Analysis of Different Classifier Using Feature Extraction in Speaker Identification and Verification under Adverse Acoustic Condition for Different Scenario

Shrikant Upadhyay
Assistant Professor, Department of Electronics & Communication Engineering
Cambridge Institute of Technology Ranchi, Jharkhand, India-835103

Sudhir Kumar Sharma
Professor (HOD), Department of Electronics & Communication Engineering
Jaipur National University, Jaipur, Rajasthan, India-302017

Aditi Upadhyay
Research Scholar, Department of Electronics & Communication Engineering
Jaipur National University, Jaipur, Rajasthan, India-302017

Abstract- Speech is one the easiest way of medium for human being to interact and exchange information in this universe. The language may be different and the way of speaking is different of each other because of their region and culture. The voice of human beings dependents on the different parameters like vocal cord diameter, pitch, formants (i.e. resonances of the vocal tract), shape and size of various articulators, vocal folds etc. In this paper, we try to analyze the performance of different mixture model like Hidden Markov model (HMM), Vector Quantization (VQ), Gaussian Mixture Model (GMM), Dynamic Time Wrapping (DTW), Euclidean Distance and Hybrid Hidden Markov Model for different feature extraction technique such as Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC) and Perceptual Linear Prediction (PLP). The adverse acoustic condition is important to consider as the factor like noise, jitter, reverberation etc. degrades the performance of recognition system to a great extent. So, we put your effort to analyze the performance of different feature extraction technique with their classifier under different scenario for real-time applications.

Index Terms- Different classifiers, Feature extraction techniques, Acoustic conditions and Real-time applications.

I. INTRODUCTION

All human beings voice exchange information about the different languages being spoken and the emotion, gender and, generally, the identity of the speaker. Speaker recognition is a method where a person is determined on the basis of his voice signals [1, 2]. The main objectives of speaker recognition are to determine which speaker is present based on the individual's utterance.
This is in contrast with speaker verification, where the main objective is to verify the person's claimed identity based on his or her utterance. Speaker identification and speaker verification fall under the general category of speaker recognition [3, 4]. In speaker identification there are two types, one is text dependent and another is text independent. Speaker identification is divided into two components: feature extraction and feature classification. In speaker identification the speaker can be identified by his voice, where in case of speaker verification the speaker is verified using database which is stored. The Pitch is used for speaker identification. Pitch is fundamental frequency of a particular person. This is one of the important characteristic of human being, which differ with each other. The speech signal is an acoustic sound pressure wave that originates by exiting of air from vocal tract and voluntary movement of anatomical structure. The human speech contains numerous discriminative features that can be used to identify speakers. Speech contains significant energy from zero frequency up to around 3-4 kHz. The objective of voice recognition is to extract its feature, characterize and recognize the information about the original speaker identity and voice taste.

II. IDENTIFICATION & VERIFICATION

According to Cui and Xue (2009), recognition is classified as talker recognition and voice recognition. The talker recognition also can be classified as relevant to text or irrelevant to text. For voice recognition system, the user needs to pronounce according to the stated contents. It is easy to build up models. For voice recognition that is irrelevant to text, the user does not need to pronounce contents of the talkers. It very difficult to build up models. Rabiner and Juang (1993) have classified speaker recognition into two areas. It is classified into identification and verification as shown in Figure (1) & (2) below:

A. Speaker Identification

In the identification task, an unknown speaker X is compared against a database of known speakers, and the best matching speaker is given as the identification result. The verification task consists of making a decision whether a voice sample was produced by a claimed person [5].

B. Speaker Verification

Speaker verification is the process of accepting or rejecting the identity claim of speakers. It authenticate that a person is who she or he claim to be. This technology can be used as a biometric feature for verifying the identity of a person in an application like banking by telephone and voice mail [6].

