Comparative Study on Feature Extraction and Classifier Techniques for Speaker Recognition

[1]Aparna Mohan, [2]Sulphikar A
[1]M.Tech Student, [2]Associate Professor
[1]Department of Computer Science and Engineering,
[1]LBSITW, Poojappura, Trivandrum, Kerala, India

Abstract— **In the real world when we hear a voice through a telephone or at an entry way out of sight we identify the person through the sound of their voice. It is based on the expectation that the sound of one's voice is sufficient for a hearer to recognize the speaker. This is considered to be a challenge for speaker recognition. Voice biometrics use the features of a person's voice to ascertain their identity like human listeners. Speaker recognition is the best known commercialized form of voice biometrics. Speaker recognition is a process that verify a person's claim of his/her identity using the features extracted from their voice. The speech signal carries information about the speaker, the message to be conveyed, and language emotion and so on. In this paper we have presented a study on speaker recognition and various feature extraction and classifier techniques. First section gives an introduction to speaker recognition. The next section provide a brief description about the key terms that is used in speaker recognition and then discusses the different feature extraction techniques and classifier techniques for speaker recognition.**

***Index Terms*— GMM, MFCC, LPC, PLP, VQ, Speaker Recognition, SVM, Voice.**

I. INTRODUCTION

Speaker recognition is the identification of a person using the characteristics extracted from their voices. It is different from speech recognition as it identifies what is being said. Speaker recognition has two phases. The enrollment phase and the verification phase. In the enrollment phase a number of features will be extracted after recording the speaker's voice and that forms a template, a model or a voice print. The verification phase compares an utterance of speech sample with a previously created voice print [1].

Speaker recognition can be done as either text independent or text dependent. In text independent speaker recognition the text spoken by the speaker is unknown to the system. This is used in forensic like applications where the user does not know that he / she is being evaluated for some recognition purpose. It requires only very little coordination by the user. Here different text is used in enrollment and verification [2]. In text dependent speaker recognition the text spoken by the person is already known to the system. Here the same speech will be used in enrollment and the verification phase. The text may be common to all speakers. This is used in applications where the dialog unit has a strong control which guides the user. Knowledge of spoken text helps in the improvement of system performance when compared to the text independent speaker recognition. Speaker recognition can be classified as speaker identification and speaker verification.

Speaker identification identifies the person who is talking from a given set of known voices of speakers. Here the speaker does not claim for his identity. If the speaker is found from a predefined set of known speakers it is termed as closed set speaker identification. If the speaker is not restricted to the one in a predefined database then it is referred to as open set speaker identification. Here a 1: N match is found where N represents the number of templates in the database. The error in speaker identification is the false identification of the speaker. Speaker verification is the verification for the identity of a person of his claim. This is referred to as an open set task as the imposters are not known to the system. In speaker verification there is only a 1:1 match which is between a speaker's voice and a template. The error here is false rejection which is the rejection of a true speaker as an imposter and false acceptance which is the acceptance of a false speaker as a true one [2].

Speaker recognition contain two major steps. Feature extraction and classification. Selection of suitable features along with methods to extract them is known as feature selection and feature extraction. During classification the test template is compared with a reference template and a similarity measurement is computed between them. If the measurement is within a threshold then the identity claim is accepted else rejected. SVM is the popularly used today because it is more robust and has got more generalization performance as it has the ability to classify the unseen data accurately [10].

The remaining section of this paper contain the key terminologies used in speaker recognition, popular feature extraction and classifier techniques. Finally it concludes with the best method and approach for speaker recognition.

II. KEY TERMINOLOGIES

- Training: The process that captures sample biometric (here voice) for extracting features from it and to generate a template.
- Training data: The data captured for training.
- Test data: The biometric data (here voice) captured for verification or identification by the biometric system.

