CHAPTER II
LITERATURE REVIEW

Speaker Recognition Systems include four steps i.e. pre-processing, feature extraction, speaker modeling and methods of assessing the speaker recognition performance. This chapter gives a brief introduction to speaker recognition. This includes description of the main components that make up a speaker identification system. At the outset, the process of speech production in humans is explained so as to get an idea of key audible characteristics used by humans in identifying speakers. A literature survey of the major techniques developed by various researchers in each stage of speaker recognition has been carried out. This gives a clear understanding of the developments that have taken place in each stage. All the available choices are analyzed on the basis of their relative merits and demerits. The major drawback of the present day speaker recognition system has been identified as slow convergence rate and scalability. Feature selection is an utmost important task in any speaker recognition system. A comparative study of different dimension reduction techniques for feature selection has been done at the end of chapter. The detailed literature survey and review of the research papers and technical papers regarding speaker recognition system in general and in the area of improving its overall performance in particular, has been presented.

2.1 INTRODUCTION
The speech signal is a complex signal and it carries various kinds of speaker specific information. The low-level properties include pitch, tone, volume, formant frequencies, spectral correlations, short-time spectrum. Also, high level properties such as dialect, speaking style, context, and emotional state of speaker are obtained from speech. Speech is a slow varying signal. The amount of data which is generated during the speech production process is quite large and not all of them contain useful information. Relatively less data is required to represent the characteristics of speech and the person who has spoken it. Hence, feature extraction can be defined as the process of reducing data while retaining the information that can be used to discriminate the speakers. Since, the physical traits obtained from the speech signal is
used for the speaker recognition task, it is very important to understand properly how voice is produced and perceived by a human being.

2.1.1 Speech Production Mechanism in Human Beings

No two individuals have similar voice. The main reason for this difference lies in the construction of their articulatory organs. The length of the vocal tract, characteristics of the vocal cord, size and shape of the mouth, teeth etc. contribute to give every individual a unique voice. The speech production organs are shown in Figure 2.1. The length of adult vocal tract is approximately 17 cm. Excitation signals are generated in the following ways – phonation, whispering, frication, compression, and vibration. Phonation is generated when vocal cords oscillate. These cords can close and stretch because of cartilages. The oscillations depend on the mass and tension of the cords, and are also governed by the Bernoulli effect of the air passing through the vocal cords. The opening and closing of the cords breaks the air stream impulses. The shape and duty cycle of these pulses will depend on loudness and pitch of the speech signal.

The V-shaped opening between the vocal cords, known as the glottis, is the most important source of sound in the vocal system. It generates a periodic excitation because of the natural frequency of vibration of vocal cords. The vocal cords may act in several different ways during speech. The rapid opening and closing of vocal cords modulates the air flow and thus produces vowels and voiced consonants. The fundamental frequency of vibration of vocal cords is 150 Hz, 250 Hz and 300 Hz for men, women, and children, respectively.

The first step in production process is the formulation of a message in mind which is to be transmitted to the destination through speech. Then the message is converted into a language code. Also each sentence is associated with duration, loudness and pitch. A number of neuromuscular commands are executed which causes the vocal cords to vibrate. This changes the shape of the vocal tract and the proper sequence of speech sounds is produced. These commands also control the movements of the other speech production organs. At the output, an acoustic signal is finally produced.
Figure 2.1: Human Speech Production Organs

As shown in Figure 2.2, through the normal breathing mechanism air enters the lungs. As air passes from the lungs through the trachea the vocal cords vibrate. It is in the form of quasi periodic pulses. During the passage through the pharynx, the mouth and the nasal cavity it gets modulated in frequency. The positions of the various articulators i.e. jaw, tongue, velum, lips, mouth is responsible for the production of different sounds.

Figure 2.2: Schematic Representation of Speech Production (Patra, 2011)
2.1.2 Source Filter Model of Speech Production

The source of excitation and the vocal tract system are assumed to be independent of each other in a speech processing system. Based on this assumption the digital model for speech production is formulated.

A speech signal consists of a voiced segment and an unvoiced segment. The impulse train generator shown in Figure 2.3 is used as an excitation when a voiced segment is produced for example, during the utterance of any vowel, including a composite vowel. When consonants, such as ‘s’, are uttered, they are generated as a result of turbulence functioning as the excitation. The unvoiced segments are generated when the random signal generator is used as the excitation. The voice model can be used for a short segment of speech (of duration of around 10 ms) for which it is assumed that the parameters of vocal tract remain almost constant.

If the vocal tract is considered as a linear system, the speech signal $S(z)$ which is obtained at the output is due to the result of filtering an excitation signal with the vocal tract impulse response $H(z)$.

![Diagram](image)

Figure 2.3: Source Filter Model of Speech Production
Based on the mode of excitation, speech sounds can be classified into three categories:

**Voiced Speech**- Vocal tract is excited by producing quasi-periodic pulses of air. Therefore voiced speech exhibits quasi-periodic behavior. Vowels are usually classified as voiced sounds. These types of sounds have high average energy levels and very distinct formant frequencies. Such sounds are produced by forcing the air from the lungs over the vocal cords. Hence, vocal cords vibrate in a periodical pattern and generate series of air pulses called glottal pulses. These glottal pulses or air pulses travel through rest of the vocal tract to mouth, where some frequencies resonate. Pitch of the sound is defined as the rate at which vocal cords vibrate. Generally in women and children, due to a faster rate of vibration of the vocal cords while producing voiced speech, pitch is believed to be higher than in men. Therefore, pitch is also an important parameter to be included for analysis or synthesis of voiced sounds. The perceived pitch differs with the gender and age of the speaker. Its range for humans lie between 50 and 500 Hz. Children have the highest pitch voices followed by females and then males with the lowest pitch. Pitch varies with time and tells about the prosody of utterance. With age, females tend to lower their pitch and male voices tend to rise in pitch.