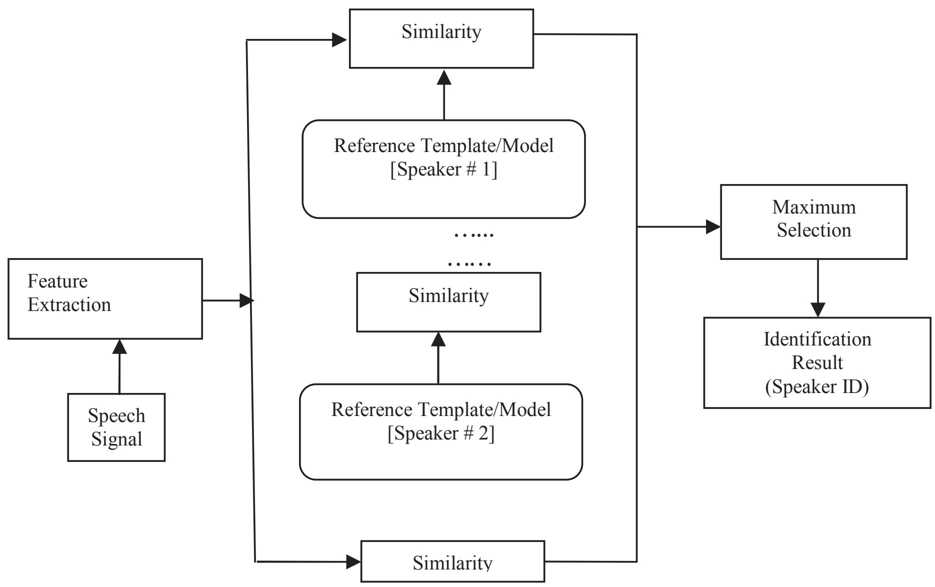

Fig. 1. Speaker identification system structure.

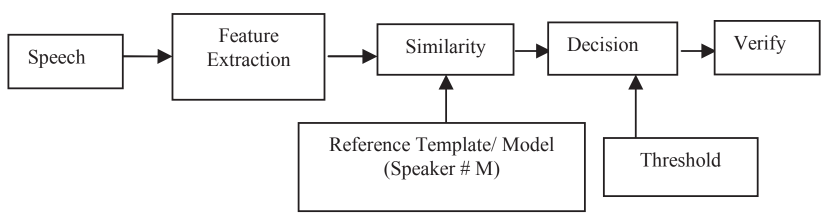

Fig. 2.Speaker verification system structure.

III. FEATURE EXTRACTION

A spoken utterance can be represented with feature vector using feature extraction method. The speech signal of a person will be similar still differently arranged vector and to identify or recognize these feature vectors modelling of voice is done using classifier algorithm by which a template of features is generated for a particular registered user and that is used as reference in recognition process. This means that every registered user will have a reference model in database and if a new user comes into picture then it will be declared as unregistered one. Feature Extraction is the most important part of speech recognition since it plays an important role to separate one speech from other [7]. Because every speech has different individual characteristics embedded in utterances. These characteristics can be extracted from a wide range of feature extraction

techniques proposed and successfully exploited for speech recognition task. But extracted feature should meet some criteria while dealing with the speech signal such as:
1) Easy to measure extracted speech features
2) It should not be susceptible to mimicry
3) It should show little fluctuation from one speaking environment to another
4) It should be stable over time.
5) It should occur frequently and naturally in speech.
The most widely used feature extraction techniques are explained below.

A. Mel-frequency cepstrum co-efficient (MFCC)

MFCC technique is basically used to generate the fingerprints of the audio files. It is based on the known variation of the human ear's critical BW frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the essential characteristics of speech signal. Here signal is divided into different overlapping frames to compute MFCC co-efficient.

Let us consider each frame consist of 'N' samples and let its adjacent frames be separated by 'M' samples where M is less than N. Hamming window is used in which each frame is multiplied. Mathematically, Hamming window equation is given by:

$$W(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (1)$$

Now, Fourier Transform (FT) is used to convert the signal from time domain to its frequency domain. Mathematically, it is given by:

$$X_k = \sum_{i=0}^{N-1} x_i \quad e^{\frac{2\pi i k}{N-1}} \quad (2)$$

Further, the frequency domain signal is then converted into Mel frequency scale, which is more suitable for human perceptions and hearing. This is achieved by taking a set of triangular filters that are used to calculate a weighted sum of spectral components so, that the output of the process approximates a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of the two adjacent filters.

Mathematically, this equation is used to find out the Mel for a given frequency response:

$$M = 2595 \log_{10}\left(1+\frac{f}{700}\right) \quad (3)$$

In the next step log Mel scale spectrum is converted to time domain using Discrete Cosine Transform (DCT). Mathematically, DCT is defined as follow:

$$X_k = \alpha \sum_{i=0}^{N-1} x_i \ (2i + 1/2N) \quad (4)$$

The result of the conversion is known as MFCC and the set of co-efficient is called acoustic vectors. Hence, each input utterance is transformed into a sequence of acoustic vectors.

