Speaker-dependent Isolated-Word Speech Recognition System Based on Vector Quantization

Yinyin Zhao
Engineering Training Center of Beihua University
Jilin, China
Zhyy8119@126.com

Lei Zhu
Beihua University College of Electrical and Information Engineering
Jilin, China
2295591145@qq.com

***Abstract*—Speaker-dependent speech recognition system requires the system should not only recognize speech, but also recognize the speaker of the segment. In this paper, two indicators are selected—short-time average zero-crossing rate and dual-threshold endpoint to test the signal endpoint through the study of speaker-dependent isolated-word speech characteristics, and MFCC parameters are taken as the characteristic parameters; based on vector quantization, template matching algorithms are designed, and one is adopted to improve LBG algorithm to increase the computing speed; speaker-dependent isolated-word speech recognition system is designed based on vector quantization technique and simulation experiments are conducted in the MATLAB platform under various backgrounds, which proves the system has better recognition effect.**

Keywords-Speech recognition; LBG; MFCC; Vector Quantization;

I. INTRODUCTION

The human speech is the most natural and easiest means of communication in the exchange and transfer of information. Therefore, it becomes a new technique to make machine able to understand human speech and make the communication between human and machine as convenient as the one between human and human, which is explored by people continuously. Speech recognition can be divided into speaker-dependent and speaker-independent speech recognition. Speaker-dependent recognition requires the system not only identify the corresponding speech signal, but also identify the speaker who issues the speech segment[1]. Compared with speaker-independent speech recognition, the speaker-dependent recognition highlights both relevant characteristics of the speech signal, and the personality of speaker. As a result, speaker-dependent speech recognition is widely used in many fields, such as network security, banking systems, stock systems, and security systems.

The application of vector quantization technique in speaker-dependent isolated-word recognition system is mainly studied in this paper, and a simulation study of speech recognition system is conducted, achieving good experimental results.

II. DESIGN OF SPEECH RECOGNITION SYSTEM

There are mainly two parts,that is parameter extraction and pattern matching[2]. The basic components of the speech recognition system are shown in Fig.1. Pretreatment and endpoint detection are to ensure the extracted speech features can reflect the characteristics of the speech signal segment. After the detection of start and end points of the voice segment, the characteristic parameters are extracted, and then the appropriate training algorithm is selected based on the eigenvalues to train them and form template library for pattern matching at the time of speech recognition.

III. EXTRACTION OF SPEECH SIGNAL FEATURES

A. Pretreatment

Pretreatment of speech signals includes three steps: pre-emphasis, framing and windowing[3].

The purpose of pre-emphasis is to emphasize the high-frequency portion of the speech, increasing the high-frequency resolution of the speech, to facilitate spectral analysis or channel parameter analysis, which is usually realized by first-order FIR high-pass filter. The function is:

$$H(z) = 1 - \lambda z^{-1} \tag{1}$$

where λ is set to 0.9375 in the experiment, and Fig.2 shows the comparison of “beihua” before and after the pre-emphasis.

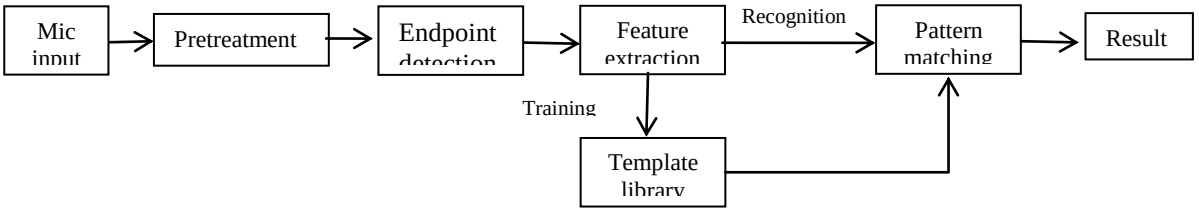

Figure 1. Pre-emphasis before and after comparison

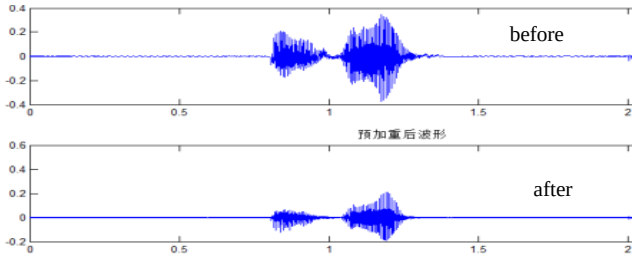

Figure 2. Pre-emphasis before and after comparison

Since the speech signal has the feature of short-time stability, the speech signal can be divided into several frames. According to the sampling frequency of the system, the frame length of system is set as 256, with an overlapped region between two adjacent frames, which makes a smooth transition between frames, ensuring the continuity of signal. This system adopts the frame-shift of 100, and applies Hamming window for the window function.

