Speaker Recognition Based on the Improved Double-Threshold Endpoint Algorithm and Multistage Vector Quantization

Jun Zhu, Jingjing Zhang, Qiang Chen, Peipei Tu
Key Lab of Intelligent Computing and Signal Processing, Ministry of Education, China
School of Electronics and Information Engineering, Anhui University, Hefei, China
e-mail: junzhu@ahu.edu.cn; zhangjingjingahu@outlook.com

***Abstract*—The traditional double-threshold endpoint detection method has the phenomenon of missing detection. Therefore, the speech recognition (SR) system based on vector quantization (VQ) in this paper proposes an improved algorithm for this phenomenon, which effectively avoids the problem of missing detection. Then, Mel Frequency Cepstral Coefficients (MFCC) is used to extract the characteristic parameters of the speech signal, and the multistage vector quantization is used to quantify the characteristic parameters. Experimental results show that, the proposed algorithm improves the recognition rate of the text-independent speaker recognition system by 8.7%, and it also confirms that the longer the training speech is, the higher the recognition rate will be.**

Keywords-the multistage vector quantization; MFCC; double-threshold endpoint detection method; missing detection; characteristic parameters

I. INTRODUCTION

Speaker recognition is to distinguish the speaker in accordance with a speaker's voice, which mainly refers to extracting one or several characteristics of parameters from the tester voice, so as to identify and verify authentication technology [1]. In order to achieve a higher recognition rate, most speaker recognition systems make improvements in recognition methods and pattern matching methods [2], such as Dynamic Time Waring (DTW) [3], Artificial Neural Network (ANN) [4], Vector Quantization (VQ), and various model recognition methods, which have made some achievements [5].

Vector Quantization (VQ), applied to speech recognition, includes two parts [6], [7]: the training generated codebook and matching recognition speaker. In the matching recognition, the traditional vector quantization algorithm needs to calculate the distortion measure about speech characteristic of the speaker to be identified and each codebook [8], but when the speaker recognition is of larger scale, the calculation of distortion measure is too much, which greatly affects the timeliness of the speaker identification. In [9], the codebook classification and reassignment vector quantization algorithm are proposed to reduce the computational complexity in quantization, however, this algorithm still uses the LBG (Linde-Buzo-Gray) to classify the codebook with large size, which is more complex. Literature [10] combines speech enhancement with traditional dual-threshold speech detection, to improve the recognition rate of speech endpoint detection under low SNR. The number of adjustment thresholds, smoothing filtering, and the minimum length of speech ending are introduced to improve the accuracy of speech endpoint detection in [11].

Aimed at the improved algorithm mentioned above, this paper proposes a speaker recognition method based on improved endpoint detection method and multistage vector quantization. In the preprocessing stage of speech signal, the improved endpoint detection method improves the veracity of speech endpoint detection, and adopts multistage vector quantization to extract the characteristic parameters obtained by MFCC, thus effectively avoiding the problem of missing detection and improving the recognition rate.

II. VECTOR QUANTIZATION RECOGNITION METHOD

A. Principle and Process of VQ

Vector quantization is based on shannon's rate-distortion theory [12]. The theory suggests that, for a given distortion D, the rate distortion function R(D) can be calculated. Conversely, for a given quantization rate, the inverse function D(R) of the rate distortion function can be calculated. Where, D is constant, the required minimum rate is satisfied, R is the number of bits to quantize the signal per unit dimension, and D is the weakest distortion that can be obtained when R is constant.

For the quantization of the semaphore, according to some rules into groups, the formation of a total of N eigenvectors can be described as:

$$X = \{X_1, X_2, \cdots, X_N\} \tag{1}$$

where X is the trained sequence of the characteristics, $X_i=\{x_1,x_2,\cdots,x_k\}, i=1,2,\cdots,N$, k is the number of eigenvectors.

Each region in Euclidean space (K dimension) must satisfy the following formula:

$$\begin{cases} \bigcup_{j=1}^{M} S_j = S^k, 1 \le j \le M \\ S_i \bigcap S_j = \varnothing, 1 \le i \le M \ \& \ i \ne j \end{cases} \tag{2}$$

Each region is represented by a representative vector Y_j, that is:

$$Y = \{Y_1, Y_2, \cdots, Y_M\} \tag{3}$$

Vector quantization includes encoding and decoding:

The encoding process: firstly, it need to find the codeword Y_j corresponding to the input vector in the codebook (I), then output the subscript j of Y_j. This process is denoted as J=C(X), where J is subscript label of the output.