B. Linear Predictive Coding Analysis (LPC)

The linear predictive coding analysis required where a speech sample can be approximated as linear combination of previous or past speech samples. It is a frame based analysis of the speech signal performed to provide observation vectors [8]. Over the finite interval and my minimizing the sum of the squared differences between the actual samples and linearly predicted one, a unique set of predictor co-efficient is achieved. During voice speech is modelled as the output of time-varying or linear system excited by quasi-periodic pulses or, random noise during unvoiced speech. The LPC techniques provide a reliable, robust and more accurate technique for estimating the parameter that characterizes the linear-time varying system representing the vocal trace. The relation between speech sample S (n) and excitation X (n) for auto regressive model (system assumes all pole modes) is explained mathematically as:

$$S(n) = \sum_{k=1}^{p} \alpha_k \, s(n-k) + G.\, X(n) \quad (5)$$

The system function is defined as:

$$H(z) = \frac{S(Z)}{X(Z)} \quad (6)$$

A linear predictor of order 'p' with prediction co-efficient (α_k) is defined as a system whose output is defined as:

$$\hat{s}(n) = \sum_{k=1}^{p} \alpha_k \, S(n-k) \quad (7)$$

The system function is p^{th} order polynomial and it follows:

$$P(z) = \alpha_k z^{-k} \quad (8)$$

The prediction error e (n) is defined as:

$$\begin{aligned} e(n) &= s(n) - \hat{s}(n) \\ &= s(n) - \sum_{k=1}^{p} \alpha_k \, S(n-k) \quad (9) \end{aligned}$$

The transfer function of prediction error sequence is:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \quad (10)$$

Now, by comparing equation (5) and (10), if $\alpha_k = \alpha_k$ then A (z) will be inverse filter for the system H (z) of equation (6).

$$H(z) = G/A(z) \quad (11)$$

The purpose is to find out set of predictor coefficients that will minimize the mean squared error over a short segment of speech waveform. So, short-time average prediction error is defined as [9].

$$E(n) = \sum_m (e_n(m))^2 = \sum_m \{s_n(m) - \sum_{k-1}^{p} \alpha_k s_n(m-k)\} \quad (12)$$

where, $s_n(m)$ is segment of speech in surrounding of n samples i.e. $s_n(m) = s(n+m)$

Now, the value of α_k minimize E_n are obtained by taking $\partial E_n / \partial \alpha_i = 0$ & i = 0, 1, 2....p thus getting the equation:

$$\sum_m s_n(m-i) s_n(m) = \sum_{k-1}^{p} \alpha_k \sum s_n(m-i) s_n(m-k) \quad (13)$$

$$\text{If } \emptyset_n(i,k) = \sum_m s_n(m-i) s_n(m-k) \quad (14)$$

Thus, equation (13) rewritten as:

$$\sum_{k=1}^{p} \alpha_k \varphi_k(i, k) = \varphi_k(i, 0), \quad \text{for } i = 1, 2, 3...p \quad (15)$$

The three ways available to solve above equation i.e. autocorrelation method, lattice method and covariance method. In speech recognition the autocorrelation is widely used because of its computational efficiency and inherent stability [9].

Speech segment is windowed in autocorrelation method as discuss below:

$$S_n = S(m+n) + w(m) \text{ for } 0 \le m \le N-1 \quad (16)$$

where, w (m) is finite window length. Then, we have

$$\varphi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i) s_n(m-k) \text{ for } 1 \le i \le p,\ 0 \le k \le p \quad (17)$$

$$\varphi_n(i,k) = R_n(i-k) \quad (18)$$

where, $R_n(k) = \sum_{m=0}^{N-1-k} s_n(m)\, s_n(m+k)$

R_k (k) is autocorrelation function then equation (15) is simplified as [11]

$$\sum_{k-1}^{p} \alpha_k R_n(|i-k|) = R_i(i) \text{ for } 1 \le i \le p \quad (19)$$

This results in P× P of autocorrelation value and it is symmetric and all elements along diagonal are equal. Thus using Durbin's recursive procedure the resulting equation is solved as:

$$E^{(0)} = R(0) \quad (20)$$

$$k_i = \{R(i) \sum_{j=1}^{i-1} \alpha_j^{(i-1)} R(i-j)\} / E^{(i-1)} \quad (21)$$

$$\alpha_i^{(i)} = k_i \quad (22)$$

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} - k_i \alpha_{(j-1)}^{(j-1)} \quad (23)$$

$$E^{(i)} = (1 - k_i^2)\, E^{(i-1)} \quad (24)$$

Then from equation (19) to (22) are solved recursively for i = 1, 2 ...p and this give final equation as:

α_j = LPC coefficient = $\alpha_j^{(p)}$

k_i = PACOR coefficients

Voiced regions of speech all pole model of LPC provides a good approximation to vocal tract spectral. But for unvoiced and nasalized region of speech the LPC model is less effective than voice region. A very essential LPC parameter set which is derived directly from LPC coefficients is LPC cepestral coefficients C_m. The recursion used for this discussed as [10]:

$$C_0 = \ln G \quad (25)$$

$$C_m = \alpha_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) C_k a_{m-k} \text{ for } 1 \le m \le p \quad (26)$$

$$C_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) C_k a_{m-k}, \text{ for } m > p \quad (27)$$

Where, G is the gain in LPC model. This method is very efficient, as it does not require explicit cepstral computation. Hence combine decorrelating property of cepestral with computational efficiency of LPC analysis.

C. Perceptually Based Linear Predictive Analysis (PLP):

PLP analysis model perpetually motivated auditory spectrum by a low order all pole function using the autocorrelation LP method. The concept of PLP method is shown in form of block diagram below in Figure (3). It involves two major steps: obtaining auditory spectrum, approximating the auditory spectrum by an all pole model. Auditory spectrum is derived from the speech waveform by critical-band filtering, equal loudness curve pre-emphasis, and intensity loudness root compression [11].

IV. CLASSIFIERS/ PATTERN MODELLING

The speech feature extraction in a categorization problem is about reducing the dimensionality of the input vector while maintaining the discriminating power of the signal. As we know from fundamental formation of speaker identification and verification system, that the number of training and test vector needed for the classification problem grows with the dimension of the given input so we need different models to determine the similarity between a reference pattern and test pattern during testing as discussed below.

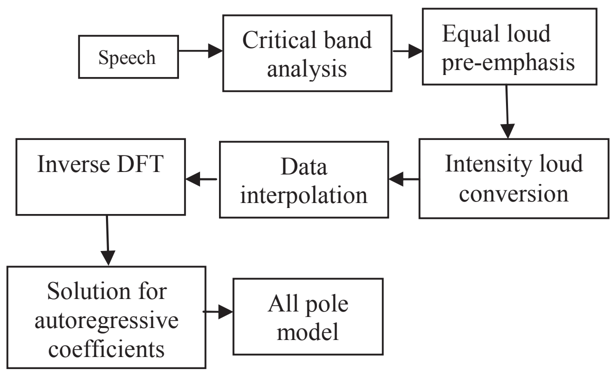

Fig. 3. PLP speech analysis process.

A. Gaussian mixture model (GMM)

The GMM is a density estimator. The distribution of the feature vector x is modelled clearly using a mixture of M Gaussians. Expectation maximization algorithm is used to estimate mean, covariance parameters. During recognition, a sequence of features is extracted from the input signal. Then the distance of the given sequence from the model is obtained by computing the log likelihood of given sequence. The model that provides the highest likelihood score is verified as the identity of the speaker [12].

B. Dynamic time wrapping (DTW)

This is used specifically to deal with variance in speaking rate and variable length of input vectors because this algorithm calculates the similarity between two sequences which may vary in time or speed. To normalize the timing differences between test utterance and the reference template, time warping is done non-linearly in time dimension. After time normalization, a time normalized distance is calculated between the patterns. The speaker with minimum time normalized distance is identified as authentic speaker [12]. The text-dependent template model is a sequence of feature vectors ($X_1, X_2, \ldots, X_N$). The DTW finds a match between the template model and the input sequence of feature vectors ($Y_1, Y_2, \ldots, Y_M$). In general N is not equal to M because of the differences in the speaking rate. The DTW match yields as score D_{total} given as

$$D_{total} = \sum_{i=1}^{M} d\left(Y_i, X_{j(i)}\right) \tag{28}$$

Where j (i) is obtained from the DTW algorithm and d (.) is the distance between the feature vectors.