- Pre-processing: It involves modulation of speech removing unwanted noise from it and diving it into voiced or unvoiced sounds and channel compensation.
- Feature extraction: The process that convert a captured biometric sample (here voice) into a biometric data (that represents the characteristics of the sample) so that it can be compared to a reference model.
- Modeling: The process that create a speaker model from feature vectors extracted from voice sample.
- Model Database: A repository of templates or speaker models that are used for recognition by biometric system.
- Matching: the process that compares a biometric sample (here voice) against a previously created template and scoring the level of similarity. An accept or reject decision is based on whether the score exceeds a particular threshold or not.
- Threshold: the point that must be reached for a speech sample to be considered a match with a previously enrolled voice print.
- Reference template: template generated using feature vectors of voice data sample. It is also called speaker model or reference model.
- Voice print: A form in which speech sample can be converted that can be analyzed by voice biometric system

III. FEATURE EXTRACTION

Feature extraction is the process that convert a captured biometric sample (here voice) into a biometric data that represents the characteristics of the sample so that it can be compared to a reference model.

Types of features

There have been proposed a vast number of features for speaker recognition [3] .They can be divided into following classes.

- Spectral features.
- Dynamic features.
- Source features.
- Suprasegmental features.
- High -Level Features

Spectral features which are the descriptors of the short term speech spectrum. They are obtained by converting the time domain signal into frequency domain by using Fourier Transform. These features help to identify the pitch, rhythm and melody. Dynamic features relate to the evolution of spectral and other features. Source features represents the features of the glottal voice source. Suprasegmental features are the specific features that are superimposed on the utterance of the speech. This include stress, duration and tone in the syllable or even nasalization and harmony are included in this category. High level features represent symbolic type of information such as characteristic word usage.

Properties of optimal feature

- High inter-speaker variation.
- Low intra-speaker variation.
- Easy to measure.
- Robust against disguise and mimicry.
- Robust against distortion and noise.
- Maximally independent of the other features.

The first two requirements require that the features should be discriminative.

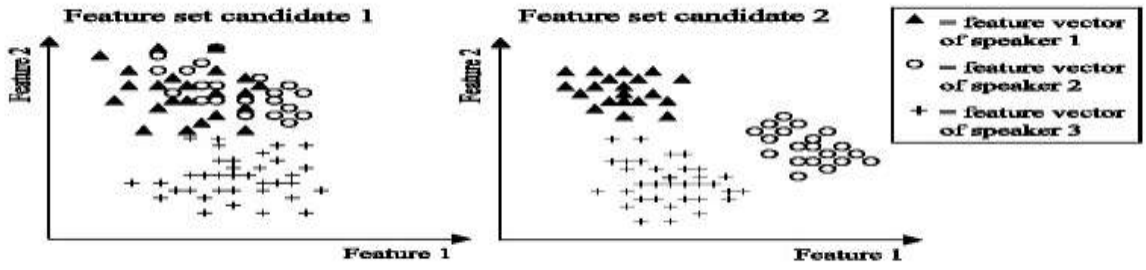

Fig.1 Examples of two-dimensional feature sets with poor and good discrimination

The feature should be easily measurable consists of two factors. Firstly the feature should occur naturally and frequently in such a way that it can be extracted from short speech samples. Secondly the feature extraction technique should be easy. A good feature should be robust against disguise, distortion, noise. Different features extracted from the speech signal should me

maximally independent. They should not have any correlation between each other. Even though we consider these requirements are necessary for optimal features all features cannot satisfy all of these requirements. So we relax some requirements. But the signal processing techniques used in speaker recognition are computationally efficient

Feature Extraction Techniques

Mel Frequency Cepstral Coefficients (MFCC)

MFCC is widely used in speaker recognition as it can extract both linear and nonlinear features. Thus it can be used to extract dynamic features as well [5]. MFCC is based on human auditory system. Human perception of sounds does not follow a linear scale which is above 1 kHz. In MFCC the frequency scales are placed on a linear scale for frequencies below 1 kHz and on a log scale for frequencies above 1 kHz. It contain both time and frequency information of the signal and thus it is more useful for feature extraction. MFCC is computation have the following steps [4]:

