	0.93
db3	0.80	0.85	0.95
db4	**0.86**	**0.92**	**0.98**
db5	0.79	0.87	0.98
db6	0.73	0.83	0.97
db7	0.69	0.76	0.95
db8	0.65	0.76	0.95

In Table 2, we could find that the NPEs of all mother wavelets reached higher than 0.9 when n was equal to 20 and db4 and db5 obtained the highest NPEs of 0.98. When n was equal to 10, only db4 obtained the NPE of 0.92 compared with other mother wavelets which obtained the NPEs of less than 0.9. Moreover, the db4 wavelet could use only 5 wavelet coefficients to obtain the NPE of 0.86, but other mother wavelets obtain the NPE of less than 0.8. Those results show that the db4 wavelet was the most suitable mother wavelet, because its DWT can capture the key information of speech. In [33], researchers suggested that the Symlet wavelets could also obtain good performance. However, the complex conjugate mirror filters of Symlet wavelets produced the complex wavelet coefficients whose imaginary parts were redundant for real signal such as speech, so we abandoned the Symlet wavelets.

5.3. The Accuracy Rate of Investigation. This experiment tested the performance of the speaker recognition model when it was used for investigation. In investigation task, the speaker recognition model was used to compare the questioned speech with all of the speeches of known criminals in police database to produce a small list of potential suspects. In our experiment, we selected 384 speakers (192 females and 192 males) to form the large criminal set. For each speaker, the 7 recordings sampled at 16 KHz are used as the known criminal's speeches (called known speeches for short) and the recordings sampled at 8 KHz are used as the questioned speech. The similarity between the criminal's speech sample and the questioned speech was defined as

$$s = \frac{1}{7}\sum_{i=1}^{7} K\left(y_i, x\right), \tag{15}$$

where y is the known speeches and x is the questioned speech. $K(\cdot, \cdot)$ is the kernel function defined in (7). For each questioned speech, we required the speaker recognition model to produce a list of top 10 potential suspects that obtained the highest similarity. If the "real criminal" is in the

Table 3: The accuracy of investigation.

Short-term feature extractor	Accuracy rate of investigation (%)
MFCCGMM	82.46
FMFCCPNN	86.81
MFCCICDS	93.75
WCCICDS	95.48

list, the speaker recognition model got one score; if the "real criminal" was not found in the list, the model got zero score. We summed the score of the speaker recognition model for the 384 questioned speeches and calculated the accuracy rate that is defined as

$$\text{ACC} = \frac{\text{score}}{384} \times 100\%. \tag{16}$$

Because the speaker recognition model is used for investigation, the likelihood ratio is not required to be calculated.

For speaker recognition, many types of models were proposed by researchers. The typical speaker recognition model is MFCCGMM [34]. This model used 14D short-term feature vectors obtained by MFCC method to directly represent the speeches and used GMM for classification. In [20], researchers proposed a model based on Fused MFCC and probabilistic neural network (PNN), which is named FMFCCPNN. This model used 52D short-term feature vectors obtained by Fused MFCC method for speech representation and used the PNN for classification. Recently, [15] proposed a speaker model based on the MFCC, i-vector, and CDS, which is named MFCCICDS. This model used i-vector for speech representation and CDS was used for classification. Moreover, i-vector was trained using 14D short-term feature vectors obtained by MFCC method. Inspired by the above model, we proposed new speaker model based on WCC, i-vector, and CDS, which was called WCCICDS. This model was similar to the above one, but we used 126D short-term feature vectors obtained by WCC to train i-vector. Moreover, the mother wavelet used in our model was db4. For comparison, we employed the above 4 models to achieve the investigation task. The accuracy rates were shown in Table 3. In this experiment, we just used the clear speeches.