The decoding process: according to the subscript J, it is necessary to find the codeword corresponding to the index in the decoding codebook (II), and then use Y_J=D(J) to represent this process and output it. Where, codebook (I) and (II) is completely the same, the capacity of the codebook is represented by the letter M, the capacity is calculated as B=$\log_2$M in bit units, and quantized by the number of bits, which is written as R=B/N=($\log_2$M)/N, R determines the transmission rate of the quantizer. The distortion between X and Y_j is expressed by d(X, Y_j). The purpose of VQ is mainly to minimize the average distortion when R is constant. That is D=min(E[d(X,Y_j)]), where E() denotes the statistical average function, and D is used to find the optimal codeword of the vector quantizer in Fig. 1.

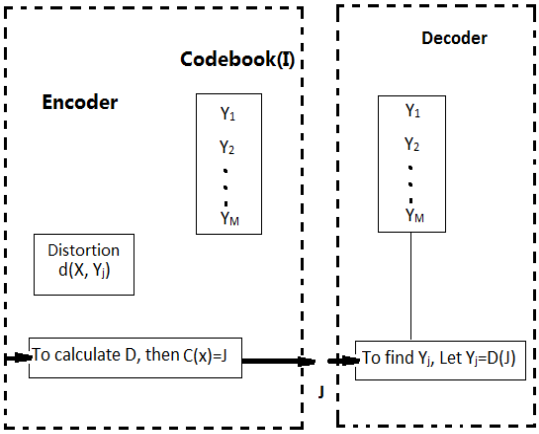

Figure 1. The schematic of vector quantization

In order to ensure that, the divided M quantized regions minimize the mean distortion D and find the optimal codebook, design should follow the following two principles:

1) Nearest Neighbor Rule (NNR), that is:

$$C(X)=J \Leftrightarrow d(X,Y_j)=\min(d(X,Y_j)) \qquad (4)$$

If the distortion between the input vector X and the code word Y_j is less than the distortion of the others, the vector X belongs to the representative region S_j of the codeword, which is also called the cell.

2) Centriod Rull (CR):

After the input vector is partitioned into the cell, the codeword Y_j of S_j should satify all the vectors X and Y_j in this cell to have the smallest average distortion. If distortion measure between X and Y_j is equal to the Euclidean distance, Y_j is the centroid of all the vectors in the cell S_j, which can be written as:

$$Y_j=\frac{1}{n}\sum_{X\in S_j} X \qquad (5)$$

where n is the number of all input vector in the cell S_j.

B. Distortion Measure of Vector Quantization

The distortion measure is the cost of the input vector X_i represented by the codebook vector Y_j, and can denote the degree of similarity between the test vector and each codebook, where the distortion measure is obtained by Euclidean distance.

Let the codeword Y and the measured vector X be K-dimension, and let the components of the same dimension correspond to X and Y be expressed as x_i,y_i ($0\le i\le K$-1), the mean-squared deviation is the Euclidean distance, which can be written as:

$$d_2(X,Y)=\frac{1}{K}\sum_{i=1}^{K}(x_i-y_i)^2=\frac{(X-Y)^T(X-Y)}{K} \qquad (6)$$

where the subscript 2 denotes the square error.

C. Linde-Buzo-Gray (LBG) Algorithm

In order to minimize the distortion measure and seek the optimal codebook, the LBG algorithm shown in Table I solves the iterative process of these two problems. The first step is to choose an initial codebook, and then iterate it. The above process needs a lot of training sequence. According to NNR and CR, the optimal can be obtained finally.

