IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 13, 2018, accepted December 4, 2018, date of publication December 17, 2018,
date of current version January 7, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2886428

A Robust Multilevel Speech Verification With Wavelet Decomposition for Inadequate Training Data Sets of Mobile Device Systems

KUO-KUN TSENG[1], YAN ZHANG[1], K. L. YUNG[2], W. H. IP[2], ZHIYE OU[1], AND QI NA[1]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China
[2]Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: W. H. Ip (wh.ip@polyu.edu.hk)

This work was supported in part by the Shenzhen Government under Grant KQJSCX20170726104033357 and Grant JCYJ20160531191837793, in part by the Department of Industrial and Systems Engineering, in part by The Hong Kong Polytechnic University under Grant H-ZG3K, in part by the Shenzhen Medical Biometric Perception and Analysis Technology Engineering Laboratory, and in part by the Shenzhen Development and Reform Commission through Shenzhen Development and Reform under Grant (2016) 889.

ABSTRACT With the development of speech signal processing, universality, easy collection and personal speech signal uniqueness, many researchers are attracted to the field of speech verification. Most of the current speech verifications are based on long training data sets in order to achieve good results, and there are no good verification schemes in case of inadequate training data sets. This paper proposes a novel architecture for speech verification using a multilevel method, which extracts feature parameters through a multiple wavelet transform for mobile phone voice. The experiments show that the multilevel wavelet authentication architecture improves performance in speech verification. The recognition rate of the mobile phone system is more robust and superior to other methods.

INDEX TERMS Biometric, speech verification, wavelet transform, mobile computing.

I. INTRODUCTION

In recent years, more and more researchers have been interested in the application of biometric technology to identification and verification. Studies of biometric technology include face recognition [1]–[3], fingerprint recognition [4], iris recognition, palm-print recognition [5] and speech recognition [6]. There are some problems associated with the use of each biometric; for example, facial features will change greatly with increasing age and with cosmetics. These traditional biometric technologies require special equipment or supporting hardware, and therefore are not conducive to popular application. In addition, the cost is high. This paper proposes a speech-based verification system on a mobile device which has several advantages compared to the previously-mentioned systems: (1) speech is a common and easily-obtained communication signal and (2) we need only a simple mobile device such as a mobile phone to collect original speech.

With the rapid development and popularity of mobile phone, the combination of biometric technology with such devices not only increases convenience for users, but also promotes the popularity of biometric technology. Although speech-based verification technology has a number of advantages, it also has some problems, such as: (1) noise interference and (2) the use of different types of mobile phones.

We propose a novel architecture for multilevel speech verification with DTW in this paper. This architecture is a new method of multilevel verification which can make use of various feature parameters in order to achieve a higher verification rate than single feature parameter methods. Theoretically, with enough complementary feature parameters, the recognition rate will be close to unity. We show experimentally that a variety of features for multilevel verification does indeed result in an improvement for the verification system, to some degree.

Our experiments show that the speech signal after the process of effective components extraction algorithm, not only can ensure the recognition rate of existing, the most important is to reduce consumption of system and database storage. Then, our architecture can make full use of the details

of various feature extraction algorithm to achieve a better verification.

This paper also put forward three contents, from different angles of speech verification to improve the overall recognition performance. The system implementation also proves the feasibility of the system at last. The whole framework based on the mobile speech verification system has higher reference value for the subsequent research.

In a summary, these research works include the following aspects:

(1) Fully studied the related work, and proposed a new architecture and application on the speech verification for mobile terminals.

(2) Aiming at each specific content, this paper builds a speech database based on mobile terminal authentication which meets the requirements of the experiment. The database covers a wide range of contents, and the speech signal acquisition follows the experimental rules. It is fully fitted to the experimental needs.

(3) Through the self-built speech database experiment, it has proved the superiority of the speech effective component extraction algorithm based on short time energy. It can effectively extract the effective voice segment in the speech signal segment, remove the invalid noise segment, reduce the error rate. It fully proves the reliability and excellent performance of the algorithm.

(4) A multilevel speech verification architecture is proposed, which effectively combines the speech features of each layer of the wavelet feature extraction algorithm to improve the overall recognition rate of speech data based on small data. Through experiments, the experimental results also confirm this improvement.

The remaining parts of this paper are organized as follows. In the second section, we introduce related work. The third section proposes a novel multilevel architecture for speech verification and describes in detail the endpoint detection algorithm. The fourth section, compares the algorithms and gives the experimental results. The final section summarizes the whole paper.

II. RELATED WORK

Related algorithms for speech verification are mainly concentrated in three areas: pre-processing for speech signals, feature extraction from the original speech signal and pattern matching. There are now many sophisticated algorithms relating to each aspect. From a large number of research papers on speech verification, we obtained taxonomy for related algorithms as shown in Figure 1.

In a speech verification system, pre-processing can be divided into two parts: speech denoising and endpoint detection. Wavelet analysis is a significant technology for denoising. Wang and Li [7] offer a comparison of the performance for denoising between the traditional discrete Fourier transform (DFT) approach and a discrete wavelet packet transform. Their experiments show that the denoising performance

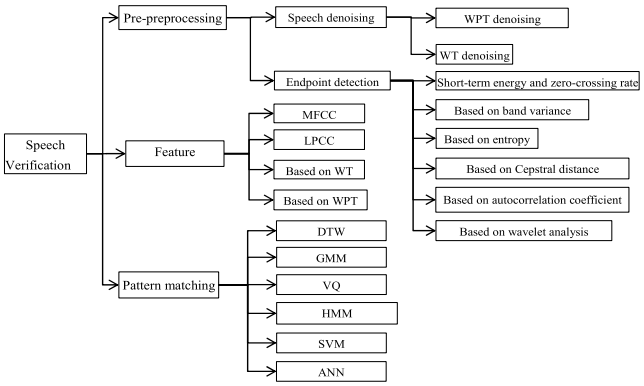

FIGURE 1. Taxonomy of speech verification.

of the discrete wavelet packet transform is much better than that of traditional DFT methods.

Speech endpoint detection is an important step in speech verification systems which correctly and effectively identifies the start and end of a given speech signal, thus greatly reducing the amount of calculation necessary and shortening the processing time. It also eliminates interference and silences noise. Many studies have shown that in a quiet environment, inaccuracy in endpoint detection can directly reduce the operating performance of speech verification systems. A number of different speech endpoint detection algorithms have been proposed. These improve the anti-noise performance mainly through a variety of new parameters such as cepstral-based measures [8], band variance [9], autocorrelation similarity distance [10] and information entropy [11], and are gradually being applied to endpoint detection. However, in many cases, an endpoint detection algorithm is mainly based on the time-domain characteristics of the speech signal. The main parameters are the short-term energy and short-term average zero-crossing rate. The method based on short-term energy and zero-crossing rate is called the double-threshold comparison method.

