Speaker Recognition using VQ and DTW

Maruti Limkar
Terna college of Engineering
Department of Electronics Engineering
Mumbai University, India

B.Rama Rao
Vidyalankar Institute of Technology
Department of Electronics and Telecommunication Engg.
Mumbai University, India

Vidya Sagvekar
K.J.Somaiya Institute of Engg. And Information Technology
Department of Electronics Engineering
Mumbai University, India

ABSTRACT
Speaker recognition is a process where a person is recognized on the basis of his/her voice signals. In this paper we provide a brief overview for evolution of pattern classification technique used in speaker recognition. Also discussed propose process to modeling a speaker recognition system, which include pre-processing phase, feature extraction phase and pattern classification phase. Linear Prediction Cepstrum Coefficient (LPCC) and Mel Frequency Cepstrum Coefficient (MFCC) are used as the features for text dependent speaker recognition in this system and the experiments compare the recognition rate of LPCC, MFCC or a combination of LPCC and MFCC through using Vector Quantization (VQ) and Dynamic Time Warping (DTW) to recognize a speaker's identity. It proves that the combination of LPCC and MFCC has a higher recognition rate.

General Terms
Pattern Recognition, Minimum Distance,Accuracy

Keywords
Speaker recognition; LPCC; MFCC; VQ; DTW

1. INTRODUCTION
Speaker recognition is a process of automatically identify who is speaking on the basis of individual information integrated in speech waves. Speaker recognition can be further broken into two categories: Speaker identification and speaker verification. Speaker identification determines from which of the registered speakers a given utterance comes whereas speaker verification is the process of accepting or rejecting the claimed identity of a speaker.

Extracting speaker's personal audio signal is the key to speaker recognition. Linear Prediction Cepstrum Coefficient (LPCC) reflects the difference of the biological structure of human vocal track and Mel Frequency Cepstrum Coefficient (MFCC) is based on the human ears' non-linear frequency characteristic[3].This paper presents a speaker recognition system based on the Vector Quantization (VQ)[8] and Dynamic Time Warping(DTW),which uses the combination of LPCC and MFCC as features and compares the recognition rate of speaker recognition which used LPCC, MFCC or the combination of LPCC and MFCC as features through a series of experiments.

2. EXTRACTING FEATURES
A. LPCC: Linear predictive cepstral coefficients (LPCC) combine the benefits of LPC and cepstral analysis and also improve the accuracy of the features obtained for speaker recognition. LPCC is equivalent to the smooth envelop of the log of the speech that allows for the extraction of speaker specific features. The block diagram of the LPCC is shown in the figure below.

LPC is transformed into cepstral coefficients using the following recursive formula

$$c_1 = a_1 \quad \text{.....(1)}$$

$$c_n = a_n + \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a_k c_{n-k} \quad \text{.....(2)}$$

Where c_i and a_i are the i th-order cepstrum coefficient and linear predictor coefficient, respectively.

B.MFCC: The MFCC which is different from other frequency cepstrum focuses on the human ears' non-linear frequency characteristic, and the size of Mel frequency corresponds to the relation of actual frequency's logarithmic distribution on the whole and accords with the human ears' characteristic. The idiographic relationship between Mel frequency and actual frequency is as follows:

$$M = 2595 \log_{10}\left(\frac{f}{700} + 1\right) \quad \text{.....(3)}$$

MFCC start with dividing the speech signal into short frame and windowing each frame to discard the effect of discontinuities at edges of the frames. In fast fourier transform (FFT) phase, it converts the signal to frequency domain and after that Mel frequency warping the frames. After Mel frequency warping the frames, logarithm of the signal is passed to the inverse DFT function converting the signal back to time domain.

As a result of the final step, 13 coefficients named MFCC for each frame are obtained. The 0^{th} coefficient is not used because it represents the average energy in the signal frame and contains little or no usable information. Figure 1 shows process of MFCC

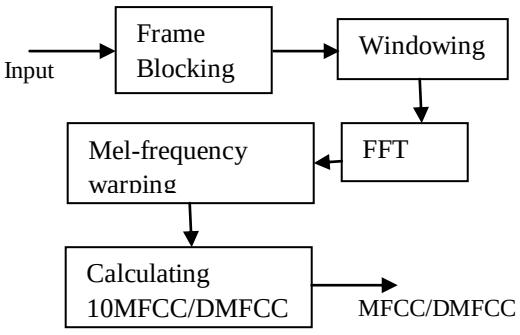

Fig.1 outline of the process of MFCC

As the output of feature extraction phase, vectors in 12 dimensions are obtained for each frame. The vectors are used in pattern matching/classification technique for compare and match the feature sets against the model already stored before hand.

