Speaker Recognition

Supriya Tripathi
ECE. Jaypee Institute of Information Technology
Jaypee Institute of Information Technology
Noida, India
email : supriya.tripathi31@gma il.com

Smriti Bhatnagar
ECE. Jaypee Institute of Information Technology
Jaypee Institute of Information Technology
Noida, India
email : smriti.bhatnagar@jiit.ac.in

Abstract— **Speech processing has emerged as one of the important application area of digital signal processing. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. This paper proposes the comparison of the MFCC and the Vector Quantisation technique for speaker recognition. Feature vectors from speech are extracted by using Mel-frequency cepstral coefficients which carry the speaker's identity characteristics and vector quantization technique is implemented through Linde-Buzo-Gray algorithm. Vector quantization uses a codebook to characterize the short-time spectral coefficients of a speaker. These coefficients are used to identify an unknown speaker from a given set of speakers. The effectiveness of these methods is examined from the viewpoint of robustness against utterance variation such as differences in content, temporal variation, and changes in utterance speed.**

Keywords- Speaker recognition, Mel Frequency Cepstral coefficients, Vector Quantization

I. INTRODUCTION

Speech recognition is a difficult task and it is still an active research area. Automatic speech recognition works is based on the premise that a person's speech exhibits characteristics that are unique to the speaker. However this task has been challenged by the highly variant of input speech signals [1]. The principle source of variance is the speaker himself. Speech signals in training and testing sessions can be greatly different due to many facts such as voice of people changing with time, health conditions (e.g. the speaker has a cold), speaking rates, etc. There are also other factors, beyond speaker variability, that present a challenge to speech recognition technology. Several other approaches may be used to implement a speaker recognition system. Other approaches like Hidden Markov Model (HMM) show heavy computational complexity.

The main aim of this paper is to design an application for speaker identification, which should be robust in nature. For this purpose, the MFCC and the Vector Quantization technique is used to extract the feature vectors and their performances are compared.

All speaker recognition systems contain two main modules : feature extraction and feature matching [8]. Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.

An ordinary speech signal is taken as an input and its acoustic vectors are extracted which characterizes that signal. These acoustic vectors are unique for each speaker. These are used to identify the speaker during the testing phase by matching the features of a known speaker with the unknown speaker.

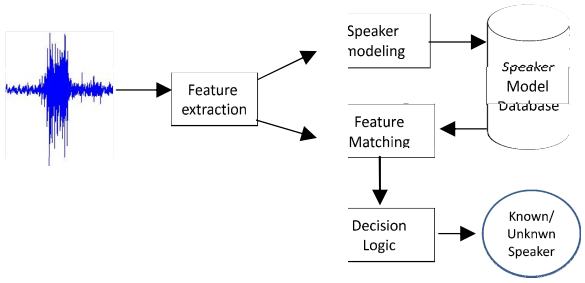

Figure 1. Feature Extraction and Feature Matching

II. MEL-FREQUENCY CEPSTRAL COEFFICIENTS

These are derived from a type of cepstral representation of an audio clip [2][3]. The difference between the cepstrum and the Mel-frequency cepstrum is that in the MFC, the frequency bands are positioned logarithmically (on the Mel scale) which approximates the human auditory system's response more closely than the linearly-spaced frequency bands obtained directly from the FFT or DCT. The following figure shows the basic steps of computing the MFCCs.

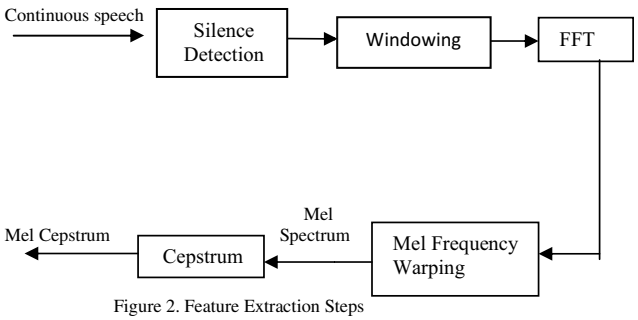

Figure 2. Feature Extraction Steps

978-0-7695-4872-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICCCT.2012.64

These steps have been explained below as follows :

A. Silence Detection

The speech signal was first stored as a 10000 sample vector. Based on observation, the actual uttered speech, eliminating the static portions, came up to about 2500 samples. A simple threshold technique can be used to extract the actual uttered speech.

