International Journal of Electronics Communication and Computer Engineering
Volume 3, Issue 4, ISSN (Online) 2249 –071X, ISSN (Print): 2278 – 4209

Speaker Identification Using MEL Frequency Cepstral Coefficients and Vector Quatization

Ms. Vidya Sagvekar
P.G.Student
Vidyalankar Institute of Technology,
Mumbai
vidya.sagvekar@gmail.com

Prof. Maruti Limkar
Assistant Professor
Terna College of Engineering,
Navi Mumbai
maruti_limkar@rediffmail.com

Prof. B. Rama Rao
Professor
of Vidyalankar Institute of Technology,
Mumbai
b.ramarao@vit.edu.in

***Abstract*— In this paper, we build a VQ-based speaker identification system. The speaker identification, which consists of mapping a speech signal from an unknown speaker to a database of known speakers, i.e. the system has been trained with a number of speakers which the system can recognize. Here developed, Text-dependent systems require the speaker to utter a phrase like digits zero to nine in an isolated way. Speaker identification has been done successfully using Vector Quantization (VQ). This technique consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. Using training data these features are clustered to form a speaker-specific codebook. In the recognition stage, the test data is compared to the codebook of each reference speaker and a measure of the difference is used to make the recognition decision. The paper shows identification rate when triangular, or rectangular or hamming window as well as codebook size increases, the identification rate for each of the three cases increases.**

***Keywords* - Mel Frequency Cepstrum Coefficient;Speaker Identificaiton;Vector Quatization.**

I. Introduction

In speaker identification, the task is to use a speech sample to select the identity of the person that produced the speech from among a population of speakers. In speaker verification, the task is to use a speech sample to test whether a person who claims to have produced the speech has in fact done so [1].

The basic structure of SI system (SIS) is shown in Figure 1. Training and Testing are two important phases of a SI system. In the training phase a model for each speaker is constructed and in testing phase an unknown voice is compared with the model of each speaker to find the identity of true speaker.

In this paper, modelling technique vector quantization has been used with MFCC based voice feature vectors. MFCC mimic the behavior of human's ear and they perform well in speaker identification system [7]. Voice database is recorded at two different frequencies to show that sampling frequency influence the identification accuracy.

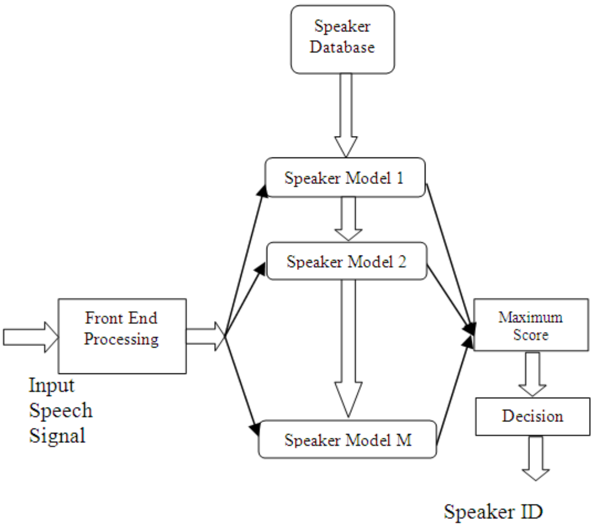

Fig.1. Basic Structure of Speaker Identification

The effect of number code vectors in VQ codebooks is also analyzed with the help of developed application.

II. Speech Feature Extraction

The purpose of this module is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate). The speech signal is a slowly time varying signal (it is called quasi-stationary). When examined over a sufficiently short period of time (between 5 and 100 ms), its characteristics are fairly stationary. However, over long periods of time (on the order of 0.2s or more) the signal characteristics change to reflect the different speech sounds being spoken.

Therefore, short-time spectral analysis is the most common way to characterize the speech signal. A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC is perhaps the best known and most popular, and this feature has been used in this paper. MFCC's are based on the known variation of the human ear's critical bandwidths with frequency. The MFCC technique makes use of two types of filter, namely, linearly spaced filters and logarithmically spaced filters. To capture the phonetically important characteristics of speech, signal is expressed in the Mel frequency scale. This scale has a linear frequency spacing below 1000 Hz and a logarithmic spacing above

1000 Hz. Normal speech waveform may vary from time to time depending on the physical condition of speakers' vocal cord. Rather than the speech waveforms themselves, MFFCs are less susceptible to the said variations [1, 4].