Feature extraction algorithms play a significant role in speech verification systems. In this paper, we list some of the more common and practical feature extraction algorithms such as MFCC, LPCC and some extraction algorithms based on wavelet or wavelet packet transforms. Algorithms based on wavelet analysis have an improved robustness for environmental noise. Sarikaya *et al.* [12] propose a subband based cepstral (SBC) using a wavelet packet transform which achieves a higher performance. The principle of SBC is to apply a wavelet packet transform to the windowed speech signal and then calculate the energy of each sub-band to give the feature parameters. Experiments show that this method does indeed have a better performance and environmental robustness.

Pattern recognition technology has been in existence for some time. Many pattern matching algorithms are equally applicable to speech verification, e.g., DTW, GMM, VQ, SVM, ANN and HMM. DTW is one of the pattern matching techniques, which uses dynamic programming ideas to successfully resolve the problem that speech signal feature

parameter sequences do not all have the same length. DTW can achieve good performance in isolated-word speech recognition. Furthermore, DTW is quick and easy to calculate, and is a very efficient choice for specific short utterance verification in this study. In 1994, Reynolds [13] began applying the Gaussian mixture model to speaker recognition and verification, and achieved a high recognition performance. Subsequently, in 2000, Reynolds *et al.* [14] proposed an adaptive Gaussian mixture model. Vector quantization (VQ) is also a common pattern matching algorithm in speech verification. The basic principle of VQ is that for each specific person, an eigenvalues training codebook is used to provide pattern matching for later reference. The process of verification is simple and fast, but cannot achieve a high recognition rate. Therefore, Chunbao *et al.* [15] propose an improved VQ algorithm with a new weighted measure, which takes into account the correlations between the interspaces and intraspaces of vectors, and the experimental results show that the new VQ algorithm enhances the differences between speakers and boosts the rate of speaker recognition. At the same time, many researchers have begun to try to achieve a higher performance using a combination of two pattern matching algorithms. For example, Kruger *et al.* propose a method using a combination of SVM and HMM. They use parallel mixtures of SVMs for classification by integrating this method into an HMM-based speech recognition system.

After much related literature study and compared algorithm experiment, we found that for a specific person and a specific sentence, the dynamic time warping (DTW) algorithm is an efficient and lightweight matching algorithm. This algorithm can effectively solve the problem of inconsistent speech length by dynamically finding the best-matching path, to derive the similarity of two short speech signals. We found that the existed approaches are applied to speech verification with good quality speech data or a large training data set. However, a small training data set for the mobile phone verification system, no efficient speech verification is observed.

Recently, the research and development trend of this aspect is to optimize the algorithm of front end device [17], integrate the two algorithms of GMM and CNN (Convolutional Neural Network) [18], and consider the research of the Multi-scenario scene [19], but at present there is not a proper combination of a variety of special certificates, and the integration framework suitable for the mobile terminal has been proposed. In this paper, we propose a new method based on a multilevel architecture, has not been used previously for this purpose.

III. BACKGROUND STUDY

In this section, we will study three important algorithms in more detail that are used in the later proposed architecture. They are Endpoint Detection (Double-Threshold Comparison Method), Feature Parameters Based on Wavelet Decomposition Dynamic Time Warping (DTW), and Multilevel Dynamic Time Warping (DTW) algorithms.

A. ENDPOINT DETECTION (DOUBLE-THRESHOLD COMPARISON METHOD)

This method integrates short-term energy and zero-crossing rate and uses the energy and zero-crossing rate as endpoint detection features. In cases where the SNR is not too low, it is assumed that the energy of the speech signal is greater than the noise energy. By comparing the energy of the input signal with the speech energy threshold, we can distinguish between speech segments and non-speech segments. The signal {x (n)} of the short-term energy is defined as:

$$E_n = \sum_{m=-\infty}^{\infty} [x\,(n) \cdot w(n-m)]^2 \tag{1}$$

where w (n) is the window function. The short-term zero-crossing rate formula can be expressed as:

$$Z_n = \frac{1}{2} \sum_{m=-\infty}^{\infty} |sgn\,[x\,(m)] - sgn\,[x(m-1)]| \cdot w\,(n-m) \tag{2}$$

In the above formula, sgn is the sign function, given by:

$$sgn\,[x(n)] = f\,(x) = \begin{cases} -1, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{3}$$

In general, the characteristics of the zero-crossing rate are as follows. Dullness has a clear cycle, so the zero-crossing rate of ambient noise and voiceless volume is greater than that of dullness; ambient noise is very similar to the voiceless volume, and therefore it is difficult to distinguish between them by their zero-crossing rate, and because the voiceless volume is generally greater than the ambient noise, it is possible to combine the zero-crossing rate and the volume to detect the endpoints.

Several other endpoint detection methods require relatively extensive calculations which are more complex than those of the double-threshold method, so these are not suitable for endpoint detection in mobile end devices. For example, although the detection method based on the wavelet transform has a high detection accuracy, the speed of detection is very slow. For 50kB of voice data, even on a CPU with a speed of 2GHz, it takes about five seconds to complete the detection. Therefore, this kind of computation is completely outside the scope of application in this study.

B. FEATURE PARAMETERS BASED ON WAVELET DECOMPOSITION

Wavelet transforms can also be understood as a filtering process for the original speech signal. The results of the decomposition will be different for different base functions. However, no matter what type of base function is selected, different decomposition scales use the fixed proportion between the filter center frequency and the bandwidth. This is also called the "constant Q" feature. Meanwhile, the smoothed signal and the detailed signal of each scale space can provide the local frequency and time information

of the original signal, in particular reflecting the composition information of the signals of different frequency bands. If we calculate the signal energy using different decomposition scales, we obtain the speech feature parameters.

The wavelet function is defined as follows: If we let $\varphi(t) \in L^2R$, then its Fourier transform is $\hat{\varphi}(\omega)$, when $\hat{\varphi}(\omega)$ satisfies the condition

$$C_\varphi = \int_R \frac{|\hat{\varphi}(\omega)|^2}{|\omega|} d\omega < \infty \tag{4}$$

Then φ(t)is the base function or mother function. Equation (1) gives tolerable conditions for the wavelet function. Scaling $\varphi(t)$ and translating gives $\varphi_{a,b}(t)$:

$$\varphi_{a,b}(t) = \frac{1}{\sqrt{a}}\varphi\left(\frac{t-b}{a}\right) a, b \in R; a > \tag{5}$$

where a is the scale factor, b is the translation factor and $\varphi_{a,b}(t)$ is a wavelet function dependent on a and b.

