IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 27, 2019, accepted January 22, 2020, date of publication February 12, 2020, date of current version February 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2972511

Adaptive Speaker Recognition Based on Hidden Markov Model Parameter Optimization

YANGJIE WEI, (Member, IEEE)
College of Computer Science and Engineering, Northeastern University, Shenyang 110819, China
e-mail: weiyangjie@cse.neu.edu.cn

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0701200, in part by the Natural Science Foundation of China under Grant 61973059 and Grant 61532007, and in part by the Fundamental Research Funds for the Central Universities under Grant N161602002.

ABSTRACT The Hidden Markov Model (HMM) is a widely used method for speaker recognition. During its training, the composite order of the measurement probability matrix and the number of re-evaluations of the initial model affect the speed and accuracy of a recognition system. However, theoretical analysis and related quantitative methods are rarely used for adaptively acquiring them. In this paper, a quantitative method for adaptively selecting the optimal composite order and the optimal number of re-evaluations is proposed based on theoretical analysis and experimental results. First, the standard deviation (SD) is introduced to calculate the recognition rate considering its relationship with Mel frequency cepstrum coefficients (MFCCs) dimension, then the composite order is optimized according to its relationship curve with the SD. Second, the composited measurement probability with different number of re-evaluations is calculated and the number of re-evaluations is optimized when a convergence condition is satisfied. Experiments show that the recognition rate with the optimal composite order obtained in this paper is 97.02%, and the recognition rate with the optimal number of re-evaluations is 98.9%.

INDEX TERMS Speaker recognition, Gaussian composite order, re-evaluation, parameter optimization.

I. INTRODUCTION

Speaker recognition refers to identifying a speaker's identity using characteristic parameters extracted from the speaker's speech signal [1]. Compared to other biometric authentication methods, speaker recognition based on speech features has advantages such as convenience and economy [2]–[7].

The Hidden Markov Model (HMM) is a stochastic model based on transition probability and output probability [8]. It considers a speech signal as a random process consisting of an observable sequence of symbols. HMM does not require time regulation, which reduces the judgment time and storage. However, certain important initial parameters, including the composite order of the observed probability density matrix and the number of re-evaluations of the initial model, still require to be manually set by a user when training an HMM for speaker recognition. This not only reduces the adaptive ability of the speaker recognition system, but also affects the recognition accuracy. In order to address the initial parameter problem, state-merging and state-splitting were implemented in some HMM algorithms. The former iteratively merges states until convergence of a model, therefore, it requires large amount of computation. The latter begins with a general HMM and successively splits states until convergence. However, it is difficult to define the stopping mechanism of splitting. Although some factors, such as maximum likelihood or minimum description length, have been introduced to stop splitting, it is still difficult to acquire a balance between accuracy and generality of an HMM model in real applications. Therefore, this paper proposes a quantization method for adaptive acquisition of the Gaussian composite order and the number of re-evaluations through theoretical analysis and experimental verification to improve the accuracy and the training speed of an HMM-based speaker recognition system.

This paper is organized as follows: The basic principles of the HMM are introduced in the next section. In the third section, HMM-based speaker recognition system is detailed. The fourth section presents our adaptive acquisition of the Gaussian composite order and the number of re-evaluations through theoretical analysis and experimental verification, and we draw conclusions in the fifth section.

The associate editor coordinating the review of this manuscript and approving it for publication was Behnam Mohammadi-Ivatloo.

II. DEFINITION OF HMM

The problem solved by the HMM has two characteristics:

(1). State-based characteristics, which includes the hidden state and the observation state.

(2). Two types of data, which consists of the observation-state sequence and the hidden-state sequence.

First, suppose Θ is the collection of all possible hidden-states and $\boldsymbol{V}$ is the collection of all possible observed-states as follows:

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots\ldots, \boldsymbol{\theta_N}\}, \quad \boldsymbol{V} = \{\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, v_{\boldsymbol{M}}\} \tag{1}$$

where N and M are the number of possible hidden-states and observed-states, respectively.

For a sequence of length T, $\boldsymbol{Q}$ and $\boldsymbol{O}$ are the state sequence and the observation sequence, respectively, and they can be obtained by the following equation:

$$\boldsymbol{Q} = \{\boldsymbol{q_1}, \boldsymbol{q_2}, \cdots\cdots, \boldsymbol{q_T}\}, \quad \boldsymbol{O} = \{\boldsymbol{o_1}, \boldsymbol{o_2}, \ldots, \boldsymbol{o_T}\} \tag{2}$$

where $\boldsymbol{q_t} \in \Theta$ and $\boldsymbol{o_t} \in \boldsymbol{V}$.

