REPORT ON THE SEMINAR TOPIC

# COMPARISON BETWEEN DIFFERENT FEATURE EXTRACTION TECHNIQUES FOR SPEAKER IDENTIFICATION SYSTEM

Delivered by Group: - 78

| Student name (s) | Exam Seat no. (s) |
|---|---|
| AYUSHI VAISHNAVA (4108) | B120203014 |
| BORSE ROSHNI RAJENDRA (4172) | B120203031 |
| SARODE PALLAVI MAROTI (4173) | B120203189 |

**in partial fulfillment for the award of the degree of**

**Bachelor of Engineering in**

**ELECTRONICS AND TELECOMMUNICATION** of
**SAVITRIBAI PHULE PUNE UNIVERSITY.**

**Under the guidance of**
Prof.(Mrs)B.V. Pathak

**Sponsored by: -** Self Sponsored
In the **Department of Electronics and Telecommunication of**

**CUMMINS COLLEGE OF ENGINEERING FOR WOMEN , KARVENAGAR,PUNE- 411052.**
**Academic year**

**2015- 16**

**Project title : -**

**COMPARISON BETWEEN DIFFERENT FEATURE EXTRACTION TECHNIQUES FOR SPEAKER IDENTIFICATION SYSTEM**

**Subject area :-**

**DIGITAL SIGNAL PROCESSING**

**Nature of the Project : -**

**SOFTWARE**

# CERTIFICATE

This is to certify that

| | |
|---|---|
| AYUSHI VAISHNAVA | B120203014 |
| BORSE ROSHNI RAJENDRA | B120203031 |
| SARODE PALLAVI MAROTI | B120203189 |

**Have successfully delivered a SEMINAR on their PROJECT TOPIC**

Comparison between Different Feature Extraction Techniques for Speaker Identification System

**In partial fulfillment for the award of the degree of**

**Bachelor of Engineering in ELECTRONICS AND TELECOMMUNICATION of SAVITRIBAI PHULE PUNE UNIVERSITY,**

**In**

**CUMMINS COLLEGE OF ENGINEERING FOR WOMEN, KARVENAGAR,**

**PUNE-52.**

| **Internal guide** | **Head of the Department** | **Principal** |
|---|---|---|
| Prof.(Mrs)..B.V.Pathak | Dr.(Mrs).P. Mukherji | Dr.(Mrs).M. Khambete |

# Acknowledgement

We would like to take the opportunity to express our gratitude to respected supervisor Prof.Mrs. B.V.Pathak for her patience, support and guidance. She not only gave us time but also proper guidance and valuable advice whenever we needed it. We faced with some difficulties. Her comments and guidance helped us a lot in preparing our project. We would also like to thank Mr. M.S. Patankar, for his support and valuable guidance.

We are also thankful to our teacher, who inspired me in every step. We are also thankful to our classmates and seniors who helped us in a number of ways by providing various resources and moral support.

# ABSTRACT :

Speech processing is emerged as one of the important application area of digital signal processing .Various fields for research in speech processing are speech recognition, speaker recognition, speech synthesis ,speech coding etc. The objective of automatic speaker recognition system is to extract, characterize and recognize the information about speaker's identity.

Our proposed work consists of truncating a recorded voice signal, framing it, passing it through a window function, calculating the Short Term FFT, extracting its features and matching it with a stored template .A comprative study of different feature extraction techniques like Perceptual Linear Prediction (PLP), Linear Predictive Coding (LPC) and Mel frequency Cepstral Coefficients (MFCC) is carried out. Vector Quantization(VQ) is used for codebook generation and Euclidian Distance are used for matching purposes.

# Index

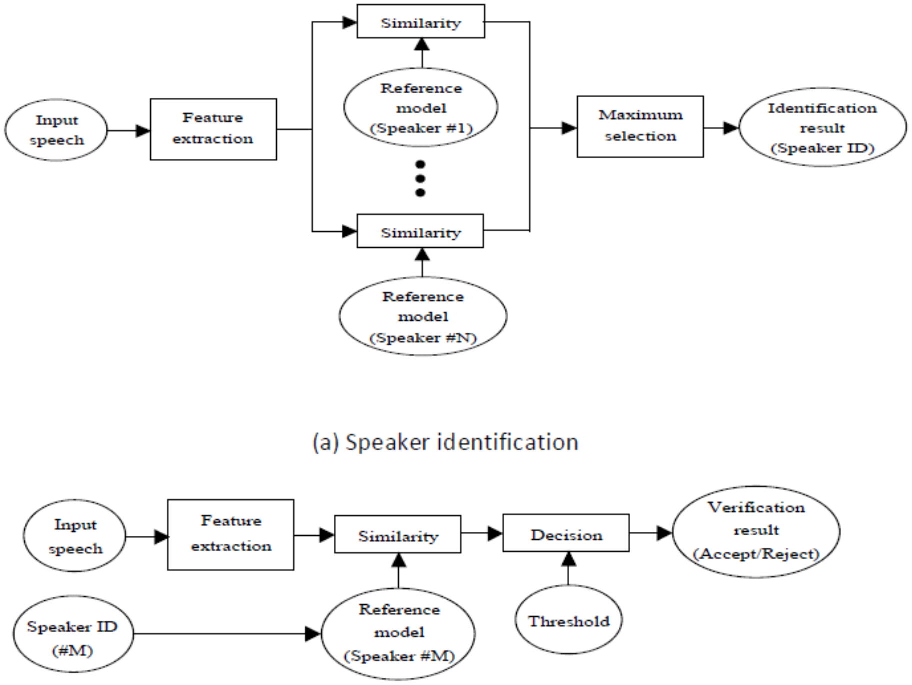
# LIST OF FIGURES:

# LIST OF TABLE:

# SPEAKER IDENTIFICATION : A REVIEW

## 1. INTRODUCTION

The SI can be classified into two ways: speaker identification & speaker verification.The speaker identification involves the comparing the incoming speaker information with speaker models that are already stored during the training phase, and identifying the required speaker by finding out the closest model to that of incoming speech. Hence, speaker identification is one-to-many process. But in the speaker verification, the incoming speaker information is directly matched to the claimed identity and match score is compared with threshold, than decision is made for verification. The speaker verification is one-to-one process. The figure 2.1 describes the speaker identification and verification basic structures of SR systems.


Input speech
Feature extraction
Similarity
Reference model (Speaker #1)
Similarity
Reference model (Speaker #N)
Maximum selection
Identification result (Speaker ID)
Input speech
Feature extraction
Similarity
Decision
Verification result (Accept/Reject)
Speaker ID (#M)
Reference model (Speaker #M)
Threshold

(a) Speaker identification

(b) Speaker verification

Fig 1 1.a Speaker Identification and 1.b Speaker Verification

Speaker identification can be further classified into closed-set or open-set modes.

Closed set speaker identification refers to the case where the speaker is a member of the set of *N* enrolled speakers. In open-set speaker identification, the speaker may also be from outside the set of *N* enrolled speakers.

## 1.1 Variants of speaker recognition :

Each Speaker recognition system has two phases: Enrollment and verification. During enrollment, the speaker's voice is recorded and typically a number of features are extracted to form a voice print. In the verification phase, a speech sample or "utterance" is compared against a previously created voice print. For identification systems, the utterance is compared against multiple voice prints in order to determine the best match while verification systems compare an utterance against a single voice print.Because of the process involved, verification is faster than identification.

Speaker recognition systems fall into two categories: text-dependent and text-independent.

### 1.1.1 Text-Dependent :

If the text must be the same for enrollment and verification this is called text-dependent recognition. In a text-dependent system, prompts can either be common across all speakers (e.g.: a common pass phrase) or unique. In addition, the use of shared-secrets (e.g.: passwords and PINs) or knowledge-based information can be employed. in order to create a multi-factor authentication scenario

### 1.1.2 Text-Independent:

Text-independent systems are most often used for speaker identification as they require very little if any cooperation by the speaker. In this case the text during enrollment and test is different. In fact, the enrollment may happen without the user's knowledge, as in the case for many forensic applications. As text-independent technologies do not compare what was said at enrollment and verification, verification applications tend to also employ speech recognition to determine what the user is saying at the point of authentication.In text independent systems both acoustics and speech analysis techniques are used .