- Pre-processing: Initially the speech signal is subjected to a preprocessing where a pre-emphasis is done. The high frequency part of the signal that was suppressed during the sound production mechanism of humans is compensated using the pre-emphasis.
- Frame blocking: Here the input signal is segmented into frames. To facilitate the use of FFT the frame size will be taken as the power of two.
- Windowing: in order to avoid the disruptions at the start and end of the frame, each frame should be multiplied with a window function. There are many window functions exist, but the commonly used one is Hamming window.
- Fast Fourier Transform: FFT is used to convert the signal from spatial domain to frequency domain.
- Mel frequency wrapping: for humans the frequency content of sounds is speech signal is not linear. Thus for each tone with an actual frequency f measured in Hz a subjective pitch will be measured on a particular scale which is called as Mel scale. During mapping for a frequency value up to 1000Hz the Mel- frequency scaling is linear frequency spacing but above 1000Hz the spacing is logarithmic. Formula to convert frequency f Hz into Mel m_f : $m_f = 2595\log\left(\frac{f}{700}+1\right)$ (1)

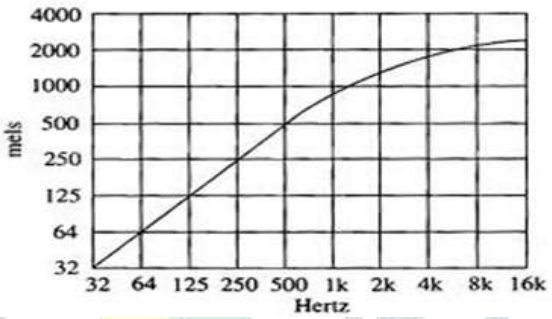

Fig.2 Frequency to Mel frequency curve

The spectrum calculated in the previous step will be converted to Mel scale to know about the approximate energy existing in each spot with the help of a triangular overlapping window also known as triangular filter bank. The filter bank is a set of band pass filters and the steady Mel frequency time decides the spacing along the bandwidth of the band pass filters. Thus the filter bank with proper spacing done by Mel scaling helps to determine the energy at each spot and the log of these energies will give Mel spectrum.

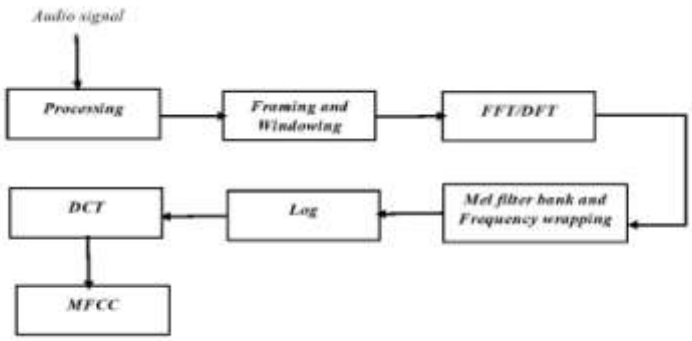

Fig.3 Block diagram of MFCC feature extraction

- Discrete Cosine Transform (DCT): DCT is applied to the log energies obtained from N triangular bandpass filters to obtain L mel –scale ceptsral coefficients. Usually N is set to 20 and L is set to 12. Since we applied FFT the DCT will transform the frequency domain into a time like domain known as quefrency domain. These features are similar to cepstrum and hence we call it as mel-scale cepstral coefficients or MFCC. We can use only MFCC as features in speaker recognition but to increase the accuracy log energy can be added and delta operation can be performed.
- Log energy: Energy within a frame has its own importance and that can be obtained without any difficulty. Thus log energy can be added as the 13th feature in MFCC.
- Delta operation: There is also an advantage if we add the time derivatives of (energy+MFCC) as new features. This time derivative indicate the velocity and acceleration of (energy+ MFCC).

Linear Predictive Coding (LPC)

LPC methods are widely used in speaker recognition, speech synthesis speech storage etc. This method can accurately estimate the speech parameters .The basic idea of linear prediction is to represent the current speech sample as a linear combination of the previous samples .this is called a linear predictor and hence the process is called linear predictive coding [6]

$$s(n) = \sum_{k=1}^{p} \alpha_k \; s(n-k) \qquad (2)$$

For some value of p and $\alpha_k{}' s$.For linear prediction the predictor coefficients $\alpha_k{}' s$ are computed (over a finite interval) by minimizing the sum of squared differences between the actual speech samples and the linearly predicted ones .These coefficients are the basis for the LPC speech.