B. Endpoint detection

Characteristics of the speech signal include short-time stability long-time change, and having instant stability, and this time period is typically less than 50ms, so classical stationary signal processing method can be adopted for the processing of speech signal. The traditional endpoint detection is to determine the end through short-time energy and short-time average zero-crossing rate point with short time-domain analysis after the pretreatment of original speech signals, to distinguish pronunciation zone and quiet zone. Short-time energy calculation is carried out based on frame, and short-time energy is defined as[4]:

$$E_n = \sum_{m=n-N+1}^{n} x^2(m) \tag{2}$$

Zero-crossing rate is an indicator reflecting the signal spectral characteristics. Short-time average zero-crossing rate is the average number of times that waveform crosses zero point within one-frame signal, defined as:

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]\right| w(n-m) \tag{3}$$

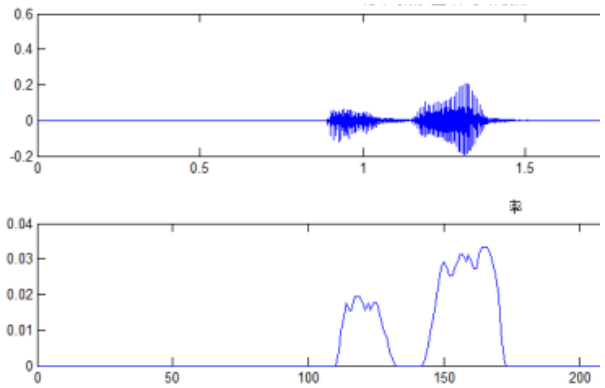

Figure 3. Short-time average zero-crossing rate of “beihua”

Because the amplitude of the speech signal can be reflected in the short-time energy, the frequency is related to the short-time zero-crossing rate, whereas these two indicators can detect sound and silence of signals. In order to detect the start and end points of the speech signal more accurately, dual-threshold endpoint detection algorithm is applied in this paper, which is to combine two indicators (short-time energy and short-time average zero-crossing rate) to detect endpoint detection. Fig.3 is the test result of speech segment “beihua” under the dual-threshold endpoint detection algorithm, in which two lines of the speech signal are the start frame and end frame of the speech segment respectively.

C. Endpoint Extraction of characteristic parameters

Extracting characteristic parameters of the speech signal is a key link in the speech recognition process. Characteristics extraction is to analyze and process the speech signal, so as to remove irrelevant redundant information for speech recognition, to obtain the basic characteristic information characterizing human in speech signal. Therefore, the feature information must be able to effectively distinguish different speakers, and keep stable for the changes of the same speaker. There are mainly two feature extraction algorithms used in speech recognition systems currently: Linear Predictive Cepstrum Coefficients (LPCC) and Mel Frequency Cepstrum Coefficients (MFCC)[5-6]. LPCC is an algorithm put forward based on the principles of the human vocalization, while MFCC is proposed based on the human auditory system. Experiments show MFCC has better result than LPCC in speaker-

dependent speech recognition, so MFCC is applied to the feature extraction of speech signals in this paper.

