International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 4 Issue 6, June-2015, pp: (65-71), Impact Factor: 1.296, Available online at: www.erpublications.com

# Speaker Recognition System: A Review

Prateek Sangwan[1], Shamsher Malik[2], Deepak Sharma[3]

[1,3]UIET MDU, Rohtak, Haryana, India
[2]Asstt. Prof., UIET MDU, Rohtak, Haryana, India

**Abstract: Speaker Recognition is a process of automatically recognizing who is speaking on the basis of the individual information included in speech waves. Speaker Recognition is one of the most useful biometric recognition techniques in this world where insecurity is a major threat. Many organizations like banks, institutions, industries etc are currently using this technology for providing greater security to their vast databases. Speaker Recognition mainly involves two modules namely feature extraction and feature matching. Feature extraction is the process that extracts a small amount of data from the speaker's voice signal that can later be used to represent that speaker. Feature matching involves the actual procedure to identify the unknown speaker by comparing the extracted features from his/her voice input with the ones that are already stored in our speech database. This paper will give a brief review of speaker recognition system, its building blocks, various techniques which can be used in feature extraction and feature matching and challenges involve in its designing.**

**Keywords: Mel-Frequency cepstral coefficients (MFCC), Gaussian mixture model (GMM), Fuzzy integral, Feature vector, Back propagation algorithm.**

## I. INTRODUCTION

Speech is the most fundamental signal that is used by human beings to convey information. Speech contains information not only about the message that is to be delivered to the other listener but it also contains information about the gender of the speaker, the language of the speaker, the age of the speaker, the ethnicity, the state of mind of the speaker (happiness, sadness and other emotions), the nature of the speaker (polite, kind, harsh etc) [1]. The speech signal is a slowly time-varying signal and when it is examined over a sufficiently short period of time, its characteristics are fairly stationary, but over long periods of time the signal characteristics change to reflect the different speech sounds being spoken. In many cases, short-time spectral analysis is the most common way to characterize the speech signal. An example of a speech signal is shown in Figure 1.

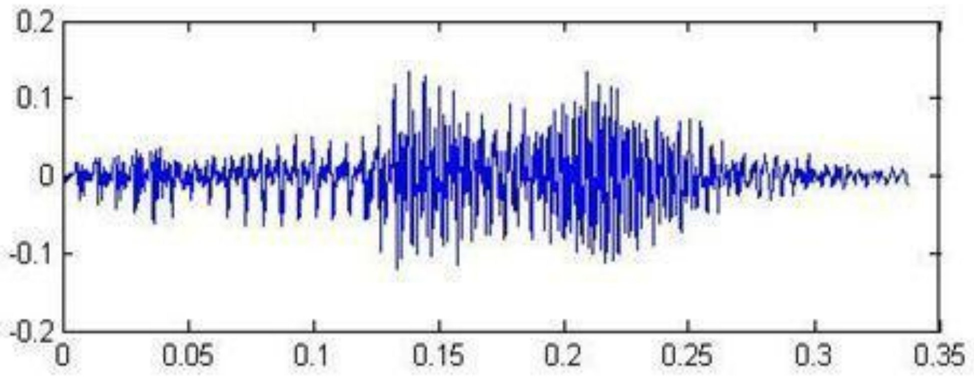

**Figure 1 : Example of a Speech signal**

Speech processing is a branch of science that deals with processing of the audio signals. It can be classified into three branches, i.e, analysis/synthesis, recognition and coding. Recognition can be further classified as speech recognition, speaker recognition and Language identification. Speaker recognition is an important branch of speech processing. It is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves as shown in figure 2 [6].

