International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-5)

Research Article | **May 2015**

Robust Speaker Recognition Biometric System a Detailed Review

Gaganpreet Kaur
Research Scholar, Dept. of CSE,
P.T.U, (Punjab) India

Dr. Dheerendra Singh
Prof. & Head, Dept. of CSE
SUSCET,Tangori, (Punjab) India

Prinka Rani
Student, Dept. of CSE,
S.G.G.S.W.U., FGS. (Punjab)India

Abstract—

This paper reviews Biometric based Speaker Recognition and presents brief about various algorithms and techniques used at various stages of Speaker Recognition and development of Attendance System as application of Speaker Recognition. The research is being carried out in this area for many years. However, the accuracy of system depends upon speaker's variability and environmental conditions. Various stages of speaker recognition include Pre processing of a signal, Feature extraction, Normalization, Post-processing, Classification or Modelling and Decision making. The main aim of this paper is to discuss and compare different approaches of different stages with main emphasis on robustness in noisy environment.

Keywords— Biometric, Signal Pre-processing, Feature Extraction, Normalization, Feature post- processing, Classification.

I. INTRODUCTION

A biometric based recognition is an automatic recognition based on physiological and/or behavioural characteristics. By the use of these biometric traits it is possible to identify individual's identity. There is no need of memorizing any password or any type of record keeping equipment (e.g. Identification cards). The physiological characteristics are related to shape of the body like Fingerprints, Palm Veins, Face recognition, DNA, Hand Geometry, Iris Recognition, Retina, etc. Whereas the behavioural characteristics are related t individual's behaviour like gait, voice/speech, typing rhythm, etc. The selection of biometric trait for any application depends upon the characteristics of trait and user requirements. There are numerous Biometric applications. This review paper discusses about an application of Speaker Recognition in Attendance System, which includes Speech Biometric Trait.

Traditionally, Fingerprint [1], Thumb impression [2], Hand geometry [3], Iris Recognition [4], Facial Recognition [5] and Voice [6] are used for attendance system. Moreover, a web based attendance system is also proposed which includes GSM/GPRS along with Radio Frequency Identification (RFID) technique in attendance system [7], attendance system using Near Field Communication (NFC) [8]. However, these systems are expensive and have limited use, there are numerous applications today. As all the other biometric traits except speech/voice are difficult to set up due to equipment costs and time consumption in querying process, this paper reviews speech/voice as Biometric trait in Attendance system which overcomes the drawbacks of other traits in this particular application.

Speech is a natural medium among human beings to communicate with each other. This biometric tool can be used to recognise an individual. Speech contains the information of an individual like spoken words, speaker's identity, expressions and emotions, accent, living region, health conditions, gender, age, language. A person's authentication and speech as biometric is collectively called Speaker Recognition (SR). The SR system can be classified into two categories: Text dependent and Text independent. Text dependent is when system knows what will be spoken by speaker and text independent is when user is free to speak anything. Relating to speaker recognition the term speech recognition is also used. The term SR is used to know 'who is speaking' and 'what is spoken'. But speaker recognition is to identify 'who is speaking' and speech recognition is to identify 'what is spoken'.

This paper is organised as follows: Section II gives various stages of Speaker Recognition; Section III contains review of algorithm of all the stages; Section IV concludes paper and defines problem.

II. VARIOUS STAGES OF SPEAKER RECOGNITION

Steps for performing speaker recognition:

1. Pre processing of speech signal
2. Feature extraction
3. Feature Post-Processing
4. Normalization
5. Classification

Pre processing of signal includes the way to extract the voiced part of speech signal and removal of silence or unvoiced part. Feature extraction is the key step in speaker recognition in which important features of person's voice are extracted. Feature post processing contains techniques to enhance the signal for example by using delta and double delta. Then there comes normalization of feature values extracted in previous stage. At last classification or modelling process is executed to classify speakers on the basis of features extracted and decision is made depending on results shown by classification model. This process identifies whether present speaker is genuine or imposter.

Next section includes summary about various proposed algorithms at different stages of speaker recognition.