TABLE I. LINDE-BUZO-GRAY (LBG)

The algorithm of LBG:
Step1: all eigenvalues X needed to form a vector-quantized codebook are denoted by the set S.
Step2: to set the size of codebook named M
Step3: to set the maximum number of iterations, which is L=15.
Step4: to set the distortion threshold written as δ =0.005
Step5: the initial value of M codewords are $Y_1^{(0)}$, $Y_2^{(0)}$, $Y_3^{(0)}$, …, $Y_M^{(0)}$
Step6: the initial value of the distortion is $D^{(0)}=\infty$
Step7: the initial iteration value is m=1
Step8: the set S is divided into M subsets, which are S_1(m),S_2(m),…,S_M(m) respectively. According to NNR, if the condition of X$\in$S,j=1,2,…,M is satisfied, then d(X,Y_j(m-1))$\leqslant$d(X,Y_i(m-1)), $\forall$ i,i$\neq$j.
Step9: the distortion value $D^{(m)}$ is calculated: $D(m)=\sum_{j=1}^{M}\sum_{x\in S_j^{(m)}} d(X,Y_j^{(m-1)})$
Step10: the relative value $\delta^{(m)}$ of the distortion improvement amount $\triangle D^{(m)}$ is written as: $\delta(m)=\frac{\Delta D^{(m)}}{D^{(m)}}=\frac{\left\lvert D^{(m)}-D^{(m-1)}\right\rvert}{D^{(m)}}$
Step11: to recalculate the new codewords of $Y_1^{(m)}$, $Y_2^{(m)}$, $Y_3^{(m)}$, …, $Y_M^{(m)}$ $Y_j^{(m)}=\frac{1}{N_j}\sum_{X\in S_j^{(m)}} X$,j=1,2, …,M N_j is the number of the jth sample in $S_j^{(m)}$.
Step12: if $\delta^{(m)}<\delta$, then to switch to Step14, otherwise, to carry out Step13.
Step13: if m<L, then m=m+1,and to carry out Step8, otherwise, to continue Step14.
Step14: the end of iteration, and to output $Y_1^{(m)}$, $Y_2^{(m)}$, $Y_3^{(m)}$, …, $Y_M^{(m)}$.

D. *Multistage Vector Quantization*

Because the codebook is small, the average error of first-order quantization is larger, the quantization error can be quantified again by using the multistage quantization as the input of the next stage, and the quantization error can be reduced with the increase of the series. As shown in Fig. 2, here is the secondary vector quantization, which first adopts the codebook in level 1, and use it to approximate the input feature vector, and then the second-level codebook is used to encode the quantization error in level 1.

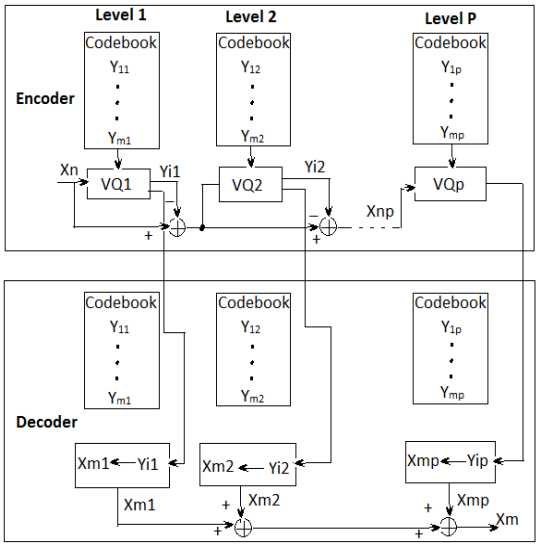

Figure 2. Multistage vector quantizer

III. MEL FREQUENCY CEPSTRUM COEFFICIENT

The analysis of MFCC is different from general cepstrum analysis, since it focuses on the human ear auditory characteristics, and feature extraction is in accordance with tone features. That is, the use of Mel frequency scale to bend the frequency axis, then the frequency cepstrum[13] of Mel will be obtained in Fig. 3.

In keeping with the human auditory characteristics of the frequency unit is generally Mel, but according to physics, the frequency unit is Hz, and the corresponding relationship of both can be described as:

$$fmel = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \tag{7}$$

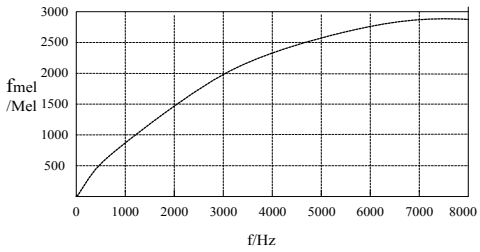

Figure 3. The corresponding relationship between Mel frequency and the actual frequency

In addition, in order to illustrate the masking effect of the noise signal on the pure tone signal, the concept of critical bandwidth is introduced. If the noise power in the critical bandwidth is larger than that of the pure tone, the pure tone is obscured by the noise.