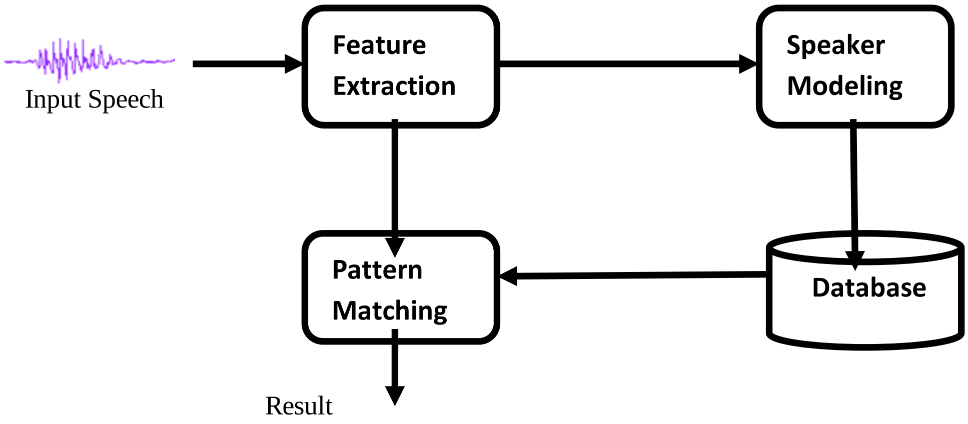

**Figure 2 : Speaker Recognition process**

Speaker recognition Systems can be broadly classified into two categories:-

- Speaker identification: In this, the goal is to determine which one of a group of known voices best matches the input voice sample.
- Speaker verification: In this, the goal is to determine from a voice sample if a person is who he or she claims to be [2].

There are further two classifications of speaker identification:-

- Open set identification: In this, a decision is made as to who is the speaker among all the already present speakers in the database (speaker data base is created by recording the sample speeches of various speakers).If the speaker database does not contain the recorded speech of the speaker, it is said that the speaker is not present in the database.
- Closed set identification: In this, a decision is made as to who is the most likely the speaker among all the already present speakers in the database. Thus, the results are that of the nearest matching features [1].

Classifications based on algorithm used for identification are:

- Text dependent: In a text- dependent system, the speech used to train and test the system is constrained to be the same word or phrase.
- Text independent: In a text-dependent system, the training and testing speech are completely unconstrained. In order to make such system an algorithm must be implemented that can extract unique features from a speaker that are independent of what is spoken [2].

Speaker recognition system works in following steps:-

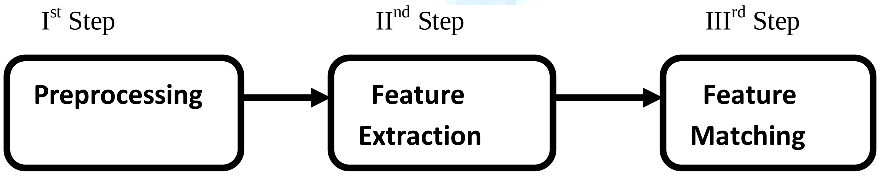

**Figure 3 : Block diagram of speech recognition system**

## II. PREPROCESSING

Before extracting the features of the signal various pre-processing tasks must be performed. The speech signal needs to undergo various signal conditioning steps before being subjected to the feature extraction methods. Pre-processing the signal reduces the computational complexity while operating on the speech signal. These tasks include:-

- Truncation
- Frame blocking

- Windowing
- Fast Fourier Transform

**(a)** TRUNCATION:- The signal is truncated by selecting a particular threshold value. We can mark the start of the signal where the signal goes above the value while traversing the time axis in positive direction. In the same we can have the end of the signal by repeating the above algorithm in the negative direction.

**(b)** FRAME BLOCKING:- In this step the continuous speech signal is divided into frames of N samples, with adjacent frames being separated by M samples with the value M less than that of N. The first frame consists of the first N samples. The second frame begins from M samples after the first frame, and overlaps it by N - M samples and so on. This process continues until all the speech is accounted for using one or more frames .

**(c)** WINDOWING:- The next step is to window each individual frame to minimize the signal discontinuities at the beginning and end of each frame. The concept applied here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame.