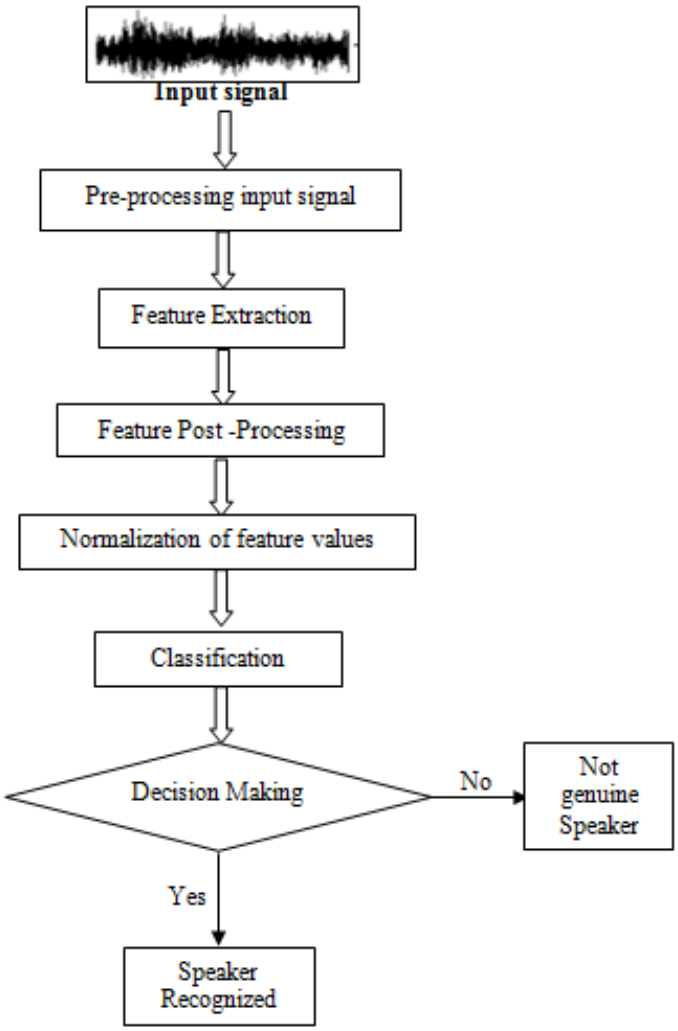


Fig. 1 Flowchart showing various steps for performing speaker recognition

III. REVIEW OF ALGORITHMS

A. Pre-processing of input speech signal

Before efficient feature attraction of signal, there is a need for detecting the end points of signal containing most of the voiced part as well as removal of silence and background disturbances. Thus this stage increases the efficiency of the recognition and also reduces the dimensionality in speech signal.

There are several ways of classifying signals according to the events. Conventionally, speech signals classified into three categories:

1. Silence(S), where there is no speech produced by the speaker,
2. Unvoiced (U), where the vocal chords are not vibrating, and hence the resulting waveform of Speech is Aperiodic or random in nature, and
3. Voiced (V), where the vocal chords are tensed and therefore vibrate physically when air flows from Lungs.

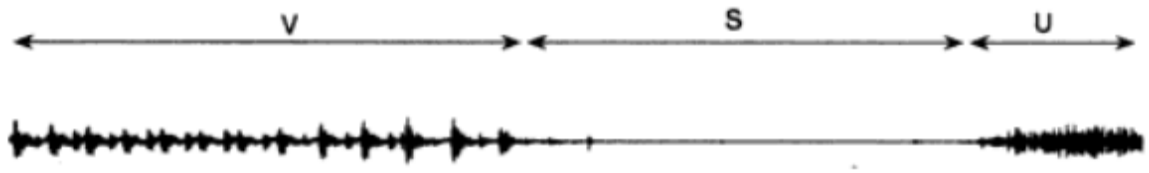
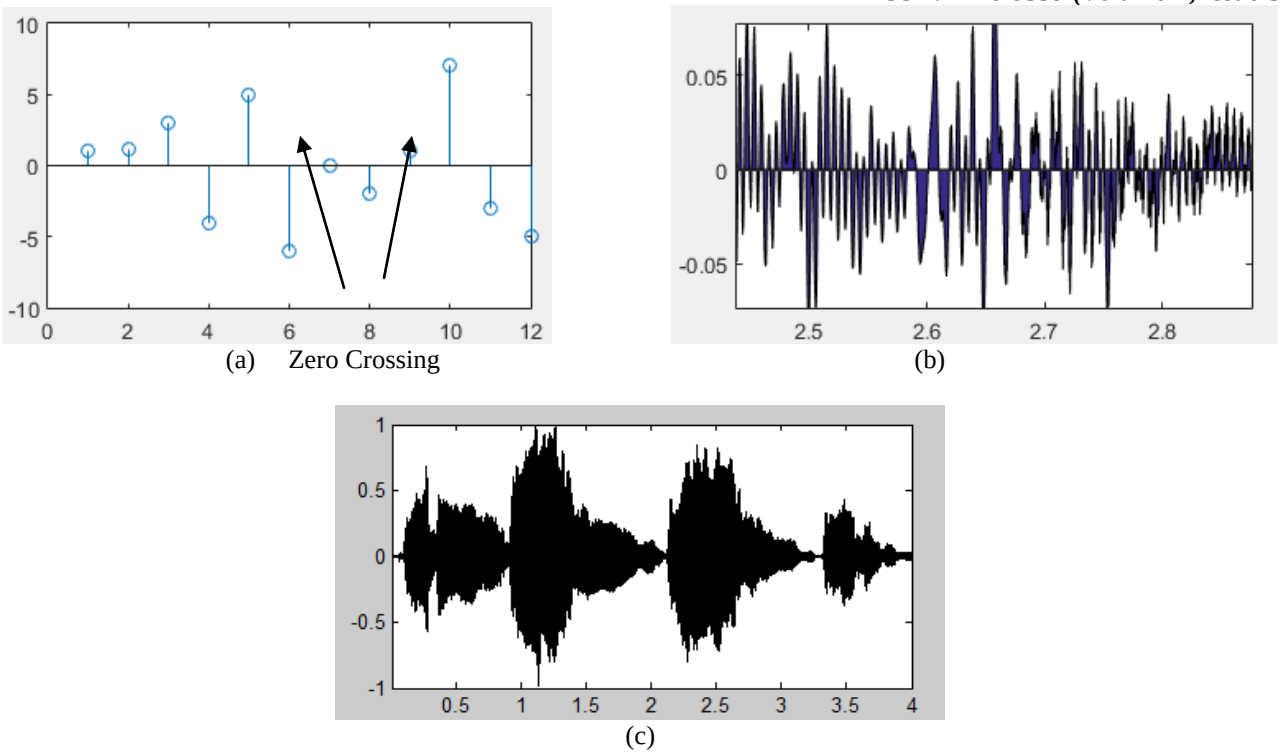