There are two assumptions in HMM:

(1). The homogeneous Markov chain hypothesis. The Markov chain is described by $\boldsymbol{\Pi}$ and $\mathbf{A}$, which determine the shape of Markov chain. The hidden state at any time only depends on the previous hidden state. If the hidden state at time t is $\boldsymbol{q_t} = \boldsymbol{\theta_i}$ and the hidden state at time $t+1$ is $\boldsymbol{q_{t+1}} = \boldsymbol{\theta_j}$, then the HMM state transition probability $\boldsymbol{a_{ij}}$ from time t to time $t+1$ can be expressed as follows:

$$\boldsymbol{a_{ij}} = P\left(\boldsymbol{q}_{t+1} = \boldsymbol{\theta}_j \,|\boldsymbol{q}_t = \boldsymbol{\theta}_i\right) \tag{3}$$

Thus, $\boldsymbol{a_{ij}}$ composites the state transition matrix $\mathbf{A}$:

$$\boldsymbol{A} = \left(\boldsymbol{a_{ij}}\right)_{N\times N} \tag{4}$$

The hidden probability matrix $\boldsymbol{\Pi}$ at $t = 1$ is given by:

$$\boldsymbol{\Pi} = \left[\boldsymbol{\pi}\ (i)_N\right] \tag{5}$$

where $\pi(i) = P(\boldsymbol{q_1} = \boldsymbol{\theta_i})$.

(2). Independence of observing states. The observation state at any time only depends on the hidden state of the current moment. If the hidden state at time t is $\boldsymbol{q}_t = \boldsymbol{\theta}_j$ and the corresponding observation state is $\boldsymbol{o_t} = \boldsymbol{v_k}$, then the measurement probability $\boldsymbol{b_j}(k)$ of observation state $\boldsymbol{v_k}$, generated from hidden state $\boldsymbol{q_j}$, at that time satisfies the following equation:

$$\boldsymbol{b}_j\ (k) = P\left(\boldsymbol{o_t} = \boldsymbol{v_k}\,|\boldsymbol{q_t} = \boldsymbol{\theta}_j\right) \tag{6}$$

where $1 \leq j \leq N, 1 \leq k \leq M$, M denotes the number of observation states.

A Gaussian function is mostly used to describe $\boldsymbol{b_j}(k)$ according to the distance between $\boldsymbol{o}_t$ and $\boldsymbol{\theta}_j$:

$$\boldsymbol{b_j}\ (k) = \frac{1}{2\pi\sigma} e^{-\frac{(\mathbf{o_t}-\boldsymbol{\theta_j})^2}{2\sigma^2}} \tag{7}$$

Then, the measurement probability of observation states generated from hidden state $\boldsymbol{q}_j$ can be calculated with a linear summation of multiple Gaussian functions as shown by the following equation:

$$\boldsymbol{b_j} = \sum_{m=1}^{M} c_{jm} G\left(j, \mu_{jm}, \Sigma_{jm}\right) \tag{8}$$

where $G(\cdot)$ denotes the Gaussian function; c_{jm}, μ_{jm} and Σ_{jm} denote the weight, the mean and the variance of the m_{th} Gaussian function, respectively.

The measurement probability matrix $\mathbf{B}$, composed by $\boldsymbol{b_j}$, is then given by:

$$\boldsymbol{B} = \boldsymbol{b_j}\ (k)_{N\times M} \tag{9}$$

Finally, the HMM can be determined by $\boldsymbol{\Pi}$,$\mathbf{A}$, and $\mathbf{B}$ as follows:

$$\boldsymbol{\lambda} = (\boldsymbol{\Pi}, \mathbf{A}, \mathbf{B}) \tag{10}$$

From (10), we can see that the number of possible hidden-states and the observed-states determines the size of HMM.

III. HMM-BASED SPEAKER RECOGNITION SYSTEM

A. PREPROCESSING

The preprocessing of a speaker recognition system mainly includes: sample quantization, pre-emphasis, frame windowing, and endpoint detection.

Sampling converts an analog signal into a discrete analog signal, and quantization divides the continuous amplitude into several levels. As researched, the speech signal is attenuated at a rate of 6 dB/octave when the frequency is greater than 800 Hz. Therefore, sampling and quantization require a pre-emphasis process to raise the high frequency components and flatten the spectrum of the signal. The general pre-emphasis filter H is expressed as follows:

$$H\ (f) = 1 - \alpha f^{-1} \tag{11}$$

where f is the frequency; α is the coefficient of the pre-emphasis process and it is normally between 0.90 and 0.97.

For the short-time stationarity analysis, a speech signal must be framed and windowed. The commonly used windows are Hamming window and Hanning window. In addition, to remove noise and mute portions in a speech signal, it is necessary to remove some valid speech parts using endpoint detection.

B. SPEECH FEATURE EXTRACTION

One of common speech features is the Mel frequency cepstral coefficients (MFCCs), which combines the auditory perception characteristics of human ears on the mechanism of speech generation and uses the Mel filter bank to mimic functions of the human cochlea. The frequency scale is close to that of the human auditory characteristics [9]. The MFCCs can be calculated as follows:

(1). Convert the actual frequency to the Mel nonlinear frequency as follows:

$$Mel\ (f) = 2595 \times \log\left(1 + f\,/700\right) \tag{12}$$

(2). Triangular filters of L channels are arranged on the Mel frequency axis, and the number of L is determined by the cutoff frequency of the signal. The center frequency $e(l)$ of each triangular filter is allocated to equal intervals on the Mel frequency axis. $s(l)$, $e(l)$, and $h(l)$ are the lower frequency, the center frequency and the upper frequency of the l^{th} triangular filter, respectively.