A speech signal is analyzed by LPC by estimating the formants in it. LPC also removes the effects of formants from the speech signal and then the frequency and the intensity of the remaining buzz is estimated. The process of removing the formants is called inverse filtering and the remaining signal is called the residue [13].

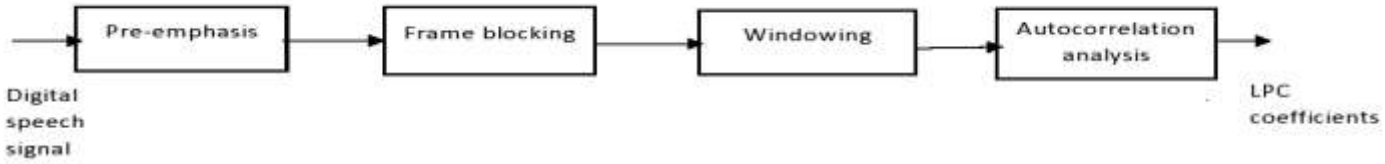

Fig.4 Block diagram of LPC algorithm

Perceptive Linear Prediction (PLP)

PLP was developed by Herman Sky in 1990. The goal of the original PLP model was to describe the psychophysics of human hearing more accurately in the feature extraction process. Like LPC, PLP is also based on the short term spectrum of speech but the difference is that PLP modifies the short term spectrum of speech by several psychophysically based transformations [6]. This technique aims to find a smooth spectra consisting of resonant peaks. The order needed by PLP based speaker recognition system is smaller when compared to the one needed by LPC based speaker recognition system.

The PLP computation steps are the following: critical-band spectral resolution, the equal-loudness hearing curve and the intensity-loudness power law of hearing [12]. After estimating the auditory-like spectrum, it is converted to autocorrelation values by doing a Fourier transform. The resulting autocorrelations are used as input to a standard linear predictive analysis routine. The resulting output is perceptually-based linear prediction coefficients. Typically, these coefficients are then converted to cepstral coefficients using a standard recursion.

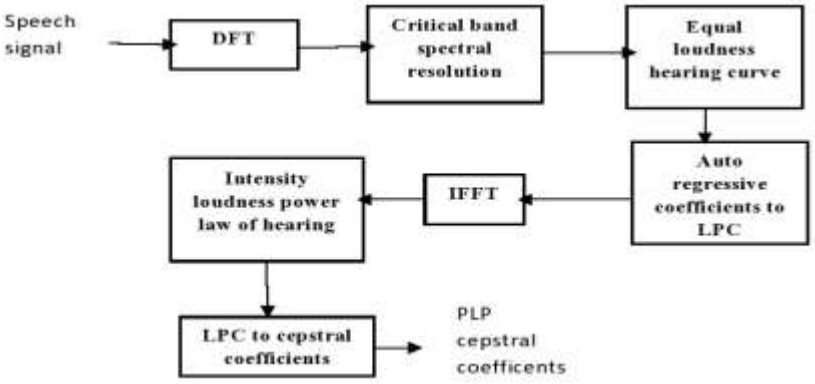

Fig.5 PLP Implementation

IV. CLASSIFICATION

The anatomical structure of vocal tract differs from person to person. That is it is unique for every person and hence every person speech has different speech. That is why the speakers can be identified using the information available in the speech. Recognizing speaker by his/her voice is termed as speaker recognition. Voice comes under the category of biometric identity

because the difference in anatomic structure is an intrinsic property of the speaker. Selection of suitable features along with methods to extract them is known as feature selection and feature extraction. These extracted features forms a test template. During classification the test template is compared with a reference template and a similarity measurement is computed between them. If the measurement is within a threshold then the identity claim is accepted else rejected. Based on threshold value two results are possible. One is False acceptances (false claim is accepted) and other is False rejection (owner of the identity is rejected). So usually the threshold is set high. But this can lead to many false rejections that is undesirable. Keeping threshold low while not giving any false rejections may result in many false acceptance. Hence the in some classification the threshold will have two values, low and high. If the measurement is below the low value then the identity claim is accepted and if it is above the high value then the claim is rejected. If it is between the low and high values then further classification is done.