**Unvoiced speech**- Unvoiced speech produces sounds which have a random behavior and are generated as a result of constriction at some point in the vocal tract towards the mouth. A turbulence is produced when air passes through the constriction at a very high velocity. Thus noise is generated to excite the vocal tract. Unvoiced speech is also referred to as fricative speech. Consonants are classified as unvoiced sounds. Unvoiced sounds have lower energy levels and high frequencies than voiced sounds. Unvoiced sounds are produced when air is forced through the vocal tract with vocal cords open until the sound is produced in a turbulent flow. There is no vibration of vocal cords taking place here and therefore pitch does not come into picture.

**Plosive Sounds**- These sounds are generated due to complete closure towards the front of the vocal tract. This causes pressure to build up behind the closure and it is released abruptly. The frequencies at which the vocal tract resonates are called formant frequencies. They depend upon the shape and dimensions of the vocal tract.
2.1.3 Short Term Analysis of Speech Signal
Speech signal has quasi-stationary nature. When it is considered for a short duration of time say 20-30 milliseconds, it has stable acoustic characteristics. Hence, short-term analysis of speech signal is done for feature extraction. A fixed length window as shown in Figure 2.4 is moved along the signal. The adjacent frames overlapping is about 30-50%. The purpose of overlapping is that it prevents loss of information. The abrupt changes at the end points of the frame are avoided by multiplying it by a window function. Hamming window and Hanning window are usually used in speech analysis.

![Figure 2.4: Short-Term Analysis of Speech Signal]

2.2 BASIC STRUCTURE OF SPEAKER RECOGNITION SYSTEM
Speaker recognition systems works in four stages, namely, pre processing, feature extraction, modeling and testing. The major components of Speaker recognition system are shown in Figure 2.5. These are explained in detail.

(i) Preprocessing: Input data is usually preprocessed before being fed into the feature extraction phase. Preprocessing techniques include operations such as filtering, normalization, transformation, trimming, alignment, windowing, offset correction, smoothing etc. and depends on the application. For instance end-point detection for a speech recognizer is a commonly used preprocessing technique.
(ii) **Feature Extraction:** Features are higher level representations compared to raw data representations, for example, frequencies instead of raw temporal samples. Speech signals are fed to the front end processing system. The speech signal is initially converted into digital form. The next step is feature extraction. The features extracted from the speech signal convey the identity of the speaker. The shape of speech spectrum carries the information about shape of vocal tract and glottal source of a speaker. Therefore, spectral based features are widely used in most of the speaker recognition systems.

(iii) **Training:** Training phase corresponds to the process of learning from labeled data, i.e., training data. It can also be considered as detection of decision boundaries which distinguish different classes in the feature space. For the unsupervised case, there is no learning from labeled data but the decision boundary detection phase can be thought together with the classification phase.

(iv) **Testing:** Testing phase is conducted after training; this is when the stochastic models for each class (speaker) have been already built. During the testing (or recognition) phase, the speaker recognition system is exposed to speech data not seen during the training phase. Speech samples from an unknown speaker or from a
claimant are used to calculate feature vectors using the same methodology as in the enrolment process. These vectors are then passed to the classifier which performs a pattern matching task determining the closest-matching speaker model. This process results in a decision making process which determines either the speaker identity (in speaker identification) or accepts/rejects the claimant identity (in speaker verification).

All the information contained in speech signal is not relevant. It includes silent or noisy regions where the speaker is not saying anything. The overall system performance can be improved by removing such segments from the input signal. VAD is a technique that segregates an input signal to speech and non-speech segments. More commonly it is also known as silence removal method. But, the main drawback of VAD is that it may also remove non-silent portions.

In speaker identification systems, the identification time is required to be minimized so that the system works in real-time. Speaker identification is a computationally exhaustive task. Identification time increases with number of speakers, N. This problem is known as the curse of dimensionality. If the number of speakers or feature vectors is reduced, identification time reduces significantly. Using low-dimensional features, computational savings are observed. Processing delay (i.e. how much time it will take before the system can give the response) is a significant factor in real-time recognition.

For spoken language processing applications like speaker recognition/verification, not only that the silence segments do not contribute any speaker specific information, but also it dilutes the already available information content in the speech segments in the audio data. It has been experimentally studied that removing silence segments with the help of a VAD from the utterance before feature extraction enhances the performance of speaker recognition systems (Pillay, 2010). In order to increase noise robustness and to achieve computational savings, the speaker recognition system has been modified and two more techniques namely VAD and Feature reduction have been included.

Flow diagram of the entire speaker identification system used in this research is shown in Figure 2.6. After using the VAD on the speech, the features are extracted, feature selection is done and then the training set is sent to the speaker modelling. After the training is finished and the database build, the speaker matching for the test set can be done. The main emphasis is on the analysis of various dimensionality
reduction techniques available so that computation time and robustness can be improved for real time speaker recognition.

Figure 2.6: Flowchart of Speaker Recognition System

A brief description along with exhaustive literature survey of various techniques used in speaker recognition system is given in the following sections.