In Table 3, MFCCGMM and FMCCPNN obtained low accuracy of 82.46% and 86.81%, respectively. However, the two models in [20] obtained the accuracy of higher than 89% and 92%, respectively. This was because the speeches in [4] were obtained from same channel and had the same duration, but the speeches in our experiment were obtained from different channels and their durations were different too. This shows that the CDS could deal with the channel variability problem and i-vector was able to model the different length speeches to improve the performance of speaker recognition. Compared with the two models based on i-vector and CDS, we found that MFCCICDS obtained lower accuracy than the WCCICDS did. This was because WCC used DWT to analyze the speech signal, but MFCC used DFT. DFT used the fixed window to decompose signal, but DWT used the variable window that could obtain high frequency resolution at low frequency and high time resolution at high frequency

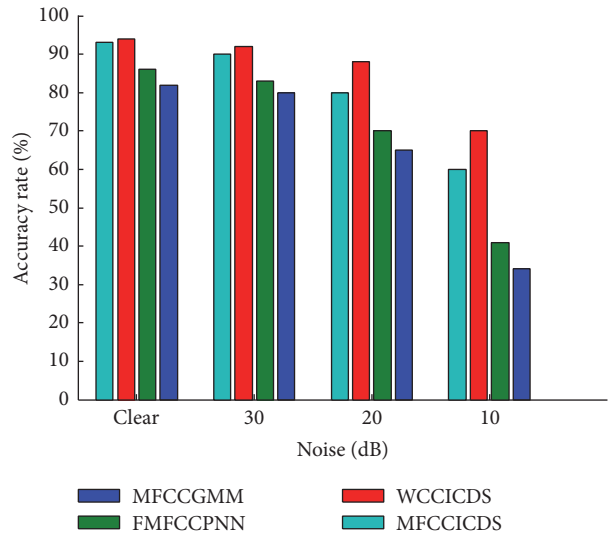

Figure 4: The accuracy rates of speaker recognition in noisy environment.

[35]. For the explodent sound that had short duration and high frequency, variable window can capture its information, but the fixed window might make the information fuzzy [22]. Therefore, the model which used WCC obtained higher accuracy.

In forensic context, the ambient noise cannot be controlled. In particular, the questioned speech is usually recorded in the condition that is full of noise. We therefore carried out an experiment to evaluate the performance of our model in noisy environment. We added 30 dB, 20 dB, and 10 dB noises to the speeches used in the above experiment. All noise was generated by the MATLAB Gaussian white noise function. We repeated the above experiment and the accuracy rates were shown in Figure 4. We could find that the accuracy rates of the four models decreased by less than 3% when the noise is 30 dB. This shows that the weak noise could not interfere in those models. However, if we enhanced the noise, we found that the accuracy rates of MFCCGMM, FMFCCPNN, and MFCCICDS dropped off rapidly. When the noise increased to 10 dB, the accuracy of the three models decreased by about 48.7%, 45.2%, and 33.4%, respectively. This shows that the two models were susceptible to noise. Compared with the above two models, the proposed model, named WCCICDS, performed better. When the 10 dB noise was added to those speeches, the accuracy of the model decreased by about 24.6%. This shows that the model using WCC was more robust against noise.

5.4. The Performance of Evidence Reporting. This experiment evaluated the performance of speaker recognition model when it was used for evidence reporting. In the evidence reporting, the suspect had been found and police used the speaker recognition model to give out the evidence that is defined as the degree of similarity between the suspect speech and the questioned speech. In this case, the speaker recognition model was required to report the strength of the

evidence as a likelihood ratio. In this experiment, we also used the above 384 speakers. For each speaker, the 7 speeches sampled at 16 KHz were also used as the suspect's speeches and the speech sampled at 8 KHz was also used as the questioned speech. We used all the speeches of the 384 speakers to estimate the probability distribution of evidence for the two hypotheses H_0 and H_1. Given a suspect, we required the speaker model to give out the evidence and report the evidence and its strength as a likelihood ratio. The evidence in this experiment is defined as

$$E = T\left[\frac{1}{7}\sum_{j=1}^{7} K\left(x, y_j\right)\right], \tag{17}$$

where x and y are the questioned speech and suspect's speeches, respectively. $K(x, y_j)$ was the CDS kernel defined in (7) and $T(\cdot)$ was a transform function defined in (9).