(3). Determine the output of each triangle filter based on the amplitude spectrum $|X_n(r)|$ of the speech signal as follows:

$$u(l) = \sum_{r=s(l)}^{h(l)} W_l(r) |X_n(r)|, \quad l = 1, 2, \ldots, L \tag{13}$$

where,

$$W_l(r) = \begin{cases} \dfrac{r - s(l)}{e(l) - s(l)} s(l) \le r \le e(l) \\ \dfrac{r - s(l)}{h(l) - e(l)} e(l) \le r \le h(l) \end{cases} \tag{14}$$

(4). The logarithm operation is performed on all filter output, and discrete cosine transform (DCT) is performed to obtain the MFCCs:

$$MFCC(i) = \sqrt{\frac{2}{N}} \sum_{l=1}^{L} \log u(l) \cos\left[\left(l - \frac{1}{2}\right)\frac{i\pi}{L}\right] \tag{15}$$

where MFCC(i) is the MFCC of the i^{th} channel or dimension and normally it is considered to reflect the static properties of a signal. To present the dynamic properties of a signal, mostly Δ MFCC is introduced, which is obtained through calculating the first order difference of the MFCCs.

C. MODEL TRAINING AND IMPORTANT PARAMETERS IN SPEAKER RECOGNITION

Training parameters of an HMM is to estimate the optimal parameters of $\boldsymbol{\lambda}$ assuring $P(\mathbf{O}|\boldsymbol{\lambda})$ is maximal based on the observation state sequence $\mathbf{O}$. In fact, this is the most complicated problem solved by the HMM because it is difficult to obtain the optimal λ due to the limit size of a given data set in practice. Therefore, the Baum-Welch algorithm is adopted to locally maximize $P(\mathbf{O}|\boldsymbol{\lambda})$ and obtain the estimated model $\lambda = (\boldsymbol{\Pi}, \mathbf{A}, \mathbf{B})$ with the concept of iterations.

First, the parameters of the initial model $\boldsymbol{\lambda_0}$ must be defined before the re-evaluation using the Baum-Welch algorithm. As known, $\mathbf{B}$, compared to $\boldsymbol{\Pi}$ and $\mathbf{A}$ in $\boldsymbol{\lambda_0}$, is closely related to the training quality of an HMM. To calculate component $\boldsymbol{b_j}$ in matrix $\mathbf{B}$, it is generally to cluster all the MFCCs of a speech signal into M clusters and obtain the mean, the variance and the weight of a Gaussian function in each cluster. Then, the Gaussian functions of each hidden state are linearly summed together to obtain component $\boldsymbol{b_j}$ in matrix $\mathbf{B}$. Here, M has a significant influence on the recognition accuracy of a system because it is closely related to the distance between $\mathbf{B}$ and the true distribution of MFCC features. Some researchers tried to capture an appropriate M with a serial of practical experiments during setting parameters for the initial model, however, the result highly depends on the data sets and there is rare theorical and quantified foundation to calculate it. Therefore, it is necessary to research an adaptive acquisition method to obtain an appropriate value for M.

Second, the Baum-Welch algorithm is used to re-evaluate the initial model, and the revaluation formula is expressed as follows:

$$\bar{\boldsymbol{\pi}}_{\mathbf{i}} = \sum_{l=1}^{L} \alpha_t^{(l)}(i)\, \beta_l^{(l)}(i) \Big/ P(\mathbf{O}|\boldsymbol{\lambda}) \tag{16}$$

$$\bar{\boldsymbol{a}}_{ij} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T_l-1} \alpha_t^{(l)}(i) \boldsymbol{a}_{ij} b_j\left(\boldsymbol{o}_{t+1}^{(l)}\right) \beta_{t+1}^{(l)}(j) / P\left(\mathbf{O}^{(l)}|\boldsymbol{\lambda}\right)}{\sum_{l=1}^{L} \sum_{t=1}^{T_l-1} \alpha_t^{(l)}(i) \beta_t^{(l)}(j) / P\left(\mathbf{O}^{(l)}|\boldsymbol{\lambda}\right)} \tag{17}$$

$$\bar{\boldsymbol{b}}_{jk} = \frac{\sum_{l=1}^{L} \sum_{o=1, o=l_k}^{T_l} \alpha_t^{(l)}(i) \beta_t^{(l)}(j) / P\left(\mathbf{O}^{(l)}|\boldsymbol{\lambda}\right)}{\sum_{l=1}^{L} \sum_{t=1}^{T_l} \alpha_t^{(l)}(i) \beta_t^{(l)}(j) / P\left(\mathbf{O}^{(l)}|\boldsymbol{\lambda}\right)} \tag{18}$$

where $1 \le i \le N$, $1 \le k \le M$.

