# A Review of Automatic Speaker Recognition System

**Tejal Chauhan, Hemant Soni, Sameena Zafar**

***Abstract*— In the recent time, person authentication in security systems using biometric technologies has grown rapidly. The voice is a signal of infinite information. Digital signal processes such as Feature Extraction and Feature Matching are introduced to authenticate person in security system. In this paper concept of speaker recognition is discussed. Several methods such as Liner Predictive Coding (LPC), Mel Frequency Cepstral Coefficients (MFCCs) etc. are utilized for feature extraction and methods like Dynamic Time Warping (DTW), Vector Quantization (VQ), Hidden Markov Model (HMM), Gaussian Mixture Models (GMM) etc are used with a view to identify voice signal.**

***Index Terms*— LPC, MFCC, DTW, VQ, HMM, GMM.**

## I. INTRODUCTION

Human voice contains various discriminative features that can be used in speaker recognition. The main goal of speaker recognition is to automatically identify a speaker by his/hers voice among a population. Recent development has made it possible to use this in voice dialling, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers etc. Speaker recognition is the task of recognizing a speaker based on the information obtained from his/her speech signal. Speaker recognition may be divided into speaker identification and speaker verification. Speaker identification is the process of determining the identity of the person that produced the speech from among a population of speakers. Speaker verification is the process of accepting or rejecting the identity claim of the speaker. Based on the text to be spoken, speaker recognition methods can also be grouped into text-dependent and text-independent cases. Text-dependent speaker recognition systems require the speaker to produce speech for the same text in both training and testing, whereas text-independent speaker recognition text in both training and testing may not be same.

**Manuscript received September, 2013.**

**Tejal Chauhan**, Research Student, E & C, Patel college of Science and Technology, Bhopal, India.

**Hemant Soni,** Assistant Professor, E & C, Patel college of Science and Technology, Bhopal, India.

**Sameena Zafar,** Assistant Professor, E & C, Patel college of Science and Technology, Bhopal, India.

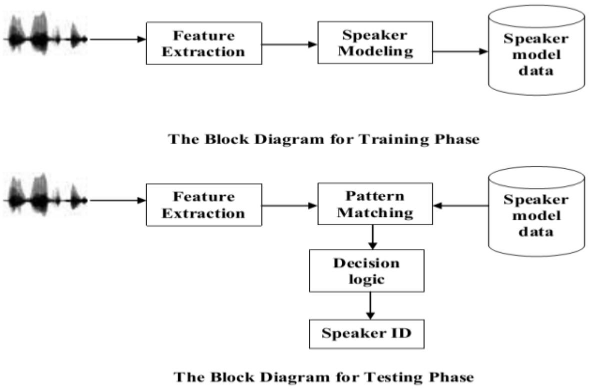

Fig.1 Training and testing phase [1][2]

## II. PRINCIPLE OF SPEAKER RECOGNITION

The system will operate in two modes as shown in fig 1: A training phase and a testing phase. The training phase will allow the user to record voice and make a feature model of that voice. The testing phase will use the information that the user has provided in the training mode and attempt to isolate and identify the speaker.

## III. SPEECH FEATURE EXTRACTION

Several feature extraction algorithms are used to this task such as; linear predictive coefficients (LPC), Mel frequency cepstral coefficients (MFCC) etc. 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.

### *A. LPC*

LPC (Linear Predictive coding) 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.

Linear predictive coding (LPC) is one of the earliest standardized coders. LPC has been proven to be efficient for the representation of speech signal in mathematical form. LPC is a useful tool for feature extraction as the vocal tract can be accurately modeled and analysed. Studies have shown that the current speech sample is highly correlated to the previous sample and the immediately preceding samples. LPC coefficients are generated by the linear combination of the past speech samples using the autocorrelation or the auto covariance method and minimizing the sum of squared difference between predicted and actual speech sample

$$y(n) = a_1x(n-1) + a_2x(n-2) + \cdots + a_Mx(n-M) = \sum_{i=1}^{M} a_i\, x(n-i)$$

