IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

# A Survey on Speaker Recognition with Various Feature Extraction Techniques

**Parmar Dharmistha R**

Department of Computer Engineering, BVM, VVnagar, Anand, Gujarat, India

DOI: https://doi.org/10.26438/ijcse/v7i2. 884887 | Available online at: www.ijcseonline.org

Accepted: 11/Feb/2019, Published: 28/Feb/2019

***Abstract*:-** Speech processing is one of the important application area of digital signal processing. For this purpose, speaker recognition is dominating today's world. Speaker recognition is a process of speaker identification and speaker verification refers to specific tasks. Speaker recognition is the process of identifying a speaker by his/her speech samples. By extracting the speaker-specific features from the speech samples, the recognition task can be done. Speaker recognition technique is one of the most helpful recognition techniques in today world. It is very important to efficiently work without fail of Recognition system and identify correct person. Speaker recognition is to extract, characterize and recognize the information about speaker identity. This system involves many stages with multiple techniques for each. In this paper, the performance of Mel Frequency Cepstral Coefficient (MFCC), VQ vector quantization and Linear Prediction Coding (LPC) speaker recognition system using method. It is found that the MFCC is offer better recognition rate as contrasted to BFCC using VQ vector quantization as speaker modeling technique. The best technique in each stage makes the system more accurate and efficient.

***Keywords*:** Speaker Recognition, Speaker identification and verification, vector quantization, Mel Frequency Cepstral Coefficient

## I. INTRODUCTION

Voice signal is the most important way of people's communication. Speech signal is play a crucial role in communication around the world. Speaker recognition is derived from the biometrics technology that has been swiftly raised in modern era in many core areas for instance information security and atomization of business. A speaker recognition system is based on two key sections, feature extraction and speaker modeling. In feature extraction unique characteristics of the speech signals such as pitch frequency, loudness etc has been fetched from the speech signal. These unique characteristics are different for each speaker. Different types of techniques used to extract the characteristics of the speech signal. All speaker recognition systems contain two main modules: feature extraction and feature matching. 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 feature of a known speaker with the unknown speaker.

## II. RELATED WORK

As we have shown detailed literature survey, we can conclude that speaker recognition is to extract, characterize and recognize the information about speaker identity. MFCC (Mel-Frequency Cepstral Coefficient) technique for speaker recognition. Feature vectors from speech are extracted by using Mel-frequency cepstral coefficients.

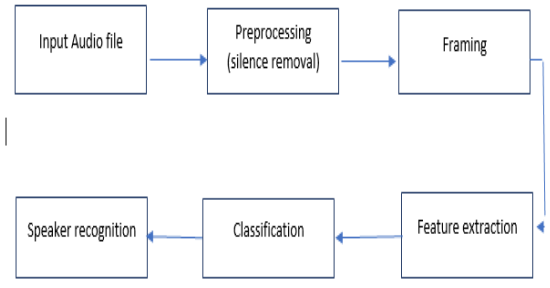


Fig.1 Proposed Model

Preprocessing (silence removal):

**Zero Crossing Rate**:
A zero crossing is said to have occurred in a signal when its waveform crosses the time axis or changes its algebraic

 

sign. For a discrete time signal with zero crossing rate(ZCR) in zero crossing/sample and a sampling frequency of Fs, the frequency Fo is given as

Fo = (ZCR*FS)/2

The speech signal contains most of its energy in voiced signals at low frequencies. For unvoiced sounds, the broadband noise excitation takes place at higher frequencies due to the short length of the vocal tract. Therefore a high and a low ZCR relates to unvoiced and voiced speech respectively.

**Short Time Average Energy and Magnitude**

The short-term processing technique provide signals in the following form

$$Q(n) = \sum_{m=-\infty}^{\infty} T[s(m)]w(n-m) \ldots\ldots\ldots\ldots (1)$$

T[s (m)] is a transformation, which is applied to the speech signal and the signal is thereafter weighted by a window w(n). the summation of T[s(m)] convolved with w(n) represents contain property of the signal that is averaged over the window duration.

The output in the equation (1) will be representing short time energy or amplitude if the transformation T is squaring or absolute magnitude operation. The energy indicates high amplitudes as the signal is squared for calculating Q (n). Such techniques enable the segmentation of speech into smaller phonetic units e.g. phonemes or syllables. There is a large variation in the amplitude between the voiced and the unvoiced segments. Also, the variation between phonemes with different manners of articulation is small.