Fig. 2 Parts of Speech signal divided into Voiced (V), Silence(S), and Unvoiced (U) signal.

It is not possible to segment the signal into well-defined regions of silence, unvoiced and voiced signal. However, small errors in detection of signal usually have no consequences in many applications.

The widely accepted methods [9] are:

1) Zero Crossing Rate (ZCR): Zero Crossing rate is the number of times an amplitude of signal passes through a value of zero in a given time interval. In unvoiced speech, most of the amplitude has high ZCR and in voiced signal there is low ZCR. This way the voiced and the unvoiced signals are discriminated by taking specific threshold value of ZCR.



(a) Zero Crossing (b)

(c)

Fig. 3 (a) Zero crossing of signal, Zero crossing of unvoiced speech (b) and voiced speech (c)

2) Short Time Energy (STE): A signal varies in amplitude with time. The amplitude of voiced signal is much higher as compared to that of unvoiced signal. The energy of signal is a representation of amplitude variation. The energy of voiced speech signal is much higher than the energy of unvoiced one. Similarly, threshold value is predetermined for energy speech signal.

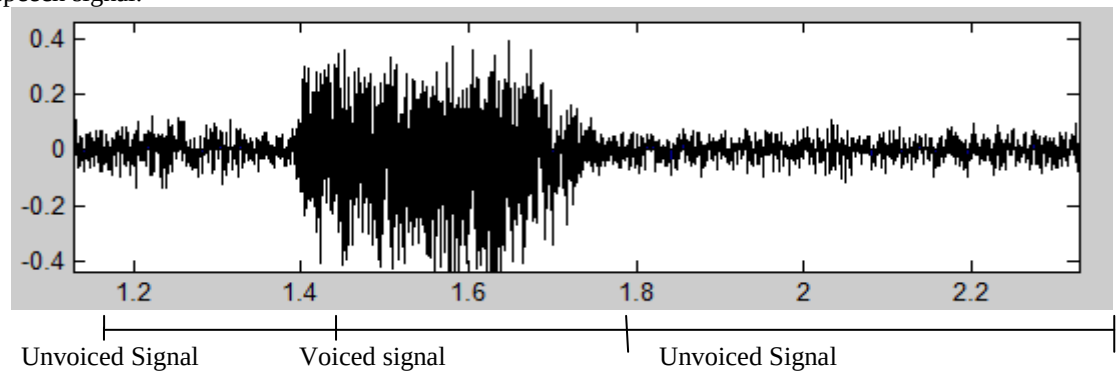


Fig. 4: Amplitude variation in voiced and unvoiced signal

When Probability Density function is used to detect background noise and Linear Pattern Classifier to classify voiced speech from silence or unvoiced part, this integration shows better results than ZCR and STE methods [10].

B. Feature Extraction

After the detection of voiced signal, feature extraction from signal is done. This stage is to extract those features of speech signal which gives maximum information related to the application. This stage is necessary in speech recognition. There exist many different algorithms. This section provides the brief summary about the existing algorithms for feature extraction. Feature extraction is basically used to remove all the reluctant information of signal depending upon user application and retaining only necessary features which give us enough information to identify the speaker. Existing algorithms are as follows:

1) Mel-Frequency Cepstral Coefficient (MFCC): MFCC is the most commonly used algorithm for the stage of feature extraction. MFCC tries to mimic the human for resolving non-linearity. Both the human and the MFCC work linearly. MFCC is used to extract those features which are used by human ears to listen. Firstly, in MFCC, the signal is divided into frames for which the feature vectors are calculated individually. Then the Hamming window is done to each frame. Further, after applying Fast Fourier Transformation (FFT), a Mel Filter Bank is generated. After Mel Frequency wrapping is done to obtain the coefficients, Inverse Discrete Fourier Transformation (IDFT) is calculated for cepstral coefficient generation.