C. Hidden Markov model (HMM)

HMM is a stochastic model. The elements of HMM are number of states, number of distinct observation symbols, the probability of going from one state to another, the observation symbol probability distribution and probability of being in a particular state initially. It assumes that the observation at some time is generated by some process whose state is hidden from the observer. It also follows markov property that the current state St does not depend upon any state prior to time t-1 and depends upon only on St-1 state [13]. The speech training features are represented by probability measures which train the HMM speaker model. For each speaker model, the metric is determined as the probability of the observation sequence. The HMM speaker model which yields the highest probability is selected [12].

D. Vector quantization (VQ)

Here a large set of vectors are divided into groups having approximately the similar number of points closest to them. Each vector is represented by its centroid point and is defined as a mapping function that maps k-dimensional vector space to a finite set CB = { C_1, C_2, C_3.... C_N} where CB is called codebook consisting of N number of code vectors and each code vector C_i = {c_{i1},c_{i2}, c_{i3} ..... c_{ik}} is of dimension k. This method is commonly used to generate codebook known as Linde-Burzo-Gray (LBG) algorithm. Feature vectors are extracted from input speech signal and the Euclidean distance between input speech signal and each code vector calculated. The input vector belongs to the cluster of code vector that yields the minimum distance [31]. The Euclidean distance is mathematically defined as:

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

(29)

where x_j is the jth component of the input vector and y_{ij} is the jth component of the codeword y_i.

E. Support vector machines (SVM)

Support Vector Machine is a supervised learning algorithm. It needs training of the tool before classification procedure gets started. This is the best tool for binary classification of the data. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the input. The hyperplane is constructed defined by set of weights W, data points X and a bias or offset b, such that: W.X + b = 0 where, W.X denotes the dot product of the data and the normal vector to the hyperplane. The parameter b determines the offset of the hyperplane from the origin along the normal vector. Figure 4. shows the partition of the input data into two classes. Points lying on the hyperplane satisfy the equation (28):

$$W.\,X + b = 0 \qquad (30)$$

Points lying on one side of this hyperplane are denoted by class C1 as positive examples satisfying: W.X + b > 0, d (i) = +1 Points lying on the other side of this hyperplane are denoted by class C2 as negative examples satisfying the above condition.

E. Hybrid hidden Markov model (HHMM)

Hybrid HMM/ANN is a potential connectionist method for developing ASR system. This method combined the respective properties of ANN and HMM with the aims to solve the limitations in ASR, such as in building continuous and speaker independent recognizer.

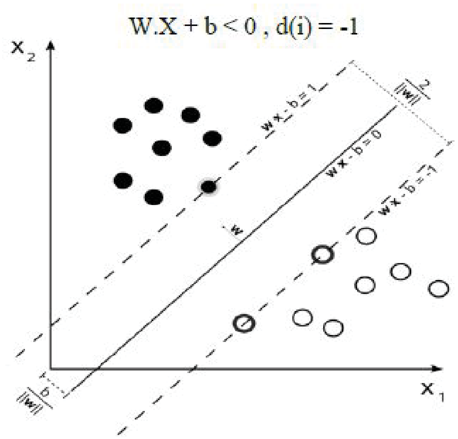

Fig. 4. Hyperplane dividing input data into two parts.

V. PREVIOUS OUTCOME

Many papers have different opinion when combining feature extraction with classifiers but they never focus on the acoustic condition and the real application. J. Kumar, O. P. Prabhakar etc. [7] shows the result of different classifiers with the feature extraction techniques MFCC, LPC, and PLP feature extraction techniques. In an average the MFCC and VQ techniques give the maximum recognition rate in graphical representation as shown in Figure 5 below.

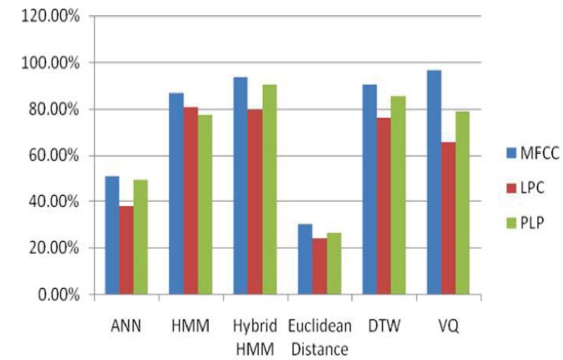

Fig. 5. Recognition rate with different feature extraction techniques.