## 1.2 Description of Speaker Identification system:

The general approach to SI consists of following steps: digital speech data acquisition, Feature extraction, pattern matching, making an accept/reject decision,and enrollment to generate speaker reference models.There are two phases for speaker recognition: training phase for taking speaker information, enrollment generating and testing phase for the verification &identification. This is described simply as below:

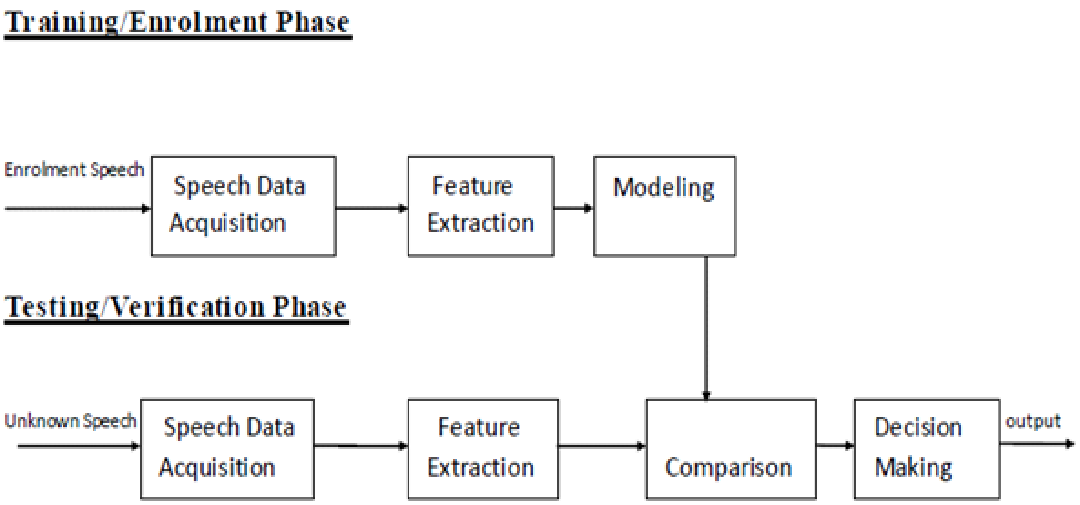


Fig 2 : Block Diagram of SI system

Speaker recognition system may be considered to consist of four stages. They include: speech analysis (digital speech acquisition), feature extraction, speaker modeling and speaker testing (pattern matching and decision making). Speech analysis involves analyzing the speech signal using suitable frame size and shift for the feature extraction. Feature extraction involves extracting speaker –specific features from the speech signal at reduced data rate. The extracted features are further combined using modeling techniques to generate speaker models. The speaker models are then tested using the features extracted from the test speech signal. The improvement in the performance can be achieved by employing new or improved techniques in one or more of these stages.

## 1.3 Vocal Tract Features:

The vocal tract is generally considered as the speech production organ above the vocal folds. Vocal tract includes the following: laryngeal pharynx (beneath epiglottis),oral pharynx (behind the tongue, between the epiglottis and velum), oral cavity (forward of the velum and bounded by the lips, tongue, and palate), nasal pharynx and the nasal cavity. An adult male vocal tract is approximately 17 cm long. The vocal folds are stretched between the thyroid cartilage and the arytenoid cartilages. The area between the vocal folds is called the glottis. As the acoustic wave passes through the vocal tract, its frequency content (spectrum) is altered by the resonances of the vocal tract. Vocal tract resonances are called formants. Thus, the vocal tract shape can be estimated from the spectral shape (e.g. formant location and spectral tilt) of the voice signal. Voice verification systems typically use features derived only from the vocal tract. Vocal tract features gets affected by the health, mood, emotional and expressive behavior of the speaker.

## 1.4 Excitation Source Features:

The human vocal mechanism is driven by an excitation source, which also contains speaker-dependent information. The excitation is generated by airflow from the lungs, carried by the trachea (also called the “wind pipe”) through the vocal folds (or the arytenoid cartilages). The excitation can be characterized as phonation, whispering, frication, compression, vibration, or a combination of these. The excitation source features are robust features against the expressive behavior of the speaker and these makes him/her unique in the world. But their extraction is more difficult than the vocal tract features. The speaker-specific resonant structure information is due to the shape and size associated with pharynx, oral and nasal cavities and also dynamics associated with the oral cavity. Thus the speaker -specific information from each of them is different. Human beings recognize people from the source characteristics such as glottal vibrations and prosodic features such as intonation and duration.

### 1.4.1. Pitch :

Pitch information also contributes to the uniqueness of the speaker's voice at different levels. Pitch frequency is the acoustic correlate of the rate of vibration of vocal folds. And this uniqueness of rate of vibrations of the vocal folds is due to the differences in the size of vocal folds among the speakers. That's why female and male voices are different due to difference in their sizes of vocal folds. The pitch information can be extracted by methods like zero-crossing, cepstral methods, group delays etc. Several methods have been proposed for making distinction among the speakers using pitch information in the past.

### 1.4.2 Intonation :

The speaking style determines the pitch pattern of the utterance or the variation of pitch frequency with time, called intonation. The intonation is used in text - dependent speaker verification. In this, the similarities of intonation pattern of reference and the utterances are captured by using DTW method.

### 1.4.3 Jitter :

Jitter is defined as the perturbation of pitch or fundamental frequency. The jitter values are expressed as a percentage deviation of the pitch period. Jitter appears very significant source in the speech signal.

## 1.10 Glottal Flow Derivative :

Glottal flow derivative is one of the glottal source features, which is important to generate natural sounding synthetic speech. It is also useful for characterizing different voices. One of one cycle of the derivative of the glottal flow using seven parameters.

The function of this feature extraction stage is to extract maximum amount of information from those samples with reduced data rate and convert it into vector form called as feature vector. These vectors are put in a k-dimensional space called as feature space, where k is the dimension of the feature vector.

# 2.LITERATURE SURVEY:

Feature extraction techniques evolved starting from the very basic techniques based on long and short term spectral averages [5], predictive coefficients (LPC, PLP) to widely used filterbank coefficients (MFCC). D.A. Reynolds (1994) has compared MFCC, LFCC, LPCC, PLPC techniques with each other [3]. MFCC proves to be better than all other techniques for lower filter orders. LPCC, PLPC gives better performance with increasing filter order but performance degrades in linear coefficients (LFCC) because it gives equal detail to entire band of the signal hence highlights the superfluous information also [3].MFCC gives more detail to lower frequencies only whereas LFCC gives equal detail to all the frequencies and captures the high formant frequencies[4].

Performance of MFCC based system has been improved by using modified window function in MFCC technique [6]. This system represents power spectrum of the original spectrum as well as its derivative and also includes the phase information.

Vector quantization (VQ) is used for comparing the trained data with new entered input data. It is a classical quantization technique that allows the modeling of probability density functions by the distribution of vectors. It divides a large set of points called vectors into groups having approximately the same number of points closest to them.The density matching property of VQ is powerful for identifying the density of large and high-dimensioned data [7].The representative codeword is determined to be theclosest in Euclidean distance from the input vector.[8]

# 3.TECHNICAL SECTION:

## 3.1. TECHNIQUES OF FEATURE EXTRACTION:

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach.. Many feature extraction techniques are available, these include

Linear predictive analysis (LPC)

Linear predictive cepstral coefficients (LPCC),

Perceptual linear predictive coefficients (PLP)

Mel-frequency cepstral coefficients (MFCC)

Power spectral analysis (FFT)

Mel scale cepstral analysis (MEL)

Relative spectra filtering of log domain coefficients (RASTA)

First order derivative (DELTA)

Few techniques generate a pattern from the features and use it for classification by the degree of correlation,other techniques use the numerical values of the features coupled to statistical classification method.