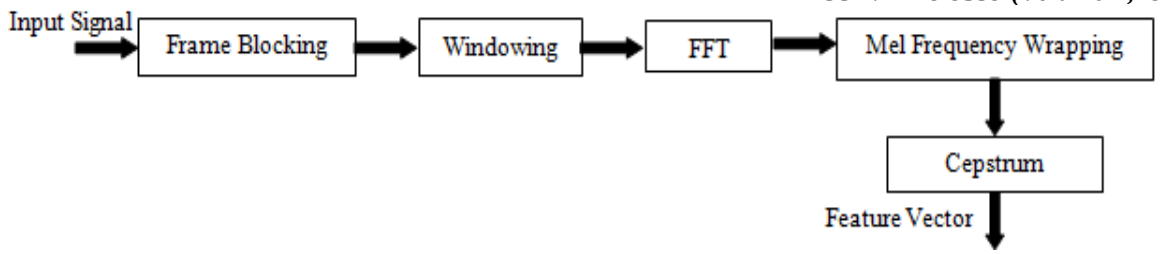


Fig. 5 Stages of MFCC

2) Linear Predictive Analysis (LPC): LPC is a technique to estimate the basic important features or characteristics. It includes the source filter model where filter is restricted to be all pole linear filters. Filtering is used for signal compression before transmission process. LPC calculates power spectrum of signal. The main idea is to generate the speech sample from the residual error (output after filtering speech signal). So this is parametric model based on the least mean squared error theory. By this minimisation of sum of squared differences, coefficients can be determined which are called Cepstral coefficients.

3) Perpetual Linear Prediction (PLP): PLP minimises the difference between speakers and includes all the important formant structure. PLP is bases upon the concept of human hearing. PLP is similar to Linear Predictive Code (LPC) method except its spectral characteristics are transformed to get matched with characteristics of human auditory system. PLP uses following three techniques:

1. Critical Band Spectral Resolution
2. Equal Loudness curve Pre-emphasis
3. Applying Intensity –Loudness power Law

Then auditory spectrum is approximated by an autoregressive all-pole model.

Comparing with LPC, PLP analysis is more consistent with human hearing [11]. PLP is efficient computationally and gives low dimensional speech representation. PP is better for speaker independent Automatic Speech Recognition (ASR) systems.

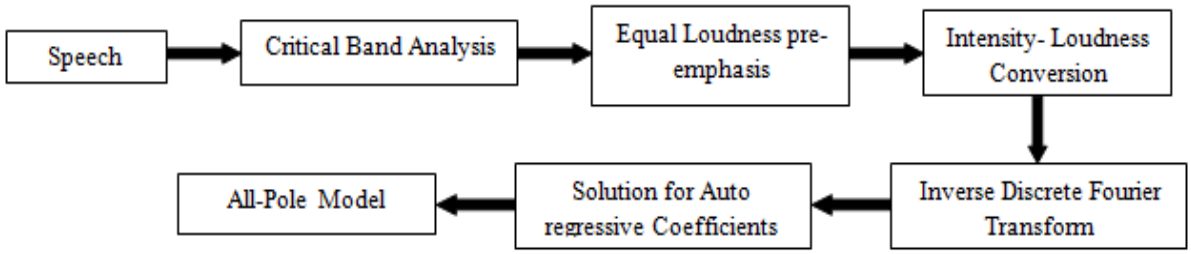


Fig. 6 Block diagram for PLP Algorithm

4) Gammatone Frequency Cepstral Coefficients (GFCC): GFCC is a FFT based technique in speaker identification system. The above mentioned techniques of LPC and PLP were used in speech recognition applications. GFCC technique includes 64 channels Gammatone filter bank (GTFB) for modelling the acoustic feature of human auditory system.

Firstly, a signal is passed through gammatone filter bank. Equal loudness is applied to each filtered output. Further, a logarithmic and Discrete Cosine transform is applied to get GFCC features of input signal.

GFCC has a fine resolution at low frequency as compared to MFCC [12]. MFCC works with a log while GFCC works with cube root. Cube roots provide more robustness to GFCC as compared to that of logs in MFCC [13]. Hence, GFCC is more robust in noisy environment than MFCC.