$y(n)$ is the predicted $x(n)$ based on the summation of past samples. $ai$ is the linear prediction coefficients. M is the number of coefficients and n is the sample. The error between the actual sample and the prediction can then be expressed by

$$\mathcal{E}(n) = x\,(n) - y(n)$$

$$\mathcal{E}(n) = x(n) - \sum_{i=1}^{M} a_i\, x(n-i)$$

$$x(n) = \sum_{i=1}^{M} a_i x(n-i) + \mathcal{E}(n)$$

The speech sample can then be accurately reconstructed by using the LP coefficients $ai$ and the residual error $\mathcal{E}(n)$. [1][7]

### B. MFCC

The Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. The difference between the cepstrum and the Mel-frequency cepstrum is that in MFC the frequency bands are equally spaced on the Mel scale, which approximates the human auditory system's response. MFCCs are commonly used as features in speech recognition system. To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. Speech signal pre-processing covers digital filtering and speech signal detection. Filtering includes pre-emphasis filter and filtering out any surrounding noise using several algorithms of digital filtering. In general, the digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range, pre-emphasis is applied. The process of dividing the total number of speech samples by number of sample in one frame. The range of one frame duration is 20 to 40 msec. In this range, the speech signal is for the most part stationary. Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines The use for hamming windows is due to the fact that MFCC will be used which involves the frequency domain (hamming windows will decrease the possibility of high frequency components in each frame due to such abrupt slicing of the signal). After windowing first FFT and then Mel spaced filter banks are applied to get the Mel-spectrum. FFT converts from time domain to frequency domain. We use Mel scale because human perception of the frequency content of sounds does not follow a linear scale. This observation is often accounted for in acoustic feature extraction by passing the power spectrum of each frame through a bank of filters that are non- uniformly spaced in frequency.

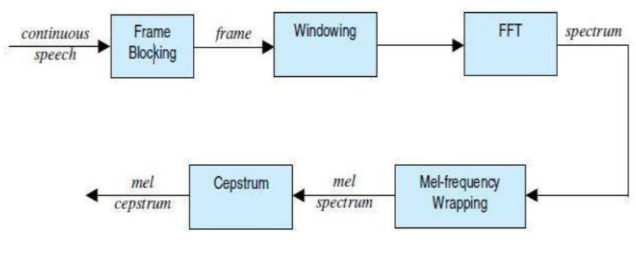

Fig 2. MFCC block diagram

Mel scale is roughly linear from 0 to 1 kHz and logarithmic after that. The Mel scale is defined as shown in Equation.

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

The natural logarithm is taken to transform into the cepstral domain and the discrete cosine transform (DCT) is finally applied to get the MFCCs. DCT de-correlates the features and arranges them in descending order of information they contain about speech signal.[2][3]

## IV. SPEAKER MODELING TECHNIQUE

The decision making process to determine a speaker's identity is based on previously stored information. This step is basically divided into two modes: training and testing. Training is a process of enrolling a speaker into the identification system database by constructing a unique model for each speaker based on the features extracted from the speaker's speech sample. Testing is a process of computing a matching score, which is a measure of similarity of the features extracted from the unknown speaker and the stored speaker models in the database. The speaker with the minimum matching score is chosen to be identified as the unknown speaker. The classification or speaker modeling techniques include, Dynamic Time Warping (DTW), Vector Quantization (VQ), Gaussian Mixture Modeling (GMM) etc.

### A. Dynamic Time Warping

DTW algorithm is based on Dynamic Programming techniques. This algorithm is for measuring similarity between two time series which may vary in time or speed. This technique also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find corresponding regions between the two time series or to determine the similarity between the two time series. Figure 5.1 shows the example of how one time series is „warped" to another. In Fig 3, each vertical line connects a point in one time series to its correspondingly similar point in the other time series. The lines have similar values on the $y$-axis, but have been separated so the vertical lines between them can be viewed more easily. If both of the time series in Fig 3 were identical, all of the lines would be straight vertical lines because no warping would be necessary to "line up" the two time series. The warp path distance is a measure of the difference between the two time series after they have been warped together, which is measured by the sum of the distances between each pair of points connected by the vertical lines in Fig 3. Thus, two time series that are identical except for localized stretching of the time axis will have DTW distances of zero. The principle of DTW is to compare two dynamic patterns and measure its similarity by calculating a minimum distance between them. [9]