Classifier Techniques

Vector Quantization (VQ)

Vector quantization is the process that take large number of feature vectors of a particular speaker and produce smaller set of feature vectors. These feature vectors forms the centroid of the distribution. Centroid is the point spaced so that there is only minimum average distance to every other point [7]. To find centroid is a time consuming task but it avoids the overhead of representing every single feature vector in the feature space which is generated from the training utterance of each speaker. A vector quantizer maps k-dimensional feature vectors R*k* to finite set of feature vectors. That is Y= $\{y_i$:1,2,…. N$\}$ .k dimension means the number of feature coefficients in the feature vector. Each such vector $y_i$is termed as the code word and set of all code-words forms the codebook.

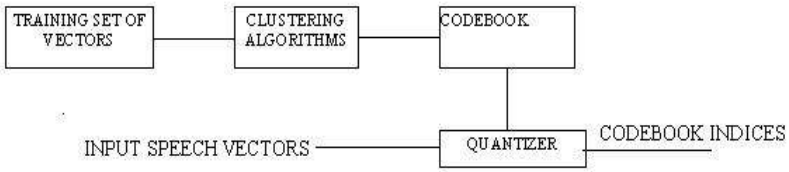

Fig 6. Block diagram of basic VQ training and classification

Fig below shows the example of vector quantization. The circles refer to the feature vectors of speaker 1 and triangles refer to the feature vectors of speaker 2. During training a speaker specific codebook is generated for each known speaker by clustering the training feature vectors.

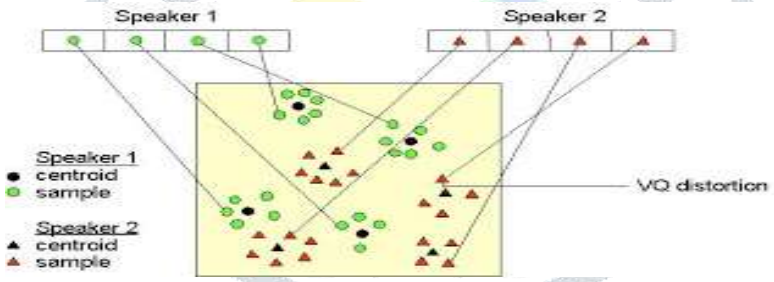

Fig 7. Vector Quantization codebook formation [8]

The black circles and triangles represent the codewords (centroids) for speaker 1 and speaker 2 respectively. Distance from a vector to the closest codeword of a codebook is called the VQ-distortion. During the recognition phase when an input utterance of an unknown speech comes it will be vector quantized using each trained codebook. That is the total VQ-distortion will be computed. The speaker corresponding to the VQ codebook with minimum total distortion will be identified as the speaker of the unknown utterance.

Issues in Vector Quantization*:*

- It can be applied only to limited vocabulary speech.
- It is affected by noise and channel degradation

Gaussian Mixture Model (GMM)

Gaussian Mixture Model is one of the most commonly used classifier and it is a density estimator [6]. A Gaussian mixture density can be considered as the weighted sum of M component densities [9]. It is denoted as:

$$p(x|\lambda) = \sum_{i=1}^{M} p_i . b_i(x) \qquad (3)$$

Where x is a D- dimensional random vector, p_i i=1,2…M are the mixture weights , $b_i(x)$ where i=1,2.. M are the component densities. Each component densities the D- variate Gaussian function of the form

$$b_i(x) = \frac{1}{2\Pi^{\left(\frac{D}{2}\right)}.|\sum_i|^{\left(\frac{1}{2}\right)}}. \text{ a} \quad (4)$$

here q represents the following:

$$\text{a} = \exp\left\{-\frac{1}{2}(x-\mu_i)'\sum_i{}^{-1}(x-\mu_i)\right\} \quad (5)$$

$\mu_i, \sum_i$ represents the mean and covariance of the i^{th} mixture respectively. During training when the data $x_1, x_2, x_3 \ldots x_n$ and the number of mixtures M are given then the $\mu_i, \sum_i$ and p_i is learned using expectation maximization.