### 3.1.1 LINEAR PREDICTIVE CODING (LPC) :

LPC is one of the most powerful speech analysis techniques and is a useful method for encoding quality speech at a low bit rate **.** LPC analyzes the speech signal by estimating the formants, removing their effects from the speech signal, and estimating the intensity and frequency of the remaining buzz. The process of removing the formants is called inverse filtering, and the remaining signal is called the residue. In LPC system, each sample of the signal is expressed as a linear combination of the previous samples. This equation is called a linear predictor and hence it is called as linear predictive coding .The coefficients of the difference equation (the prediction coefficients) characterize the formats.

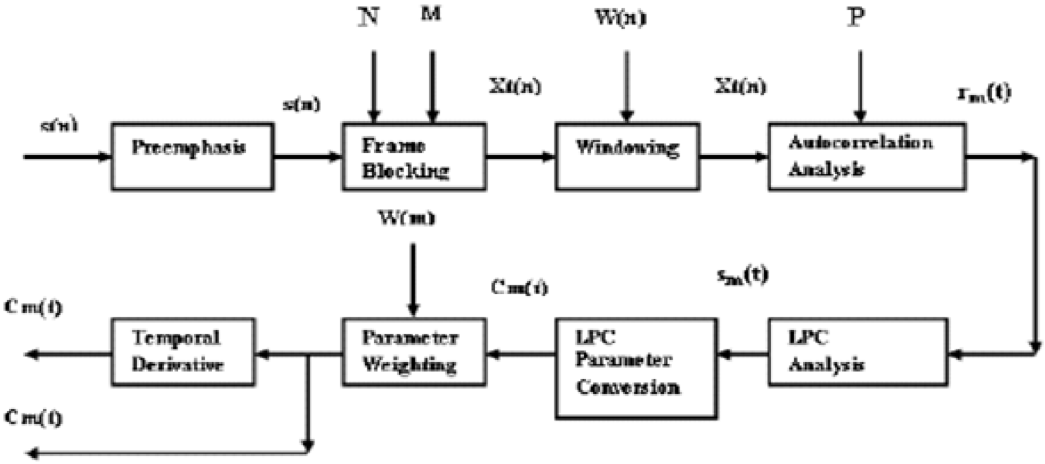


**Fig3 : Block Diagram of LPC**

**Methodology :**

LPC is a model based on human speech production. It utilizes a conventional source-filter model, in which the glottal, vocal tract, and lip radiation transfer functions are integrated into one all-pole filter that simulates acoustics of the vocal tract

The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech signal and the estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients.

These predictor coefficients are estimated every frame, which is normally 20 ms long.The predictor coefficients are represented by ak. Another important parameter is the gain(G).Levinson-Durbin recursion will be utilized to compute the required parameters for the auto-correlation method (Deller et al., 2000). The LPC analysis of each frame also involves the decision-making process of voiced or unvoiced. A pitch-detecting algorithm is employed to determine to correct pitch period / frequency. It is important to re-emphasis that the pitch, gain and coefficient parameters will be varying with time from one frame to another.

In reality the actual predictor coefficients are never used in recognition, since they typical show high variance. The predictor coefficient is transformed to a more robust set of parameters known as cepstral coefficient.

### 2.1.2 MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC):

The extraction and selection of the best parametric representation of acoustic signals is an important task in the design of any speech recognition system; it significantly affects the recognition performance. A compact representation would be provided by a set of MFCC, which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale. The MFCCs are proved more efficient. The calculation of the MFCC includes the following steps

*Mel-frequency wrapping* :

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus 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 frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz .As a reference point ,the pitch of a 1 KHz tone ,40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency *f* in Hz.

Mel($f$) = 2595*log10(1 + $f$/700)

Ours approach to simulate the subjective spectrum is to use a filter bank, one filter for each desired mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel-frequency interval. The mel scale filter bank is a series of l triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds toband pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale

*Cepstrum:*

In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC).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 (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform (DCT). In this final step log mel spectrum is converted back to time. The result is called the Mel Frequency Cepstrum Coefficients.

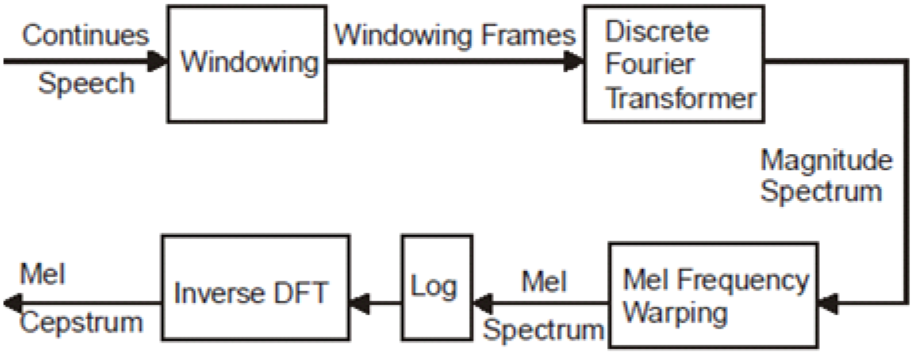


Fig 4 : Complete Pipelining for MFCC

## 2.1.3 PERCEPTUAL LINEAR PREDICTION (PLP):

The PLP model developed by Herman sky 1990. The goal of the original PLP model is to describe the psychophysics of human hearing more accurately in the feature extraction process. PLP is similar to LPC analysis, is based on the short-term spectrum of speech. In contrast to pure linear predictive analysis of speech, PLP modifies the short-term spectrum of the speech by several psychophysically based transformations.

*PLP Algorithm*

In the PLP technique, several well-known properties of hearing are simulated by practical engineering approximations, and the resulting auditory like spectrum of speech is approximated by an autoregressive all-pole model. A block diagram is shown in figure

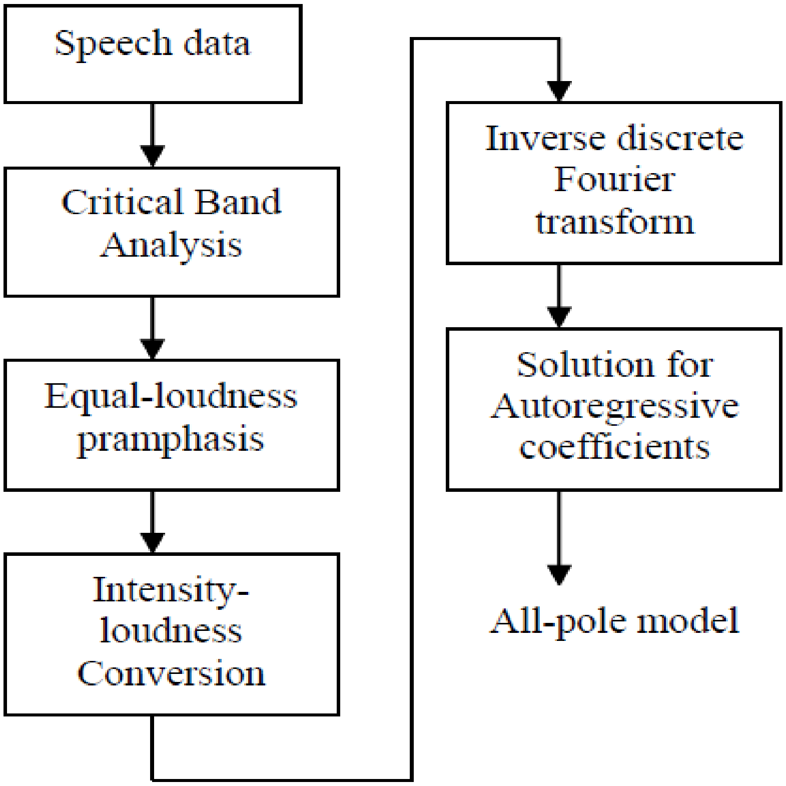

Speech data
Critical Band Analysis
Equal-loudness pramphasis
Intensity-loudness Conversion
Inverse discrete Fourier transform
Solution for Autoregressive coefficients
All-pole model

Fig 5. Block diagram of PLP speech analysis (hermansky)

*Spectral analysis:*

The speech segment is weighted by the Hamming window

$$w(n) = 0.54 + 0.46\cos[2\pi n/(N-1)]$$

Where N is the length of the window.