The re-evaluation terminates until $P(\mathbf{O}|\boldsymbol{\lambda_e})$ with estimated model λ_e converges. Therefore, more algorithms must be enrolled, such as expectation maximization (EM), to calculate $P(\mathbf{O}|\boldsymbol{\lambda_e})$ of evaluation step i and judge whether $P(\mathbf{O}|\lambda_i)$ converges. In theory, only when the number of re-evaluations is infinite, $P(\mathbf{O}|\lambda)$ can reach the optimal convergence value. This is not feasible in practice; therefore, the number of re-evaluations is generally set according to experimental results when different data sets are used. However, if the number of re-evaluations is too small, the model will deviate too much far from the ideal value. Too many re-evaluations will increase the complexity and the time burden of the algorithm. Therefore, setting the number of re-evaluations of an initial HMM is an important research topic.

IV. SPEAKER RECOGNITION BASED ON IMPROVED HMM

A. ADAPTIVE ACQUISITION FOR COMPOSITE ORDER

In this section, a mathematical relationship between the composite order and recognition accuracy based on HMM is established based on experiments and theoretical analysis. Then, an adaptive acquisition method for the composite order is proposed.

First, we select a total of 168 speakers based on the TIMIT dataset and test the relationship between the speaker recognition accuracy and the composite order. The pre-processing and MFCCs extraction feature methods used in the training and identification phases are the same as Reference [9]. Our experiment is conducted on a 3.40 GHz machine with 8GB random access memory (RAM) using Matlab implementation. The specification of our experiment is shown in Table 1. In our experiment, the number of hidden states is set to one and the composite order is M. The speaker recognition accuracy of the system based on HMM is calculated by increasing M, and the result is shown in Fig.1, where we can see that when the composite order M is gradually increased from 2 to 32, the recognition rate of the system rapidly increases

TABLE 1. The specification of our experiment.

Parameter	Values
Sampling frequency	16 kHz
Endpoint detection	Double threshold method
Pre-emphasis coefficient	0.9375
Frame length	1024 points
Frame is shifted	512 points
Windowing function	Hamming window
Feature extraction	12 dimensions MFCC+ΔMFCC

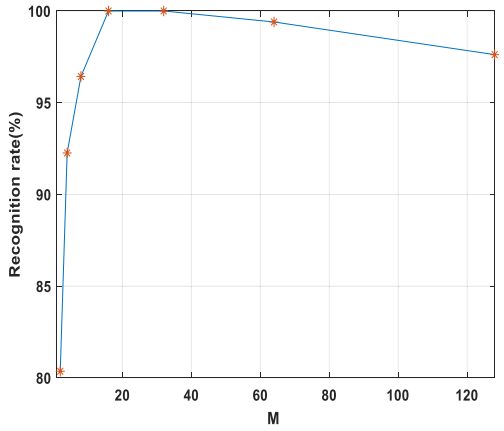


FIGURE 1. Speaker recognition rates of the system with different *M*.

from 80.36% to 100%; however, as M continues to increase to 128, the recognition rate gradually decreases to 97%.

The experimental result is consistent with the theoretical analysis, because when the composite order is small, it means that the measurement probability of each hidden state is calculated by the linear summation of a small number of Gaussian functions. Although the total number of variables in these Gaussian functions is small, the gap between the summated measurement probability and the true distribution of features is also large, resulting in low recognition accuracy. As the number of Gaussian functions is increased, the gap between them is gradually reduced, and the recognition rate of the recognition system is also improved. On the other hand, when the number of Gaussian functions is too large, the number of variables required for parameter identification of Gaussian functions is also significantly increased. As the size of the training data-set is not increased, the limited training data-set limits the accuracy of parameter identification of Gaussian functions, resulting in a gradual increase in the gap between the measurement probability and the true feature distribution. Thus, the recognition rate then begins to decline. Therefore, in practical applications, the magnitude of M affects the recognition rate of the system, the training complexity, and the computation time. The recognition rate is also limited by the size of the data-set. Theoretically, when training an HMM, the composite order represents the number of clusters based on the MFCCs of speeches and the Gaussian functions describe the feature distribution within each cluster of MFCCs. Therefore, M is related to the clustering quality of MFCCs. In clustering evaluation, the standard deviation (SD) is often used as a quality evaluation factor [10-11], therefore, in this paper, SD is introduced to evaluate the appropriate M.