From the theoretical basis, we can see that when the center frequency is less than 1000Hz, the critical bandwidth is almost unchanged, about 100Hz. When the center frequency is more than 1000Hz, the critical bandwidth increases with the z-index of the center frequency, as shown in Fig. 4.

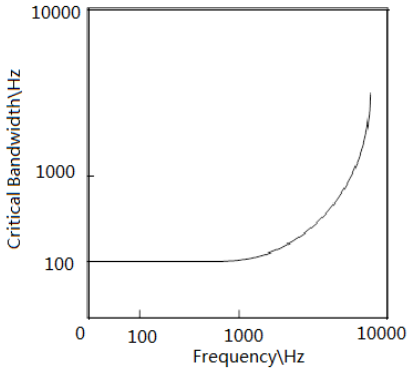

Figure 4. Critical bandwidth diagram

MFCC [14] performs the feature extraction process as follows:

1) The voice time-domain signal x(n) after preprocessing and endpoint detection should be carried out Fast Fourier Transform (FFT) as follows:

$$X(K) = \sum_{n=0}^{N-1} x(n) e^{-j2nk\pi/N}, 0 \le n \le N-1 \tag{8}$$

In the above formula, N is the number of FFT, whose value is set as 256.

2) Mel frequency is obtained by x(k) through the triangle filter, and then the logarithmic spectrum wil be received through the logarithmic transformation, the expression is:

$$S(m) = \ln\left(\sum_{k=0}^{N-1} \left|X(K)\right|^2 H_m(k)\right), 0 \le m \le M \tag{9}$$

where M is the number of filters, and the filter is the frequency response, which is defined as:

$$H_m(k) = \begin{cases} 0 & ,k < f(m-1), k \ge f(m+1) \\ \dfrac{2(k - f(m-1))}{(f(m+1) - f(m-1))(f(m) - f(m-1))} & ,f(m-1) \le k \le f(m) \\ \dfrac{2(f(m+1) - k)}{(f(m+1) - f(m-1))(f(m+1) - f(m-1))} & ,f(m) \le k \le f(m+1) \end{cases} \tag{10}$$

where f(m) denotes the center frequency of the triangular filter, $m=1,2,\cdots,M$.

3) Then discrete cosine transform is performed on S(m) to obtain MFCC:

$$c(n)=\sum_{m=0}^{m} S(m)\cos\left(\frac{n\pi(m+0.5)}{M}\right), 0\le m\le M \tag{11}$$

where $n=1,2,\cdots,p$, p is the dimension of the MFCC.

IV. Double-Threshold Endpoint Detection and Its Improved Method

There would be at a standstill when people talk, if the pause time is too long, and it will be into the termination section in advance, and results in the missing detection, which will lead to inaccurate endpoint detection [15], and finally, the recognition effect of the system is affected.

The implementation procedure of double-threshold endpoint detection algorithm is that, the first is to set a high value amp1 of En parameter, then a rough decision is performed to get the two corresponding locations of N1 and N2, and the initial position to determine the voice is out of N1 and N2. The mean value of the noise energy is set to a lower En parameter value, which is further positioned by the amp2 to find the start position N1 and the end position N2 of the voiced segment in the speech. Finally, according to zcr of Zn parameter value, the surd segment is judged to find the start position and end position of the useful section, thereby, the latest useful section of the start position is N3, so the useful segment is defined as N3 to N2. As shown in Fig. 5.