**(d)** FAST FOURIER TRANSFORM:- The next step is the application of Fast Fourier Transform (FFT),which converts each frame of N samples from the time domain into the frequency domain. The FFT which is a fast algorithm to implement the Discrete Fourier Transform(DFT) is defined on the set of N samples $\{x_n\}$, as follows:-

$$X_K = \sum_{n=0}^{N-1} x.\, e^{-j2\Pi kn/N}$$

In general $X_k$'s are complex numbers and we consider only their absolute values. Here, k=0, 1, 2,..., N-1 The resulting sequence $\{X_k\}$ is interpreted as follows: positive frequencies $0 \leq f < F_s / 2$ correspond to values $0 \leq n \leq N / 2 -1$, while negative frequencies $- F_s / 2 < f < 0$ correspond to $N / 2 +1 \leq n \leq N-1$. $F_s$ denotes the sampling frequency. The result after this step is often referred to as spectrum or periodogram.

## III. FEATURE EXTRACTION

Feature extraction is the process that extracts a small amount of data from the speaker's voice signal that can later be used to represent that speaker. The purpose of this module is to convert the speech waveform into a set of features or rather feature vectors (at a considerably lower information rate) for further analysis. Extracted features should have some criteria in dealing with the speech signal such as:

- Stable over time
- Should occur frequently and naturally in speech
- Should not be susceptible to mimicry
- Easy to measure extracted speech features
- Shows little fluctuation from one speaking environment to another
- Discriminate between speakers while being tolerant of intra speaker variabilities

Many feature extraction techniques are available, some of them are:-

- Mel-frequency cepstral coefficients(MFCC)
- LPC-based cepstral parameters
- Relative spectra filtering of log domain coefficients (RASTA)
- Local discriminant bases (LDB)

**(a)** Mel-frequency cepstral coefficients (MFCC): After conditioning the speech signal i.e. after pre-processing, the next step is to extract the features in the form of mel frequency cepstral coefficients. MFCC's are coefficients that represent audio, based on perception. It can also derived from the Fourier Transform or the Discrete Cosine Transform of the audio clip. The basic difference between the FFT/DCT and the MFCC is that in the MFCC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system's response more closely than the linearly spaced frequency bands of FFT or DCT. This allows for better processing of data, for example, in audio compression. The main purpose of the MFCC processor is to mimic the behaviour of the human ears. The MFCC process is subdivided into five phases or blocks as shown in figure 4. In the frame blocking section, the speech waveform is more or less divided into frames of approximately 30 milliseconds. The windowing block minimizes the discontinuities of the signal by tapering the beginning and end of each frame to zero. The FFT block converts each frame from the time domain to the frequency domain. In the Mel frequency wrapping block, the signal is plotted against the Mel-spectrum to mimic human hearing.

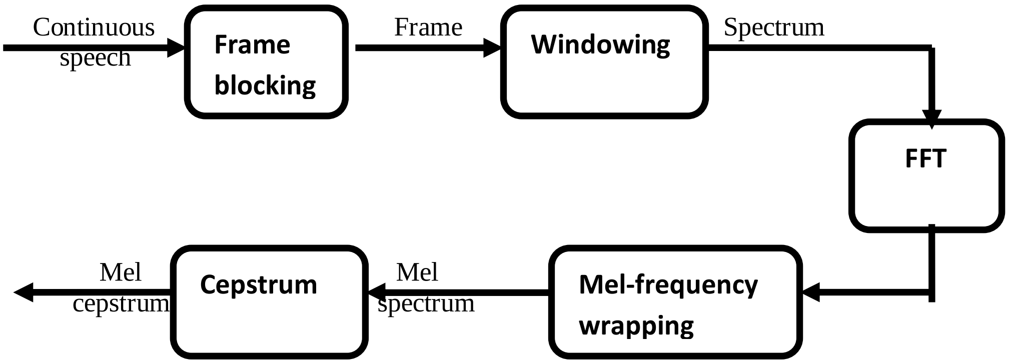

**Figure 4 : Block diagram of MFCC Processor**

Studies have shown that human hearing does not follow the linear scale but rather the Mel-spectrum scale which is a linear spacing below 1000 Hz and logarithmic scaling above 1000 Hz.This is done by using the following equation:

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

In the final step, we convert the log Mel spectrum to time domain. The result is called the MFCC (Mel Frequency Cepstral Coefficients). This representation of the speech spectrum provides a good approximation of the spectral properties of the signal for the given frame analysis. The Mel spectrum coefficients being real numbers are then converted to time domain using Discrete Cosine Transform (DCT). we can calculate the MFCC's $C_n$ as:

$$C_n = \sum_{k=1}^{K} (\log S_k).\cos[\frac{n\left(k-\frac{1}{2}\right)\Pi}{k}]$$

K:- the number of mel spectrum coefficients
$S_k$:- Mel power spectrum coefficients

**(b)** LPC-based cepstral parameters: The LPC analysis is based on a linear model of speech production. The model usually used is an auto regressive moving average (ARMA) model, simplified in an auto regressive (AR) model. The speech production apparatus is usually described as a combination of four modules: (1) the glottal source, which can be seen as a train of impulses (for voiced sounds) or a white noise (for unvoiced sounds); (2) the vocal tract; (3) the nasal tract; and (4) the lips. Each of them can be represented by a filter: a low pass filter for the glottal source, an AR filter for the vocal tract, an ARMA filter for the nasal tract, and an MA filter for the lips. Globally, the speech production apparatus can therefore be represented by an ARMA filter. The principle of LPC analysis is to estimate the parameters of an AR filter on a windowed (pre-emphasized or not) portion of a speech signal. Then, the window is moved and a new estimation is calculated. For each window, a set of coefficients (called predictive coefficients or LPC coefficients) is estimated and can be used as a parameter vector. Finally, a spectrum envelope can be estimated for the current window from the predictive coefficients [3].

**(c)** The relative spectral analysis technique (RASTA): It is based on the idea that the rate of changing of the short-term spectrum for linguistic and non-linguistic components in speech is different. This means that the spectral components of the communication channel vary more quickly or more slowly than the spectral components of the speech and they could be separated (filtered). The core part of RASTA processing is a band-pass filtering of the spectral parameters trajectories by an IIR filter. The convoluted (in the time domain) distortions in the communication channel can be reduced by using the RASTA filtering in the logarithmic domain (spectral or cepstral). The RASTA approach can be combined with the perceptually linear prediction method (so called PLP-RASTA approach) or can directly be applied to the cepstral trajectories [3].

**(d)** Local discriminant bases (LDB): It is an audio feature extraction and a multi group classification scheme that focuses on identifying discriminatory time-frequency subspaces. Two dissimilarity measures are used in the process of selecting the LDB nodes and extracting features from them. The extracted features are then fed to a linear discriminate analysis based classifier for a multi-level hierarchical classification of audio signals [3].

## IV. FEATURE MATCHING

Feature matching is a classification procedure to classify objects of interest into one of a number of classes. The objects of interest are called patterns are sequences of feature vectors that are extracted from an input speech using the MFCC processor. Each class here refers to each individual speaker. There are a lot of feature matching techniques used in speaker recognition such as:

- Vector Quantization (VQ)
- Support Vector Machine (SVM)
- Gaussian Mixture Model (GMM)
- Hidden Markov Modeling (HMM)

**(a)** Vector Quantization (VQ): It is also known as centroid model. It is one of the simplest text-independent speaker models. It was introduced to speaker recognition in 1980s and its roots are originally in data compression. Even though VQ is often used for computational speedup techniques and lightweight practical implementations, it also provides competitive accuracy when combined with background model adaptation [4].

**(b)** Gaussian mixture model (GMM): It is a stochastic model which has become the de facto reference method in speaker recognition. The GMM can be considered as an extension of the VQ model, in which the clusters are overlapping. A GMM is composed of a finite mixture of multivariate Gaussian components [4].

**(c)** Support vector machine (SVM): It is a powerful discriminative classifier that has been recently adopted in speaker recognition. It has been applied both with spectral, prosodic, and high-level features. Currently SVM is one of the most robust classifiers in speaker verification, and it has also been successfully combined with GMM to increase accuracy. One reason for the popularity of SVM is its good generalization performance to classify unseen data [4].