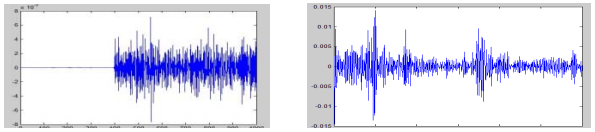

Figure 3. (a) Utterance of the word "hello' (b) After Silence detection

B. Windowing

The Hamming window offers the bell-shaped [13] weighting function but does not bring the signal to zero at the edges of the window. It minimizes the spectral distortion. This is done in order to eliminate discontinuities at the edges of the frames. If the hamming window is given by W_n, then the resulting signal will be; $Y[n] = x[n] * W_n$.

$$W_n = 0.54 - 0.46 \cos (2\pi n / N-1) \quad (1)$$

Here, $x[n]$ refers to the n^{th} speech sample in the frame and N is the number of samples in each frame. A speech sample is shown in Figure 5(a) and Figure 4(b) shows the signal after the multiplication with the Hamming window.

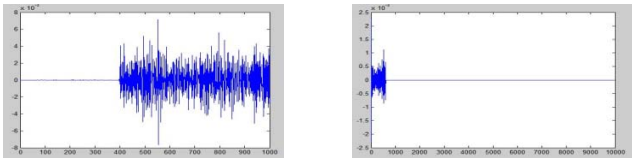

Figure 5. (a)without (b)with hamming window

C. Fast Fourier Transform

Fast Fourier Transform (FFT) converts each frame of N samples of speech from time domain to the frequency domain [6]. The FFT is defined on the set of N samples, as $Y_1[n]$ as follows:

$$Y_2[n] = \sum_{k=0}^{N-1} Y_1[k] e^{-\frac{2\pi jkn}{N}} \quad (2)$$

Where $n = 0, 1, 2, ..., N-1$. Note that we use j here to denote the imaginary unit, i.e. $j = \sqrt{-1}$. In general, the $Y_2[n]$'s are complex numbers. To compute the real numbers, the square of the magnitude for each frequency component is taken by using:

$$Y_3[n] = (\text{real } (Y_2[n]))^2 + (\text{imag } (Y_2[n]))^2 \quad (3)$$

Result of above equation is referred to as a spectrum.

D. Mel Frequency Warping

The human ear perceives the frequencies non-linearly. Researches show that the scaling is linear up to 1 kHz and logarithmic above that. The Mel-Scale (Melody Scale) filter bank characterizes how the human ear perceives frequencies. The signal is passed through Mel-Scale band pass filter to mimic the human ear.

For each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the *Mel-scale*. The *Mel-scale* is linear frequency spacing below 1000 Hz and logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. The following approximate formula is used to compute the Mels for a given frequency f in Hz [2][8][10] :

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \quad (4)$$

One approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired Mel-frequency component [5][7]. That filter bank has a triangular band-pass frequency response. Figure 5 shows the typical Mel-filter banks.

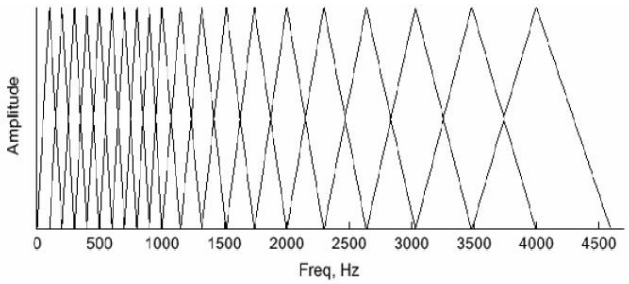

Figure 6. Filter Banks on Mel frequency Scale

E. Cepstrum

Cepstrum name was derived from the spectrum by reversing the first four letters of Spectrum. The speech signal is represented as a convolution between slowly varying vocal tract impulse response and quickly varying glottal pulse .The goal is to separate these two parts. In the time domain, convolution becomes multiplication. Hence, by taking the inverse FFT or DCT of the logarithm of the magnitude

spectrum, the glottal pulse and the impulse response can be separated [9].