2.3 VOICE ACTIVITY DETECTION

A voice activity detector, VAD separates the portions containing the speech signal from the portions corresponding to silence and background noise. VAD increases the robustness of speaker recognition systems. Traditional methods for VAD make voiced/unvoiced decision based on the following:
- energy level
- autocorrelation function
- linear predictor coefficient
- zero crossing rates
In some cases the output from feature extraction algorithms is used whereas in other cases non-processed signal is used. Recent methods suggest that background noise can be modelled to distinguish whether signal frame contains speech or not. Kinnunen and Rajan (2013) suggested a method for VAD that is based on likelihood ratio. It trains speech and non-speech models on an utterance–by-utterance basis from mel-frequency cepstral coefficients (MFCCs). The training labels are obtained from enhanced energy VAD. Sadjadi and Hansen (2013) proposed an unsupervised speech activity detection scheme based on four different speech voicing measures which are combined with a perceptual spectral flux feature. Sreekumar et al. (2014) used spectral matching (SM) to distinguish between silence and speech segments. It outperforms the VAD using signal energy and spectral centroid. VAD is still not a solved problem and more research is required in this area as the performance degrades in adverse conditions i.e. at very low signal-to-noise ratios (SNRs).

2.4 FEATURE EXTRACTION METHODS USED IN SPEAKER RECOGNITION

The process of converting a raw speech signal into a sequence of acoustic feature vectors which carries speaker specific information is called feature extraction. Speech signal includes many features. But all of them are not important for speaker discrimination. An ideal feature should (Rose, 2002; Wolf, 1972):

- have large inter-speaker variations and small intra-speaker variations.
- be robust against noise and distortion.
- occur naturally in speech.
- be easy to measure from speech signal.
- be difficult to impersonate/mimic.
- not be affected by the speaker’s health or ageing.

The choice of features is application specific. It also depends on resources required for computation, amount of speech data available and whether the speakers are cooperative or not (Kinnunen and Li, 2009). Based on the domain in which the analysis is done, the characteristic features can be divided into:

a) **Spectral features** – It represents the short-term speech spectrum. The spectral features represent physical characteristics of the vocal tract.

b) **Dynamic features** – It describes the time variations of spectral features.
c) **Prosodic features** - It gives the fundamental frequency and energy contours. The prosodic features can be further divided into source features or suprasegmental features depending on the time duration of the analyzed speech segment. Source features span over a single glottal period whereas suprasegmental features span over a few glottal periods.

d) **High-level features** – It is a long time feature and span over the time duration of a word or utterance.

Reynolds et al. (2003) stated that the spectral features are easy to compute and yield good performance. Prosodic and high-level features are more robust (Ashour and Gath, 1999; Kitamura, 2008) but the main drawback is that they are less discriminative and easier to impersonate. More complex front-end is required for high-level features, i.e. automatic speech recognizer. These features are discussed in detail in the subsequent paragraphs.

### 2.4.1 Spectral Features

The speech signal is a time varying signal. Therefore, the signal need to be analyzed in short frames. The frame duration may lie between 20–30 ms. For this time interval, the signal is assumed to remain stationary. Pre-emphasis and windowing is done prior to feature extraction stage. A feature vector is extracted from each frame.

Harrington and Cassidy (1999) found that the attenuation that arises from voice source is about -12 dB / octave. Pre-emphasis boosts the higher frequencies of voiced sounds. Discrete Fourier transform (DFT) assumes that the signal is periodic. Windowing reduces the effect of the spectral artefacts (spectral leakage/smearing) that arise from discontinuities at the frame endpoints. Typically, Hamming window is used.

Oppenheim et al. (1999) used the fast Fourier transform (FFT). It transforms a signal into its frequency components. FFT is a fast implementation of DFT. Alternatives to FFT-based signal decomposition such as non-harmonic bases, data-driven bases and aperiodic functions were given by Gopalan et al. (1999), Jang et al. (2002). They were derived from independent component analysis (ICA). The DFT are more simple and efficient.

Davis and Mermelstein (1980) used the so-called Mel-frequency cepstral coefficients (MFCCs) which have become popular features in speech processing. Kinnunen et al.
(2007); Thian et al. (2004) studied various alternative features, such as spectral subband centroids but MFCCs still remains the best. Charbuillet et al. (2006); Miyajima et al. (2001) studied some alternative features that give more importance to speaker-specific information.

Mammone et al. (1996) studied linear prediction (LP) as an alternative spectrum estimation method to DFT. The predictor coefficients are not used as features. They are transformed into less correlated and robust features such as linear predictive cepstral coefficients (LPCCs) (Huang et al., 2001), line spectral frequencies (LSFs) (Huang et al., 2001), and perceptual linear prediction (PLP) coefficients (Hermansky, 1990).

Atal (1972); Kinnunen (2004) gave comparisons of these features. Campbell et al., (2006); Kinnunen (2004) suggested that different spectral features can complement each other and accuracy can be enhanced by combining them. In summary, any of the following features: MFCC, LPCC, LSF, PLP can be used.

Murty and Yegnanarayana (2006); Prasanna et al. (2006); Zheng et al.(2007) stated that the glottal features cannot be measured directly due to the vocal tract filtering effect. An alternative method was suggested by Gudnason and Brookes (2008). It makes use of closed-phase covariance during the portions when the vocal folds are closed. This gives improved estimate of the vocal tract but it requires exact detection of closed phase. It is difficult to find in noisy conditions.

From the literature, it is found that vocal tract features are more discriminative as compared to voice source features. These two complementary features can be combined to improve accuracy (Murty and Yegnanarayana, 2006; Zheng et al., 2007).

The voice source features requires less train and test data as compared to the amount of data needed for the vocal tract features (Chan et al., 2007; Prasanna et al., 2006). The phonetic content is required for the vocal tract features and large amount of phonetic data is required for both the train and test utterances. On the other hand, voice source features depend comparatively less on phonetic factors.