This section used Tippett plots [36] to evaluate the performance of our model. It was originally used for the forensic DNA analysis and then was used for evaluating the forensic speaker recognition model. Assume that there were m questioned speeches and M suspects that had been found by police. The speaker recognition model reported the likelihood ratio (LR) for each of questioned speeches and each suspect, so we obtained $m \times M$ LRs. The Tippett plot was defined as

$$T(t) = \frac{n}{N} \times 100\%, \quad 0.1 < t < 10, \tag{18}$$

where $T(\cdot)$ was the Tippett plot; n was the number of LRs that were greater than the threshold t. $N = m \times M$ was the total number of LRs. We varied the threshold t from 0.1 to 1 to obtain different Tippett plot. To evaluate the performance of our model, we calculated two types of Tippett plot. The first one was calculated in the assumption that the questioned speech and the suspect speech were spoken by same speaker and we call it T_1. The second one was calculated in the assumption that the questioned speech and the suspect speech were spoken by different speakers and we called it T_2. LR presented the strength of the evidence that the suspect was the criminal. For good speaker recognition model, the reported LR would be very high if the questioned speech and the suspect's speeches were spoken by same speaker. On the other hand, the LR would be very low if the questioned speech and the suspect's speeches were spoken by different speakers. In other words, the separation between the two types of Tippett plots is an indication of the performance of the model [37]. Given small t, a larger separation implied better performance than a smaller one. The results of this experiment were shown in Figure 5.

In Figure 5, the two types of Tippett plots of WCCICDS separated from each other when t is less than 0.5 and greater than 0.1, but the two Tippett plots of MFCCICDS stuck together. The two curves of MFCCICDS separated from each other until the threshold increased to about 1. This shows that the separation between the two types of Tippett plots of WCCICDS was slightly larger than MFCCICDS', so the WCCICDS performed better than the MFCCICDS. When

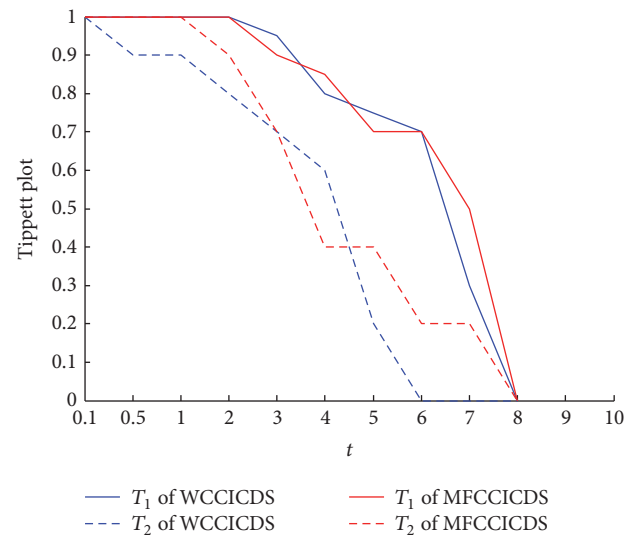

Figure 5: Tippett plot curve of the speaker recognition models.

the threshold t increased to 8, all plots decreased to 0. This was because of the fact that the threshold became too high and no LR was greater than it.

5.5. The Time Cost of Speaker Recognition Models. I-vector in many cases can improve the accuracy of recognition at cost of increasing the computational complexity, so in this experiment we test the time cost of the four speaker recognition models, MFCCGMM, FMCCPNN, MFCCICDS, and our WCCICDS. We used 200 5-second-long speeches to test their time cost and calculated the average time cost. The time cost of inputting those speeches was discarded. The result was shown in Table 4.

In Table 4, MFCCGMM and FMFCCPNN did not employ i-vector for speech representation, so they did not cost time to train i-vector. FMFCCPNN, MFCCICDS, and WCCICDS used the PNN and CDS for speaker classification. Because the PNN and CDS were the unsupervised learning models, the three models need not cost time during the training step of speaker classification. The time cost of short-term feature extraction of the proposed WCCICDS was the highest. This was because this model used WCC to calculate 126 cepstral coefficients from 9 local frequency domains compared with other three models that used MFCC or FMFCC to calculate 14 cepstral coefficients from a global frequency domain. Training i-vector was also a time-consuming process, so i-vector-based model costs more time than the model that did not employ i-vector. In all, WCC and i-vector could slightly improve the performance of the speaker recognition model at cost of increasing the time cost, so selection of speaker recognition model was the process that found the balance between the performance and time cost.