For each basic wavelet, the continuous wavelet transform of f(t) is

$$WT_f(a, b) =< f(t), \quad \varphi_{a,b}(t) > \frac{1}{\sqrt{a}} \int_R^{f(t),\ \varphi^*} (\frac{t-b}{a})d \tag{6}$$

In practical applications, the continuous wavelet must be discrete, therefore letting $a = a_0^j, b = ka_0^j b_0, \in Z$, and assuming $a_0 > 0$, the corresponding discrete wavelet transform function is:

$$\varphi_{j,k}(t) = a_0^{\frac{-j}{2}} \varphi(a_0^{-j} t - kb_0) \tag{7}$$

The discrete wavelet coefficients can be expressed as

$$C_{j,k} = \int_{-\infty}^{+\infty} f(t)\varphi_{j,k}^*(t)\, dt =< f, \varphi_{j,k} > \tag{8}$$

and the reconstruction formula is:

$$f(t) = \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} C_{j,k}\varphi_{j,k} \tag{9}$$

However, here we do not need to reconstruct the signal, but only to extract the energy characteristics of each decomposition scale:

$$energy = \sum_{start}^{end} |coef(i)|^2 \; start \le i \le end \tag{10}$$

where, coef is the discrete wavelet coefficient, and start and end are the coordinates for starting and ending in the corresponding scale. The implementation process is as shown in Figure 2.

C. DYNAMIC TIME WARPING (DTW)

Dynamic time warping (DTW) is a typical optimization problem which describes the time correspondence based on the Euclidean distance between input templates and the reference template with the time warping function $W(n)$.

Suppose we have two speech time series Q and C, with lengths n and m, respectively.

$$Q = q_1, q_2 \ldots, q_i, \ldots, q_n$$

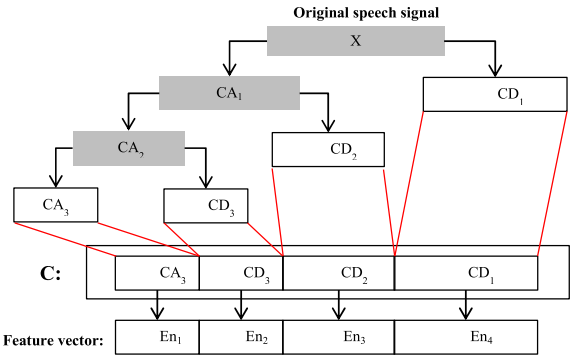

FIGURE 2. Parameter extraction process based on wavelet decomposition characteristics.

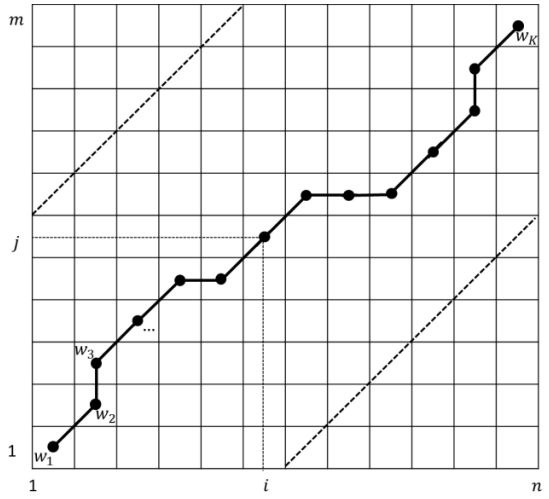

FIGURE 3. An example warping path.

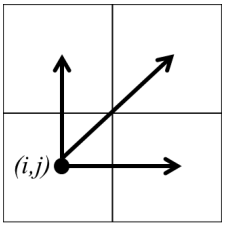

FIGURE 4. Three directions for a point.

$$C = c_1, c_2, \ldots, c_j, \ldots, c_n$$

If n is equal to m, we can calculate the distance between the two sequences directly. If n is not equal to m, they must be aligned.

First, we construct a matrix $n * m$, where the matrix element (i, j) represents the distance between q_i and c_j, with the general Euclidean distance, $d(q_i, c_j) = (q_i - c_j)^2$. Each matrix element (i, j) represents the alignment between points q_i and c_j.

We define the best path as the warping path with the minimum distance, represented by W. The kth element of W

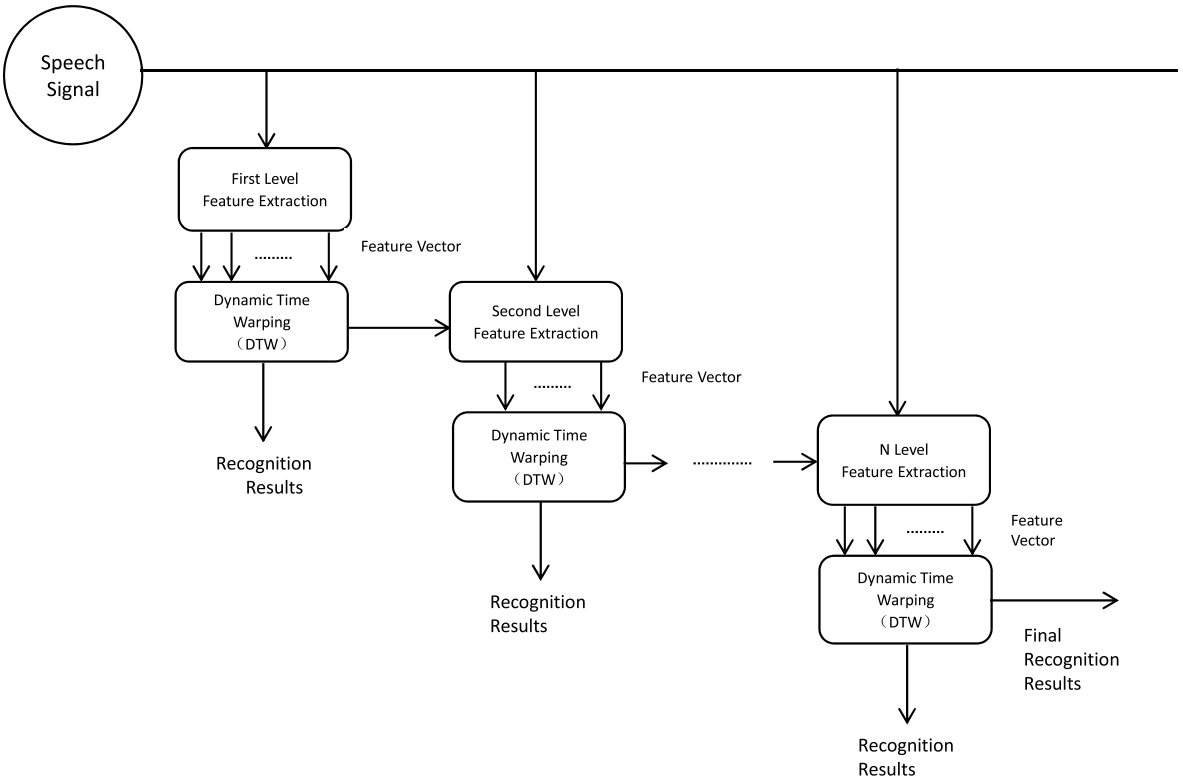

FIGURE 5. The architecture of the multilevel algorithm.

is defined as $w_k = (i, j)_k$. Thus, we have:

$$W = w_1, w_2, \ldots, w_k, \ldots, w_K \max(m, n) \leq K < m + n - 1$$

The selection of the path must meet the following constraints:

Boundary conditions: $w_1 = (1, 1), w_K = (m, n)$. Speech pronunciation speed is likely to change, but the order of the various parts cannot be changed, so the path we choose must be from the lower left corner to the upper right corner.