2.4.2 Dynamic Features

Speaker specific information is contained in the dynamic signals such as formant transitions and energy modulations. Furui (1981); Huang et al. (2001); Soong and Rosenberg (1988) suggested that temporal information can be extracted from features
through first and second order time derivatives. They are called delta and double-delta coefficients, respectively. Rabiner and Juang (1993) gave an alternative method which was more robust. They suggested a method that fits a regression line or an orthogonal polynomial to the temporal trajectories. But simple differentiation gives equal or better performance. Kinnunen (2004), Magrin-Chagnolleau et al. (2002) studied time-frequency principal components and Malayath et al. (2000) worked on data-driven temporal filters. Shifted Delta Cepstrum was used for the speaker recognition and has shown good results. A method called multivariate auto-regression (MAR) model is applied to the time series of cepstral vectors and is used to characterize speakers. The combination of the cepstra and delta cepstra features gives good results for the task of speaker recognition. The speaker recognition system performance can be increased by adding time derivatives to the static features. The first order derivatives referred to as delta features can be calculated using Equation 2.1.

\[
d_t = \frac{\sum_{\theta=1}^{6}(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{6} \theta^2}
\]

where \(d_t\) is the delta coefficient at time \(t\), computed in terms of the corresponding static coefficients \(c_{t-\theta}\) to \(c_{t+\theta}\) and \(\theta\) is the size of delta window.

Kinnunen et al. (2006); Kinnunen et al. (2008) studied reduced-dimensional spectro-temporal features. DCT was used instead of DFT. It has an advantage because the relative phases of the feature coefficient trajectories are retained. Both phonetic and speaker-specific information is preserved. However, more research is required to substantiate this finding.

### 2.4.3 Prosodic Features

The most important prosodic parameter is the fundamental frequency, \(F_0\) (Peskin, 2003). When \(F_0\)-related features are combined with spectral features it improves performance in noisy environments. Prosodic features for speaker recognition have included duration, speaking rate, and energy distribution / modulations (Adami et al., 2003; Bartkova et al., 2002; Shriberg et al., 2005). Without actual \(F_0\) extraction, multidimensional pitch and voice-related features can be extracted using the auto-correlation function as is suggested in
27

(Laskowski and Jin, 2009). An alternative way to improve accuracy is by using both the local and long-term temporal variations of F0 for modeling (Mary and Yegnanarayana, 2006). To capture local F0 dynamics the delta features with the instantaneous F0 value are added. F0 contour can be segmented and represented by a set of parameters associated with each segment for long term modelling. The segments may be syllables which is obtained from automatic speech recognition (ASR) system. An alternative is to divide the speech sample into syllable-like units using vowel onsets (Mary and Yegnanarayana, 2008) or F0/energy inflection points (Adami, 2007; Dehak et al., 2007) as the segment boundaries. Prosodic feature statistics and their local temporal slopes within each segment are used to get parameters of the segments. Shriberg et al. (2005) used N-gram counts of discretized feature values as features with an SVM classifier. Dehak et al. (2007) used prosodic features which were extracted using polynomial basis functions.

2.4.4 High-level Features
Speakers differ in their voice timbre, accent/pronunciation and in the vocabulary used. Vocabulary refers to the collection of words used by speakers in their conversations. Doddington (2001) did pioneering work on such high-level conversational features. The speaker’s characteristic vocabulary was used to characterize speakers. The high-level modeling converts each utterance into a sequence of tokens where the co-occurrence patterns of tokens characterize speaker differences. The information used for modeling is in discrete form.

2.5 SPEAKER MODELING - CLASSICAL APPROACHES
For speaker recognition, speaker models are generated using speaker-specific feature vectors (Jain et al., 2000; Mami and Charlet, 2006). By using feature vectors extracted from a speaker’s speech samples, a speaker model is trained and stored into the system database. In text-dependent speaker recognition, the model includes the temporal dependencies between the feature vectors. It is utterance specific.

Classical speaker models can be divided into template models and stochastic models (Campbell, 1997). Vector quantization (VQ) (Soong et al., 1988) and dynamic time warping (DTW) (Furui, 1981) are examples of template models for text-independent and text-dependent recognition, respectively. In template models, training and test
feature vectors are directly compared with each other. The amount of distortion between them represents their degree of similarity.

In stochastic models, each speaker is modeled as a probabilistic source with a fixed but unknown probability density function. Matching is usually done by evaluating the likelihood of the test utterance with respect to the model.

Models can also be divided into generative and discriminative models (Furui, 1997). The generative models i.e. GMM and VQ estimate the feature distribution within each speaker whereas discriminative models such as artificial neural networks (ANNs) (Farrell et al., 1994; Yegnanarayana and Kishore, 2002) and support vector machines (SVMs) (Campbell et al., 2006) model the boundary between speakers.

In short, a speaker is characterized by a speaker model using any modeling technique i.e. VQ, GMM or SVM. At run-time, an unknown voice is first represented by a collection of feature vectors and then evaluated against the stored speaker models. As per the literature, various speaker modeling techniques are discussed in the subsequent paragraphs.

2.5.1 Template Models
The speaker recognition consists of matching the incoming speech template with the stored templates. The template which is at minimum distance from the input pattern is the recognized speaker. The distance between the template \( x \) of the claimed speaker and an input feature vector \( y \) from an unknown speaker is given by \( d(x, y) \).

\[
\begin{align*}
    x = \frac{1}{N} \sum_{k=1}^{N} y_k \\
    d(x, y) = (x - y)^T W (x - y)
\end{align*}
\]  -------- (2.2)

The distance \( d(x, y) \) is given in Equation 2.3.

\[
    d(x, y) = (x - y)^T W (x - y)
\]  -------- (2.3)

where, \( W \) is the weighting matrix. The distance is called Euclidean, if \( W \) is an identity matrix. All the elements of the vectors are equally treated. If \( W \) is a positive and a definite matrix, the distance is called as Mahalanobis. It permits desired weighting of the template features.