Parallel computation was an effective way to reduce the time cost, because many loops in the linear computation could be finished at once using a parallel algorithm. For example, we used DWT to decompose a signal at M levels. In the linear algorithm, we had to run a filtering process

Table 4: Average time cost of the speaker recognition models.

Model	Feature extraction (s/speech)		Speaker classification (s/speech)	
	Short-term feature extraction	I-vector training	Training	Recognition
MFCCGMM	0.45	—	1.92	0.81
FMFCCPNN	1.29	—	—	0.83
MFCCICDS	0.51	2.21	—	0.85
WCCICDS	2.91	2.19	—	0.89

whose time cost was $O(\log N)$ N times for each level, so the time complexity of DWT was $O(MN \log N)$. However, if we used a parallel algorithm to implement the DWT, we could use N independent cores to compute N filtering processes at once, and therefore the time cost reduced to $O(M \log N)$. In a further study, the parallel computation may be used to reduce the time cost of the proposed model.

6. Conclusions

In the forensic context, the speaker recognition model is usually used for investigation and evidence reporting. In the investigation, police assume that the real criminal is in a large set of known criminals and the speaker recognition model is used to produce a small list of potential suspects from a large set of known criminals. In the evidence reporting, the suspect is found, and the speaker recognition model is used to report the evidence that supports the fact the suspect was the real criminal. In this case, speaker recognition model also should report the strength of evidence as a likelihood ratio.

The forensic scene is very challenging for speaker recognition, because the ambient noise cannot be controlled and the much-change speech data are usually obtained from different channels. In this paper we propose a new speaker recognition model based on WCC, i-vector, and CDS. WCC has good performance on antinoise, because the DWT employed by WCC is a local analysis approach that can prevent the noise interfering in whole frequency domain. I-vector is a robust way to represent a speech utterance using a signal i-vector, so it can model the much-change speech data effectively. CDS employs the LDA and WCCN to compensate the channel to deal with the channel variability problem. Our experiments simulate the investigation and evidence reporting tasks. The result of our experiments shows that the proposed WCCICDS obtained high accuracy rate in the investigation task and also obtained good performance in the evidence reporting task, but its time cost was higher compared to other models. The result also shows that the parallel algorithm could effectively reduce the time cost of the model based on i-vector.

In the future, we will use the parallel algorithm to reduce the time cost of the proposed model. Moreover, we will combine audio and visual features to improve the performance of the forensic speaker recognition system.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The authors thank Professor Kun She for assistance in preparation of the manuscript and acknowledge support by Key Technology Support Program of Sichuan Province (Grant no. 2015GZ0102).

References

[1] M. Mandasari, M. McLaren, and D. A. V. Leeuwen, "Evaluation of I-vector speaker recognition systems for forensic application," in *Proceedings of the INTERSPEECH*, pp. 21–24, August 2011.

[2] A. Eriksson, "Aural acoustic vs. Automatic methods in forensic phonetic case work," in *Forensic Speaker Recognition*, pp. 41–69, 2012.

[3] A. Drygajlo, "From speaker recognition to forensic speaker recognition," *Biomedical Authentication*, vol. 8897, pp. 93–104, 2014.

[4] J. R. Conzalez, A. Drygajilo, D. R. Castro, and M. G. Garcia, "Robust estimation interpretation and assessment of likelihood ratios in forensic speaker," *Computer Speech & Language*, vol. 20, pp. 331–355, 2006.

[5] 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.

[6] C. Champod and D. Meuwly, "Inference of identity in forensic speaker recognition," *Speech Communication*, vol. 31, no. 2, pp. 193–203, 2000.