The typical length of the window is about 20ms.The discrete Fourier transform(DFT) transforms the windowed speech segment into the frequency domain. Typically, the fast fourier transform (FFT) is used here.

The real and imaginary components of the short-term speech spectrum are squared and added to get the short term power spectrum.

$$P(w) = \mathrm{Re}[s(w)]^2 + \mathrm{Im}[s(w)]^2$$

*Critical-band spectral resolution:*

The spectrum P(w) is warped along its frequency axis w into the bark frequency by

$$\Omega(w) = 6\ln\left\{w/1200\pi + [(w/1200\pi)^2 + 1]^{0.5}\right\}$$

The resulting warped power spectrum is then convolved with the power spectrum of the simulated critical-band masking curve $\Psi(\Omega)$ . This step is similar to spectral processing in mel cepstral analysis, except for the particular shape of the critical-band curve. In PLP technique, the critical-band curve is given by

$$\Psi(\Omega) = \begin{cases} 0 & \mathit{for}\ \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & \mathit{for}\ -1.3 \le \Omega \le -0.5, \\ 1 & \mathit{for}\ -0.5 \le \Omega \le 0.5, \\ 10^{-1.0(\Omega-0.5)} & \mathit{for}\ 0.5 \le \Omega \le 2.5, \\ 0 & \mathit{for}\ \Omega \succ 2.5. \end{cases}$$

The discrete convolution of $\Psi(\Omega)$ with (the even symmetric and periodic function) P(w) yields samples of the criticl-band power spectrum.

The convolution with the relatively broad critical-band masking curves $\Psi(\Omega)$ significantly reduces the spectral resolution of $\theta(\Omega)$ in comparison with the original P (w).This allows for the down-sampling.

*Equal-loudness pre-emphasis:*

The sampled $\Theta[\Omega(w)]$ is pre-emphasized by the simulated equal-loudness curve:

$$\Xi[\Omega(w)] = E(w)[\Theta(w)]$$

The function E (w) is an approximation to the non equal sensitivity of human hearing at different frequencies and simulates the sensitivity of hearing at about the 40-dB level.The particular approximation is adopted from Makhoul and Cosell(1976) and is given by:

$$E(w) = \left[(w^2 + 56.8*10^6)w^4\right] / \begin{bmatrix} (w^2 + 6.3*10^6)^2 * \\ (w^2 + 0.38*10^9) \end{bmatrix}$$

Finally, the values of the first (0bark) and the last (Nyquist frequency) samples (which are not well found) are made equal to the values of their nearest neighbors. Thus $\Xi[\Omega(w)]$ begins and ends with two equal-valued samples.

*Intensity-loudness power law:*

The last operation prior to the all-pole modelling is the cubic-root amplitude compression.

$$\Phi(\Omega) = \Xi(\Omega)^{0.33}$$

This operation is an approximation to the power law of hearing (Stevens1957) and simulates the nonlinear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-loudness pre-emphasis, this operation also reduces the spectral amplitude variation of the critical band spectrum so that the following all-pole modeling can be done by a relatively low model order.

**Table 1: Comparison of feature extraction techniques:**

| LPC | MFCC | PLP |
|---|---|---|
| Calculate power spectrum. | Calculate power spectrum. | Calculate power spectrum |
| Integration of power spectrum using Linear filter bank. | Integration of power spectrum using mel-filter bank. | Integration of power spectrum using Bark scale filter bank. |
| Pre-emphasis done before spectrum analysis. | Pre-emphasis done before spectrum analysis. | Pre-emphasis is done using equal loudness pre-emphasis curve. |
| LPC coefficients is Converted into cepstral coefficients compulsory. | No such option is present. | Coefficients converted to cepstral coefficients optionally. |
| Takes IDFT,gets predictors coefficients. | Takes IDFT, gets cepstral coefficients. | Takes IDFT,gets autocorrelation. |

## 3.2 Modeling Techniques for Speaker Identification :

The modeling is an important stage is SI systems. The modeling of speakers in SR systems is responsible for enrolment of speakers in the training phase. The modeling of speakers refers to the making, storing the copy of speaker specific information contained in feature vectors extracted by feature extraction stage in SR tasks. So, the modeling plays an important role in speaker identification & verification. And the importance of modeling lies in the storing the most robust speaker –specific information as reference for future testing phases. In modeling of speakers, nearest feature vectors are clumped together and assigned with representative vector. A model is formed, because at the time of testing or comparison, instead of comparing to all the speakers, comparison is done with only one speaker model. The distribution of speaker feature vectors obtained from the speaker voiced phrase using mathematical models is modeling. A large set of feature vectors of a speaker is grouped into its representative vector by several modeling methods. The speaker models obtained by the modeling methods are parametric (stochastic) and non-parametric (template). In template models, training & test feature vectors are directly compared with each other with assumption that one is imperfect replica of other. Vector Quantization (VQ) and Dynamic Time Warping (DTW) are examples of template models for text-independent & text-dependent respectively. In stochastic models, each speaker is modelled as probabilistic source with unknown & fixed probability density function (PDF). The training phase is to estimate the parameters of the PDF from the training sample. Matching is usually done by evaluating the likelihood of the test utterance with respect to the model. The popular modeling methods are text-independent Gaussian Mixture Model (GMM), text-dependent Hidden Markov Model (HMM), text-independent Vector Quantization (VQ). The mostly used models are Artificial Neural Networks (ANN), Support Vector Machines (SVM), Gaussian Markov Models (GMM) and GMM-universal back ground model (GMM-UBM). State-of-art SR systems uses VQ, GMM-UBM etc.

## 3.2.1 Vector Quantization (VQ) :

A speaker recognition system must able to estimate probability distributions of the computed feature vectors. Storing every single vector that generate 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. VQ is a process of taking a large set of feature vectors and producing a smaller set of measure vectors that represents the centroids of the distribution.The technique of VQ consists of extracting a small number of representative feature vectors as an efficient means of characterizing the speaker specific features. By means of VQ, storing every single vector that we generate from the training is impossible .By using these training data features are clustered to form a codebook for each speaker. In the recognition stage, the data from t he tested speaker is compared to the codebook of each speaker and measure the difference. These differences are then use to make the recognition decision .The problem of speaker recognition belongs to a much broader topic in scientific and engineering so 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 previous section. The classes here refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be also referred to as feature matching .Furthermore, if there exists some set of patterns that the individual classes of which are already known, then one has a problem in supervised pattern recognition. This is exactly our case since during the training session, we label each input speech with the ID of the speaker (S1 to S8).These patterns comprise the training set and are used to derive a classification algorithm. The remaining patterns are then used to test the classification algorithm; these patterns are collectively referred to as the test set. If the correct classes of the individual patterns in the test set are also known, then one can evaluate the performance of the algorithm .The state-of-the-art in feature matching techniques used in speaker recognition include Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ). In this project, the VQ approach will be used, due to ease of implementation and high accuracy.

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 codewords is called a codebook .

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.

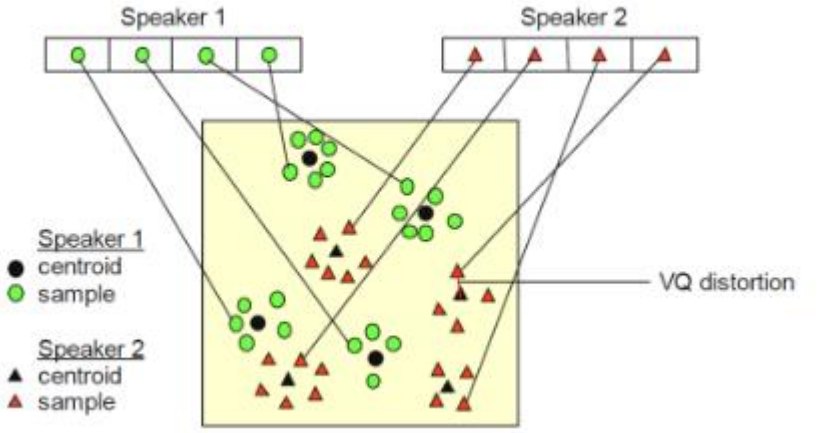


Fig 6 : Conceptual diagram illustrating vector quantization codebook formation.