**Dynamic Time Warping**

The lengths of speech utterances will not be the same and the number of segments for the reference and input templates will be different. Different utterances should be aligned in time otherwise a small shift would lead to incorrect identification. This can
be done using dynamic time warping technique. The distance $\beta$ on comparison of an input frame $y$ of $M$ samples and the template sequence $x$ is given in Equation 2.4.

$$\beta = \sum_{i=1}^{M} d(y_i, x_{j(i)})$$

(2.4)

The template indices is given by $j(i)$. Both the signals are aligned using a piece-wise linear mapping of the time axis. The dynamic time warp of two energy signals is shown in Figure 2.7. The warp has no effect for two identical signals. In this case the warp path is a diagonal line. The Euclidean distance between the two signals is given by the sum of all deviations from the dashed diagonal warp path. The parallelogram around the warp path acts as boundary to prevent excessive warping.

![Figure 2.7: DTW of Two Energy Signals](image)

### 2.5.2 VQ Source Modeling

Vector quantization is a text dependent template method. Codebook is generated for each speaker with the help of his/her training data. Clustering methods are used for the formation of the codebook (Srinivasan, 2012; Furui, 1991). These methods average out the temporal content from the codebook. It does not require time
alignment. The distance between the input vector and the minimum distance code word in the codebook is obtained. This gives the match score.

**Nearest Neighbors**

This method is better than both the dynamic time warping and vector quantization methods (Linde et al., 1980). It does not cluster data to obtain the codebook. It retains all the data obtained from training phase. Therefore, it takes into account the temporal information present in the input speech. The inter-frame distance matrix is obtained by calculating the distance between the input frame and the stored frames. The nearest neighbor distance is given by the minimum distance between the input and the stored frames. The average of nearest neighbor distances for all input frames gives the match score. The match scores are combined to form an approximation of the likelihood ratio.

**2.5.3 Hidden Markov Model**

Recent speech recognition systems use probabilistic / stochastic modeling methods, such as Hidden Markov model (Rosenberg et al., 1994; Matsui and Furui, 1994). The probabilistic process has optimized the recognition accuracy. In a stochastic model, a pattern is matched by measuring the likelihood of a feature vector in a given speaker model. These methods can model the time variations of the spectral characteristics of speech signals. The model parameters can be found automatically from the training data.

![Five State Markov model](image)

**Figure 2.8: Five State Markov model**
A Hidden Markov Model (HMM) as shown in Figure 2.8 is a finite state machine having the following specifications- a set of hidden states, an output that is visible from the state, transition probabilities between the states, emission probabilities and initial state probabilities. It consists of a set of transitions between a set of states. The transitions are allowed to the next right state or the same state, and $a_{ij}$ represents the transition probabilities. The HMM parameters are generated from the speech during the training phase and for identification, the likelihood of the input feature sequence is computed with respect to the speaker HMMs.

2.5.4 Neural Networks
Artificial neural networks (ANNs) have been widely used in various pattern recognition problems, including speaker recognition (Farrell et al., 1994; Yegnanarayana and Kishore, 2002). The strength of neural networks lies in discriminating patterns of different classes. Neural network consists of an input layer, one or more hidden layers and an output layer. Other forms of neural networks which are used for speaker recognition task are Multi-Layer Perceptron Neural Network (MLPNN) and Radial Basis function (RBF). The Self-Organizing Map (SOM) is a special class of neural network. It is based on competitive learning. SOM does not use class information for modeling speakers. Speaker model obtained is poor and hence the performance decreases. Linear vector quantization (LVQ) is a supervised learning technique. The class information is used to optimize the positions of code vectors obtained by SOM. The quality of the classifier-decision region is improved.

Auto-associative neural network (AANN) was developed as an alternative to GMM for pattern recognition task. AANN is a feed forward neural network. The number of units in the input and output layers is kept same as the size of the input vectors. But, the number of nodes in the middle layer is less than that in the input or output layers. The advantage of AANN over GMM is that, it does not impose any distribution. However, there is no significant proof to support that AANN is superior to GMM in terms of computational efficiency or recognition performance. The probabilistic neural network (PNN) is also a feed forward network. It has been derived from the Baye’s decision method. Based on the training samples, probability density function for each class is derived. The training time required for PNN is very less as compared to other classifiers. In this case, the training is done in a single step of each training
But, PNN requires large amount of memory. The reason for this lies in its structure. The problem gets aggravated when the train and test data sets are large. PNN are highly sensitive to noisy data.

2.5.5 Support Vector Machines
Support vector machine (SVM) was developed by Vapnik (1998) and is a new technique in the field of statistical learning theory. In recent years, SVMs have been widely used to solve binary classification problems. It constructs a hyperplane in a multidimensional vector space, which is then used to separate vectors that belong to two different classes. SVM searches for the hyperplane with the largest margin, i.e., maximum margin hyperplane. Support vector machine is a powerful discriminative classifier. It has been used with spectral (Campbell et al., 2006), prosodic (Shriberg et al., 2005, Ferrer et al., 2007), and high-level features (Campbell et al., 2004). SVM has shown robust performance in speaker identification. When used in combination with GMM the accuracy increases (Campbell et al., 2006). The increased performance is obtained with SVM due to its ability to classify unseen data. The original training data is mapped into a higher dimension using nonlinear mapping. The mapping from the input space to the high dimensional feature space is achieved using kernel functions. It searches for the linear optimal separating hyperplane i.e. decision boundary to separate data from two classes. SVM finds this hyperplane using support vectors. During testing, a classification score is obtained by calculating the distance of the test sample in relation to the hyper-plane.