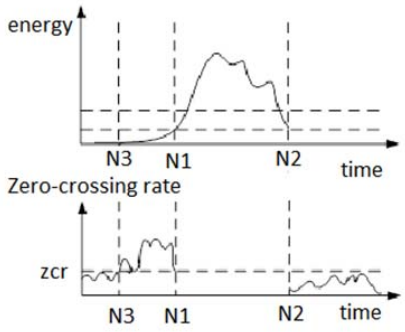

Figure 5. Threshold setting of double-threshold method

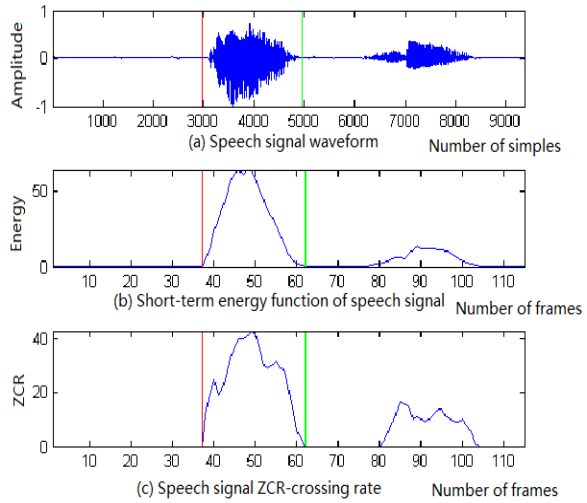

Figure 6. The chinese phonetic "Probabilities" after double-threshold detection

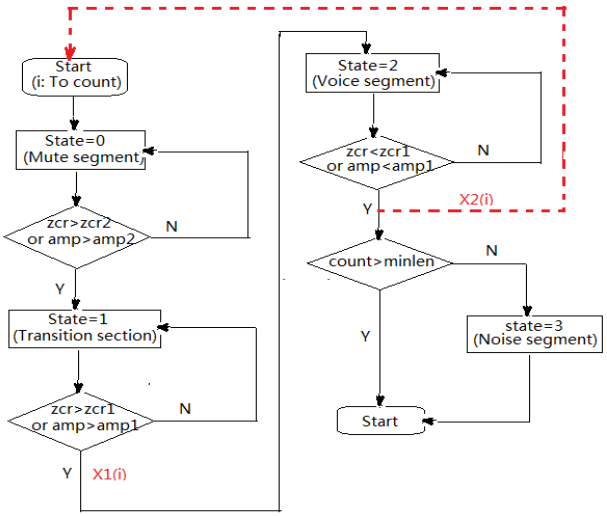

Figure 7. Improved flow chart for endpoint detection

A simulation of the Chinese phonetic "probabilities" with a short pause is shown in Fig. 6. The red vertical line indicates the position of the beginning of the speech signal, and the right vertical line indicates the end of the speech signal. Obviously, the detection has finished early without "rate" word. In order to solve this problem, this paper improves the algorithm, which adds a loop based on the above algorithm, and uses the variables X1(i) and X2(i) to express the start and end positions of the voice signal found in the ith time, respectively. Its improvement flow diagram is shown as Fig. 7.

As shown in Fig. 8, compared with the traditional double-threshold detection method, the improved algorithm effectively determines the start and end positions of two words. This method is used to detect the end-point of the voice "I am a student of the college of Electronic and Information Engineering of Anhui University", and find the start and end points of this voice, thus the speech between the two points will be extracted, the waveform is shown in Fig. 9.

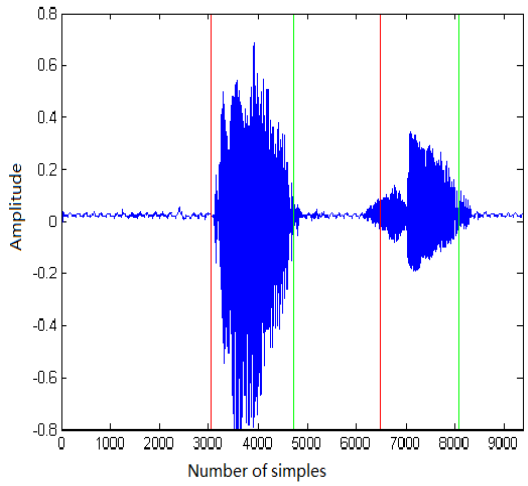

Figure 8. The chinese phonetic "Probabilities" through the improved double-threshold detection method