During recognition for an unknown utterance set of feature vectors will be extracted that is $x_1, x_2, x_3 \ldots x_l$ and distance of the given sequence from the model is obtained by computing the log likely hood of the given sequence of feature vectors when the data is given [6]. The speaker will be corresponding to the model with the highest score.

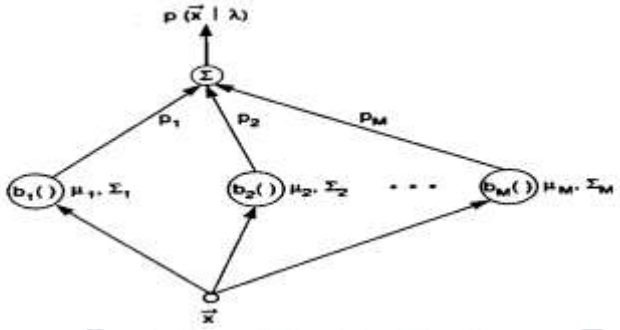

Fig 8. Depiction of M component Gaussian Mixture Density

Issues in GMM:

- Speaker verification accuracy decrease as the size of training data increases.

Support Vector Machine (SVM)

SVM is a powerful discriminative classifier which is popularly used today. SVM can be used with prosodic, spectral and high level features. In order to increase the accuracy it has been combined with GMM. SVM maps the given input to a high dimensional plane and it separate the classes with a hyper plane [10]. Currently this is considered as a robust method for speaker verification .this method has got more generalization performance as it has the ability to classify the unseen data accurately. SVM is a two class classifier which is constructed from the sums of a kernel function K (. , .) that is :

$$f(x) = \sum_{i=1}^{N} \alpha_i t_i K(x, x_i) + \text{d} \quad (6)$$

Where t_i represents the ideal outputs $\sum_{i=1}^{N} \alpha_i t_i = 0$, for $\alpha_i > 0$. x_i denote the support vectors that are obtained after optimization process from the training set. Depending on the class in which the vector belongs the output will be either 1 or -1. The class will be class 0 for +1 output and class 1 for -1 output. The value of f(x) determine the class to which the vector belongs. That is whether its value is above or below a particular threshold.

Properties of SVM:

- Duality: SVM is a linear learning machine that is expressed in a dual fashion. That is data appear only in the inner product as shown :

$$f(x) = \langle w, x\rangle + b \quad (7)$$
$$f(x) = \sum \alpha_i y_i \langle x_i, x\rangle + b \quad (8)$$

- Kernels: Kernel is a function that returns the value of the dot product between two images of two arguments. SVM operate in a feature spaced induced by a kernel. That is f(x) can be written as

$$f(x) = \langle w, x\rangle + b \quad (9)$$
$$\textit{That is, } f(x) = \sum \alpha_i y_i \langle \phi(x_i), \phi(x)\rangle + b \quad (10)$$

Using kernel this can be expressed as:

$$\text{K}(x_1, x_2) = \langle \phi(x_1), \phi(x_2)\rangle \quad (11)$$

where K (x_1, x_2) is the kernel function.

- Margin: In SVM the objective is to maximize the margin. The hyerplane that maximizes the margin should be considered.

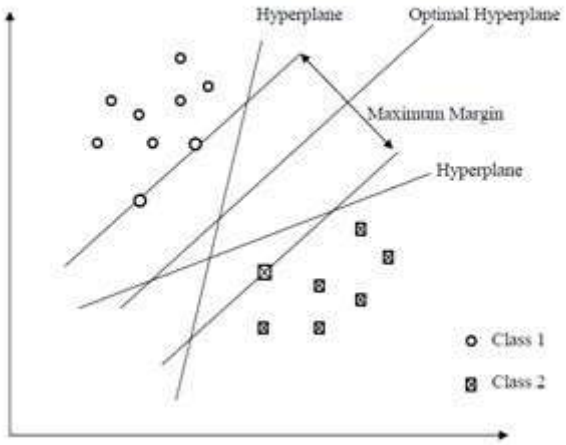

Fig.9 Support Vector Machine

In SVM the kernel should satisfy the Mercer condition. That is the kernel can be expressed as

$$K(x, y) = b(x)^t . b(y) \quad (12)$$

b(x) indicate the mapping from the input space where the x lies to a possibly infinite dimensional space. To satisfy the margin concept and optimization rule the kernel should satisfy the Mercer condition. The optimization rule relies on the margin concept which means that for a separable data set to be placed in a high dimensional space the system should ensure that the hyperplane with the maximum margin is taken [11]. The vectors in the boundary of the margin forms the support vectors. SVM training process should focus to model the boundary which is opposed to the GMM concept which model the probability distributions of two classes.