3. METHOD OF SPEAKER RECOGNITION

Vector Quantization (VQ) is an important method of digital signals processing. VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a cluster and can be represented by its centre called a codeword. The collection of all codeword's is called a codebook. Training and recognizing are two steps of VQ. Training is namely establishing N codebooks of N speakers and these codebooks are not superposing each other in the feature space. In the step of recognizing, firstly extract a group of vectors from speech waiting to be recognized, then use N codebooks founded in the system to quantize these vectors with VQ to grain O={o1,o2,…,ol},namely judge that group of vectors in accordance to the codebook in feature space. Assume that the number of code words of these N codebooks is M. Dynamic Time warping (DTW) is based on dynamic programming and can resolve the matching problem of the difference of speech's length. This paper stores LPCC and MFCC distilled by speech processing according to frames. Referenced template and test template was compared with DTW arithmetic for template matching; calculate the distance between referenced template and test template by DTW and the template of the minimum distance is the best matching result [5].

B. Speaker Recognition with Combination of Features

Mel frequency cepstrum reflects the human ears' non linear frequency characteristic, Linear Prediction Cepstrum reflects the differences of biological structure of human vocal track, and their one-order differential coefficients both describe their own dynamic characteristics. When use MFCC parameter to recognize, the system tends to judge this speaker as a legal speaker of the system if a speaker embezzles the password. So use LPCC which reflects the difference of the biological structure of human vocal track and its one order differencing a feature to enhance the security of the system. First the template matching by DTW arithmetic is applied to two combined feature vectors in the process of recognition. Then set the distance threshold p of template matching in order to reduce miscarriage of justice. If the distance of template matching d is larger than p, the speaker is considered as an illegal speaker, even if he was a legal speaker. Compare the script code i,j of the minimum distance. If i equal to j, the speaker is judged as a legal speaker, even if he was an illegal speaker.

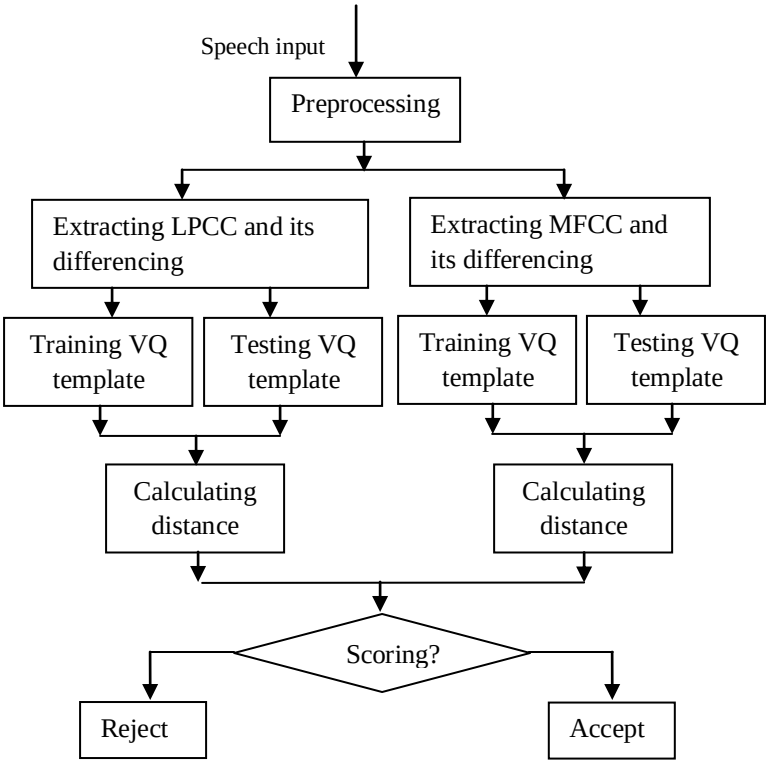

Fig.2 Speaker Recognition based on LPCC and MFCC

4. RESULT

A. Experiment setup

For this experiment we use database includes 50 voice samples of different speakers (20 males and 30 females) and recorded in two different sessions with a gap of two weeks is used for the training and testing of the developed system. Voice samples recorded in the first session are used as training data and those of second session as testing data. Both sessions have been recorded with microphone at sampling frequencies 8000Hz and 11025Hz.