**(d)** Hidden Markov Model (HMM): It provide more flexibility and produce better matching score. In this, the process of pattern matching is carried out by measuring the likelihood of a feature vector in a given speaker model.The Hidden Markov Model is widely used for modeling of sequences. This technique efficiently models the statistical variations of the features and provides a statistical representation of the manner in which a speaker produces sounds [3].

## V. TRAINING AND TESTING

Like any other pattern recognition systems, speaker recognition systems also involve two phases:

- Training
- Testing

Training is the process of familiarizing the system with the voice characteristics of the speakers registering. Testing is the actual recognition task. The block diagram of training phase is shown in Figure 4. Feature vectors representing the voice characteristics of the speaker are extracted from the training utterances and are used for building the reference models [14].

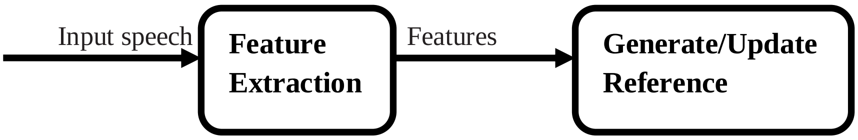

**Figure 5 : The block diagram of training phase**

During testing, similar feature vectors are extracted from the test utterance, and the degree of their match with the reference is obtained using some matching technique. The level of match is used to arrive at the decision. The block diagram of the testing phase is given in Figure 5 [14].

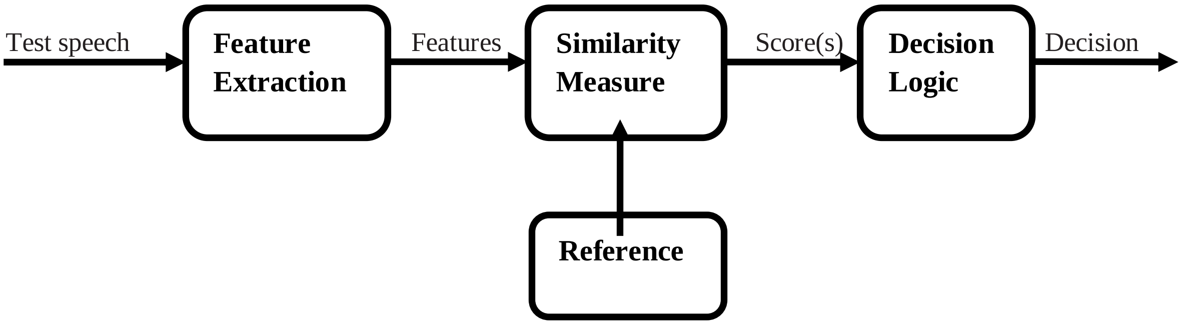

**Figure 6 : The block diagram of testing phase**

## VI. CHALLENGES IN THE DESIGNING OF ASR SYSTEM

The major technique used for speaker recognition is based on MFCC and GMM (Gaussian Mixture Model). The use of Gaussian mixture models for modelling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities. But the approaches based on MFCC and GMM are known to perform very well for small population speaker identification under low noise conditions. They also have some drawbacks:

- The first drawback is that they suffer from the mismatch between training and testing caused by noisy conditions. The noisy conditions can severely degrade the identification performance.
- The second drawback is actually a common problem of almost all existing speaker identification techniques. The success of almost all existing identification systems (including GMM-based systems) lies in the fact that they are trained on datasets with only a relatively small population.

However, it is pretty straightforward that when the population has a significant increase (e.g., thousands of registered speakers or even more), the probability of identification errors will significantly increase, accordingly. Unfortunately, there are not much existing research work studying this problem. Some papers mainly focused on reducing the computational complexity in large population cases at the cost of a very slight accuracy loss. In some other papers which claimed to deal with large population identification, the experiments were actually carried out on datasets with only hundreds of registered speakers [5].