## Clustering the training vector :

After the enrolment session, the acoustic vectors extracted from input speech of a speaker provide a set of training vectors. As described above, the next important step is to build a speaker-specific VQ codebook for this speaker using those training vectors. There is a well-known 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 thefollowing 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 accordingly to the rule where n varies from 1 to the current size of the codebook, and e is a splitting parameter(we choose e =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.

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 5.10 shows, in a flow diagram, the detailed steps of the LBG algorithm. "Cluster vectors.

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.

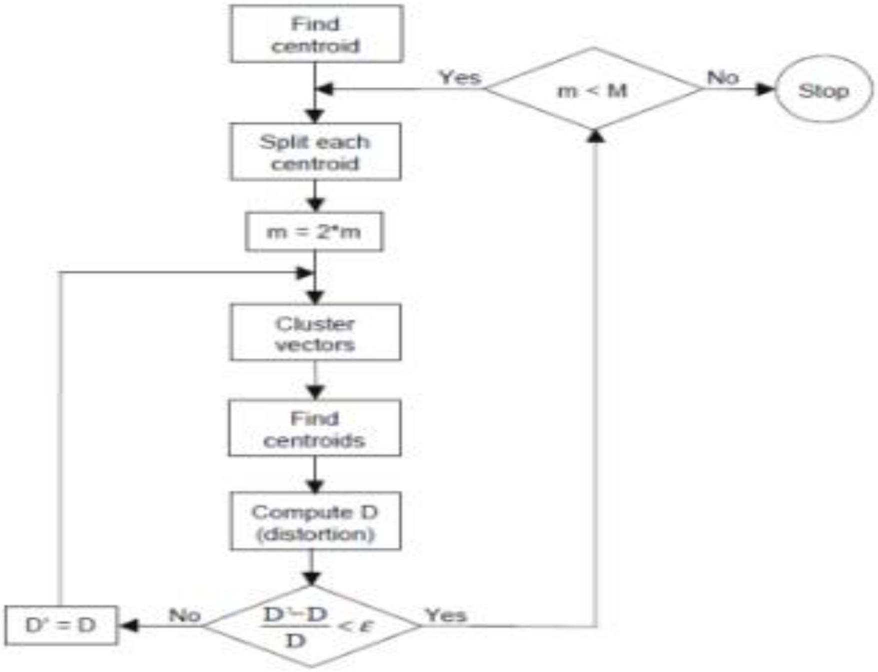


Fig 7: Flow diagram of the LBG algorithm

*Advantages of VQ:*

-Reduced storage for spectral analysis information.

-Reduced computation for determining similarity of spectral analysis vectors.

-In Speech recognition, a major component of the computation is the determination of Spectral similarity between pair of voice.

-Discrete representation of speech sounds.

*Disadvantages of VQ modeling :*

-VQ based methods suffer from the problem of outliners, i.e., the vectors lying at the boundaries of the clusters creates problems.

-A large amount of training data is required to have a good estimate of cluster representatives.

-Degradation of performance occurs if the training and testing environments are different.

-VQ is lossy compression, lossy correction and lossy template-based modeling
method etc.

### *Applications of Vector Quantization (VQ):*

-Used for lossy data compression, lossy data correction, pattern recognition,density estimation and clustering.

-Audio codecs are based on VQ.

-Used in pattern recognition- VQ is used for speech & speaker recognition.

-Main advantage of VQ in pattern recognition is its low computation burden when compared burden with other technique such as dynamic time warping.(DTW) & Hidden Markov Model (HMM).

-Main drawback when compared to DTW & HMM is that it does not take into account the temporal evolution of signals (speech, signature) because all the vectors are mixed up.

-Used as clustering algorithm.

-Used in data stream mining

## 3.3 TECHNIQUES OF FEATURE MATCHING-:

### 3.3.1 Distance Measure :

In the speaker recognition phase, an unknown speaker’s voice is represented by a sequence of feature vector {x1, x2 ….xi), and then it is compared with the codebooks from the database. In order to identify the unknown speaker, this can be done by measuring the distortion distance of two vector sets based on minimizing the Euclidean distance.The Euclidean distance is the "ordinary" distance between the two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem.

The formula used to calculate the Euclidean distance can be defined as following:

The Euclidean distance between two points P = (p1, p2…pn) and Q = (q1, q2...qn),

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

The speaker with the lowest distortion distance is chosen to be identified as the unknown person

# 4.RESULTS:

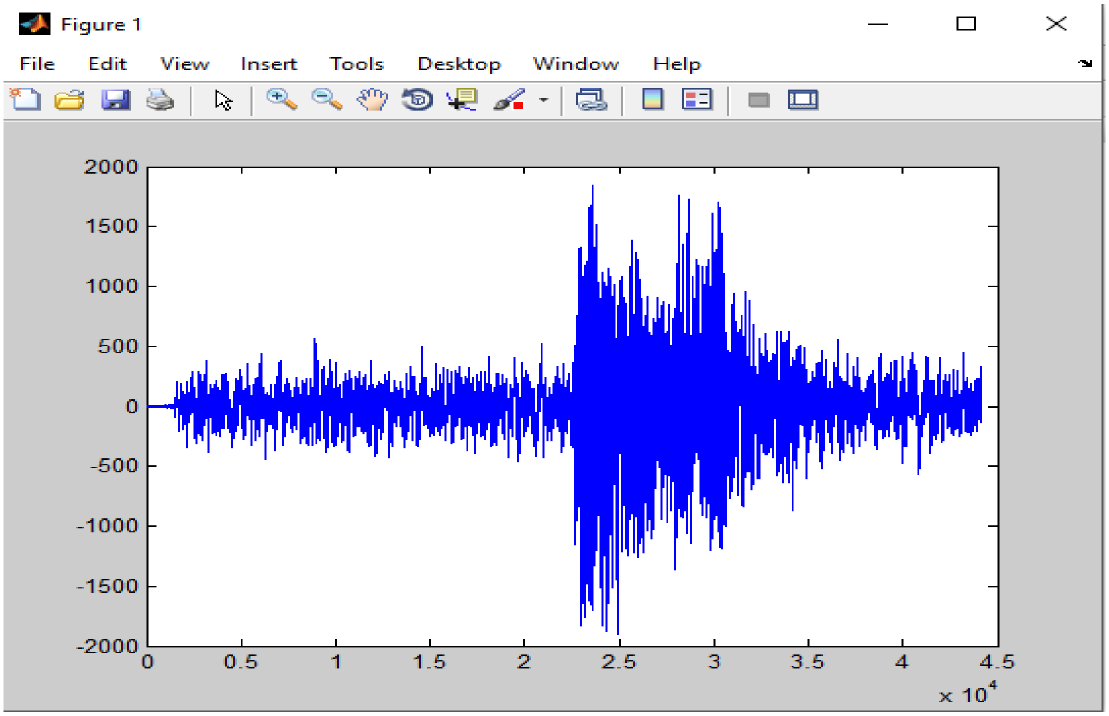
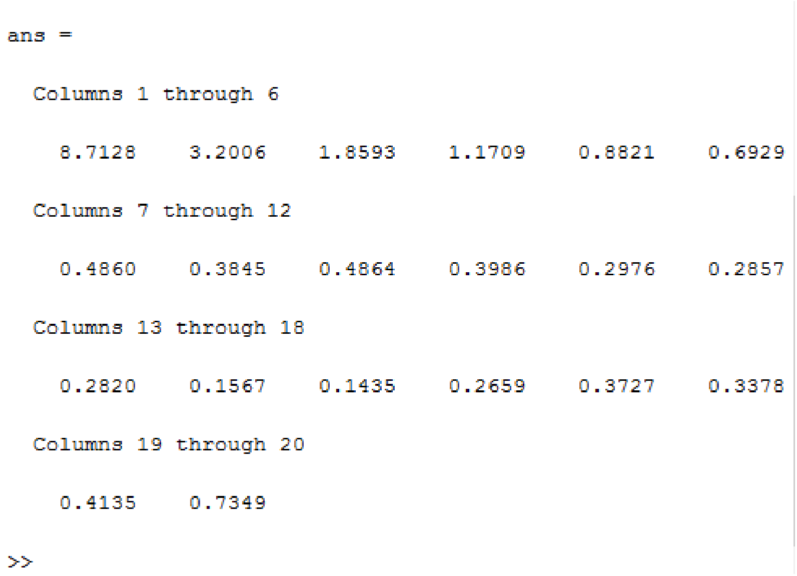