Continuity: $w_{k-1} = (a, b)$, then the next point in the path with $w_k = (a, b)$, must satisfy $(a - a) \leq 1$ and $(b - b) \leq 1$. This is unlikely to match with another point, and each point can only align with adjacent dots. This ensures that both Q and C are present in W.

Monotonicity: $w_{k-1} = (a, b)$, then the path for the next point $w_k = (a, b)$ must satisfy $0 \leq (a - a)$ and $0 \leq (b - b)$.

For continuity and monotonicity, the path of each grid point has only three possible directions. For example, if the path has passed the grid point (i, j), the next grid point must be one of the following three: $(i + 1, j)$, $(i, j + 1)$ or $(i + 1, j + 1)$.

Starting from point $(0, 0)$, we begin to match the two sequences Q and C. For each point, all the distances calculated previously will be accumulated to calculate the total distance. The cumulative distance $\gamma(i, j)$ can be expressed as in equation (11).

$$\gamma(i,j) = d\left(q_i, c_j\right) + min\gamma\left(i-1, j-1\right), \gamma\left(i-1, j\right), \gamma\left(i, j-1\right) \quad (11)$$

After reaching the endpoint (n, m), the cumulative distance represents the similarity between Q and C.

IV. MULTILEVEL-BASED ARCHITECTURE FOR SPEECH VERIFICATION

The multi-feature fusion is our research direction, the proposed multilevel concept as Figure 5 shows, after the endpoint detection and feature extraction; we compute the distance using DTW algorithm in multiple iteration. Only when the distance is less than a given threshold will the sample enter the next level. After performing the entire process, we obtain an optimized threshold and corresponding level where the best verification rate is achieved.

The detailed algorithm of our proposed multilevel architecture for speech verification is shown in Figure 6.

The second step is using *featureExtraction()* to extract feature. For speech denoising, due to the complexity and diversity of environmental noise, it is difficult to achieve a perfect method which can handle all types of noise. Therefore, in this paper we extract feature parameters based on a wavelet transform with robustness for the ambient noise.

In this algorithm, the first step is using *epd()* for endpoint detection. The device is susceptible to ambient noise while obtaining speech signals. The tone, the device used and the speech rate of the speaker will affect the quality of original speech. Another concern is that useless speech signals commonly exist at the beginning or the end of original

Multilevel Architecture for Speech Verification

INPUT: original *train data and test data*
OUTPUT: *optimized threshold, rate and level-number*

```
train data←epd(original train data)
test data←epd(original test data)
feature_of_train←featureExtraction(train data)
feature_of_test←featureExtraction(test data)
threshold←0
level_num←1
Rate←0
For each i from 1 to N
  dist←DTW(feature_of_train[i], feature_of_test[i])
  temp_thr←max(dist)
  temp_rate←compuRate(dist,Temp_thr)
  If temp_rate>rate then
    rate←temp_rate
    threshold←temp_thr
    level_num←i
  end
end
```

FIGURE 6. The pseudocode for a multilevel architecture for speech verification.

speech. Therefore, endpoint detection is necessary before the feature parameters are extracted, in order to reduce the amount of calculation required. In this paper, we included endpoint detection for the architecture proposed. Our architecture selects the double-threshold comparison with short-term energy and zero-crossing rate. This method combines the advantages of the both; this method is simple and efficient.

We make use of the wavelet transform for its good localization properties in the time domain and the frequency domain, and we proceed via a multilayer wavelet decomposition of each frame of the speech signal, to obtain multiple sets of characteristic parameters for later pattern matching. That the feature extraction parameters based on wavelet decomposition are used to extract multi-layer wavelet parameters of speech signals as characteristic parameters for later DTW matching algorithm.

In this algorithm, we need to use multilayer wavelet decomposition to extract a seven-dimensional feature parameter. One of the seven dimensions is the SBC parameter mentioned above. The others are the energy feature parameters of the third-, fourth-, fifth-, sixth-, seventh- and eighth-order wavelet decomposition. These feature parameters are assigned to each level of the architecture in order to seek the best results. At each level, we can remove some samples which have larger differences compared with the training data. We must also record the recognition rate and threshold in this level of the system. Following this, the rest of the test samples enter a later stage up to the final level.

The third step is using DTW() for similarity matching, DTW can meet the requirements of specific speech verification. However, using simple DTW does not result in a better performance than before, and therefore we propose a multilevel speech verification architecture based on DTW. At each level, we calculate the distance between test samples and train samples with DTW for multiple level N. Using this novel architecture based on DTW to obtain a better performance. Two optimized parameters used in this algorithm which are rate and threshold. They are obtain from computRate().

V. COMPARED ALGORITHMS AND EXPERIMENTAL RESULTS

In this section, we will briefly introduce some compared algorithms and our personal speech database. In this paper, we also give some experimental results for related feature extraction algorithms based on our personal speech database. The experimental results demonstrate that the architecture for multilevel speech verification using wavelet transforms on mobile devices does indeed improve speech verification.

A. COMPARED ALGORITHMS

In this subsection, we compare various algorithms related to this study. These algorithms can be divided into two types: feature extraction algorithms and pattern matching algorithms.

1) FEATURE EXTRACTION ALGORITHMS

a: Mel-Frequency Cepstral Coefficients (MFCC)

MFCC is a feature parameter that focuses on the auditory perception of the human ear; the process of extraction involves the concept of the critical band. According to the divisions of the critical band, the speech frequency can be divided into a series of triangles in the domain filter sequence. This is often called a mel filter bank. It is necessary to extract all the weighted sums of the signal amplitude as the output of a critical band filter for each critical band, and then construct a logarithmic arithmetic for all filter outputs, by drawing a vector. Finally, we obtain the MFCC parameters using a discrete cosine transform (DCT).

MFCC parameters are commonly calculated according to the following processes:

Step 1:Confirming the number of points in each frame of the speech sample. This paper selects $N = 256$ and then through the discrete FFT (fast Fourier transform) transforms each frame sequence $S(n)$ to obtain the power spectrum $S(n)$ which is squared for the modulus.