2.5.6 Gaussian Mixture Models
Gaussian mixture model (GMM) (Reynolds and Rose, 1995) is a stochastic model and it can be considered as an extension of the VQ model. The Gaussian mixture model (GMM) method models speech as a weighted sum of multivariate normal probability density functions (pdf) (Reynolds et al., 2000). Each pdf is called a component of the GMM. A GMM with M components is said to be a GMM of order M. For an R-dimensional feature vector $x$, the posteriori probability for M component GMM and the probabilistic model $\lambda$ is defined as in Equation 2.5.

$$p(x / \lambda) = \sum_{m=1}^{M} w_m p_m(x)$$

\-------- (2.5)
The Equation 2.6 corresponds to the weighted linear combination of M unimodal Gaussian densities. The probability density function \( p_m(x) \), is given by

\[
p_m(x) = \frac{1}{2\pi|\Sigma_m|^{1/2}} e^{-\frac{1}{2} (x-\mu_m)^T \Sigma_m^{-1} (x-\mu_m)}
\]

---------- (2.6)

Each of the pdf is parameterized by a R–dimensional mean vector \( \mu_m \), RxR-dimensional covariance matrix \( \Sigma_m \) and a mixture weight \( w_m \), also known as a priori probability. The a priori probability satisfies the constraint of Equation 2.7.

\[
\sum_{m=1}^{M} w_m = 1
\]

---------- (2.7)

The set of parameters, \( \lambda = \{ \mu_m, \Sigma_m, w_m, 1 \leq m \leq M \} \) completely define a GMM. The use of single covariance matrix for the entire set of components is adapted in some of the applications; however for speaker recognition technology this practice is not common.

Given a GMM \( \lambda = \{ \mu_m, \Sigma_m, w_m, 1 \leq m \leq M \} \) and a set of feature vectors \( X= \{ x_1, x_2, \ldots, x_T \} \), the log likelihood of the model is computed as in Equation 2.8.

\[
\log P(X / \lambda) = \sum_{t=1}^{T} \log P(x_t / \lambda)
\]

---------- (2.8)

The constraints applied to the GMM include: a priori-probability, covariance matrix, and initialization of the GMM parameters (Reynolds, 1995).

**Priori probability**

The a priori probabilities of the Gaussian components maintain the requirement that it should be summed to 1. This constraint keeps the reliability of posteriori probability estimate of the GMM. The parameter \( w_m \) represents the apriori probability of each Gaussian component so it maintains the condition \( 0 \leq w_m \leq 1 \). In other words, a minimum value except zero may be enforced so that each Gaussian density may have a reliable share in optimization of probabilistic model. This approach would ultimately avoid singularities or over-fitting to the training data.

**Covariance Matrix**

For speaker recognition technology a local covariance matrix for each Gaussian density is adapted, thus it can lead to a substantial computational burden. A number of careful practical restrictions are then applied for selection of covariance matrix. The covariance matrix is a matrix of size R x R. Typically, in speaker recognition applications the covariance matrix is restricted to being a diagonal matrix. Based on
the empirical evidence it can be stated that diagonal covariance matrices outperform full covariance matrices. This restriction reduces the trainable covariance parameters to R parameters per Gaussian component. The covariance matrix for R=4, is given in Equation 2.9.

\[ \sum_{m,D=i} = \{\sigma_m (1) \} \]  

-----------------

(2.9)

**Initialization of GMM Parameters**

The initial distribution of the training data before tuning by EM procedure would have significant impact on the overall training procedure of the speaker models. K means clustering procedure is most commonly used technique to initialize the training speaker data. The K-means initializes the clusters for the feature vectors, each cluster would then become a single component of the GMM (Roch, 2006). The initial values of weights, means and covariances for each of the Gaussian densities are calculated using conventional statistics. The weights are determined by reciprocating the total number of Gaussian densities.

**Universal Background Modeling (UBM)**

GMM needs large amount of data to model the speaker. GMM-universal background model (UBM) was introduced by Reynolds et al. for the speaker recognition task to overcome this drawback. For UBM-GMM system, data collected from the enrolled speakers is grouped and then training of UBM is done (Hasan and Hansen, 2010). It acts as a speaker-independent model. The speaker-dependent model is then created from the UBM by using maximum a posteriori (MAP) adaptation technique from speaker-specific training speech. The advantage of the UBM-based modeling technique is that it provides good performance for small amount of speaker-dependent data.

Recently a new technique namely joint factor analysis (JFA) has been proposed. JFA uses correlations between Gaussians during speaker modeling. The speaker model is decomposed into three components namely a speaker-dependent component, a session-dependent component and lastly both speaker and session-independent component. It is a computationally challenging method. It requires a well-balanced training set recorded under a variety of channel conditions. Dehak, 2009 proposed a novel method that does not model inter-speaker and intra-speaker variations in a high dimensional supervectors space. It uses a low-dimensional supervector subspace that
represents both the channel and speaker variabilities. This space is called total variability space. It is also called as i-vector method and front-end factor analysis.

2.6 DIMENSIONALITY REDUCTION TECHNIQUES
In a speaker identification system, it is necessary to determine the optimum number of coefficients that can best distinguish different speakers. If fewer features are used then there is a loss of the discrimination power of the classifier leading to decrease in the recognition accuracy. But, if large number of features is used then the processing time and storage required for classification increases. Hence, it is important to reduce the dimensionality of the data in speaker identification. MFCC and their derivatives is widely used feature in the field of speaker recognition. However, depending on the application and the available computational resources, this dimension is required to be reduced. The $K$ features most relevant to the recognition task are extracted. Hence, the $D$-dimensional feature space is transformed into a $K$-dimensional subspace ($K < D$).