Procedures of MFCC in implementation are as follows:

- Actual frequency is converted into Mel frequency according to the Eq. (4).

$$f_{Mel} = 2595\log_{10}(1+ f / 700) \tag{4}$$

- The output of E (mel) on Mel coordinates passing through this Mel filter group is calculated in Mel frequency;
- The results of the output of E(mel) are transformed into the logarithms by calculation to get logarithmic spectrum S(mel) ;
- The logarithmic spectrum in S(mel) is subject to discrete cosine transform, and then the corresponding MFCC parameters can be obtained. Transformation formula is as follows[7]:

$$C(n) = \sum_{mel=1}^{M} S(mel)\cos\left(\frac{\pi n(mel-0.5)}{M}\right), \quad 0 \le n < M \tag{5}$$

C =

Columns 1 through 8

-39.7223	-31.5947	-29.6053	-29.1472	-30.4809	-30.2553	-30.3434	-29.9795
12.7233	13.6380	12.6410	12.0431	13.6327	12.2813	10.7685	9.1535
0.5474	-3.1691	-3.0699	-3.4973	-5.2782	-4.5714	-2.5129	-0.6298
3.0628	2.6689	1.7899	1.1456	3.4962	4.2581	3.9945	3.8833
-0.9859	-2.7486	-2.9701	-1.7778	-3.2985	-4.4408	-4.5609	-4.6157
1.2912	0.9610	0.1087	-2.3530	-2.3167	-1.9318	-2.4382	-2.3294
-1.4374	-2.8386	-2.4237	-1.5932	-1.5981	-2.1183	-1.2931	-1.0677
0.9777	2.3940	3.0293	3.3342	3.7892	3.9878	3.5160	3.2233
-0.5716	-0.2306	-0.8451	-0.8571	-1.6350	-2.3001	-2.3215	-2.3756
0.0490	3.1060	2.3746	0.4470	0.9915	1.3683	1.3860	1.9169
-0.3871	-2.2715	-2.2714	-0.6664	-0.5468	-0.4407	0.0959	0.0866
0.1082	0.1533	0.2889	-0.3425	-0.0766	-0.0152	-0.1032	-0.1920
1.2428	1.1648	0.6774	0.1863	-0.6744	-1.4538	-1.8430	-1.6102
-1.3987	-0.6018	-1.6985	-2.4139	-1.8054	-1.3070	-0.9069	-0.7278

Figure 4. Partial MFCC characteristic parameters of the "beihua"

Where M is the order of the Mel filter. M = 24 in this system, and Fig. 4 is a 24-order Mel filter group.

IV. MATCHING AND RECOGNITION OF TEMPLATE

Template matching principle is generally applied for a speaker-dependent small-vocabulary speech recognition system. Firstly, template library is created by the trained speech data, and then feature vectors obtained from the input speech are compared with the templates in the template library, to get the recognition result. Speech recognition algorithms mainly include Hidden Markov Model (HMM), Dynamic Time Warping (DTW), Vector Quantization (VQ)[8], Artificial Neural Networks (ANN) and so on, in which Vector Quantization is a very efficient technique in data compression and coding, which can significantly reduce the computational complexity without reducing the accuracy of identification, simple and practical. Considering the characteristics of the system, the vector quantization method is adopted in the paper to establish the template, which is more suitable for small-vocabulary, isolated-word speech recognition.

A. *Vector Quantization*

Vector quantization is based on Shannon rate-distortion theory. The theory is that: for a given distorted D, the rate-distortion function R(D) can be calculated.

Where in the average distortion :

$$Q(Y) = \sum_{X} P(X)P(Y / X) \tag{6}$$

Satisfies the condition:

$$\sum_{X}\sum_{Y} P(X)P(Y \mid X)d(X;Y) \le D \tag{7}$$

The inverse function R(D) is the distortion-rate function D(R), which indicates the smallest distortion that the system can achieve under the condition that the given rate is no more than R. Vector quantization is used to increase vector dimension k, and coding performance can be arbitrarily close to the rate-distortion function.

If the input vector is X, of which the dimension is K, then $X = [x_1, x_2, \cdots, x_k]$. The system has two identical codebooks; each codebook contains M codewords Y_i , $i = 1 \sim M$, and each codeword is a K-dimensional vector. VQ encoder principle is to select a corresponding vector Y_i from the encoder codebook according to the input vector X, where the output v is equal to the subscript of the vector, namely, $v = r(X)$. VQ decoder is to select a codeword with the corresponding subscript as output Y according to v by look-up, namely $Y = \beta(v)$.