Step 2: Calculating the sum of the products of $S(n)$ and M filters $H_m(n)$ on each discrete frequency point, we can obtain $p_m, m = 0, 1, \ldots, M-1$.

Step 3: By calculating the natural logarithm of p_m, we obtain $L_m, m = 0, 1, \ldots, M-1$.

Step 4: Finally, we construct a DCT for $L_m, m = 0, 1, \ldots, M-1$, and obtain $D_m, m = 0, 1, \ldots, M-1$.

We must delete the DC component D_0; other components can be used as MFCC parameters. Finally, we calculate the

first-order differential for MFCC to obtain a new set of MFCC difference coefficients, to use as the final feature parameters.

b: LPCC

Here, we will briefly introduce the linear predictive cepstral coefficient (LPCC). LPCC is a speech feature belonging to the frequency domain. It uses a homomorphic processing algorithm to extract the impulse response of the speech channel. LPCC parameters can be obtained by recursion for the linear prediction coefficient (LPC), using the following recursive formula:

$$\begin{cases} c_1 = a_1 m = 1 \\ c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} 1 < m \leq p \\ c_m = \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k} m > p \end{cases} \tag{12}$$

In equation (12), $a_1, a_2, a_3, \ldots a_m$ is a p-order LPC feature vector and $c_m, m = 1, 2, 3, \ldots m$ are the first p values of the cepstral. When n is less than or equal to p, we use the second equation. If n is greater than p, we use the third equation to compute the LPCC. Therefore, formula 12 can be used to obtain LPCC from LPC. The required LPCC feature parameters can thus be obtained.

2) PATTERN MATCHING ALGORITHMS

a: Hidden Markov Model(HMM)

A hidden Markov model is a kind of Markov chain, whose state cannot be observed directly. However, we can observe the observation vector sequence, and each observation vector is generated by a state sequence with a corresponding probability density distribution. Therefore, HMM is a double random process. Firstly, it is a hidden Markov chain with a certain number of states, and secondly, it has random function sets for display.

There are three basic types of problems with regard to HMM:

Learning problems:

- The HMM model parameters are $\lambda = (A, B, \pi)$, but how to adjust these parameters to make the probability of the observation sequence $O = o_1, o_2, o_3, \ldots, o_t$ as large as possible is not known.
- Assessing problems: Given the observation sequence $O = o_1, o_2, o_3, \ldots, o_t$ and the model parameters $\lambda = (A, B, \pi)$, how can the probability of a particular observation sequence be effectively calculated?
- Decoding problems: Given the observation sequence $O = o_1, o_2, o_3, \ldots, o_t$ and the model parameters $\lambda = (A, B, \pi)$, how can the best state sequence be found?

For speech verification based on HMM, we only need to solve the first two problems, i.e., the learning problem and the assessing problem for the final testing set and the training model. However, after much research, we found that large training data sets are needed to train an effective model. Therefore, this approach is not suitable for mobile devices.

TABLE 1. Details of the first group database.

	S1	S2	S3	S4	S5	S6	S7	S8	S9
Positive	10	10	10	7	10	10	10	10	10
Negative	9	9	9	9	9	8	8	8	9

b: Gaussian Mixture Model(GMM)

A probability density function in the Gaussian mixture model is obtained through calculating the weighted sum of a number of Gaussian probability density functions, as shown in the following formula:

$$p(x \mid \gamma) = \sum_{i=1}^{M} p(\omega_i) N\left(X; \mu_i; \sum_i\right) \tag{13}$$

Here, $p(x \mid \gamma)$ is the probability density function for observation x of a GMM model, X is a random vector of dimension d, $p(\omega_i)(i = 1, 2, \ldots M)$ are the weights of the GMM functions to be summed. $N\left(X; \mu_i; \sum_i\right), i = 1, 2, \ldots M$ is the probability density function of the ith single Gaussian distribution. It is necessary to calculate γ from $p(\omega_i)$, μ_i and $\sum_i$.

The algorithm steps can be described as follows:

Given speech training data, we need to estimate the parameter γ using EM (estimate maximization) to ensure that $p(X \mid \gamma)$ is maximized. We then calculate $p(X \mid \gamma)$ as follows.

$$p(X \mid \gamma) = \prod_{t=1}^{T} p(x_t \mid \gamma) \tag{14}$$

When assessing a new $p(X \mid \gamma)$, make $p(X \mid \bar{\gamma}) \geq p(X \mid \gamma)$. Then the new model parameters emerge as the initial model parameters are trained by iteration until the model converges.

However, when using the GMM model, some problems should be considered. Firstly, the order M of the model must be moderate but large enough to fully express the distribution space. However, the order cannot be too large, otherwise there is insufficient data to accurately describe the distribution space, and furthermore, it may have a non-singular correlation matrix. Therefore, we cannot make an effective model with limited training data for a mobile phone.

In summary, the experiments in the latter part of the paper are all based on dynamic time warping (DTW).

B. SPEECH DATABASE

According to needs of the research, we collected the speech information around us in order to establish the experimental database. The following tables show the details of the speech database collected, including 20 speakers which are assigned to one of three groups. The first group in the speech database was collected in a variety of noisy environments, and the recording environment for the second group of voices was relatively quiet, for later comparison experiments.

TABLE 2. Details of the second group database.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Positive	12	11	13	10	12	12	11	10	13	12
Negative	12	12	12	9	12	12	9	12	12	12

TABLE 3. Details of the third group database.

Positive/ Negative	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Note2	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9
5s	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9
M4	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9	12/9

The third group in the speech database was collected from different types of mobile phones with some environmental noise. Each speech signal was collected by three mobile phones at the same time. From Table 3, it can be seen that in this speech collection, 3 types of mobile phones, 10 experimental S subjects collected 12 positive examples and 9 negative examples respectively, which were used for later experiment. First, three positive speech signals are taken from each individual speech data for training use, and the rest is used as positive example of test data. At the same time, each individual also has 9 counterexample of test data, so that the positive and negative test data in this experiment are approximately equal, and the evaluation of experimental results is more scientific and credible.

C. EXPERIMENTAL RESULTS

1) EXPERIMENT ON PROCESSING

In the previous section, we pointed out that the preliminary design of endpoint detection in this topic is mainly based on the short-term energy and the zero-crossing rate, and we discussed previous research and our own work showing the specific results achieved by the basic algorithms. Details are given in Figure 3.

Figure 7 shows the endpoint detection based on the same speech data, and the two horizontal lines that we refer to as the double threshold. The direct impact of the relative values of these can be seen. We now consider the importance of endpoint detection in the verification rate of our entire speech database.