F. Mel Frequency Cepstral Coefficients

The log Mel spectrum is converted back to time. The result is called the Mel frequency cepstrum coefficients (MFCCs). Because the Mel spectrum coefficients are real numbers (and so are their logarithms), they may be converted to the time domain using the Discrete Cosine Transform (DCT). The MFCCs may be calculated using this equation [10][11].

$$C_n = \sum_{k=1}^{K} (logS_k) \left[n \left(k - \frac{1}{2} \right) \frac{\pi}{2} \right] \quad (5)$$

Where n=1,2,…K

The number of Mel cepstrum coefficients, K, is typically chosen as 20. The first component C_0 is excluded from the DCT since it represents the mean value of the input signal which carries little speaker specific information. This set of coefficients is called an *acoustic vecto*r. These acoustic vectors can be used to represent and recognize the voice characteristics of the speaker [4]. Therefore each input utterance is transformed into a sequence of acoustic vectors. These acoustic vectors can be used to represent and recognize the voice characteristic of a speaker.

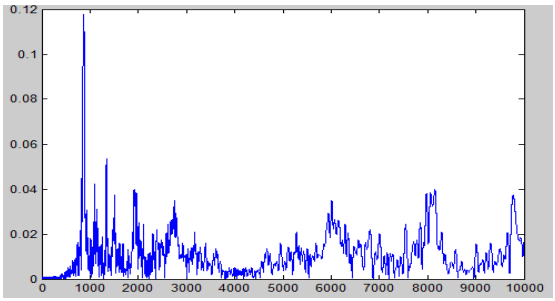

Figure 7. Mel Warped Signal

III. VECTOR QUANTIZATION

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that is generated from the training mode is impossible, since these distributions are defined over a high-dimensional space. It is often easier to start by quantizing each feature vector to one of a relatively small number of template vectors, with a process called vector quantization. The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features.

The training features are clustered to form a codebook for each speaker [13]. In the recognition stage, the data from the tested speaker is compared to the codebook of each speaker and the distance is measured to identify the speaker. These differences are then use to make the recognition decision.

The problem of speaker recognition belongs to a much broader topic in scientific engineering called *pattern recognition*. The goal of pattern recognition is to classify objects of interest into one of a number of categories or classes. The objects of interest are generically called *patterns* and in our case are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the above. The classes here refer to individual speakers. Since the classification procedure in this case is applied on extracted features, it can be also referred to as *feature matching*.

A. Codebook Formation

Figure 8 shows a conceptual diagram to illustrate the recognition process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from speaker 1 while the triangles are from the speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors. The centroids are shown in the figure by circles and triangles for speaker 1 and 2, respectively. The distance from a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition phase, an input utterance of an unknown voice is "vector-quantized" using each trained codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

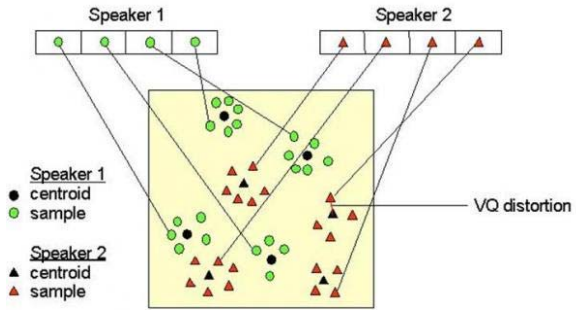

Figure 8. Codebook formation

B. Clustering the Training Vectors

The LBG (Linde, Buzo and Gray 1980) algorithm is used to cluster the training vectors. The algorithm is formally implemented by the following recursive procedure [12]:

- Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).