The original analogue signal which to be used by the system in both training and testing is converted from analogue to discrete. The sample rate, Fs used was 16KHz.An example of a signal in waveform sampled is given in given figure 2.

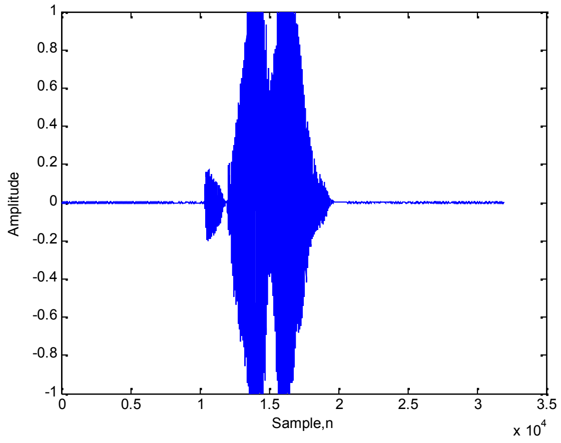

Fig.2. Sampled signal, utterance of 'zero' in waveform

A. *Preprocessing*

There is a need for spectrally flattern the signal. The preemphasizer, often represented by a first order high pass FIR filter is used to emphasize the higher frequency components. The composition of this filter in time domain is described in Eq.1

$$h(n)=\{1,-0.95\} \qquad (1)$$

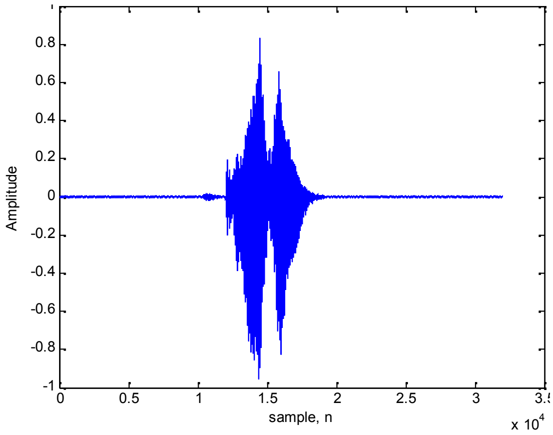

Fig.3. Preemphasized signal

B. *MFCC*

In previous section discussed pre-processing steps. In this section computation of MFCC is explained as follow:

Frame blocking

The objective with frame blocking is to divide the signal into a matrix form with an appropriate time length for each frame. Signal is blocked into frames of N samples, with adjacent frames being separated by M. The first frame consists of the first N samples. The second frame begins M samples after the first frame and overlaps it by N-M samples and so on. This process continues until all the speech is accounted for within one or more frames. Due to the assumption that a signal within a frame of 30ms is stationary and a sampling rate at 16000Hz will give the result of a frame consists of N=256 and M=100.

$$\mathrm{Yi(n) = Xi(n)w(n)}, \qquad 0 \le n \le N-1 \;\ldots\ldots(2)$$

In this paper for windowing used different window such as rectangular, triangular and hamming window. Typically the Hamming window is used, which has the form.

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \ldots\ldots(3)$$

Fast Fourier Transform (FFT)

The next processing step Windowing using Hamming window is Fast Fourier transform, which converts each frame of N samples {Xn}, as follow:

$$\mathrm{Xk} = \sum_{n=0}^{N-1} \mathrm{Xn}e^{-j2\pi kn/N}, \qquad k=0, 1, 2,\ldots., N-1\ldots\ldots.(4)$$

In general Xk's are complex numbers and consider their absolute values(frequency magnitudes).The resulting sequence {Xk} is interpreted as follow: positive frequencies 0 ≤ f ≤ Fs/2 correspond to values $0 \le n \le \frac{N}{2} - 1$,while negative frequencies –Fs/2 < f < 0 correspond to $\frac{N}{2}+1 \le n \le N-1$Here,Fs denotes the sampling frequency.