We can clearly see that there is an approximately five percent difference in recognition rate arising from the differences in threshold rate (volRatio). Therefore, the endpoint detection results directly affect the performance of the entire speech verification system. Following many experiments with volRatio = 3.5, we found that we could achieve the best performance based on our database. Here, the principle of threshold selection is:

$$volTh = (volMax - volMin)/volRatio + volMin \quad (15)$$

2) EXPERIMENTS ON FEATURE EXTRACTION

This paper is mainly focused on the study of LPCC, MFCC, an algorithm based on the wavelet transform and

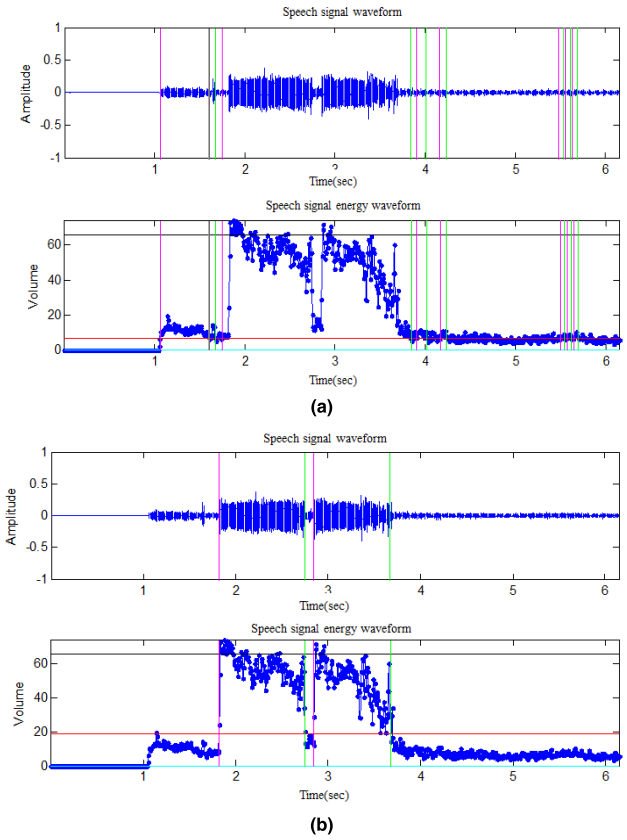

FIGURE 7. The performance of different threshold rates (volRatios). (a) Non-ideal endpoint detection. (b) Ideal endpoint detection.

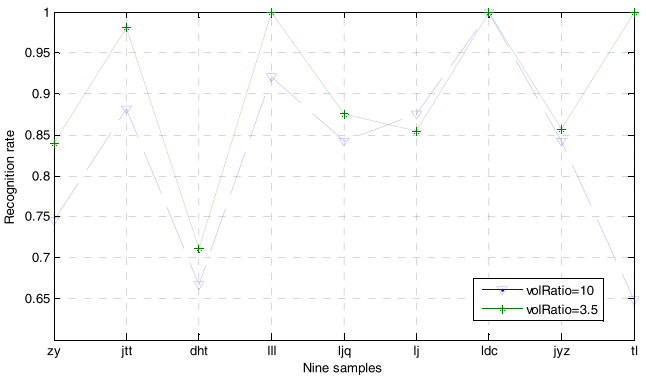

FIGURE 8. The performance of the whole database under different volRatios.

the similarity-matching algorithm of dynamic time warping (DTW). Related algorithms have been developed and verified in MATLAB. The illustrations below show the performances of the main feature extraction algorithms in different situations.

From a large number of experiments, and in conjunction with Figures 6 and 7, it can be concluded that feature extraction based on wavelet analysis has strong robustness in a noisy environment, and that the performances of LPCC and MFCC still have some deficiencies compared with SBC in

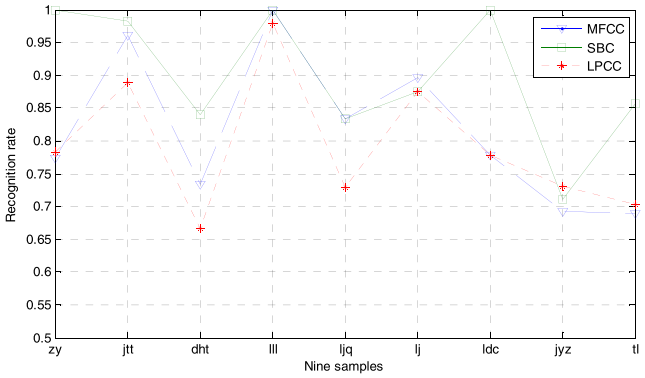

FIGURE 9. Performance comparison of MFCC, LPC and SBC in the first group(more noise).

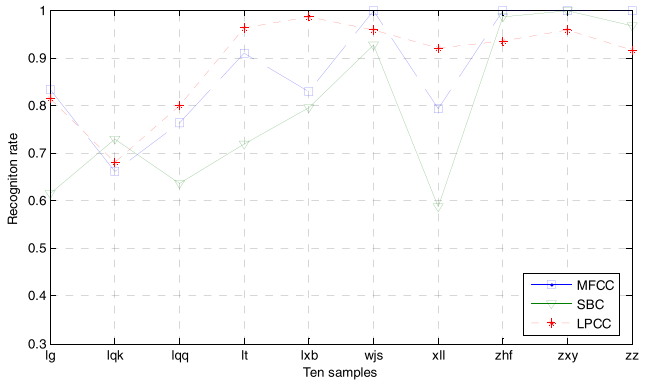

FIGURE 10. Performance comparison of MFCC, LPC and SBC in the second group.

TABLE 4. Results of multi-feature combination.

Set	MFCC	SBC	LPCC	MFCC+SBC
Test1	77.4	100	78.2	100
Test2	96.1	98.2	88.9	98.2
Test3	73.4	84.0	66. 7	88.2
Test4	100	100	97.9	100
Test5	83.3	83.3	72.9	85.5
Test6	89.6	87.5	87.5	87.5
Test7	77.8	100	77. 8	100
Test8	69.3	71.1	73.1	69.2
Test9	69.0	85.7	70. 4	83.6

noisy environments. However, we cannot ignore the fact that if the SNR is relatively large, then LPCC and MFCC perform better than SBC. This is an issue for further study.

At the same time, our approach via a variety of types of feature fusion and increasing the eigenvalue dimensions, attempts to integrate the advantages of a variety of feature extraction algorithms. Details are shown in Table 4.

This experiment combined MFCC with SBC as a single group of two-dimensional characteristic parameters for voice authentication, though the results did not meet our expectations.