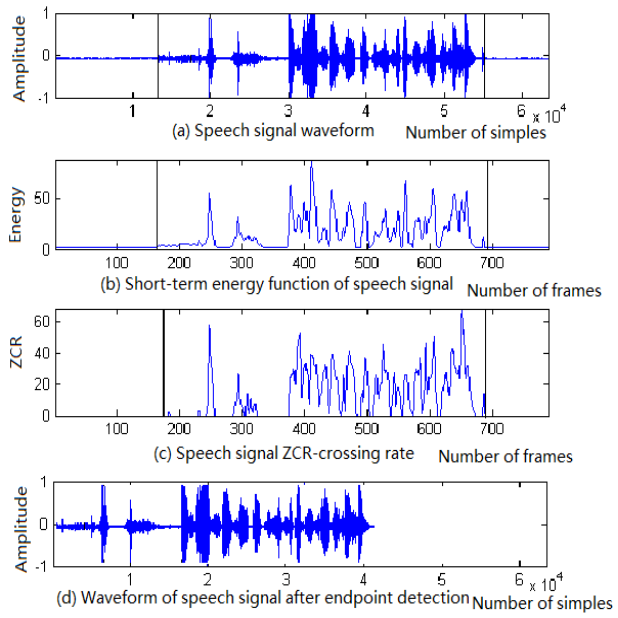

Figure 9. The waveform through the improved double-threshold detection method

V. Experimental Simulation Results

The experiment results were simulated based on Matlab 2011b. In a room with relatively low background noise, 46 people were recorded with Cool Edit Pro2.1 software. The sampling frequency was 8KHz, the quantization precision was 16bit, the channel selection was mixed with mon, and the required training voice and test voice are through the software to intercept. The experimental result is that "the speaker Si of the test set is the same person as the speaker S_j of the training set".

In order to facilitate the contrastive analysis with the follow-up experiments, here are the self-recorded voice database to correctly identify the corresponding relationship, as shown in Table II.

TABLE II. The Correct Corresponding Relationship of Identification

Test set	*Train set*	*Test set*	*Train set*
S1	S3	S24	S25
S2	S4	S25	S26
S3	S5	S26	S24
S4	S6	S27	S27
S5	S7	S28	S28
S6	S8	S29	S29
S7	S2	S30	S30
S8	S1	S31	S31
S9	S10	S32	S32
S10	S9	S33	S33
S11	S12	S34	S34
S12	S11	S35	S35
S13	S13	S36	S36
S14	S16	S37	S46
S15	S14	S38	S45
S16	S15	S39	S37
S17	S17	S40	S39
S18	S18	S41	S42
S19	S20	S42	S41
S20	S19	S43	S44
S21	S22	S44	S38
S22	S23	S45	S43
S23	S21	S46	S40

Compared with Table II, the experimental results show that, S14, S18, S22, S30, S35, S40 in the test set are not identified and the recognition rate is 86.95%.

But the experimental results obtained by the improved algorithm show that, only S14 and S40 are not identified in the test set, and the recognition rate is improved to 95.65%. The simulation results are shown in Fig. 10.

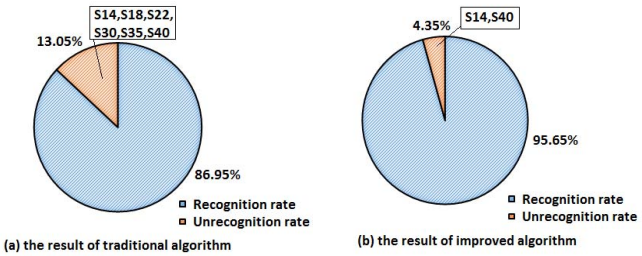

Figure 10. The recognition rate of the traditional and improved algorithm

According to the experimental requirements, different lengths of speech are taken as the training speech, and MFCC is used to extract the feature parameters. The results show that the recognition rate is affected by different lengths of speech.

TABLE III. Effect of Training Speech Length on Recognition Rate

Voice text length	Recognition rate
1s	82.60%
4s	89.10%
6s	95.65%
10s or more	97.82%

It can be seen from Table III that, with the increase of the training speech length, the recognition rate is obviously improved, and the training time is longer, the acquired codebook is more able to characterize the speaker's personal characteristics, the recognition effect of the system is better. Moreover, when the training time is more than 10s, ideal recognition effect is easier to achieve.