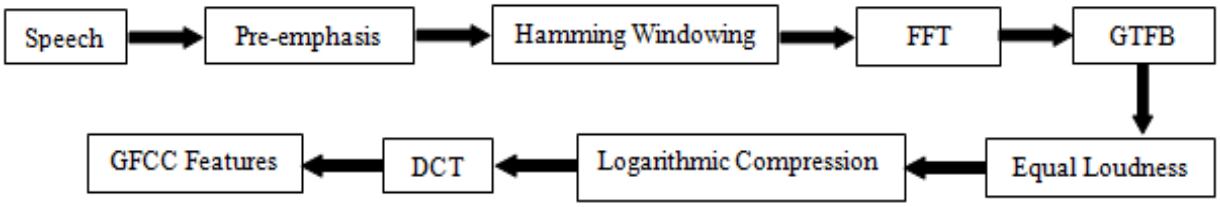


Fig. 7 Block diagram for GFCC Algorithm

5) Zero Crossing with Peak amplitude (ZCPA): This is also based on human auditory system. Zero Crossing intervals are used for detecting frequency of signal and amplitude is used to represent intensity or energy of signal. It uses band pass filters (FIR filters). The speech is divided into frames and each frame moves to both detectors. Further, non-linear compression of peak value is done. After compression, log result is taken. Frequency block is divided into several sub bands (interval histogram). Then both the frequency and intensity information are combined by the receiver.

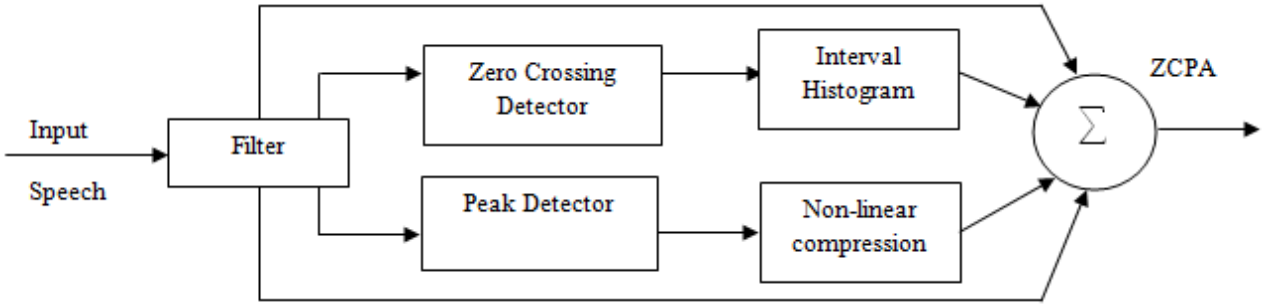


Fig. 8 Block diagram for ZCPA Algorithm

6) Power Normalised Cepstral Coefficients (PNCC): PNCC is based on an auditory processing. It utilises medium time processing to remove or suppress noise while MFCC uses log non-linearity. Firstly a power spectrum of input signal is integrated using gammatone frequency integration. Then based on medium time power analysis, filtering and temporal masking is done to remove noise. Then power law non-linearity and DCT is applied to get the features.

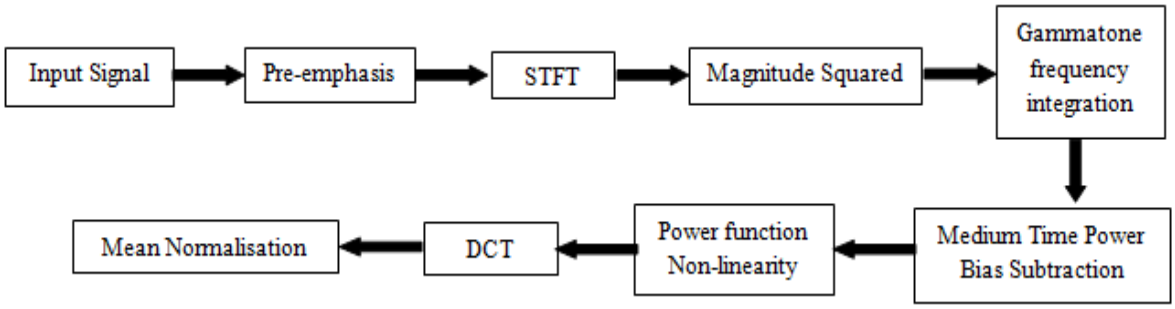


Fig. 9 Block diagram showing PNCC Algorithm steps

Features are roughly categorised as following:
1. Gammatone Features like GFCC
2. Mel Features like MFCC
3. Linear Prediction Features like LPC, PLP
4. Medium Temporal Features like PNCC
5. Zero Crossing features like ZCPA

Reference [14] shows that GFCC performs better than any other feature extraction method. GFCC works efficiently in noisy environment as compared to most widely used MFCC. MFCC can be efficiently used in noise-free environment.