Fig8:Input Audio Signal

MFCC :



```
ans =

  Columns 1 through 6

    8.7128    3.2006    1.8593    1.1709    0.8821    0.6929

  Columns 7 through 12

    0.4860    0.3845    0.4864    0.3986    0.2976    0.2857

  Columns 13 through 18

    0.2820    0.1567    0.1435    0.2659    0.3727    0.3378

  Columns 19 through 20

    0.4135    0.7349

>>
```

Fig9:Command window for MFCC

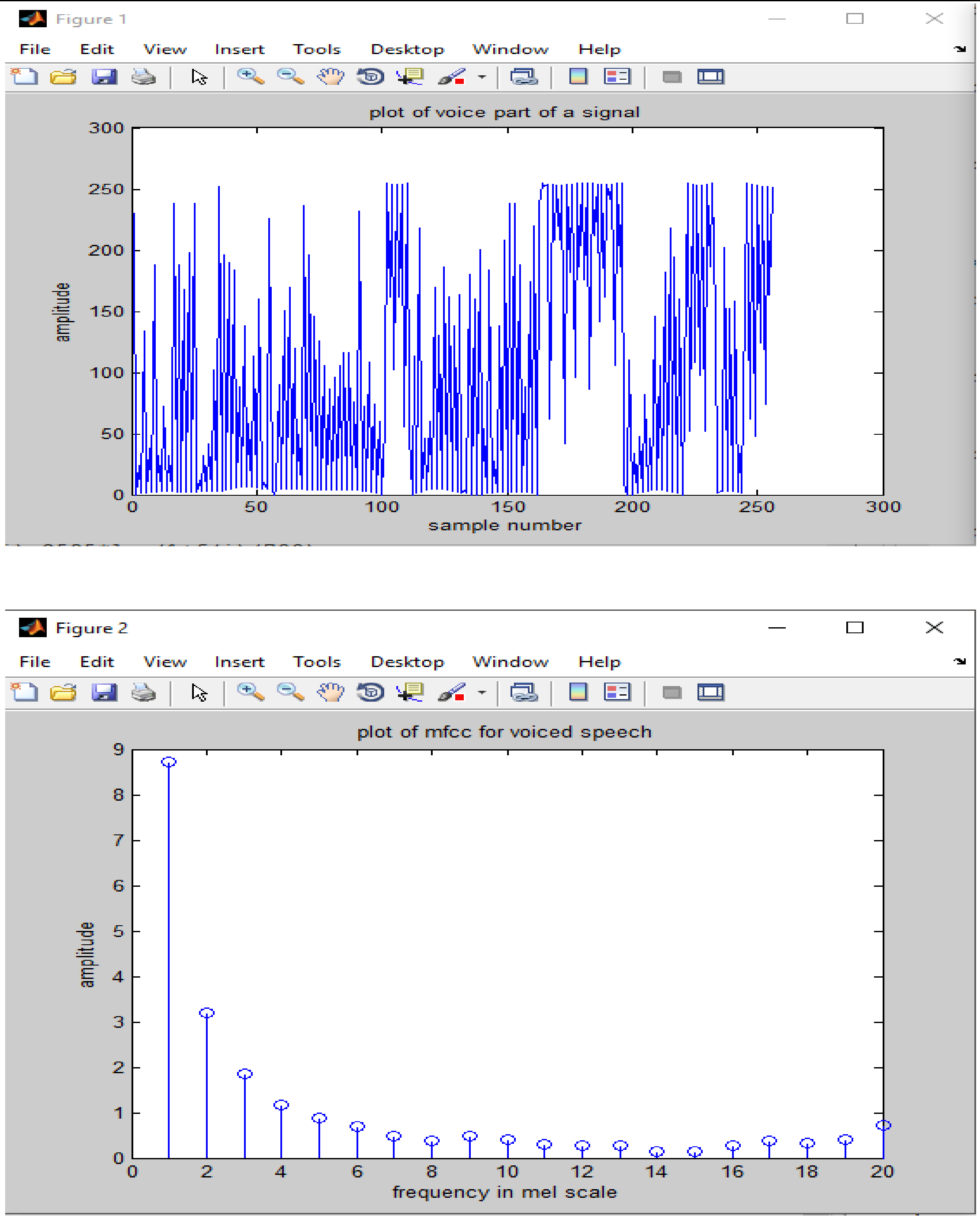
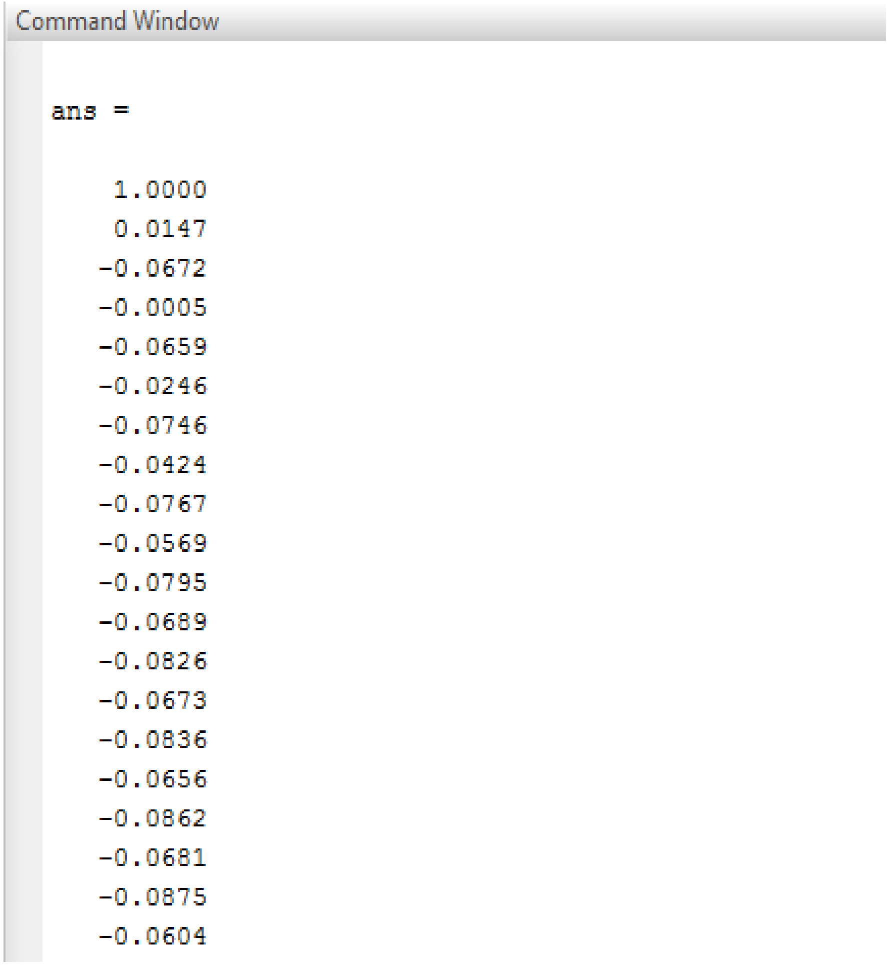


Figure 1
File Edit View Insert Tools Desktop Window Help
plot of voice part of a signal
amplitude
0
50
100
150
200
250
300
0
50
100
150
200
250
300
sample number
Figure 2
File Edit View Insert Tools Desktop Window Help
plot of mfcc for voiced speech
amplitude
0
1
2
3
4
5
6
7
8
9
0
2
4
6
8
10
12
14
16
18
20
frequency in mel scale