B. *LBG algorithm*

LBG algorithm is an efficient and intuitive design algorithm for vector quantization codebook[9]. After the MFCC parameters are extracted, characteristic parameters are trained be the application of basic LBG algorithm to get the corresponding codebook[10]. The basic LBG algorithm is as follows:

All the necessary reference vectors X for VQ codebook training are given, and the set of X is represented by S; quantization levels, distortion control threshold β, maximum number of iterations of the algorithm L and initial codebook Y are established, and the total distortion $D^{(0)} = \infty$; the number of iterations is initialized to 1; the final training codebook is obtained $Y_1^{(m)}, Y_2^{(m)}, \cdots, Y_N^{(m)}$, and the total distortion measure is output as $D^{(m)}$.

The maximum number of iterations distortion L and control threshold β in the algorithm are established in order to avoid infinite loop of iterative algorithm. The value of β is much less than one, and when $\beta^{(m)} \le \beta$, it indicates that the reduction in further iteration calculation distortion is limited, so the calculation can be stopped. L is the parameter to limit the number of iterations in order to prevent the excessive number of iterations when β is set lower.

However, in the basic LBG algorithm, due to the arbitrariness of the initial codebook selection, the problem of empty cell cavity may appear. In order to solve this problem, LBG algorithm of empty cell splitting is adopted in this paper. The main steps of empty cell splitting method is as follows:

- Removing the centroid Y_X from the empty cell;
- Splitting the maximum cell SM, multiplying the centroid YM of SM by the disturbance coefficient $1\pm\partial$ respectively, to get two codewords, YM1 and YM2, which are taken as references for Voronoi tessellation of two small cells SM1 and SM2。

The advantage of this method is the using of two smallest cells to replace the original large cell, to reduce the quantization distortion, thus improving quantization performance.

C. *Matching of codebook*

As described above, we use cell splitting LBG algorithm to train each of the speech signals to be recognized, and then store the resulted training codebooks separately, to get the corresponding template library. The signals to be identified are subject to a series of treatments to get a codebook when recognition is required, and the codebook is matched with each codebook in this template library, to calculate the distortion measure respectively. If there is a small degree of distortion within a predetermined threshold value Ľ (i.e. smaller than the predetermined threshold value Ľ), it is considered that the speech segment to be recognized matches the template, and the recognition result is obtained to be output in the recognition result site. If all the distortions are greater than Ľ, there are no matching results in the matching template for the segment of the input speech, and then "No matches" is output in recognition result site. Wherein, considering the complexity of the distortion measure calculation and the implementation simplicity of subsequent hardware, the classical absolute-value average error Euclidean distance measure is applied to calculate the corresponding distortion measure. The absolute-value average error Euclidean distance measure is as follows.

Let x be k-dimensional feature vector of unknown model, y be k-dimensional code vector in the codebook, and x_i and y_i be the components of x and y of the same dimension respectively, then absolute-value average error Euclidean distance measure is defined as:

$$d_1(x,y)=\frac{1}{k}\sum_{i=1}^{k}|x_i-y_i| \qquad (8)$$

D. *Simulation experiment*

According to the theory above, systematic experiment is conducted on the MATLAB platform in the paper. There are totally 6 recording persons in the experiment (three men and three women), who are numbered as Speaker 1 to Speaker 6, and among whom the 4th and the 5th speakers are close in speech feature, with audio sampling frequency 11.025KHz; the speech segments to be measured belong to isolated vocabulary. In the testing process, each speaker pronounces the same word twice, in which one is used for training, and the other for recognition. Fig.5 and Fig.6 shows partial recognition results. Test results shows the statistical result of the recognition accuracy.

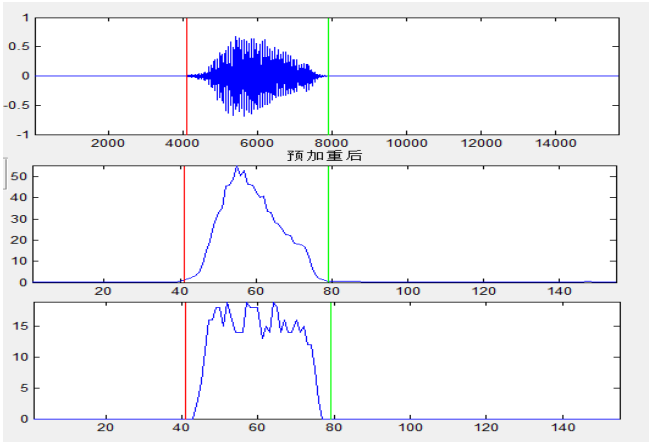

Figure 5. The recognition result of voice "1"