- Double the size of the codebook by splitting each current codebook **y***n* according to the rule where *n* varies from 1 to the current size of the codebook, and e is a splitting parameter (we choose ε = 0.01).

$$Y_n^+ = y_n(1+\varepsilon) \quad (6)$$
$$Y_n^- = y_n(1-\varepsilon) \quad (7)$$

- Nearest-Neighbor Search: for each training vector, find the codeword in the current codebook that is closest (in terms of similarity measurement), and assign that vector to the corresponding cell (associated with the closest codeword).
- Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.
- Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold.
- Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed.

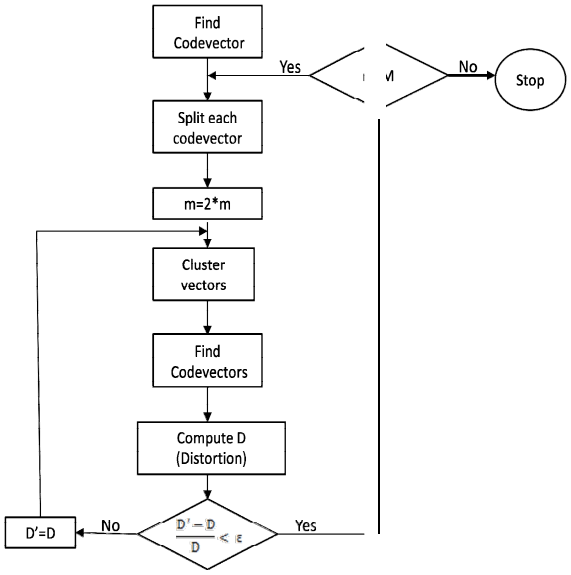

Figure 9. Flow diagram of LBG algorithm

IV. VOICE DATABASE

The voice database includes 20 samples (10 male 10 female) which are recorded for 2 speaker recognition module.

Module 1 is for a *single user* wherein a user trains his voice with an utterance and his voice is tested on the basis of the same utterance to identify the user.

Module 2 is for *multiple users* wherein multiple users train their voices and during the stage of identification the system recognizes the particular user from amongst those users.

Utterance 1: Hello.
Utterance 2: Speakers record their name and enrolment number.

Module 3 is for 'Vector Quantization' wherein the user records the utterance 'Hello' and a codebook of 10 instances of the utterance is formed. During the testing phase, the same utterance is used to identify the unknown speaker on the basis of vector quantization distortion.

V. EXPERIMENTS

A. MFCC Approach

The performance of the MFCC based speaker identification system is evaluated by performing two experiments. Following are the speech samples used for evaluation.

Single User: Hello
Multiple Users: Speakers record their name and enrolment number.

The speaker recognition system is prone to 'false rejects' and 'false accepts'. When a speaker trains his voice and the testing is carried out on a different speaker who hasn't trained his voice, the system may recognize him as an authentic speaker and validate that speaker. This is a false accept. This may happen due to environment noise, system processing noise etc. Similarly, when a user trains his voice and tests his own voice in order to validate oneself, the system might reject him on the basis of non authenticity. This is a false reject wherein the system does not recognize a valid user.

A minimum MSE threshold was maintained in both cases to calculate the number of false accepts and false rejects. If the threshold value of MSE is too large, number of false accepts may be high but at the same time, if it is too low, then the value of false rejects would be high. The following table gives the experimental results of 20 speakers.

TABLE 1 .IDENTIFICATION ACCURACY OF MFCC TECHNIQUE

Type	False accepts	False Rejects	Identification Accuracy
Single user	2	1	85%
Multiple user	2	2	80%

B. *Vector Quantization Approach*

In this approach, the number of users was gradually increased to monitor the performance of the system. The system is most accurate with the least number of speakers since the training database is small and the probability of noise being the predominant factor of non-recognition is least.