3) EXPERIMENTS ON A MULTILEVEL ARCHITECTURE FOR SPEECH VERIFICATION

In order to better evaluate the performance of this verification architecture. The differences in the voice collection quality

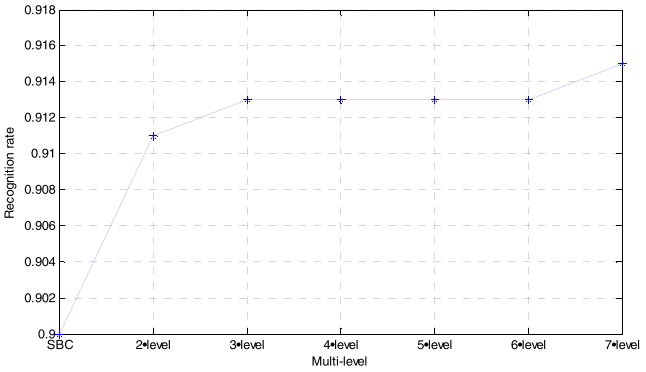

FIGURE 11. The results of the multilevel architecture for speech verification in the first group.

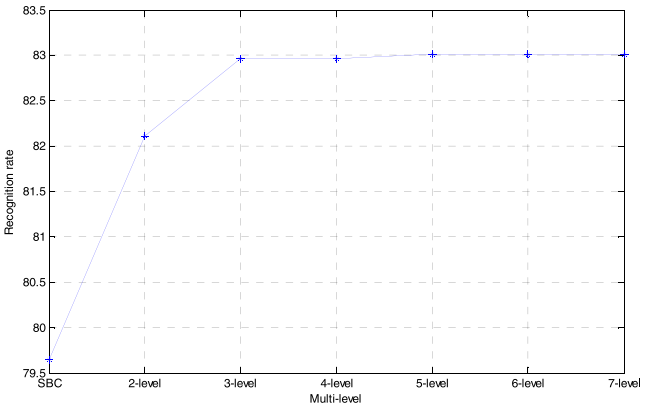

FIGURE 12. The results of the multilevel architecture for speech verification in the second group.

of different mobile handheld devices are taken into account in the study of this experiment. At the same time, it also makes a comprehensive consideration of multi-level verification architecture.

First of all, for the robust research, we selected three more representative brand mobile phone, including Iphone 5S, Samsung Note2 and Xiaomi mobile phone M4, and used the these mobile device to collect the same speech for the different devices. Moreover, in order to make the experimental data more persuasive, this speech acquisition is to comply with the same collection environment, to ensure the fair experiment, excluding other external factors in the experimental data collection.

As Figure 11 shown, the experimental results based on multi-level speech authentication architecture in real databases, the verification results of the first group of voice databases. From the overall recognition rate, the verification architecture has achieved a 1.5% improvement in the first set of database experiments without processing of noise; a speech verification system based on small training sets can achieve a recognition rate of more than 90%, which is already an expected result.

Figure 12 shows the results of the verification based on the second group of voice databases. It can be seen that the

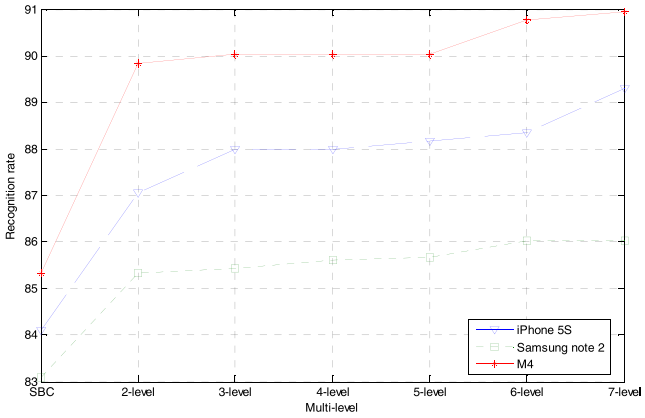

FIGURE 13. The results of the multilevel architecture for speech verification in the third group.

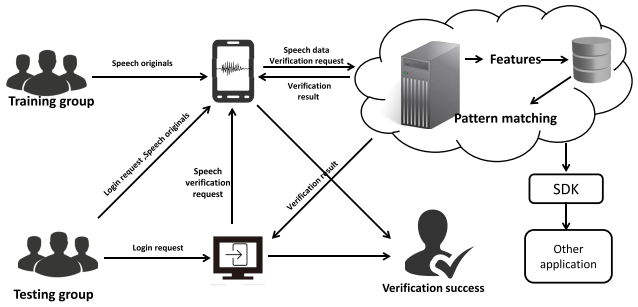

FIGURE 14. Architecture of third-party certification with speech verification.

overall recognition rate has been greatly improved and the improvement has exceeded 3%. From this, it can be seen that in a relatively quiet situation, multi-level voice authentication architecture will have better performance.

Figure 13 shows the results of speech authentication based on three mobile phone platforms. The third speech group is used here. First of all, it can be seen that regardless of which type of mobile device is based on, this architecture can achieve a certain degree of improvement, with an average increase of nearly 5 percentages. Especially on the Xiaomi M4 mobile phone, it has achieved an astonishing 6 percentage improvement.

At the same time, we can see from Figure 13 that a multilevel architecture can improve the recognition rate regardless of the type of mobile device used. According to this architecture, if we can obtain enough complementary feature parameters, the recognition rate will be close to unity.

VI. APPLICATION ARCHITECTURE

Today, many biometric technologies are used to make people's lives more convenient. There are also some real-life application scenarios for speech verification technology. In this section, a brief introduction is given to the actual application architecture of speech verification, as shown in Figure 14.

From Figure 14, it is clear that we can divide the verification system into two parts. The first is the training period. In this period, speech originals must be collected and transmitted to the server. The server is the most important part of this system because the process of speech verification is performed by the server. The second part is the verification. Users can log in to applications through mobile phones or computers, but the speech originals for verification must be sent through a mobile phone due to the inconvenience of collection from computers. If the result of the verification is success, the user will enter the application. For further promotion of speech verification technology, we can package the certification process on the server side into an SDK for other applications, and these applications can easily access speech verification.

VII. CONCLUSION

In this paper, we proposed a novel multilevel architecture for speech verification on mobile devices. For limited training data, the architecture was shown to have a better performance using our personal database than the method with single feature extraction. We can obtain improved performance from different five least three percentage points improvement compared with the SBC approach. This paper makes a large number of experiments directed to the difference on the sensitivity of verification between the different speech groups.

According to the results of the experiment, this topic make a optimization and improvement for the speech verification system based on the Android client and server side which is developed. It is possible that speech verification can achieve practical implementation on mobile phones with the proposed architecture.

Because of this paper, there are unique views on several aspects of speech verification system. Therefore, there are still some challenges, including the following two aspects:

(1) The noise in the speech signal is complicated because of the diversity of the acquisition equipment and the variability of the environment. This research does not address the denoising of speech signals, if we can effectively remove all kinds of unpredictable noise, the recognition rate can be improved to certain amount.