Short Time Autocorrelation:

The autocorrelation function for a discrete time signal is given as

$$\phi(k) = \sum_{m=-\infty}^{\infty} s(m)y(m-k)$$

This function measures the similarity of two signals s (n) and y (n), by summing the product of a signal sample and a delayed sample of another signal. The short time autocorrelation function is obtained by windowing s (n) and applying the autocorrelation given by equation (3), which results in

$$R\square(k) = \sum_{k=1}^{p} s(m)w(n-m)s(m-k)w(n-m+k)$$

This short time auto correlation function provides information about the harmonic and formant amplitudes of s (n) and also indicates its periodicity. Thus pitch estimation and voiced/unvoiced speech detection can be carried out using this feature.

Framing:

The speech signal is segmented into small duration blocks of 20-30 ms knowns as frames. Voice signal is divided into N samples and adjacent frames are being separated by M(M<N) M=100,N=256

## III. METHODOLOGY

**FEATURE EXTRACTION**

There are different techniques are used for feature extraction like Linear Prediction Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC), BFCC (Bark Frequency Cepstral Coefficient), and VQ (Vector Quantization).

**MEL-FREQUENCY CEPSTRAL COEFFICIENTS**

There are derived from a type of cepstral representation of an audio clip. The difference between the cepstrum and Mel-frequency cepstrum is that in the MFC, the frequency bands are positioned logarithmically (on the mel scale) which approximatcs 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 of computing the MFCCs.

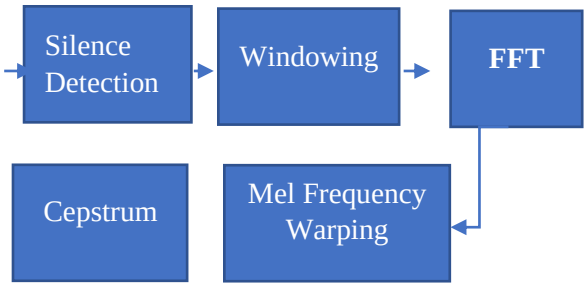


Fig. 2 Feature Extraction Steps

**BARK FREQUENCY CEPSTRAL COEFFICIENT (BFCC)**

Bark frequency Cepstral Coefficient (BFCC) is another approach used in speech signal for feature extraction. The bands of frequency in BFCC are almost linear to 500 Hz and which is also in logarithmic illustration is computed via mentioned formula.

Bark(f) = {13 atan(0.76f/100)+ 3.5}

 

The Methods of computing BFCC and MFCC are similar in nature as revealed by given block diagram.

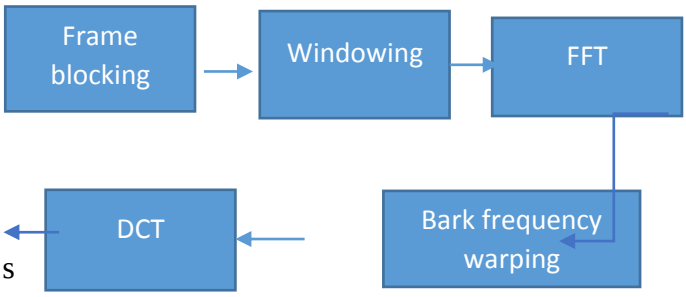


Fig. 3 Feature Extraction Steps

**SPEAKER MODEL CREATION**

**Vector Quantization (VQ)**

It is the ability of a speaker recognition system to evaluate possibility allocations of the calculated feature vectors. Also it is not conceivable to keep every single generated vector by training-mode; meanwhile such allocations are well-defined over a high-dimensional space. It seems easy to initiate that every single feature vector is quantized to one of smallest part of template vectors. Vector-quantization is the techniques of mapping vectors from large spaced vector to different regions in same space. Every single region is termed as a cluster which symbolized through its center known as a code word . Further the codebook is the sum of individual code words, which is estimated by training material.

Here the clustering of individuals training acoustic vectors generates the VQ code for every single identified speaker for this the K-Mean algorithm is utilized, in which speech samples is divided into clusters (K).

Following recursion steps are needed to implement the k-Mean algorithm:

Step_1: k points are divided into space to make k clusters

Step_2: After making k clusters, we compute the centroids of each cluster by computing the mean of feature vectors

Step_3: Place the feature vectors near each centroid by computing minimum distance and make their groups.