TABLE 2 .IDENTIFICATION ACCURACY OF VECTOR QUANTIZATION TECHNIQUE

Number of Speakers	Identification Accuracy
2	98%
5	96.3%
8	95.1%
12	93.8%
15	91.5%
20	90.2%

VI. RESULTS

It was seen from the experiments that because of the prominence given to energy, this approach sometimes failed to recognize the same word uttered with different energy.
The Vector Quantization technique is however, more accurate than the MFCC technique. The accuracy of Vector Quantization ranges from 90.2%-98% and may be lesser when tested on a larger database. Since there are several applications of speaker recognition systems, a comparison helps in identifying which technique is more efficient, less time consuming and has lesser computational complexity.

The MFCC approach uses only a single instance (utterance) of the speaker compared to the VQ technique where for each speaker, 10 instances are recorded in the training file. Also, as the MFCC approach takes the summation of the energy within each triangular window it would essentially give the same value of energy irrespective of whether the spectrum peaks at one particular frequency and falls to lower values around it or whether it has an equal spread within the window. Noise is a random signal and is inevitable. Hence, a 100% accurate system cannot be developed.

VII. CONCLUSION

The speaker recognition system is designed for a single user which authenticates the validity of the speaker by matching the features of his speech and for multiple users which creates a database and then takes an unknown speech as its input and investigates the extracted features of this unknown speech. Hence, the features of this unknown speech are compared to the stored extracted features for each different speaker in order to identify the unknown speaker.

The feature extraction is done by using MFCC (Mel Frequency Cepstral Coefficients) and Vector Quantization. The accuracy of VQ technique is seen to be higher than the MFCC approach and hence, better identification is carried out using the VQ technique.
This system may be extended to larger set of users and database.

REFERENCES

[1] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[2] Md. R. Hasan, M. Jamil, Md. G. Rabbani, Md. S. Rahman, “Speaker Identification using Mel Frequency Cepstral Coefficients,” Third International Conference on Electrical & Computer Engineering ICECE, Dhaka, 2004.

[3] W. Yutai, J. Xiaoqing, L. Feng,“Speaker Recognition Based on Dynamic MFCC Parameters, ”School of Information Science and Engineering, University of Jina , 2002.

[4] A. Zulfiqar, T. Enriquez, “A Speaker Identification System Using MFCC Features with VQ Technique,” Third International Symposium on Intelligent Information Technology Application, vol.3, pp.115 – 118, Mar. 2009.

[5] W.Han, C.F. Chan, C.S. Choy and K.P. Pun, “An Efficient MFCC Extraction Method in Speech Recognition,” Department of Electronic Engineering, The Chinese University of Hong Kong, Hong, IEEE – ISCAS, 2006.

[6] L. D. Alsteris and K. Paliwal, “ASR on Speech Reconstructed from short- time Fourier Phase Spectra’, School of Microelectronic Engineering Griffth University, Brisbane, Australia, ICLSP – 2004.

[7] P. Kumar and P. Rao, “A Study of Frequency-Scale Warping for Speaker Recognition”, Dept of Electrical Engineering, IIT- Bombay, National Conference on Communications, NCC 2004, IISC Bangalore, Jan 30 -Feb 1, 2004.

[8] V. Tiwari, “ MFCC and its applications in speaker recognition”, International Journal on Emerging Technologies”,vol.1, pp.19-22,Feb.2010.

[9] K. Samudravijaya, R. Madan,“ A novel approach to speaker verification”, http://speech.tifr.res.in

[10] S.H. Chen and Y.R. Luo, “Speaker Verification Using MFCC and Support Vector Machine”, Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol.1 IMECS 2009, March 18 - 20, 2009, Hong Kong.

[11] H. Lei, E.L. Gonzalo-“ Mel, Linear, and AntiMel Frequency Cepstral Coefficients in Broad Phonetic Regions for Telephone Speaker Recognition” The International Computer Science Institute, Berkeley, CA.

[12] Y. Linde, A. Buzo, R. M. Gray, “An algorithm for Vector Quantizer Design”. *IEEE Transaction on Communications*, 28: 1980, pp 84-95.

[13] Kinnunen T.and Kärkkäinen I., "Class-Discriminative Weighted Distortion Measure for VQ-Based Speaker Identification". *Joint IAPR Int. Workshop on Statistical Pattern Recognition (SPR'2002)*, Windsor, Canada, 681-688, August 2002.