## VII. CONCLUSION

Speaker recognition system uses the voice of speaker to verify their identity and control access to services such as voice dialing, mobile banking, database access services, voice mail or security control to a secured system. But even the major technique for ASR system can achieve superior performance only for small population under low noise conditions. This paper gives a brief review of speaker recognition system, its building blocks, techniques involved in it and challenges in its designing. In future, we intend to implement the ASR system in MATLAB, which uses fuzzy integral for feature matching and back propagation algorithm for training the network to obtain more accuracy and less computational complexity.

## REFERENCES

[1]. Amit Kumar Singh, Rohit Singh,Ashutosh Dwivedi "Mel Frequency Cepstral Coefficients BasedText Independent Automatic Speaker Recognition Using Matlab" 2014 International Conference - ICROIT 2014, India, Feb 6-8 2014.

[2]. D. A. Reynolds, W. M. Campbell " Text- Independent Speaker Recognition" Springer Handbook of speech Processing Benesty, Sondhi, Huang (Eds.) · © Springer 2008.

[3]. Nisha.V.S , M.Jayasheela "Survey on Feature Extraction and Matching Techniques for Speaker Recognition Systems" International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 2, Issue 3, March 2013.

[4]. Hemlata Eknath Kamale, Dr.R. S. Kawitkar "Vector Quantization Approach for Speaker Recognition" International Journal of Computer Technology and Electronics Engineering (IJCTEE) Volume 3, Special Issue, March-April 2013.

[5]. Yakun Hu, Dapeng Wu, Fellow, IEEE, and Antonio Nucci "Fuzzy-Clustering-Based Decision Tree Approach for Large Population Speaker Identification" IEEETransaction on Audio, Speech, and Language processing, VOL. 21, NO. 4, APRIL 2013.

[6]. Izuan Hafez Ninggal & Abdul Manan Ahmad "The Fundamental of Feature Extraction in Speaker Recognition : A Review" Proceedings of the Postgraduate Annual Research Seminar 2006.

[7]. Vibha Tiwari “MFCC and its applications in speaker recognition” International Journal on Emerging Technologies 1(1): 19-22(2010).

[8]. Nisha.V.S , M.Jayasheela “Speaker Identification Using Combined MFCC and Phase Information” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 2, February 2013.

[9]. Jung-Hsien Chiang, Member, IEEE “Choquet Fuzzy Integral-Based Hierarchical Networks for Decision Analysis” IEEE Transactions on Fuzzy Systems, VOL. 7, NO. 1, February, 1999.

[10]. James M. Keller and Jeffrey Osborn “Training the Fuzzy Integral” International Journal of Approximate Reasoning 1996; 15:1-24 © 1996 Elsevier Science Inc. .

[11]. Douglas A. Reynolds “Automatic Speaker Recognition Using Gaussian Mixture Speaker Models” Volume B, Number 2,1995 The Lincoln Laboratory Journal.

[12]. Kashyap Patel, R.K. Prasad “Speech Recognition and Verification Using MFCC & VQ” International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-7, May 2013.

[13]. Rivarol Vergin, Douglas O’Shaughnessy, Senior Member, IEEE, and Azarshid Farhat “Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition” IEEE Transactions on Speech and Audio Processing, VOL. 7, NO. 5, SEPTEMBER 1999.

[14]. Sejal Shah, Archana Bhise “Fast Speaker Recognition using Efficient Feature Extraction Technique” IJCSN International Journal of Computer Science and Network, Vol 2, Issue 1, 2013.

[15]. S G Bagul* & Prof R.K.Shastri “Text Independent Speaker Recognition System using GMM” International Journal of Scientific and Research Publications, Volume 2, Issue 10, October 2012.

[16]. Douglas A. Reynolds, Member, IEEE, and Richard C. Rose, Member, IEEE “ Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models” IEEE Transactions on Speech and Audio Processing,VOL. 3, NO. 1. JANUARY 1995.

[17]. Bhupinder Singh, Rupinder Kaur, Nidhi Devgun, Ramandeep Kaur “The process of Machine Interaction with Humans: A Review” International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 2, February 2012.

ER PUBLICATIONS