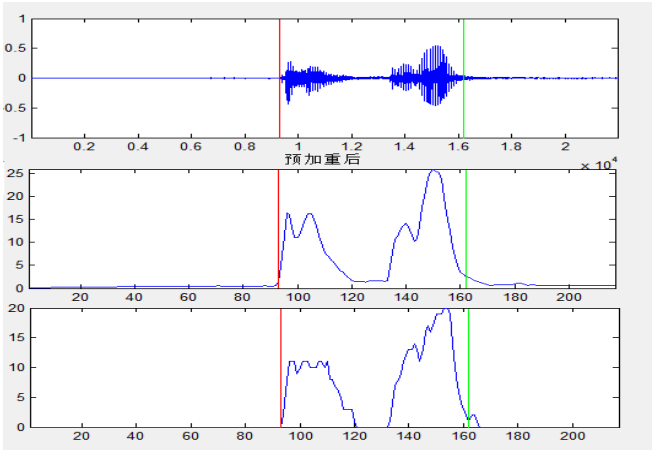

Figure 6. The recognition result of voice "beihua"

From the experimental results, it can be obtained that the system can realize the speaker-dependent isolated-word speech recognition, and the recognition rate can reach 95% in a relatively quiet environment. However, when the speech features of both speakers are close, or environmental noise is great, the accuracy of the recognition will be impacted, and the influence of noise on accuracy is more prominent.

V. CONCLUSION

Vector quantization method is applied in this paper to train the speech to be recognized, and establish the appropriate recognition template library, greatly reducing the amount of computation and data storage, which achieve ideal recognition performance in speaker-dependent isolated-word speech recognition experiment. The treatment of environmental noise is not favorable in this paper, so the next step will be on how to eliminate the influence of ambient noise to improve signal to noise ratio, thus improving recognition accuracy.

ACKNOWLEDGMENT

The authors acknowledge the Science and technology projects in Jilin Province Department of Education (Grant: JJH20170031KJ), the Science and technology projects in Jilin Province Department of Education (Grant: JJH20170035KJ).

REFERENCES

[1] Y. Wang , F. Tang ,J Zheng . Robust Text-independent Speaker Identification in a Time-varying Noisy Environment[J].Journal of Software,2012, 7(9):1975-1980.

[2] Satyanand Singh, E.G Rajan. MFCC VQ based Speaker Recognition and Its Accuracy Affecting Factors[J]. International Journal of Computer Applications. 2011(6) :1-5.

[3] Zhang Y, Long H, Shen S, et al. A novel codebook design with the LBG algorithm in precoding systems under spatial correlated channel[C]. Communications Circuits and Systems (ICCCAS), 2010 International Conference on . 2010:770-775.

[4] Ashkan Parsi, Ali Ghanbari Sorkhi, Morteza Zahedi. Improving the unsupervised LBG clustering algorithm performance in image segmentation using principal component analysis[J].Signal, Image and Video Processing, 2016, 10 (2):301-309.

[5] Chen Chen Huang, Wei Gong, Wen Long Fu, Dong Yu Feng.A Research of Speaker Recognition Based on VQ and MFCC[J]. Applied Mechanics and Materials, 2014, 3468 (644):4325-4329.

[6] M Sahidullah, G Saha. A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition[J]. IEEE Sinal Processing Letters, 2013, 20(2):149-152.

[7] Woo Yong, ChoiHwa Jeon, SongHoon Chung.I-vector Based Utterance Verification for Large-Vocabulary Speech Recognition System[C]. 2016 First IEEE International Conference on Computer Communication and the Internet,2016:334-337.

[8] Lin CY, Prangjarote P, Yeh CH, et al. Reversible joint fingerprinting and decryption based on side match vector quantization[J]. Signal Processing, 2014, 98(1):52-61.

[9] Jian BS.,Robust Multiple Antennas Cooperative Spectrum Sharing Design With Random Vector Qquantization[J]. 2014, 62(4): 486-492.

[10] A Chaudhari,A Rahulkar, SB Dhonde. Combining dynamic features with MFCC for text-independent speaker identification[J]. International Conference on Information Processing,2016:160-164.