[7] M. A. Silveria, C. P. Schroeder, J. P. C. L. da Costa et al., "Convolutive ica-based forensic speaker identification using mel frequency cepstral coefficients and gaussian mixture model," *The International Journal of Forensic Computer*, vol. 1, pp. 27–34, 2013.

[8] S. Srivastava, S. Bhardwaj, and A. Bhandari, "Wavelet packet based mel frequency cepstral features for text independent speaker identification," *Advances in Intelligent Systems and Computing*, vol. 182, pp. 237–247, 2013.

[9] 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.

[10] D. A. Leeuwen and N. Brümmer, "The distribution of calibrated likelihood-ratios in speaker recognition," in *Proceedings of the*

14th Annual Conference of the International Speech Communication Association (INTERSPEECH '13), pp. 1619–1623, Lyon, France, August 2013.

[11] N. Dehak, R. Dehak, and P. Keney, "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)*, vol. 9, pp. 1559–1562, September 2009.

[12] H. Arsikere, S. M. Lulich, and A. Alwan, "Estimating speaker height and subglottal resonances using MFCCs and GMMs," *IEEE Signal Processing Letters*, vol. 21, no. 2, pp. 159–162, 2014.

[13] T. Becker, M. Jessen, and C. Grigoras, "Forensic speaker verification using formant features and Gaussian mixture models," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 1505–1508, September 2008.

[14] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: a review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[15] 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, September 2015.

[16] 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, May 2013.

[17] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, and M. Mason, "Weighted LDA techniques for I-vector based speaker verification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '12)*, pp. 4781–4784, Kyoto, Japan, March 2012.

[18] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[19] S. K. Kopparapu and M. Laxminarayana, "Choice of Mel filter bank in computing MFCC of a resampled speech," in *Proceedings of the 10th International Conference on Information Sciences, Signal Processing and Their Applications (ISSPA '10)*, pp. 121–124, Kuala Lumpur, Malaysia, May 2010.

[20] K. S. Ahmad, K. J. Thosar, A. S. Nirmal, and J. H. 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*, pp. 1–6, Kolkata, India, January 2015.

[21] N. Lulla and N. Purohit, "An improved algorithm for efficient computation of MFCC," in *Proceedings of the 11th IEEE India Conference (INDICON '14)*, pp. 1–4, Pune, India, December 2014.

[22] 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, November 2006.

[23] 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.

[24] Y. Zhao, L. Zhang, J. Hu, and T. Liao, "Mallat wavelet filter coefficient calculation," in *Proceedings of the 5th International Conference on Computational and Information Sciences (ICCIS '13)*, pp. 963–965, Chengdu, China, June 2013.

[25] 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.

[26] Y. Lei, L. Burget, and N. Scheffer, "A noise robust I-vector extractor using vector taylor series for speaker recognition," in *Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '13)*, pp. 6788–6791, Vancouver, Canada, May 2013.

[27] 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, May 2011.

[28] A. Alexander, D. Dessimoz, F. Botti, and A. Drygajlo, "Aural and automatic forensic speaker recognition in mismatched conditions," *International Journal of Speech, Language and the Law*, vol. 12, no. 2, pp. 214–234, 2005.

[29] 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.

[30] S. Mallat, *A Wavelet Tour of Signal Processing*, Elsevier, 2012.

[31] I. Deubechies, *Ten Lectures on Wavelets*, Society for Industrial and Applied Mathematics, 1992.

[32] D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis*, vol. 4, Cambridge University Press, 2006.

[33] I. S. Prabha, M. K. Rao, and B. Manjusha, "Voice verification system using symlet wavelet," *International Joural of Research in Computer Applications and Robotics*, vol. 1, pp. 24–30, 2013.

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

[35] 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 Journal*, vol. 27, pp. 231–239, 2015.

[36] G. A. Morrison, "Vowel inherent spectral change in forensic voice comparison," in *Vowel Inherent Spectral Change*, pp. 263–282, Springer, 2013.

[37] G. Zadora, A. Martyna, and D. Ramos, *Statistical Analysis in Forensic Science: Evidential Value of Multivariate Physicochemical Data*, John Wiley & Sons, New York, NY, USA, 2013.