Fig10:Output of MFCC

LPC :



```
ans =

    1.0000
    0.0147
   -0.0672
   -0.0005
   -0.0659
   -0.0246
   -0.0746
   -0.0424
   -0.0767
   -0.0569
   -0.0795
   -0.0689
   -0.0826
   -0.0673
   -0.0836
   -0.0656
   -0.0862
   -0.0681
   -0.0875
   -0.0604
```

Fig11:Command window for LPC

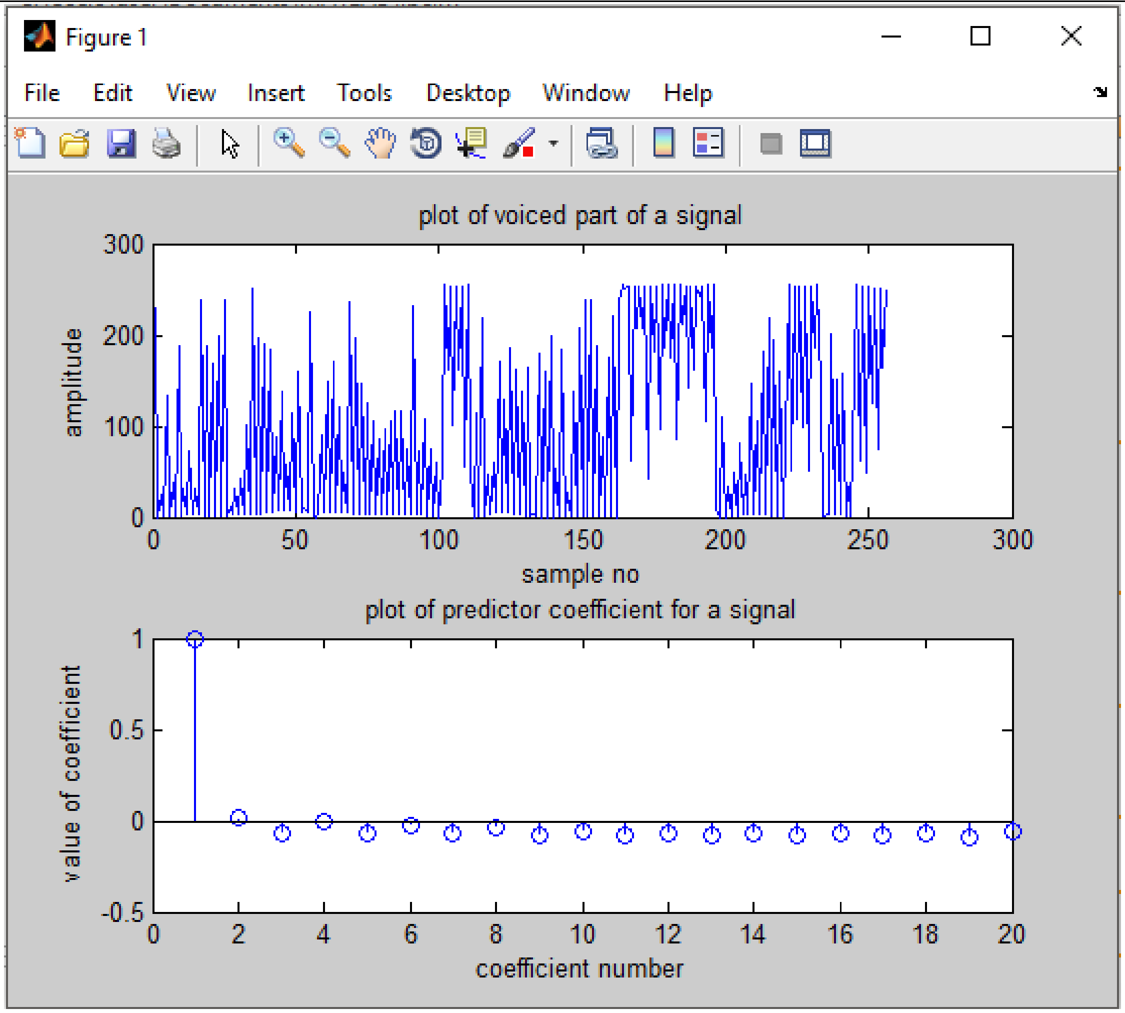
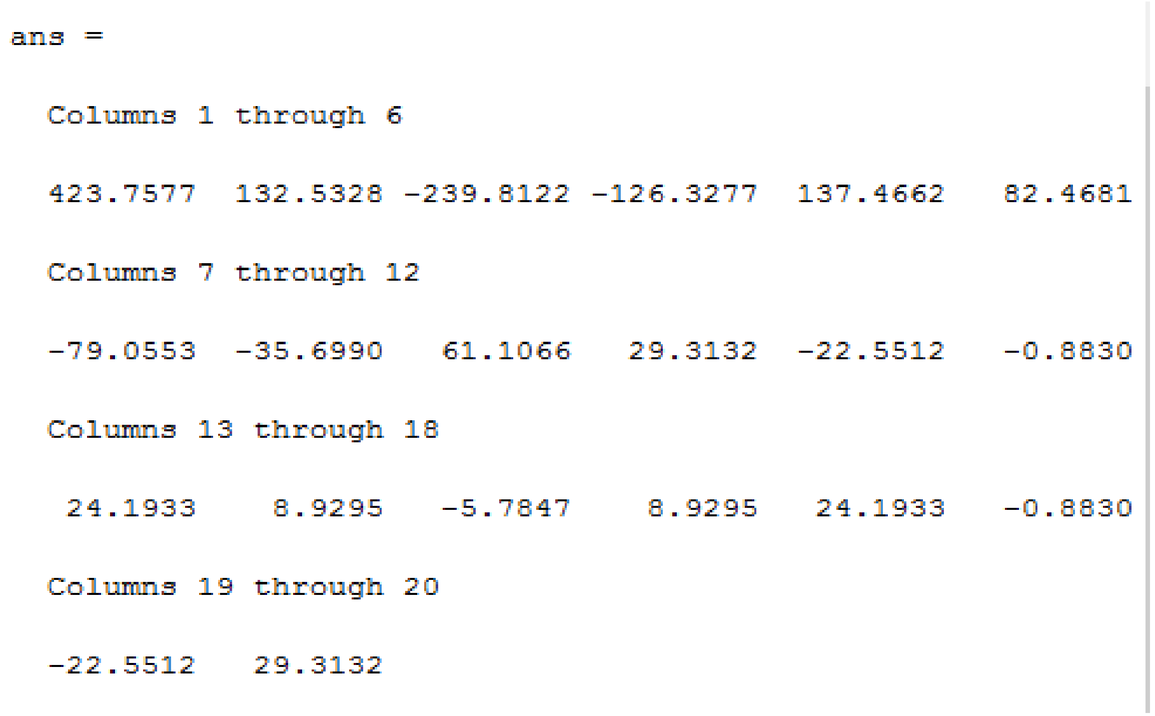


Fig12:Output of LPC

**PLP :**



```
ans =

  Columns 1 through 6

  423.7577  132.5328 -239.8122 -126.3277  137.4662   82.4681

  Columns 7 through 12

  -79.0553  -35.6990   61.1066   29.3132  -22.5512   -0.8830

  Columns 13 through 18

   24.1933    8.9295   -5.7847    8.9295   24.1933   -0.8830

  Columns 19 through 20

  -22.5512   29.3132
```