VI. Conclusion

Aim at the missing detection in speech signal preprocessing stage, an improved algorithm is proposed. Based on the traditional double-threshold endpoint detection method, this algorithm adds a loop, the end position of the last detected speech signal is adopted as the starting position of this cycle, then the detection of the loop starts until the last frame of speech is detected. The experimental results show that, the improved algorithm effectively avoid the phenomenon of missing detection. In addition, this paper uses the multistage vector quantization to quantify the feature parameters, and intercepts different lengths of speech for training. Compared with traditional algorithm, the proposed algorithm improves the recognition rate, and with

the increase of training speech length, recognition effect is better.

ACKNOWLEDGMENT

This research was supported by National Natural Science Foundation of China (No. 61501002), Anhui province quality project (2013zjjh003).

REFERENCES

[1] T Kinnunen, H Li, "An overview of text-independent speaker recognition: From features to supervectors," Speech Communication, vol. 52, pp. 12-40, Jan. 2010.

[2] C Yu, G Liu, S Hahm and JHL Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," IEEE International Conference on Acoustics, pp. 4017-4021, 2014.

[3] LIU Jing-Wei, XU Mei-Zhi, ZHENG Zhong-Guo and CHENG Qian-Sheng, "DTW-Based Feature Selection for Speech Recognition and Speaker Recognition," Pattern Recognition and Artificial Intelligence, vol. 18, pp. 50-54, Feb. 2005.

[4] Li Xiangping, "Research and experiment of speaker identification algorithms based on artificial neural network," Electronic Measurement Technology, vol. 30, pp. 170-172, Nov. 2007.

[5] V Hantamaki, AL Kong, DV Leeuwen, and Rahim Saeidi, et al. "Automatic Regularization of Cross-entropy Cost for Speaker Recognition Fusion," Interspeech, pp. 1609-1613, Aug. 2013.

[6] S Singh, EG Rajan, "Vector Quantization Approach for Speaker Recognition using MFCC and Inverted MFCC," International Journal of Computer Applications, vol. 17, pp. 1-7, Mar. 2011.

[7] CHEN Shanxue, YIN Xuejiao and ZHANG Yan, "Codebook design based on improving particle swarm algorithm," Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition), vol. 25, pp. 221-225, Apr. 2013.

[8] YANG Shu-Ying, LIU Xu-Peng, TAO Chong and LIU Ting-Ting, "Vector Quantization Codebook Design and Speech Recognition Based on Immune Cat Swarm Optimization Algorithm," Pattern Recognition and Artificial Intelligence, vol. 27, pp. 577-583, Jul. 2014.

[9] LI Fenglian, ZHANG Xueying, Zizhong John Wang and LI Hongchun, "Codebook classification rearrangement vector quantization," Journal of Tsinghua University(Science and Technology), vol. 53, pp. 893-897, Jun. 2013.

[10] LIU Yu-zhen and TIAN Jin-bo, "Double threshold Endpoint Detection Algorithm Based on Speech Enhancement," Measurement & Control Technology, vol. 35, pp. 33-35, Jan. 2016.

[11] XUE Sheng-yao. "Research on speech endpoint detection based on the improved dual-threshold algorithm," Electronic Design Engineering, vol. 23, pp. 78-81, Feb. 2015.

[12] M Stner, B Hammer, M Biehl, T Villmann, "Functional relevance learning in generalized learning vector quantization," Neurocomputing, vol. 90, pp. 85-95, Aug. 2012.

[13] Hu Feng-song and ZHANG Xuan, "Speaker recognition method based on Mel frequency cepstrum coefficient inverted Mel frequency cepstrum coefficient," Journal of computer Applications, vol. 32, pp. 252-254, Sept. 2012.

[14] OC Ai, M Hariharan, S Yaacob and LS Chee, "Classification of speech dysfluencies with MFCC and LPCC features," Expert System with Applications, vol. 39, pp. 2157-2165, Jul. 2011.

[15] Peipei Tu, "The Research of Speaker Recognition Based on Vector Quantization," Master's Degree Thesis of Anhui university, May. 2016.