C. Feature Post-processing

In speech signal processing, delta Δ and double delta ΔΔ features are commonly used to collect time dynamics. For instance, Delta + Double Delta + GFCC will give better results than only GFCC. Recently, post processing with delta and double delta improves speech preparation result [15]. Delta computes the first order derivative of an input feature vector and double delta again computes the first derivative of first delta feature vector.

D. Normalisation

Normalisation can be done either on the feature value or score value after classification step. Normalisation can be classified as either model based or data distribution based [16].

i) Model based normalisation techniques include mean, variance, moments to reduce mismatch in feature vector. It includes for instance, Cepstral Mean Normalisation (CMN), Mean Variance Normalisation (MVN).
ii) Data Distribution based techniques emphasise on feature distribution, for instance Histogram Equalisation Normalisation Method.

Cepstral Mean Normalisation (CMN) is performed on all speech signals with assumption that channel effect is uniform throughout the while speech signal recording. CMN is used for noisy environment for reducing distortion due to noise. So, there is no relevant information in mean as reduction of this reduces irrelevant information.

Mean and Variance Normalisation (MVN) is extension of CMN with assumption that mean and variance of coefficients should not vary. MVN is also known as Cepstral Variation Normalisation as it also contains Cepstral Mean Normalisation (CMN).

Histogram Equalisation (HEQ) is used in image processing. This assumes that shape of distribution of cepstral coefficients is invariant. In HEQ, any detail of cepstral distribution is regarded as irrelevant and is to be removed [17]. CMN gives better verification rate than MVN for the same database [18].

E. Classification

After extracting the features and removing irrelevant information, there comes classification or modelling or pattern matching. This is also an important and necessary step in speaker recognition. There are some existing methods like VQ, HMM, GMM, SVM, DTW, etc. This section gives a brief about these classification methods.

1) Vector Quantization (VQ): Vector Quantization technique is to split a large set of data points or set of points into clusters containing approximately same number of point's density in each cluster. Data points are represented by their closest centroid index called code book. If there are 'n' numbers of clusters then it will be having 'n' numbers of code books. VQ is lossy data compression method based on the principle of block coding. Following figure shows clusters of datapoints and code book or centroid containing within each cluster

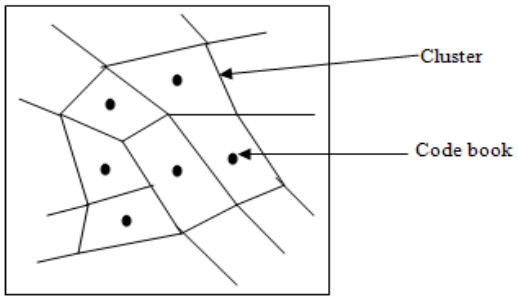


Fig. 10 Clusters and code book formed after Vector Quantization Classification of datapoints

Cluster includes subset of all datapoints. Datapoints are allotted to particular cluster if it lies within cluster region and have parametric value near to code book of that particular cluster. All clusters so formed by VQ have equal density approximately.

2) Hidden Markov Model (HMM): It is a popular tool for huge amount of data. This model is used to generate the pattern when the states are hidden. It uses the first order Markov process of observed states and unobserved states (hidden). In Markov models, the state is directly visible to the observer; therefore transition probability is the only parameter. In HMM, the state is not directly visible. As the output depends upon visible states, the output of HMM gives same information about the sequences of states. Recently, HMM uses second ordered and third ordered Markov processes for more complex data structures.

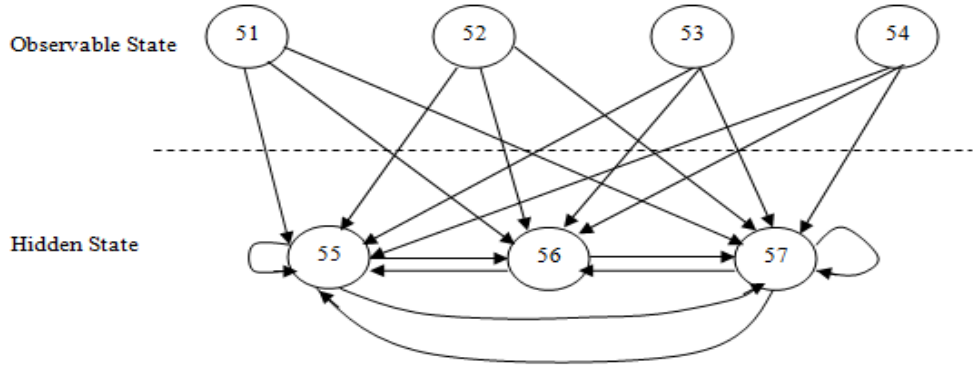


Fig. 11 Markov transition diagram showing observable and hidden states

HMM model contains two sets of states and three sets of probabilities.
1. Hidden States: The states of system which may be evaluated by a Markov process
2. Observer States: These are the visible states of a person.
3. State Transition Probability distribution
4. Observation Symbol Probability distribution
5. Initial State Probability distribution