The basic principle of application of SD is based on the concept of average scattering and total separation for clusters. The SD can be obtained by the following equation,

$$SD(M) = aScat(M) + Dis(M) \tag{19}$$

where $Scat$ denotes the average scattering for clusters to evaluate the compactness and Dis denotes the distance among all cluster centers to evaluate the separation. They are defined as follows,

$$Scat(M) = \frac{1}{M}\sum_{i=1}^{M} \|\sigma(d_i)\| \Big/ \|\sigma(\mathbf{X})\| \tag{20}$$

$$Dis(M) = \frac{D_{\max}}{D_{\min}} \sum_{c=1}^{M} \left(\sum_{z=1}^{M} \|d_c - d_z\| \right)^{-1} \tag{21}$$

where $D_{\max} = \max(||d_i - d_j||)\ \forall i, j \epsilon \{1, 2, \ldots, M\}$ is the maximum distance between cluster centers. $D_{\min} = \min(||d_i - d_j||)\ \forall i,\ j \epsilon \{1, 2, \ldots, M\}$ is the minimum distance between cluster centers and a is a weighting factor equal to $Dis(M_{\max})$, where $M_{\max}$ is the maximum number of input clusters.

The variance of the p^{th} dimension in a data set $\boldsymbol{X}$ is defined as follows,

$$\sigma_x^p = \frac{1}{n}\sum_{c=1}^{n} \left(x_c^p - \bar{x}^p\right)^2 \tag{22}$$

where $\bar{x}^p$ is the p^{th} dimension of

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{c=1}^{n} \mathbf{x}_c, \quad \forall \mathbf{x}_c \in \mathbf{X}.$$

The variance of cluster i is called $\sigma(d_i)$ and its p_{th} dimension is given as follows,

$$\sigma_{d_i}^p = \sum_{c=1}^{n_i} \left(x_c^p - d_i^p\right)^2 \Big/ n_i \tag{23}$$

We calculate the SD relationship of different dimension of MFCCs, different M and SD. The result is shown in Fig. 2, where we can observe that for each dimension of the MFCCs, as M increases, the SD value has two peaks, one at $M = 10$ and the other $M = 32$ and has a distinct trough near $M = 20$. When M is greater than 40, the SD value decreases. When the M value is fixed, the SD value corresponding to the low dimensional MFCCs is comparatively smaller, and the high dimension is reversed. That is, with a fixed M, the SD value increases as the dimension number of MFCCs increases. Here, the distinction position between high and low dimensions is probably around the 9^{th} dimension.

SD is theoretically minimal with an infinite M; however, for a fixed-size dataset, selecting an M value between the two peaks of SD shown in Fig. 2 is a practical and efficient solution. To select the most distinctive dimension among the 24 dimensions of MFCCs, we divide the speech signal of a speaker into 24 frames. Then we calculate their corresponding MFCCs and intercept multiple sections of different

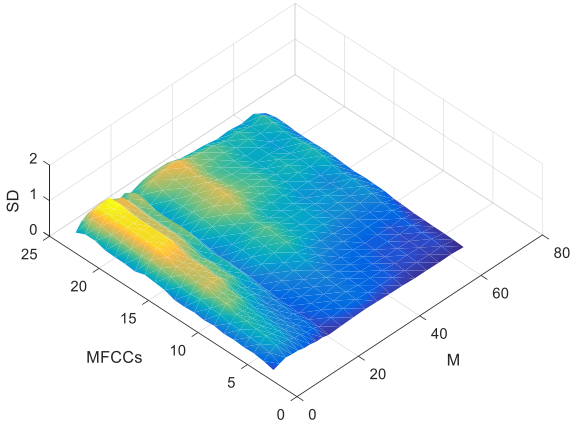


FIGURE 2. SD value with different dimension of MFCCs and different *M*.

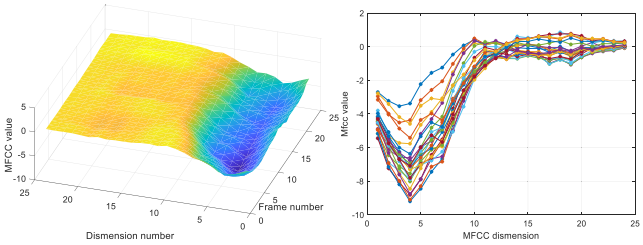


FIGURE 3. MFCC value of different dimension MFCCs in different frames and it's cross-section along the axis of dimension.

speech frames along the axis of dimension. The calculation result is shown in Fig. 3, where Fig.3 (a) is the MFCCs in different frames and Fig.3 (b) is the cross-section along the axis of dimension. It can be seen from Fig. 3 that the MFCC value of all speech frames has a significant step edge near the 9th dimension. Specifically, the MFCC value clearly increases at this dimension. Although the MFCC values of different speech frames are not equal in magnitude, their variation curve along the dimension axis is the same. Therefore, we can easily locate the step edge position of the MFCC value using the cross-sectional view. Then, this dimension of MFCCs is selected to calculate the SD value. In this paper, we select the 9th dimension MFCCs.