Feature selection can be used to reduce features. It consists of determining the best possible subset of features by searching all the possible $2^D$ combinations. The easiest way is to evaluate all the $D$ features individually and selecting the $K$ most discriminative features. It does not take into consideration dependencies among features.

Zhou et al. (2010) proposed a method to reduce feature dimension based on CCA and PCA. Experimental model based on GMM technique was used. LPC (16-dimensional) and MFCC (13-dimensional) are selected as speaker features. Compared with the conventional dimension reduction method i.e. CCA, PCA and other manual methods, experimental results clearly prove that combination of CCA and PCA method is more effective in dimension reduction.

Patra and Acharya (2011) compared the effect of dimension reduction of feature vectors using Principal Component Analysis (PCA) and Weighted Principal Component Analysis (WPCA) for speaker identification in a noisy environment. MFCC feature vectors are used as original features and their dimension is reduced by PCA and WPCA techniques and then evaluated by GMM classifier. Speaker identification rate is calculated under different SNR to test the robustness of the speaker identification system. Zergat et al. (2012) has proposed the use of hybrid GMM-SVM model in real conditions. It is important to extract features that best
characterizes the speaker for speaker verification task. Mel Frequency Cepstral Coefficients (MFCC) and their first and second derivatives were used as features and a dimension reduction method called Principal Component analysis (PCA) was used in front end step. Joshi et al. (2013), proposed a new feature extraction technique. Radon transform is used for feature extraction after obtaining the spectrogram from the speech sample. PCA has been used to achieve dimension reduction and to reduce the computational complexity. Sarkar et al. (2014) showed that the phonetically discriminative MLP features retain speaker-specific information and is complementary to the short-term cepstral features. The performance improvement is obtained with both score domain and feature domain fusion and the speaker verification equal error rate (EER) is reduced up to 50% compared to the best i-vector system using only cepstral features.

2.7 PERFORMANCE TERMS FOR SPEAKER RECOGNITION TASK

The classification stage in Speaker identification compares the test utterance against the claimed speaker’s model and computes a score value which is a measure of similarity. Then, the Speaker identification system is required to make accept or reject decision by comparing the score value with a predefined threshold.

An ideal speaker identification system accepts all identity claims made by true speaker and reject those made by impostors. Two types of errors occur in speaker identification systems namely, false acceptance (false alarm / false positive) and miss (false rejection / false negative). False acceptance error happens when the identity claim of an impostor is accepted while false rejection error occurs when true identity claim is rejected.

False acceptance and false rejection error rates, which are denoted as \( P_{fa} \) and \( P_{fr} \) respectively, are measured by counting the number of errors of both types. Given the number of false acceptance \( (N_{fa}) \) and number of false reject \( (N_{fr}) \) occurrences, false alarm probability \( (P_{fa}) \) and false reject probability \( (P_{fr}) \) are given by Equation 2.10 and 2.11 respectively.

\[
P_{fa} = \frac{N_{fa}}{N_{impostor}} \quad \text{------------- (2.10)}
\]

\[
P_{fr} = \frac{N_{fr}}{N_{true}} \quad \text{------------- (2.11)}
\]

where \( N_{true} \) and \( N_{impostor} \) are number of target trials and impostor trials, respectively.
Looking at the utterance scores that a speaker identification system yields for the two types of trials, it is possible to plot the probability distribution function (PDF) of the scores. As shown in Figure 2.9, the first PDF represents scores obtained during target trials while the other represents scores obtained during non-target trials. As shown, these score distributions overlap. To make accept or reject verification decision, a threshold must be set such that it attempts to minimize the number of errors (false acceptance and false rejection) made by the system. When a low threshold is set, the system has a tendency to accept more identity claims and as a result it makes few false rejections and many false acceptances. On the other hand, if the threshold is set to a higher value, the system makes more rejections and hence makes very few false acceptances and many false rejections. The operating point of the system is given by the two error rates. Thus, setting the operating point or the decision threshold is a trade-off between the two types of errors.

![Figure 2.9: Distribution of Target and Impostor Scores](image)
Based on the two error types i.e. false acceptance and false rejection, performance of speaker identification system can be evaluated using equal error rate (EER), relative operating characteristics (ROC) curve, detection error trade-off (DET) curve and detection cost function (DCF). The overall performance of the system for all operating points or threshold values is represented by ROC and DET curve. EER and DCF give the system performance at specific operating point.

The overall performance of the system for all operating points or threshold values is represented by ROC and DET curve. EER and DCF give the system performance at specific operating point.

The performance terms which are commonly used in measuring the speaker identification performance are discussed below.

a) Relative Operating Characteristics (ROC) curve
ROC (Relative Operating Characteristics) curve is used to represent the trade-off between false acceptances and false reject errors. This curve graphically compares Speaker identification systems over the entire operating point. The false alarm rate is plotted on the horizontal axis while correct detection rate is plotted on the vertical axis as the threshold is adjusted (Figure 2.10). When comparing Speaker identification systems, the area under the curve gives better measure of the performance than the relative position of the line, which makes ROC curves difficult to interpret. Due to this they are rarely used as performance terms in Speaker identification.

b) Detection Error Trade-off (DET) Curve
The variant of ROC curve is called DET (Detection Error Trade-off) curve (Martin et al., 1997). The two error rates are functions of the decision threshold. It is therefore possible to represent the performance of a system by plotting $P_{fa}$ as a function of $P_{fr}$ over a range of operating thresholds. The closer to the origin the curve will be, the system is said to be better. The resulting curve is monotonous and decreasing. Rather than plotting the probabilities themselves, plotting the normal deviates that correspond to the probabilities will give the curve known as the detection trade-off (DET curve).