Issues in SVM:

- Computation is complex as the output is not probabilistic.
- Choice of appropriate kernel for implementation is a question.

V. CONCLUSION

Several types of feature extraction and classifier techniques are available for speaker recognition. MFCC is the most popular acoustic feature used in speaker recognition. Because it reduces the frequency information of speech signal into a small number of coefficients. It is relatively fast and easy to compute. In the case of classifier techniques SVM has got more attention as it can give more accurate and robust classification result even for a non-monotone and non-linearly separable data. It uses the kernel concept. We also discussed the issues in each of the classifier techniques.

VI. ACKNOWLEDGMENT

I am thankful to my guide Mr. Sulphikar A, Associate Professor of Computer Science and Engineering, for his guidance and encouragement for the paper work.

REFERENCES

[1] Zia Saquib, Nirmala Salam, Rekha P. Nair, Nipun Pandey, and Akanksha Joshi, "A Survey on Automatic Speaker Recognition Systems," Procedia Computer Science, 00(2009)000-000.

[2] Marcos Faundez-Zanuy, Enric Monte-Moreno,"State- Of-the- Art in Speaker Recognition," IEEE A&E Systems Magazine, May 2005.

[3] Mr.Yoghesh Dawande , Dr. Mukta Dhopeshwarkar, Dr. Babasaheb Ambedkar, "Analysis Of Different Feature Extraction Techniques for Speaker Recognition", International Journal Of Advanced Technology & Engineering Research (IJATER) ISSN 2250-3536 ,vol.5, issue 1, Jan.2015.

[4] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad and Arpit Bansal, "Feature Extraction using MFCC," Signal & Image Processing: An International Journal (SIPIJ), vol., no.4, Aug.2013.

[5] Nilu Singh, R. A. Khan, Raj Shree," MFCC and Prosodic Feature Extraction Techniques," International Journal of Computer Applications, vol.54, no.1, Sept.2012.

[6] Jeet Kumar, Om Prakash Prabhakar , Navneet Kumar Sahu , "Comparative Analysis of Different Feature Extraction and Classifier Techniques for Speaker Identification Systems: A review," International Journal of Innovative Research in Computer And Communication Engineering , vol.2, issue 1, Jan.2014.

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

[8] Prof. Ch.Srinivasa Kumar, Dr. P. Mallikarjuna Rao "Design of an Automatic Speaker Recognition System Using MFCC, Vector Quantization and LBG algorithm" International Journal on Computer Science and Engineering (IJCSE) vol. 3 No. 8 Aug. 2011.

[9] Douglas A. Reynolds and Richard C. Rose, "Robust Text Independent Speaker Identification using Gaussian Mixture Speaker Models," IEEE Trans. On Speech and Audio Processing. Vol.3, No:1 , Jan.1995.

[10] W.M.Campbell, J.P.Campbell, D.A.Reynolds, E.Singer, P.A.Torres-Carrasquillo "Support Vector Machines for Speaker and Language Recognition" , ELSEVIER, Computer speech and language Aug.2005

[11] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," IEEE Signal Processing Letters, Vol.13, no.15, May 2006.

[12] Varsha Singh , Vinay Kumar Jain, Dr. Neeta Tripathi , "A Comparative Study on Feature Extraction Technique for Language Identification", International Journal of Engineering Research and General Science , ISSN 2091-29330, Vol.2, April-May 2014.

[13] Parvati J Chaudhary, Kinjal M Vagadia ,"A Review Article on Speaker Recognition with Feature Extraction" ,International Journal of Emerging Technology and Advance Engineering ,ISSN 2250-2459, Vol.5 , Issue 2, Feb.2015.