To observe the relationship between SD and M with one-dimensional MFCCs, we use the 4th, 9th, 18th and 23rd dimension MFCCs of nine speakers to calculate the SD value as M varied. The result is shown in Fig. 4. As seen in Fig. 4, where Fig. 4(a)-(d) is the experimental result of the 4th, 9th, 18th and 23rd dimension MFCCs. From Fig. 4, we can see that when using the 4th and 9th dimension MFCCs, the minimum SD value mostly appears at $M = 16$ or $M = 32$. In comparison, the SD value calculated using the 9th dimension MFCCs changes more gradually with M. The M-SD curves calculated using the 18th and 23rd MFCCs are too complicated to find the position where the minimum SD value appears. Therefore, if the MFCCs of the 9th dimension is used as the evaluation factor of the clustering quality, the M value with the smallest SD value could be selected as the optimal composite order of

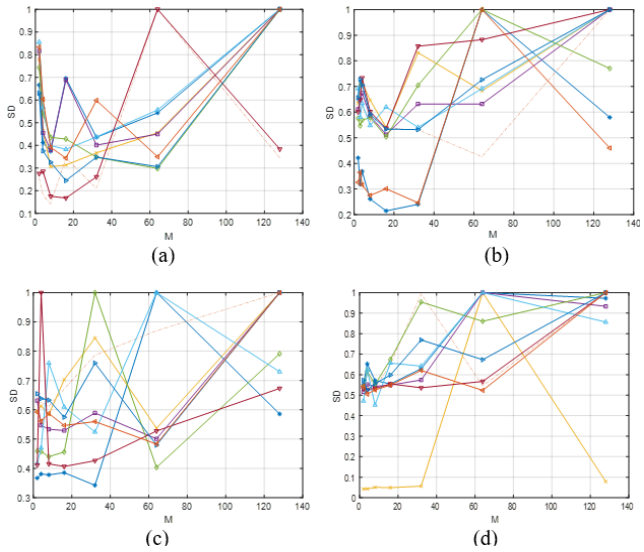


FIGURE 4. SD value variation with different *M* and different dimension MFCCs.

TABLE 2. Recognition rate with different dimension MFCCs.

Dimension of MFCCs	4	9	18	23
Recognition rate with Adaptive M (%)	95.24	97.02	95.83	90.48

the system, and this result is consistent with the case where the recognition rate is the highest in Fig. 1.

Therefore, we propose an adaptive acquisition method for the composite order based on HMM in a speaker recognition system. First, the MFCCs is selected with a step edge, the SD value is then calculated, and the value of Gaussian composite number, M, is selected, when the SD value is the smallest, as the optimal composite order of the speaker system. Then, the recognition rate of the speaker system is calculated. For comparison, we also calculate the M value that is adaptively selected using the 4th, 18th, and 23rd MFCCs, and apply them for modeling and identification. The experimental result is shown in Table 2, where we can see that when the M value corresponding to the minimum SD value is chosen as the system composite order with the 9th dimension MFCCs, the speaker system has the highest recognition rate. The number of correctly identified speakers is 163, and the recognition rate is 97.02%. The recognition rate using the 4th and 18th dimension MFCCs is slightly lower at approximately 95%, and the difference between them is small. The 23rd dimension MFCCs yield the lowest recognition rate, however, the recognition rate is still above 90%. Therefore, we select the 9th dimension MFCCs as the representative feature for calculating the SD value and the adaptive M value.

B. RE-EVALUATION ADAPTIVE OPTIMIZATION

In this section, the optimal number of re-evaluations of the initial model in HMM training is discussed. First, suppose the number of observation states in the HMM model is M =3 and

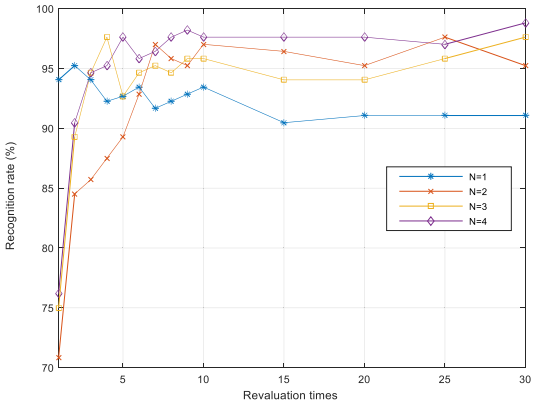


FIGURE 5. Relationship between the number of re-evaluations and speaker recognition rate.

the number of hidden states is N. The initial model uses the segmentation k-means algorithm [12].

In this study, we calculated the system identification for different number of re-evaluations of the initial model with respect to the number of hidden states N =1, 2, 3, and 4, respectively. The results are shown in Fig. 5, where the horizontal axis represents the number of re-evaluations, and the vertical axis represents the recognition rate of the system.

From Fig. 5, the following conclusions can be made:

(1). When the number of hidden states is one, as the number of revaluation initial models increases, the recognition rate of the system decreases.

(2). When $N = 2$ and $N = 3$, the two curves vary in a similar pattern: As the number of re-evaluations increases, the recognition rate of the system gradually increases from approximately 70%, and finally stabilizes between 95% and 97%.

(3). When the number of hidden states is $N = 4$, the recognition rate of the system is rapidly increased from 76% to 97% when the number of re-evaluations increases from one to five. After that, the recognition rate stabilized at approximately 98%.