Mel-Frequency Wrapping

The speech signal consists of tones with different frequencies. For each tone with an actual frequency, f ,measured in Hz , a subjective pitch is measured on the Mel scale. The mel frequency scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. Following formula to compute the mels for a given frequency f in Hz[5].

$$\mathrm{mel(f)=2595*log(1+f/700)} \qquad (5)$$

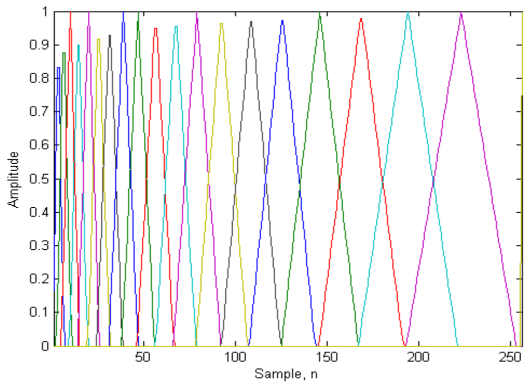

Fig.4. Mel Scale

The practical warping is done by using a triangular Mel scale filter bank according to figure 4 which handles the warping from Frequency in Hz to frequency in mel scale.

Cepstrum

In the final step, the log mel spectrum has to be converted back to time. The result is called the mel frequency cepstrum coefficients (MFCCs).The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers they may be converted to the time domain using the Discrete Cosine Transform (DCT).The MFCCs may be calculated using this equation 6.

$Cn = \sum_{k=1}^{K}(\log Sk)\left|n\left(k-\frac{1}{2}\right)\frac{\pi}{k}\right]$ Where n=1, 2,.., K (6)

The number of mel cepstrum coefficients, K, is typically chosen as 20. The first component, C_0 is excluded from DCT since it represents the mean value of the input signal which carries little speaker specific information. By applying the procedure described above, for each speech frame of about 30 ms with overlap, a set of mel frequency cepstrum coefficient is computed. This set of coefficients is called an acoustic vector. These acoustic vectors can be used to represent and recognize the voice characteristic of the speaker. The next section describes how these acoustic vectors can be used to represent and recognize the voice characteristic of a speaker.

III. Feature Matching

Feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modelling (HMM), and Vector Quantization (VQ). The VQ approach has been used here for its ease of implementation and high accuracy.

VECTOR QUATIZATION:

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 center called a codeword. The collection of all codeword is called a codebook.

Clustering The Training Vectors

After the enrolment session, the acoustic vectors extracted from input speech of each speaker provide a set of training vectors for that speaker. As described above, the next important step is to build a speaker-specific VQ codebook for each speaker using those training vectors. There is a well-know algorithm, namely LBG algorithm [Linde, Buzo and Gray, 1980], for clustering a set of *L* training vectors into a set of *M* codebook vectors. The algorithm is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors (hence, no iteration is required here).
2. Double the size of the codebook by splitting each current codebook $\mathbf{y}_n$ according to the rule

$$\mathbf{y}_n^+ = \mathbf{y}_n(1+\varepsilon)$$
$$\mathbf{y}_n^- = \mathbf{y}_n(1-\varepsilon)$$

Where *n* varies from 1 to the current size of the codebook, and ε is a splitting parameter (we choose ε=0.01).

3. 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).
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.
5. Iteration 1: repeat steps 3 and 4 until the average distance falls below a preset threshold
6. Iteration 2: repeat steps 2, 3 and 4 until a codebook size of *M* is designed.

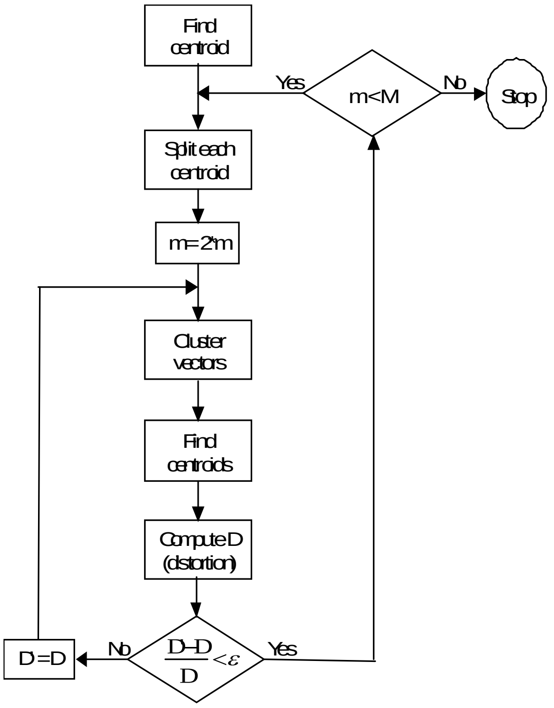

Fig.5. Flow diagram of the LBG algorithm (Adapted from Rabiner and Juang, 1993)

Intuitively, the LBG algorithm designs an *M*-vector codebook in stages. It starts first by designing a 1-vector codebook, then uses a splitting technique on the codewords to initialize the search for a 2-vector codebook, and continues the splitting process until the desired *M*-vector codebook is obtained.