Repeat step 2 and 3 until no centroids are moveable

Find the distance from a centroid of test signal to closest object in our training database. Find the shortest distance to recognize the speaker.

**Formant extraction through LPC**

Formant frequency is defined as the resonance frequencies of the vocal-tract. The formant structure of the vowel is directly related to the unique shape of the vocal tract and supplies important information about the speaker's identity. Formant frequencies are obtained from the Frequency spectrum of the voiced speech. Discrete Fourier Transform [DFT] and Fast Fourier Transform [FFT] algorithms are used for computing the frequency spectrum of the sampled data.

LPC based formant estimation method is more accurate and efficient. Hence in the proposed speaker verification system LPC based formant estimation method has been employed. The block diagram of LPC method is shown in below figure 2.2. The speech signal is multiplied with the hamming window to reduce the edge effect present in the start and end of the speech signal. The windowed speech signal then used to compute the linear predictive coefficients using the LPC method. Then these coefficients are used to find the LP smoothed spectrum which is has spectral peak amplitude and frequency. Formants are extracted by detecting the peaks from the LP smoothed spectrum.

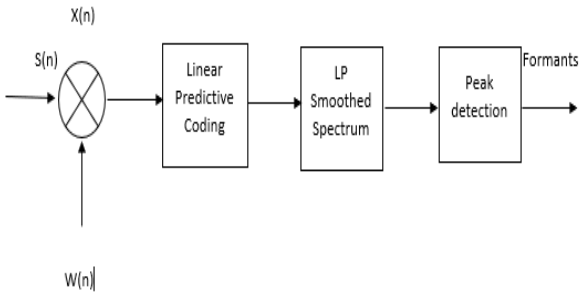


Fig 4 Block diagram of formant extraction using LPC

Speaker recognition is the process of identifying a speaker by his/her speech samples. By extracting the speaker-specific features from the speech samples, the recognition task can be done. The formant estimation of speech sample of specific speaker is important for feature extraction in speaker recognition, because the formants are unique and reflect the vocal tract information of a speaker. In the proposed work Linear Predictive Coding (LPC) based formant estimation method is employed as it is more efficient and accurate. The formant frequencies of each speaker are evaluated and stored in the database and test speaker speech sample is also processed in the same way to find the formants and finally they are compared with the formants stored in the database to perform the speaker verification and recognition.

**RESULTS**

**MFCC Approach:**

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 | 83% |

 

| | | | |
|---|---|---|---|
| Multiple user | 2 | 2 | 80% |

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

Single User: Hello

Multiple User: Speakers record their name and enrolment number.

The speaker recognition system is '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 where in 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.

**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% |

LPC: In the result speaker verification and the speaker recognition rate for each speaker is calculated. The CVC stimuli such as 'bat', 'cat', 'rat', 'hat' and 'sat' are recorded for the 4 different male speakers and stored in the database. The recognition rate for each speaker is shown in below table

| | Total No of CVC | No of CVC recognized | Recognition Rate(%) |
|---|---|---|---|
| | words | | |
| Speaker 1 | 15 | 10 | 66.66% |
| Speaker 2 | 15 | 7 | 46.66% |
| Speaker 3 | 15 | 7 | 46.66% |
| Speaker 4 | 15 | 9 | 60% |

Recognition rate for each speaker using LPC based formant estimation methods

## CONCLUSION

In this paper, there is discussion on the speaker recognition that can be used for many speech processing applications specially for security and authentication. There are most commonly used feature extraction techniques are discussed from that MFCC are widely used. Also discuss different feature classification techniques for speaker recognition.

## REFERENCES

[1] Mahaveer Chougala1' Novel Text Independent Speaker Recognition Using LPC Based Formants' 978-1-4673-9939-5/16/$31.00 ©2016 IEEE

[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] 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.

[4] 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.

[5] Dorra Gargouri, Med Ali Kammoun, "A Comparative Study of Formant Frequencies Estimation Techniques", Proceedings of the 5th WSEAS International Conference on Signal Processing, Istanbul, Turkey, May 27-29, 2006.

**Authors Profile**.

Miss Parmar Dharmistha R.Pursued Bachelor of Engineering in Computer Science engineering from Parul institute of technology Limda, Waghodia, Gujarat, India in 2017. She is currently pursuing in master of Technology in course of computer engineering(Software engineering) from Birla Vishvakarm Mahavidhyalaya (BVM),VVnagar, Anand, Gujarat, India. Her main research work focuses on image processing, speaker recognition.

 