3) Gaussian Mixture Model (GMM): GMM is a density estimator and is commonly used in classification step. In this method, a feature vector (x)'s distribution is modelled by a mixture of 'M' Gaussians.

$$P(x|M) = \sum_{i=1}^{M} a_i/(2\pi)^{D/2}|\Sigma_i|^{1/2}\exp\left(-\frac{1}{2}(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)\right)$$

Here, μ_i , Σ_i are mean and covariance of i^{th} mixture, respectively. μ_i is location parameter, which determines the centre of bell-shaped density. Σ_k in 1D is the variance and controls how spread out of density is, thus called shape parameter. In high dimensionality, it controls stretching in different directions. Different Gaussian components have different shapes. μ_i, Σ_i and a_i are learned using Expectations Maximization algorithm. During recognition, input speech's features are extracted, distance between input speech and stored data is obtained by log likelihood. Model with highest log likelihood is verified as identity of the speaker [19].

4) Support Vector machine (SVM): SVM is a kind of margin in which the goal is to find the decision boundary between two classes, which is maximally away from any point in the training data. SVM is based on the principle of risk minimization. SVM works well for the linear data but cannot classify linearity to non-linear data, so there is need for transforming data to higher dimensionality space and construct a linear binary classifier in this space. The steps needed to determine SVM decision boundary are: Firstly Find the closest points in convex hulls. Make hyperplane bisecting close points. Then find maximum margin, planes parallel to hyper plane.

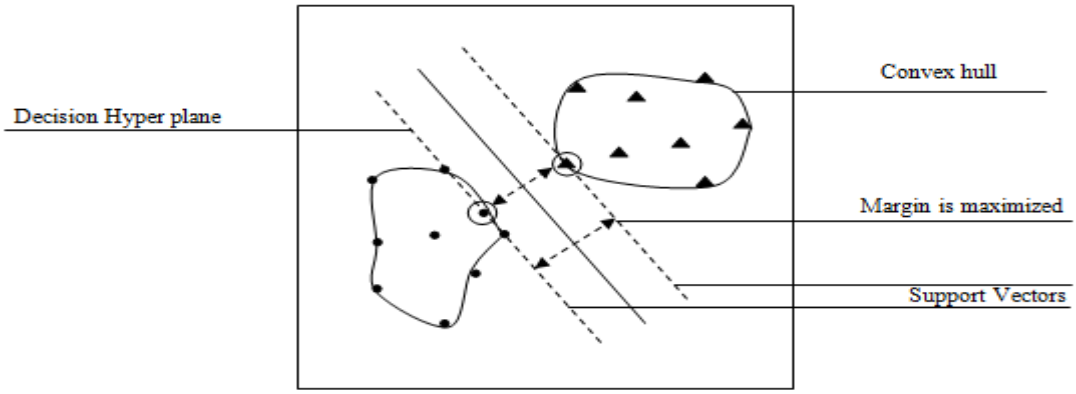


Fig. 12 Support vector and margin classifies datapoints in two classes.

Reference [20] shows results that in speaker recognition, classifying the feature after transformation will not work with SVM. The reason behind that is SVM can process on fixed length input, whereas speech signals are of variable length.

5) Dynamic Time Wrapping (DTW): This is also known as Dynamic Programming. When applied to template based speech recognition, it is called DTW. DTW is an algorithm to measure the distance between input speech and stored speech. In this algorithm firstly, the test data is converted to templates (The simplest form of template is sequence of feature vectors.).Then the process consists of making a match (lowest distance) between input and the stored templates. The input template is matched against all templates in system's repository. The best matching template is the one that has the lowest distance path aligning the input pattern to the template. The distance measurement between two feature vectors is calculated using Euclidean distance metric. If x and y are two feature vectors then the local distance is given by:

$$d(x, y) = \sqrt{\sum_j (xj - yj)2}$$

This algorithm works on some constraints like the mandatory path can't go backward, every frame of input vector should be used in matching, and local distance is combined to get global distance.

IV. CONCLUSION

In this review paper, all the fundamental steps for developing speaker recognizing system are discussed briefly along with widely used algorithms at each stage. Various alternatives for each stage are compared. Based on this review paper feature extraction and classification are key steps for speaker recognition and other stages enhance system in terms of efficiency and reducing dimensionality. Based on this review paper MFCC work well only for clean environment and its alternative GFCC work efficiently in noisy environment. Both MFCC and GFCC work equally well in clean noise free environment.

The performance of the system can be improved for additional noise signal and then can be compared on different cases like noise corrupted signal as future work. CMN and GMM model can be used for classification for better speaker recognition as these are considered efficient in noise environment.