(2) The problem of speech feature extraction algorithm, this paper uses the wavelet based feature extraction algorithm, although it can extract the special components in the short speech signal as much as possible, but the wavelet computation is more complex than other feature extraction algorithms. Therefore, the time efficiency is low, based on our multilevel architecture, the more the number of wavelets will be extracted, the slower the time will be. Therefore, we need to find out another way to balance the contradiction between efficiency and recognition rate.

REFERENCES

[1] M. Sharif, A. Khalid, M. Mudassar, and S. Mohsin, "Face recognition using Gabor filters," *J. Appl. Comput. Sci. Math.*, vol. 5, no. 11, pp. 53–57, 2011.

[2] X.-M. Wang, C. Huang, and J.-G. Liu, "Gabor-2DLDA: Face recognition using Gabor features and 2D linear discriminant analysis," in *Proc. 2nd Int. Conf. Intell. Comput. Technol. Automat.*, Oct. 2009, pp. 608–610.

[3] Q.-Y. Zhao, B.-C. Pan, J.-J. Pan, and Y.-Y. Tang, "Facial expression recognition based on fusion of Gabor and LBP features," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, Aug. 2008, pp. 362–367.
[4] Z. Jinhai, "Fingerprint image enhancement based on Gabor function," in *Proc. Cross Strait Quad-Regional Radio Sci. Wireless Technol. Conf. (CSQRWC)*, Jul. 2011, pp. 1414–1417.
[5] M. Ekinci and M. Aykut, "Gabor-based kernel PCA for palmprint recognition," *Electron. Lett.*, vol. 43, no. 20, pp. 1077–1079, Sep. 2007.
[6] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Commun.*, vol. 99, pp. 62–79, May 2018.
[7] Z. Wang and S. Li, "Discrete Fourier transform and discrete wavelet packet transform in speech denoising," in *Proc. 5th Int. Congr. Image Signal Process. (CISP)*, Oct. 2012, pp. 1588–1591.
[8] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in *Proc. IEEE Region 10th Int. Conf. Comput., Commun. Automat. (TENCON)*, Oct. 1993, pp. 321–324.
[9] S.-J. Choi and J. W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 155–167, Feb. 1999.
[10] F. Chen and J. Zhu, "A new method of endpoint detection based on distance of auto-correlated similarity," *J. Shanghai Jiaotong Univ.*, vol. 33, no. 9, pp. 1097–1099, 1999.
[11] I. Abdallah, S. Montresor, and M. Baudry, "Robust speech/non-speech detection in adverse conditions using an entropy based estimator," in *Proc. 13th Int. Conf. Digit. Signal Process. (DSP)*, Jul. 1997, pp. 757–760.
[12] R. Sarikaya, B. L. Pellom, and J. H. Hansen, "Wavelet packet transform features with application to speaker identification," in *Proc. 3rd IEEE Nordic Signal Process. Symp.*, Jun. 1998, pp. 81–84.
[13] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Commun.*, vol. 17, nos. 1–2, pp. 91–108, 1995.
[14] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digit. Signal Process.*, vol. 10, nos. 1–3, pp. 19–41, 2000.
[15] C. Huo, Y. Shao, and X. Gao, "The improved VQ algorithm for speaker recognition," in *Proc. 4th Int. Conf. Innov. Comput., Inf. Control (ICICIC)*, Dec. 2009, pp. 997–1000.
[16] S. E. Kruger, M. Schaffoner, M. Katz, E. Andelic, and A. Wendemuth, "Mixture of support vector machines for HMM based speech recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 326–329.
[17] S. E. Eskimez, P. Soufleris, Z. Duan, and W. Heinzelman, "Front-end speech enhancement for commercial speaker verification systems," *Speech Commun.*, vol. 99, pp. 101–113, May 2018.
[18] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN hybrid method for short utterance speaker recognition," *IEEE Trans. Ind. Informat.*, vol. 14, no. 7, pp. 3244–3252, Jul. 2018.
[19] S. Srivastava, G. Chaudhary, and S. Bhardwaj, "Multi-scenario dataset for speaker recognition," *J. Intell. Fuzzy Syst.*, vol. 34, no. 3, pp. 1385–1392, 2018.

KUO-KUN TSENG was born in 1974. He is currently an Associate Professor and a Shenzhen Peacock B-level Talent. He received the Ph.D. degree in computer information and engineering from National Chiao Tung University, Taiwan, in 2006. He has many years of research and development experience, since 2004, and has been long engaged in biometric systems and algorithms research. His current research results are published in 65 articles of which about 20 is a high-impact factor of the SCI magazine or the famous ACM/IEEE series of journals.

YAN ZHANG received the master's degree from the Harbin University of Technology, Shenzhen. His expertise are speech analysis and biometrics.

K. L. YUNG received the B.Sc. degree in electronic engineering, in 1975, the M.Sc. degree in DIC in automatic control systems, in 1976, and the Ph.D. degree in microprocessor applications in process control at U.K., in 1985. He became a Chartered Engineer (MIEE) in 1981. After graduation, he was with companies at U.K., such as BOC Advanced Welding Co., Ltd., the British Ever Ready Group, and the Cranfield Unit for Precision Engineering. In 1986, he returned to Hong Kong to join the Hong Kong Productivity Council as a Consultant and subsequently switched to academia to join The Hong Kong Polytechnic University, where he is currently the Chair Professor of precision engineering with the Department of Industrial and Systems Engineering. He has authored over 200 journal papers and conference publications in his research related areas. His research interests include precision motion control and system aspects of computer integrated manufacturing and management, aerospace engineering, logistic planning and optimization, and computer vision. He has received many international awards.

W. H. IP received the M.Sc. degree from Cranfield University, in 1983, the M.B.A. degree from Brunel University, in 1989, and the Ph.D. degree from Loughborough University, in 1993. He is currently an Adjunct Professor of mechanical engineering with the University of Saskatchewan, and a Principal Research Fellow with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University. He is also an Industrial Fellow of the Warwick Manufacturing Group, University of Warwick. He has published over 300 papers with over 150 papers in SCI's journals. His current research interests include space systems, healthcare systems, smart city, logistics and supply chain management, wireless sensor networks, and evolutionary computation. He is a member of the Institution of Engineering and Technology, the Institution of Mechanical Engineers, and the Hong Kong Institution of Engineers. He serves as an Editor-in-Chief for the journal of *Enterprise Information Systems*, and the Founding Editor-in-Chief of the *International Journal of Engineering Business Management*.

ZHIYE OU is currently pursuing the master's degree with the Harbin University of Technology, Shenzhen. Her research interests include smart contracts and biometrics.

QI NA is currently pursuing the master's degree with the Harbin University of Technology, Shenzhen. Her research interests include watermarking and biometrics.