Figure 5 shows, in a flow diagram, the detailed steps of the LBG algorithm. “Cluster vectors” is the nearest-neighbor search procedure which assigns each training vector to a cluster associated with the closest codeword. “Find centroids” is the centroid update procedure. “Compute D (distortion)” sums the distances of all training vectors in the nearest-neighbor search so as to determine whether the procedure has converged.

IV. Result

The system has been implemented in Matlab7 on windows7 platform. The result of the study has been presented in Table 1. The speech database consists of 15 speakers. Here, identification rate is defined as the ratio of the number of speakers identified to the total number of speakers tested.

Table 1: Identification rate (in %) for different windows [using mel scale]

Code book size	Rectangular	Triangular	Hamming
1	54.55	63.64	72.73
2	54.55	81.82	81.82
4	72.73	81.82	81.82
8	81.82	90.91	100
16	90.91	100	100
32	90.91	100	100

V. Conclusion

The result obtained using MFCC and VQ are appreciable. MFCC for each speaker were computed and vector quantized for efficient representation. The table clearly shows that as codebook size increases, the identification rate for each of the three cases increases, and when codebook size is 16, identification rate is 100% for both the triangular and hamming windows.The study reveals that as number centroids increases, identification rate of the system increase. It has been found that combination of Mel frequency and Hamming window gives the best performance. It also suggests that in order to obtain satisfactory result, the number of centroids has been increased as the number of speaker increases.

References

[1] Lawrence Rabiner and Biing-Hwang Juang, “fundamental of Speech Recognition”, Prentice-Hall, Englewood Cliffs, N.J., 1993.

[2] “F.Soong, E. Rosenberg, B.Juang and L. Rabiner,” A Vector Speech Recognition”, AT & T Technical Journal,vol.66,March/April 1987,pp.14-26

.[3] Jr.J.D. Hansen, J. and Proakis, J. Discrete Time Processing of speech signals, second ed. IEEE Press, New York,2000.

[4] Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md., Saifur Rahan,” Speaker Identification using Mel Frequency Cepstral coefficiens”, Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology,3 rd International conference on Electrical and computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.

[5] R. M. Gray, ``Vector Quantization,'' IEEE ASSP Magazine, pp. 4--29, April 1984.

[6] Y. Linde, A. Buzo & R. Gray, “An algorithm for vector quantizer design”, IEEE Transactions on Communications, Vol. 28, pp.84-95, 1980.

[7] O. D. Richard, E. H. Peter, G. S. David, “Pattern Classification”, 2nd ed, John Willey Co., Nov 2000.

[8] S. Nakagawa, K. Asakawa and L. Wang., "Speaker Recognition by Combining MFCC and Phase Information", Interspeech, 2007.

Author’s Profile

Vidya Sagvekar

has received B.E. in Electronics Engg. from Pune University in 2004, presently she is pursuing M.E. (Electronics and Telecom) from Mumbai University. She has a teaching experience of more than 4 years. She is Assistant Professor in Electronics Department in KJSIEIT, Mumbai University. Her area of interest includes Speech processing: Speech and Speaker Recognition .
Email ID : vidyarsagvekar@gmail.com

Maruti Limkar

has received B.E. in Electronics Engg. from Marathwada University in 1990, M.E. (Power Electronics) from Gulbharga University in 2003. He has a teaching experience of more than 7 years. He is Assistant Professor in Electronics Department in Terna College of Engg., Mumbai University. His area of interest include Speech processing: Speech and Speaker Recognition
Email ID : maruti_limkar@rediffmail.com

Prof. Rama Rao

has received B.S, M.S. in Electronics Engg., Ph.D. He has a teaching experience of more than 16 years. He is Professor in Electronics and Telecommunication Department in Vidyalankar Institute of Technology, Mumbai University. His area of interest Digital Signal Processing, filter theory, Speech Processing: Speech and Speaker Recognition.
Email ID : b.ramarao@vit.edu.in