REFERENCES

[1] Nur Izzati Zainal, Khairul Azami Sidek, Teddy surya Gunawan, Hasmah Mansor, and Mire Kartiwi, "*Design and development of portable classroom attendance system based on Arduino and fingerprint Biometric*", IEEE international conference on information and communication Technology, 2014.

[2] Engr. Imran Anwar Ujan and Dr. Imdad Ali Ismaili, "*Biometric Attendance System*", IEEE International Conference on Complex Medical Engineering, 2011.

[3] Tsai-Cheng Li, Huan-Wen Wu, and Tiz-Shiang Wu1, "*The study of Biometrics Technology Applied in Attendance Management System*", IEEE International Conference on Digital Manufacturing & Automation, pp. 943 – 947, 2012.

[4] Teh Wei Hsiung and Shahrizat Shaik Mohamed, "*Performance of Iris Recognition using Low Resolution Iris Image for Attendance Monitoring*", IEEE International Conference on Computer Applications and Industrial Electronics, 2011.

[5] Mashhood Sajid, Rubab Hussain, and Muhammad Usman, "*A Conceptual Model for Automated Attendance Marking System Using Facial Recognition*", IEEE International Conference on Digital Information Management, 2014.

[6] Subhadeep Dey, Sujit Barman, Ramesh K. Bhukya, Rohan K. Das, Haris B C, S. R. M. Prasanna, and R. Sinha, "*Speech Biometric Based Attendance System*", IEEE National Conference on Communications, 2014.

[7] Aamir Nizam Ansari, Arundhati Navada, Sanchit Agarwal, Siddharth Patil, and Balwant A. Sonkamble, *"Automation of Attendance System using RFID, Biometrics, GSM Modem with .Net Framework*", IEEE International Conference on Multimedia Technology, pp. 2976 – 2979, 2011.

[8] *Balazs Benyo, Balint Sodor, Tibor Doktor, and Gergely Fordo, "Student attendance monitoring at the university using NFC", IEEE,* pp. 1 – 5, 2012.

[9] Madiha Jalil, Faran Awais Butt, and Ahmed Malik, "*Short-Time Energy, Magnitude, Zero Crossing Rate and Autocorrelation Measurement for Discriminating Voiced and Unvoiced segments of Speech Signals*", IEEE International Conference on Electronics and Computer Engineering, pp. 208 – 212, 2013.

[10] G. Saha, Sandipan Chakroborty, and Suman Senapati, "*A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications*", Department of Electronics and Electrical Communication Engineering Indian Institute of Technology, Khragpur, Kharagpur-721 302, India.

[11] H. Hermansky, *"Perceptual Linear Predictive (PLP) Analysis of Speech"*, in J. Acoust. Soc. Am., vol. 87, no. 4, pp. 1738-1752, 1990.

[12] Zhao X., Shao Y., and Wang D.L., *"CASA-based robust speaker identification*", IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, pp. 1608-1616, 2012.

[13] X Zhao, and DL Wang, "*Analyzing noise robustness of MFCC and GFCC features in speaker identification*", IEEE International conference on acoustics, speech and signal processing, pp. 7204–7208, 2013.

[14] Jitong Chen, Yuxuan Wang, and DeLiang Wang, "*A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios*", IEEE Transactions on audio, speech, and language processing, vol. 22, pp. 1993 – 2002, 2014.

[15] Y. Wang, K. Han, and D. L. Wang, "*Exploring monaural features for classification-based speech segregation,"* IEEE Trans. Audio, Speech, and Language Processing, vol. 21, pp. 270–279, 2013.

[16] Md Jahangir Alam , Pierre Ouellet, Patrick Kenny, Douglas O'Shaughnessy, "*Comparative Evaluation of Feature Normalization Techniques for Speaker Verification*", Nonlinear Speech Process., pp. 246–253, 2011

[17] Yasunari Obuchi, "*Delta-Cepstrum Normalization for Robust Speech Recognition*", Proc. International Congress on Acoustics, pp.2587-2590, Kyoto, Japan, 2004.

[18] Jelil S, Kachari G, and Joyprakash Singh, "*Comparative evaluation of feature normalization techniques for voice password based speaker verification*", IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics, pp. 1-4, 2014.

[19] Douglas A. Reynolds, and Richard C. Rose, "*Robust text-independent speaker identification using Gaussian mixture speaker models*", IEEE Transaction Speech and Audio Processing, Vol. 3, pp 72–83, 1995.

[20] W.M. Campbell, J.P. Campbell, D.A. Reynolds, and E.Singer, "*Support vector machines for speaker and language recognition*," Computer Speech Language, vol.20, pp.210–229, 2006.

[21] Anil K. Jain, Arun Ross, and Salil Prabhakar, Member, "*An Introduction to Biometric Recognition*", IEEE Transactions on circuits and systems for video technology, vol. 14, 2004.