The DET curve (Figure 2.11) is approximately linear for speaker identification systems which have Gaussian distribution for true and impostor score. A curve which is more close to the origin represents better system performance. As the DET curve is approximately linear it is easier to visualize differences between classifiers and compare their performance. In speaker identification DET curve is widely used for
comparing different systems without setting a threshold value. In this thesis, DET plot is adopted as one of the evaluation metrics.

![ROC Curve](image.png)

**Figure 2.10: ROC Curve**

The DET plot shows how the decision threshold can be changed in its entire range as a tradeoff between the two error rates. However, it is also possible to indicate a fixed operating point on the curve when making decisions at a fixed point. Hence, the false alarm (acceptance) rate can be reduced to a low level by setting a high detection threshold. This makes the system to operate in the upper-left corner of the plot. For a system whose false reject is high, a lower threshold is chosen which makes the system to operate at the opposite end.
c) Equal Error Rate

Equal Error Rate (EER) is an overall measure that gives the performance of a system in a single figure. It is the value that corresponds to the operating point at which the false acceptance rate is equal to the false rejection rate. On the DET curve, EER is the value at the intersection of the DET curve with the diagonal.

The EER corresponds to a threshold which may not be a realistic operating point for a given application. But still, it shows the ability of a system to separate impostors from target speakers.


d) Detection Cost Function

Plotting the two error rates as a function of the decision threshold is a good way of comparing the performance of different speaker identification systems. But this method cannot be applied for evaluation of systems with fixed threshold. In such cases, systems are rather evaluated by use of a cost function which assigns a cost to
each of the errors and takes into account the prior probabilities of target trial. Equation 2.12 gives the decision cost function.

\[
DCF = P_{tar}C_{fr}P_{fa} + (1 - P_{tar})C_{fa}P_{fa}
\]

--------- (2.12)

where \( C_{fr} \) and \( C_{fa} \) are the costs associated with false rejection and false alarm respectively. \( P_{tar} \) and \( P_{non\ tar} = (1 - P_{tar}) \) represents the probabilities of an utterance being spoken by the target speaker and impostor.

The three application-dependent parameters \( C_{fr}, C_{fa} \) and \( P_{tar} \) which form the detection cost function DCF gives a single scalar performance measure of a speaker identification system.

### 2.8 GAPS IN THE STUDY

Robust speaker recognition in noisy environments still remains a challenge inspite of the various techniques available to tackle it in past. The increasing trend towards use of hand-held devices has made new demand for robust speaker recognition applications. Despite being well explored in past, new methods keep unfolding in this field. These are either suggested improvements or alternatives of the existing ones. Hence, robust speaker recognition is a very challenging but an interesting area to work in.

The performance gets affected by the mismatch in recording conditions and speaker-generated variations. Robustness of automatic speaker recognition is important for real-world applications. In practical acoustic environments background noise, room reverberation causes considerable problems to such systems. The problem of distortion in the channels and handsets is also a cause of concern and it is necessary to find better techniques. The state of art research in speaker identification area is more concerned about the development of fast and robust algorithms, which gives good performance in noisy conditions and can withstand the cross media problem. Therefore, it is required to find more stable features that do not change due to the noise and physical or emotional state of the speaker.

The amount of computations required and the amount of data required for training grows exponentially with the increase of dimensionality of the feature vectors. The reduction in dimension of feature vectors causes the computational complexity to reduce. But with reduced feature set, speaker identification rate may decrease significantly and the system will become useless for practical application. So, it is
required to reduce dimension in such a way so that important information present in the features is not lost. Hence, computational efficiency is an important issue that needs to be resolved in speaker recognition systems working in real time.

Earlier, PCA and LDA have been used as optimization tool for achieving better results in terms of processing time and storage. There is still a scope to further improve the performance of speaker recognition system in noisy environment by investigating other optimization techniques like GA.

2.9 CONCLUSIONS

This chapter offers literature overview of the speaker recognition technology. It begins with a brief survey of the various speech features that are commonly used in the speaker recognition tasks. Then, a brief description of the various speaker modeling techniques as used by the various researchers for speaker recognition tasks are presented. In the end, various dimension reduction techniques for optimization purpose are discussed.

The choice of feature is application-specific and depends on resources required for computation, amount of speech data available and whether the speakers are cooperative or not. The spectral features are more popular because they can be easily computed and give good performance. Prosodic and high-level features are more robust, but less discriminative and can be easily impersonated. More complex front-end i.e. automatic speech recognizer is required for high-level features. There does not yet exist globally “best” feature but the choice is a balance between speaker discrimination, robustness, and practicality.

The choice of modeling technique also depends on the intended application. In case of text-dependent speaker recognition, still the most preferred technique is DTW. In case of text-independent speaker recognition, VQ technique is preferred. GMM is the most accepted modeling technique from among the Gaussian classifiers. The commonly used neural network techniques for speaker modeling are the MLP, RBF and AANN models. More recently, SVM has also been demonstrated to be a potential discriminatory-type classifier for speaker modeling. Its performance for limited data condition is good.
The time required by the conventional classifier should not be very high to make the identification system practical. Lots of research has been done by researchers on dimensionality reduction in the field of pattern classification but still there is scope for improvement. Principal Component Analysis, Linear Discriminate Analysis, Independent Component Analysis, GA etc. are some of the common techniques which have been used by researchers. The main motivation of this research is to analyze the available dimension reduction techniques and to suggest ways for improving the performance.

The next chapter deals with the cepstrum based features for speaker recognition task. The concept of mel-scale mapping is introduced for the design of MFCC filter bank for feature extraction. Various language and data related, speaker dependant and technical factors which can affect the performance of MFCC based systems have been investigated.