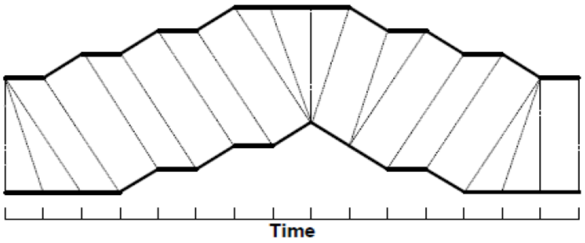

Fig 3 A Warping between two time series

### B. Vector Quantization

Vector Quantization is used to compress the information and manipulate the data in such a way as to maintain the most prominent characteristics. Vector Quantization is the classical quantization technique from signal processing which allows the modelling of probability density functions by the distribution of prototype vectors. It works by dividing a large set of points into groups having approximately the same number of points closest to them. Each group is represented by its centroid point. The density matching property of vector quantization is powerful, especially for identifying the density of large and high-dimensioned data. Since data points are represented by the index of their closest centroid, commonly occurring data have low error, and rare data high error. Hence, Vector Quantization is also suitable for lossy data compression. A vector quantizer maps k-dimensional vectors in the vector space $R^k$ into a finite set of vectors Y = {yi : i = 1, 2, ..., N}. Each vector yi is called a code vector or a codeword and the set of all the code words is called a codebook. Associated with each codeword, yi, is a nearest neighbor region called Voronoi region, and it is defined by

$$V_i = \{ x \ \epsilon \ R^k : || x - y_i || < || x - y_j || , \text{ for all } j \neq 1 \}$$

Input vectors are marked with an x, codeword are marked with circles, and the Voronoi regions are separated with boundary lines. The representative codeword is determined to be the closest in Euclidean distance from the input vector. The Euclidean distance is defined by

$$d(x,y_i) = \sqrt{\sum_{j=1}^{k} (x_j - y_{ij})^2}$$

Where xj is the jth component of the input vector, and yij is the jth is component of the codeword yi. VQ is applied on the set of feature vectors extracted from speech sample and as a result the speaker codebook is generated. There are a number of algorithms for codebook generation such as: K-means algorithm, Generalized Lioyd algorithm (GLA) (also known as Linde-Buzo-Gray (LGB) algorithm), Self Organizing Maps (SOM) and Pairwise Nearest Neighbor (PNN).[1]

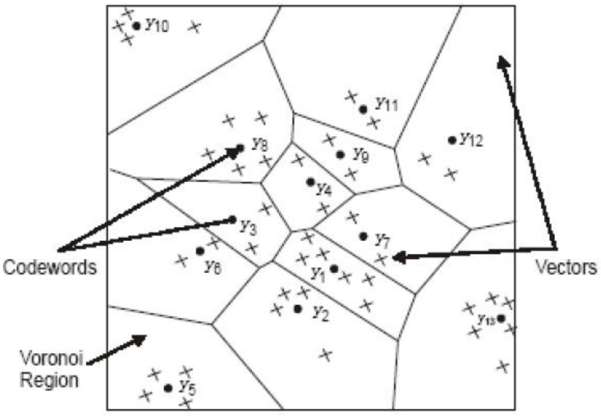

Fig 4 Codewords in 2-dimensional space

### C. *Gaussian Mixture Model*

Another type of speaker modeling techniques is Gaussian mixture modeling (GMM). This method belongs to the stochastic modeling and based on the modeling of statistical variations of the features. Therefore, it provides a statistical representation of how speaker produces sounds. The principle of GMM is to abstract a random process from the speech, then to establish a probability model for each speaker. It is relatively independent between the various probability models. Assuming the variable *M* in the M-order GMM probability density function is the number of Gaussian probability density functions. And set *X* as the feature vector from feature extraction block of the speech

$$P(X/\lambda) = \sum_{i=1}^{M} \omega_i \ b_i(X)$$

Where *bi* (*X)* is the sub-distribution and is given by the following equation

$$b_i(x) = \frac{1}{(2\pi)^{N/2} \, |\Sigma_i|^{1/2}} \exp(-\frac{1}{2}(X\text{-}\mu_i)^T \Sigma_i^{-1}(X-\mu_i))$$