In a word, compared to the number of re-evaluations, the number of hidden states does not influence the recognition rate very much, especially when N is greater than 2. The recognition rate when N is equal to 2 to 4 is becoming stable with increasing of the number of re-evaluations. Theoretically, the recognition rate could be improved if we increase the number of re-evaluations, however, it will result in a large computational load. Therefore, an adaptive method to obtain an appropriate number of re-evaluations is proposed as follows.

In theory, stabilization of the recognition rate indicates that the training model has converged. To quantify the relationship between the system recognition rate and the number of re-evaluations, we set $N = 4$ and calculate the measurement probability $\boldsymbol{b_j}$ for different number of re-evaluations. Consider $\boldsymbol{b_1}$ as an example. The calculation results are shown in Fig. 6, where $\boldsymbol{b_1}$ obtained by the linear summation of M Gaussian functions is nearly stable after ten re-evaluations.

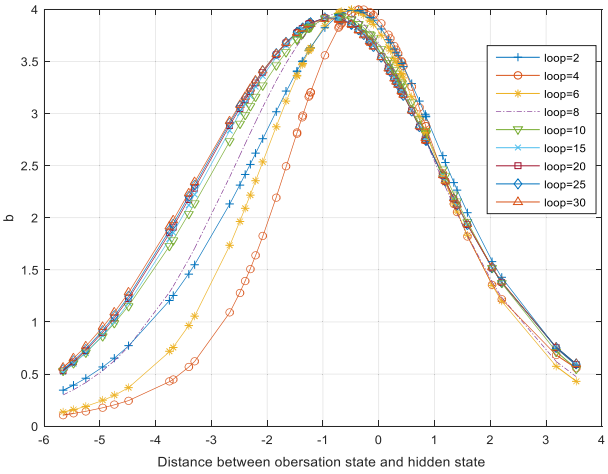


FIGURE 6. Measurement probability under different number of re-evaluations.

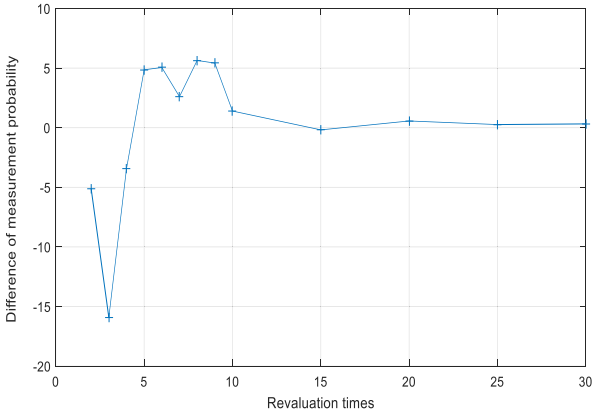


FIGURE 7. Difference of b_1 between two consecutive numbers of re-evaluations.

We can draw the conclusion that the measurement probability $\boldsymbol{b_j}$ is consistent with the law that the system recognition rate varies with the number of re-evaluations. In this paper, we adaptively optimize the number of re-evaluations based on the variation of the curve of the measurement probability as a quantization criterion.

Theoretically, the optimal number of re-evaluations is obtained when the curve of the measurement probability becomes smooth with increasing of the number of re-evaluations. Therefore, the optimal number of re-evaluations is calculated with the following equation,

$$NRE_{op} = \underset{1\le r\le\infty}{\operatorname{argmin}} \left| \sum_{i=1}^{N} \boldsymbol{\pi}_i^r \boldsymbol{b}_i^r - \sum_{i=1}^{N} \boldsymbol{\pi}_i^{r1} \boldsymbol{b}_i^{r-1} \right| \tag{24}$$

where NRE_{op} is the optimal number of re-evaluations.

In practices, it is not necessary to increase r to an infinite value due to the result in Fig. 7. In this paper we choose a threshold ε to evaluate the optimal r with the following equation.

$$\Delta \boldsymbol{b}\,(r) = \left| \sum_{i=1}^{N} \boldsymbol{\pi}_i^r \boldsymbol{b}_i^r - \sum_{i=1}^{N} \boldsymbol{\pi}_i^{r1} \boldsymbol{b}_i^{r-1} \right| \tag{25}$$

TABLE 3. Number of re-evaluation and its occurrence number in the adaptive acquisition process with 168 trained speakers.

Number of re-evaluations	3	4	5	6	7	8	>=9
Number of recognized speakers	37	42	32	24	15	9	9

where Δb is the distance between two consecutive re-evaluations.

When Δb is less than ε, the training model is approaching to converge, and the number of re-evaluations is approximately equal to the optimal value. Therefore,

$$NRE_{op} \approx r, \quad if \ \Delta \boldsymbol{b}(r) \leq \varepsilon \tag{26}$$

To verify the proposed acquisition method of adaptive number of re-evaluations, we calculated the number of re-evaluations with speech data in the database. First, let the number of observation states be three, and the number of hidden states be four. Then, we calculate the number of re-evaluations and the number of speakers adaptively re-evaluating their models for corresponding number when the consecutive measurement probability difference is minimum. The experimental result is shown in Table 3, where 135 ($37 + 42 + 32 + 24 = 135$) of 168 speakers have adaptively set the number of re-evaluations to three to six. After adaptively setting the number of re-evaluations, the number of speakers correctly identified by the system reaches 166, and the recognition rate reaches 98.9%.