Fig13: Command window for PLP

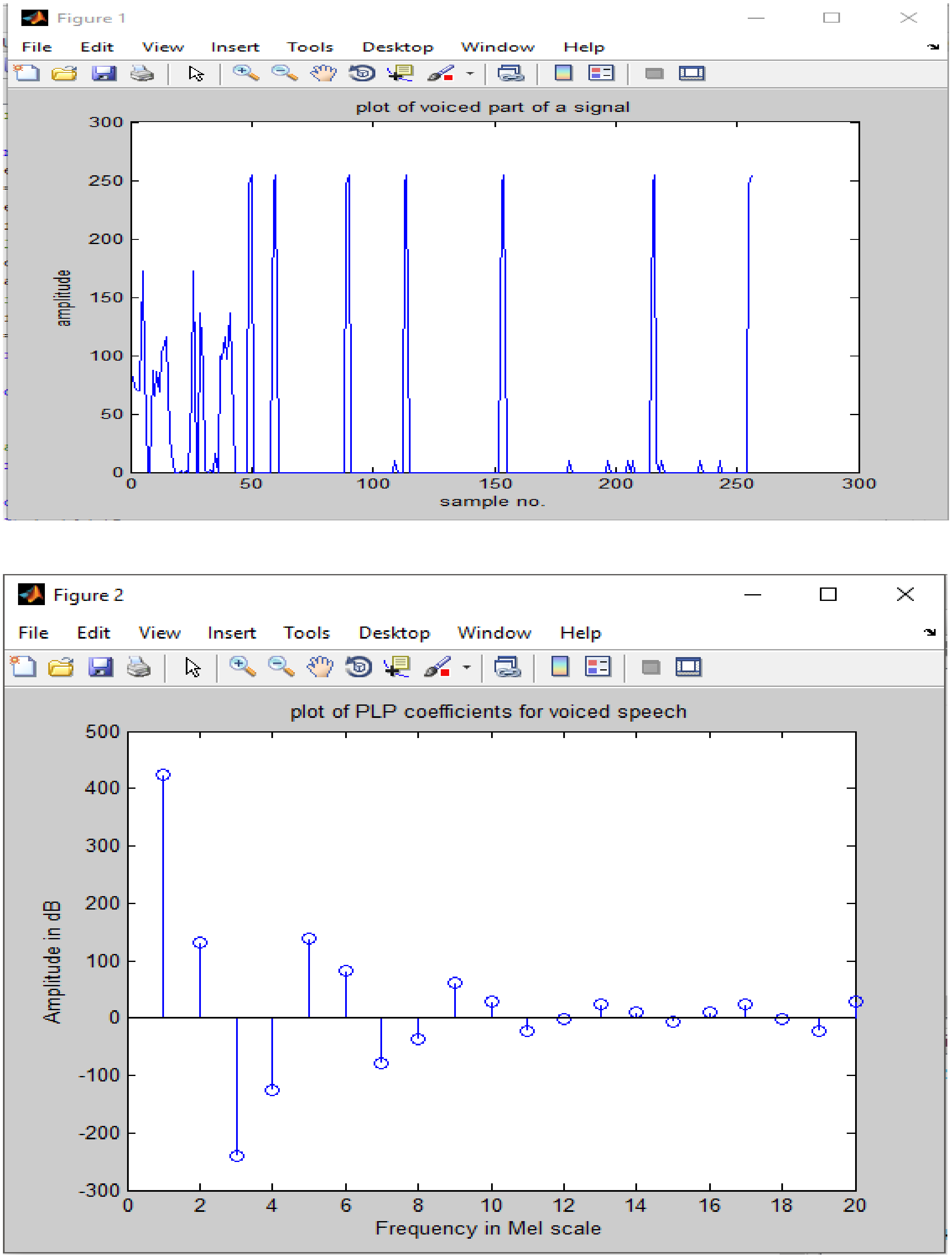


Fig14:Output of PLP

# 5.APPLICATIONS :

## 3.1 BIOMETRICS :

A biometric system is a pattern recognition system, which makes a personal identification by determining the authenticity of a specific physiological or behavioral characteristics possessed by the user. It comprises methods for uniquely recognising humans based upon one or more intrinsic physical or behavioral traits. Biometrics is used as a form of identity access management and access control. It is also used to identify individuals in groups that are under surveillance voice is also a physiological trait because every person has a different vocal tract, but voice (speaker) recognition is mainly based on the study of the way a person speaks, commonly classified as behavioral. Among the above, the most popular biometric system is the speaker (voice) recognition system because of its easy implementation and economical hardware Transaction authentication .

-Toll fraud prevention, telephone credit card purchases, telephone brokerage (e.g.stock trading) Access control

-Physical facilities, computers and data networks ,Monitoring

-Remote time and attendance logging, home parole verification, prison telephone usage Information retrieval

-Customer information for call centers, audio indexing (speech skimming device), speaker diarisation.

-Applicable services include voice dialing, banking over a telephone network, telephone shopping, database access services, information and reservation services, voice mail, security control for confidential information, and remote access
to computers.

# 6.CONCLUSION :

The goal of this project was to create a speaker Identification system, and apply it to a speech of an unknown speaker. By investigating the extracted features of the unknown speech and then comparing them with the stored extracted features for each different speaker in order to identify the unknown speaker.The feature extraction is done by using LPC ,MFCC and PLP . The speaker was modeled using Vector Quantization (VQ). A VQ codebook is generated by clustering the training feature vectors of each speaker and then stored in the speaker database. In this method, the' LBG' algorithm is used for clustering. In the recognition stage,a distortion measure which is based on the minimizing the Euclidean Distance was used when matching an unknown speaker with the speaker database.We have concluded that MFCC is more efficient than the other two techniques used for feature extraction with efficiency of 96.5% then PLP with 78.5% and last is LPC with 65.8%.

# 7.FUTURE SCOPE :

Although many recent advances and successes in speaker recognition have been achieved, there are still many problems for which good solutions remain to be found. Most of these problems arise from variability, including speaker-generated variability and variability in channel and recording conditions. It is very important to investigate feature parameters that are stable over time, insensitive to the variation of speaking manner, including the speaking rate and level, and robust against variations in voice quality due to causes such as voice disguise or colds. It is also important to develop a method to cope with the problem of distortion due to telephone sets and channels, and background and channel noises.

From the human-interface point of view, it is important to consider how the users should be prompted, and how recognition errors should be handled. Studies on ways to automatically extract the speech periods of each person separately from a dialogue involving more than two people have recently appeared as an extension of speaker recognition technology.

This project focused on "Isolated Word Recognition". But we feel the idea can be extended to" Continuous Word Recognition" and ultimately create a Language Independent Recognition System based on algorithms which make these systems robust. The use of Statistical Models like HMMs, GMMs or learning models like Neural Networks and other associated aspects of Artificial Intelligence can also be incorporated in this direction to improve upon the present project. This would make the system more tolerant to variations like accent and extraneous conditions like noise and associated residues and hence make it less error prone.

The size of the training data i.e. the code book can be increased in VQ as it is clearly proven that the greater the size of the training data, the greater the recognition accuracy .This training data could incorporate aspects like the different ways via the accents in which a word can be spoken, the same words spoken by male/female speakers and the word being spoken under different conditions say under conditions in which the speaker may have a sore throat .

# 8.BIBLOGRAPHY:

[1]Dr.Shaila D.Apte-Speech and Audio Processing,Wiley India Edition.

[2] J.P. Campbell, “Speaker recognition:A tutorial”, Proceedings of the IEEE, vol.85, issue-9, pp. 1437-1462, September, 1997.

[3] D.A. Reynolds,” Experimental Evaluation of Feature for Robust Speaker Identification”, IEEE Trans. on Speech and Audio Processing, vol. 2, issue-4, pp. 639-643, October, 1994.

[4] X. Zhou, D. G. Romero, R. Duraiswami, C.E. Wilson, S. Shamma, “Linear versus Mel Frequency coefficients for speaker recognition”, IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Waikoloa, HI, pp. 559 – 564, 11-15 December, 2011.

[5] R. E. Wohiford, E. H. Wrench, Jr., and B. P. Landell, “Comparison of four Techniques for Automatic Speaker Recognition”, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol.5, pp. 908-911, April, 1980.

[6] Md Sahidullah, G.Saha, “A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recogniton, IEEE Signal Processing Letters, vol.20, issue-2, pp. 149-152, February, 2013.

[7] Tarun Pruthi, Sameer Saksena, Pradip K Das,”Isolated Word Recognition for Hindi language using VQ and HMM” ,Journal Of Computing and Business Research, 1993

[8] Atal, B.S. and S.L. Hanauer, ‘Speech analysis and synthesis by linear prediction of the speech wave’,” Journal of the Acoustical society of America”, **50:** 637-655(1971).