Where $\mu i$ is the mean vector, Σi is the full covariance matrix. The mixed weights $\omega i$ should satisfy the following condition

$$\sum_{i=1}^{M} \omega_i = 1$$

The full GMM is constituted of the mean vectors, the covariance matrix and the mixed weights, which could be expressed as: λ = { $\omega i$ , $\mu i$, , Σi ,}, i=1,2,……M. For a given time series X = {*X*t}, t = 1,2,...,T. where T is the total number of frames of the feature vectors. The logarithm likelihood obtained from the GMM can be defined as the following equation

$$L(X/\lambda) = \frac{1}{T} \sum_{t=1}^{T} \log p(X_t / \lambda)$$

Set the diagonal covariance matrix as $Sigma_i$ , which can be defined as:

$$Sigma_i = \begin{bmatrix} \sigma_1 & 0 & 0\ldots0 \\ 0 & \sigma_2 & 0\ldots0 \\ & \ldots & \\ 0 & 0 & \ldots \quad \sigma_N \end{bmatrix}$$

Where σ1, σ2, σ3….. σN, are the main diagonal elements of Σi, i=1,2,3……M. Assuming we have S speakers in a closed-set which is different from each other. For a given speech feature vector {*X*t}, t = 1,2,...,T. The purpose of speaker recognition is to find the speaker k in the closed-set k ϵ {1,2,...,S}, whose corresponding model $\lambda_k$ will obtain the largest posterior probability *P*( $\lambda_k$ / *X* ). [2][3][5]

## V. CONCLUDING REMARK

In this paper, techniques for feature extraction like LPC and MFCC were discussed. Various speaker modeling techniques like VQ, DTW, GMM, HMM were also discussed. Speaker can be identifying efficiently using the technique of feature extraction and modeling discussed. These techniques are able to authenticate the particular speaker, based on the individual information that is included in the voice signal.

## REFERENCES

[1] Yuan Yujin, Zhao Peihua, Zhou Qun,, "Research of speaker recognition based on combination of LPCC and MFCC", Intelligent Computing and Intelligent Systems (ICIS), IEEE International Conference , vol.3, 29-31 Oct. 2010, pp.765-767.

[2] Sinith, M.S., Salim, A., Gowri Sankar, K., Sandeep Narayanan, K.V. Soman, V., "A novel method for Text-Independent speaker identification using MFCC and GMM", Audio Language and Image Processing (ICALIP), 2010 International Conference, Nov. 2010, pp.292-296.

[3] Zufeng Weng, Lin Li, Donghui Guo , "Speaker recognition using weighted dynamic MFCC based on GMM", Anti-Counterfeiting Security and Identification in Communication (ASID), International Conference, china, July 2010 IEEE, , pp.285-288

[4] R. Hasan, M. Jamil, G. Rabbani, S. Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients", 3rd International Conference on Electrical & Computer Engineering, Dhaka, Bangladesh, December 2004.

[5] A. Shende, S. Mishra, Shiv Kumar,"Comparision of different parameters used in GMM based Automatic Speaker Recognition" International Journal of Soft Computing and Engineering(IJSCE), Volume 1, Issue 3 July 2011.

[6] Douglas A. Reynolds and Richard C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models", Ieee Transactions On Speech And Audio Processing, Vol. 3, January 1995

[7] Jiang Hai, Er Meng Joo, "Improved linear predictive coding method for speech recognition", Information, Communications and Signal Processing and Fourth Pacific Rim Conference on Multimedia. Proceedings of the Joint Conference of the Fourth International Conference , vol.3, Dec. 2003, pp. 1614- 1618

[8] Bin Amin, T., Mahmood, I., "Speech Recognition using Dynamic Time Warping", Advances in Space Technologies, 2nd International Conference, Nov. 2008, pp.74-79

[9] Ashish Kumar Panda, Amit Kumar Sahoo, "Study Of Speaker Recognition Systems", Rourkela 2011.