This experiment used MATLAB 7.0 as the development environment and did three kinds of experiments towards different feature vector with DTW arithmetic. Accuracy [6] is used as the standard of evaluating the recognition performance of the system in this paper.

Experiment one: Train the template and test the system to form the whole system of speaker recognition with 12 order LPCC coefficient and its one-order differencing *LPCC.

Experiment two: Train the template and test the system to form the whole system of speaker recognition with 16-order MFCC coefficient and its one-order differencing *MFCC.

Experiment three: Use the combination of two kinds of feature vector of experiment one and experiment two and adjust the result of recognition with DTW arithmetic.

B. Result of Experiments

Results of Experiment 1, Experiment 2 and Experiment 3 are shown in Table 1 and 2 respectively. The effect of using solely LPCC and *LPCC or MFCC and *MFCC is not as that

of using the combination of LPCC, MFCC,*LPCC and *MFCC.LPCC makes up for MFCC's failure in describing the characteristics of vocal track, moreover *LPCC and *MFCC reflect the dynamic characteristic of speech and vocal track, so the combination of these feature vectors better reflect the individual characteristic of a speaker.

Table 1: Results of First Experiment at sampling Frequency 8000Hz and 11025Hz

Modeling Technique	Feature	Experiment #	Sampling Frequency	Identification Accuracy
VQ+DTW	LPCC,*LPCC	1st	8000Hz 11025Hz	87.65% 95.52%

Table 2: Results of Second Experiment at sampling Frequency 8000Hz and 11025Hz

Modeling Technique	Feature	Experiment #	Sampling Frequency	Identification Accuracy
VQ+DTW	MFCC, *MFCC	2nd	8000Hz 11025Hz	91.25% 96.27%

Table 3: Results of Third Experiment at sampling Frequency 8000Hz and 11025Hz

Modeling Technique	Feature	Experiment #	Sampling Frequency	Identification Accuracy
VQ+DTW	Combi-nation	3rd	8000Hz 11025Hz	95.85% 98.52%

5. CONCLUSION

The paper used various pre-processing stages prior to feature extraction were studied and implemented for the speaker recognition. The prototype was developed to analyze and evaluate various voice feature extraction methods such as LPCC and MFCC for their suitability in speaker recognition. The paper used VQ and DTW method to recognize a speaker's identity through extracting the combination of LPCC, MFCC,*LPCC and *MFCC and compare strengths and weaknesses of using LPCC, MFCC,*LPCC,*MFCC and their combination as speech features. The experiment showed the combination of LPCC, MFCC,*LPCC and *MFCC improved the performance in aspect of the recognition rate.

6. REFERENCES

[1] Campbell J.P,"Speaker Recogniton :ATutorial", Proc.of the IEEE,vol.85,no.9,pp.1437-1462,sep.1997.

[2] Sadaoki Furui., "Recent advances in speaker recognition",Pattern Recognition Letters. 1997,18 (9): 859-72.

[3] ZhiyouMa,"Further Extraction for Speaker Recognition", IEEE International Conference on Systems, Man and Cybernetics,153- 158,2003.

[4] Lawrence R. Rabiner., "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 77 (2), 1989, p. 257–286.

[5] DengHaojiang,Wangshoujie,XingCangju,LiuQian,"Research of Text-Independent Speaker Recognition Using Clustering Statistic",Jumal of Circuits and Systems,2001.

[6] Reynolds, D.A. and Rose, R.C. "Robust text independent speaker identification using Gaussian mixture speaker model", IEEE Trans. Speech Audio Process,3,1995,pp 72-83.

[7] Sadoki Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Trans. ASSP 29,1981, pages 254-272.

[8] F.K. Soong,A.E.Rosenberg,L.R. Rabiner and B.H.Juang, "A Vector Quantization approach to Speaker Recognition", Florida: ICASSP Vol. 1, 1985, pp.387-390.