V. CONCLUSION

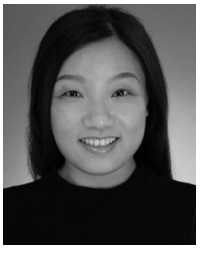
In this paper, a quantitative method for adaptively selecting the optimal composite order and the number of re- evaluations is proposed based on the detailed theoretical and experimental analysis of the HMM. For training of the HMM-based speaker recognition system, the number of observation states is closely related to the recognition rate, which depends on the user's practical experience. Herein, the clustering evaluation factor SD is introduced, and the relationship between the SD value, the MFCC dimension, and the system recognition rate is compared. Then, an acquisition method for the optimal composite order based on single-dimension MFCC feature is proposed. Given that the number of re-evaluations of the initial model directly affect the training speed and recognition accuracy of a speaker recognition system, this paper compares the impact of the number of re-evaluations on the system recognition rate by varying the number of hidden states. According to the theoretical analysis, the mathematical relationship between the measurement probability and the number of re-evaluations, as well as number of hidden states, is established, and an adaptive acquisition method for the number of re-evaluations of the initial model is proposed. Finally, a series of text-independent speaker recognition experiments are performed.

The results from these experiments show that the recognition rate with the optimal composite order obtained in this paper is 97.02%, and the recognition rate with the optimal number of re-evaluations is 98.9%.

In this paper, in order to validate our proposed method, series of text-independent speaker recognition experiments with English speeches are conducted. However, the characteristics of different language speeches may be different, even though they are from the same speaker. Therefore, different language speeches should be researched in the future to improve the robustness of our method. The other possible research direction is the influence of noises in a practical environment, especially there are different noises existing in the training data set and the recognition data set.

REFERENCES

[1] R. Chakroun, L. Beltaïfa, and M. Frikha, "An improved approach for text-independent speaker recognition," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, pp. 343–348, 2016.
[2] H. Aronowitz, "Speaker recognition using matched filters," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5555–5559.
[3] L. Chen, K. A. Lee, B. Ma, W. Guo, H. Li, and L.-R. Dai, "Exploration of local variability in text-independent speaker verification," *J. Signal Process. Syst.*, vol. 82, no. 2, pp. 217–228, Feb. 2016.
[4] K. Yu, "Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation," *IEE Proc.-Vis., Image Signal Process.*, vol. 142, no. 5, p. 313, 1995.
[5] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 2, Apr. 1992, pp. 157–160.
[6] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Comput. Speech Lang.*, vol. 20, no. 2, pp. 210–229, 2006.
[7] M. M. Hossain, B. Ahmed, and M. Asrafi, "A real time speaker identification using artificial neural network," in *Proc. 10th Int. Conf. Comput. Inf. Technol.*, Dec. 2007, pp. 1–5.
[8] J. Ashraf, N. Iqbal, and N. S. Khattak, "Speaker independent Urdu speech recognition using HMM," in *Proc. Int. Conf. Appl. Natural Lang. Inf. Syst.* Berlin, Germany: Springer-Verlag, 2010, pp. 140–148.
[9] V. Tiwari, "MFCC and its applications in speaker recognition," *Int. J. Emerg. Technol.*, vol. 1, no. 1, pp. 19–22, 2010.
[10] Y. Li, Z. Li, and H. Xiong, "Understanding of internal clustering validation measures," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2010, pp. 911–916.
[11] M. Halkidi, M. Vazirgiannis, and Y. Batistakis, "Quality scheme assessment in the clustering process," in *Proc. Eur. Conf. Princ. Data Mining Knowl. Discovery* in Lecture Notes in Computer Science, 2000, vol. 1910, no. 1, pp. 265–276.
[12] L. R. Rabiner, J. G. Wilpon, and B. H. Juang, "A segmental k-means training procedure for connected word recognition," *Bell Labs Tech. J.*, vol. 65, no. 3, pp. 21–31, 1986.



YANGJIE WEI (Member, IEEE) received the B.S. degree in electronics information engineering from the University of Jilin, Jilin, China, in 2002, and the M.S. and Ph.D. degrees in recognition and intelligent system from the Shenyang Institute of Automation, Chinese Academy of Sciences, in 2005 and 2013, respectively.

From 2010 to 2012, she was supported by the Joint Doctoral Program of Fraunhofer and CAS, and studying in Fraunhofer IZFP. She currently serves as a Professor with Northeastern University, China. Her current research interests are image processing and signal processing. Prof. Wei received the Best Student Paper Award from the ICMA 2009 in 2009, and the Award of Excellent Doctoral Dissertation of Chinese Academy of Science in 2014.