In this analysis only the recognition rate is well discussed and not clearly mention about the effect of acoustic behavior and the application which is very crucial for the real data handling. So, keeping in mind we try to improve the rate and the performance of difference feature extraction technique along with the classifier under different scenario (conditions) like surrounding condition, weather condition and crowd condition. And this will help to select the proper feature extraction technique with exact classifiers that will help for different application while training and testing is done for speaker identification and the authentic person is verified far away sitting in remote location or in rural areas.

VI. DATA ANALYSIS

The data has been analysis on “HINDI” language which is stored in the data base and the voice sample of these data have been analysis using different feature extraction along with the classifiers and simulated using MATLAB version 2012a using 2.0 Gb RAM, dual core 2 processor with Window 7 OS professional and the application chosen for voice in bank transaction, official activity, online transaction etc. The table I, table II and table III shows the simulated outcome. Taking the help of sound proof room with audio sharing devices all the verification is done.

TABLE I: SURROUNDING CONDITION (n-number of person talking nearby)

Classifiers	**Feature Extractor (MFCC) in percentage**	**Feature Extractor (PLP) in percentage**	**Feature Extractor (LPC) in percentage**
Hidden Markov Model	88.15	79.56	80.81
Hybrid HMM	94.23	91.76	79.76
Dynamic time wrapping	91.34	87.98	76.21
Vector quantization	96.14	79.91	66.01
Euclidean distance	32.23	28.43	25.34
Gaussian mixture model	98.67	93.34	95.89
Support vector machine	86.89	69.67	83.07

TABLE II: WEATHER CONDTION (bad weather like rainy, foggy, hazy)

Classifiers	**Feature Extractor (MFCC) in percentage**	**Feature Extractor (PLP) in percentage**	**Feature Extractor (LPC) in percentage**
Hidden Markov Model	83.23	75.78	83.05
Hybrid HMM	90.30	90.65	69.17
Dynamic time wrapping	89.34	85.19	72.02
Vector quantization	96.14	70.23	63.10
Euclidean distance	23.26	29.12	18.19
Gaussian mixture model	95.45	92.15	96.34
Support vector machine	79.10	74.18	90.68

TABLE III. CROWDY (more than 15 person involved)

Classifiers	**Feature Extractor (MFCC) in percentage**	**Feature Extractor (PLP) in percentage**	**Feature Extractor (LPC) in percentage**
Hidden Markov Model	90.99	73.39	82.05
Hybrid HMM	95.45	90.10	73.89
Dynamic time wrapping	92.56	88.13	73.09
Vector quantization	97.10	75.19	62.10
Euclidean distance	43.21	37.58	29.08
Gaussian mixture model	99.10	90.49	89.20
Support vector machine	81.23	81.09	77.23

VII. SIMULATION RESULT

The simulation result and the representation in the graph form are shown below and the different colours represent its variation under different scenario.

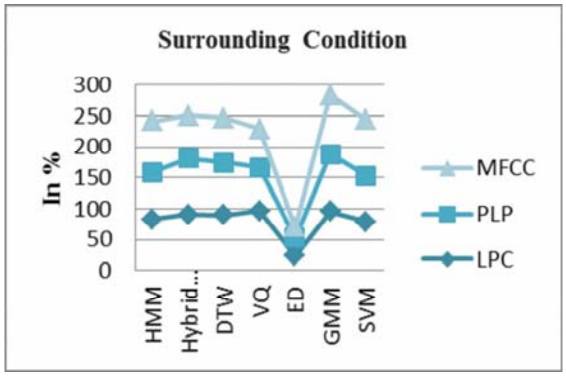

Fig. 6. Surrounding condition for different extraction techniques.

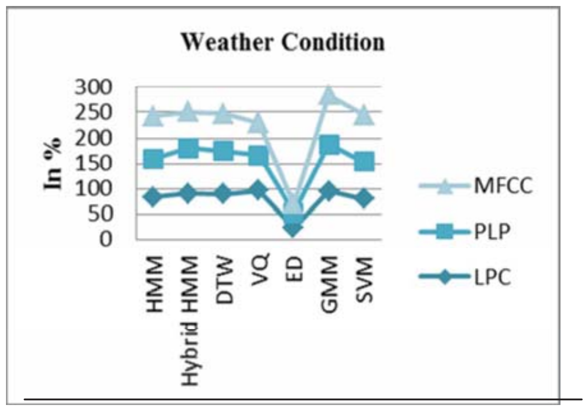

Fig. 7. Weather condition behavior of different feature extraction techniques.

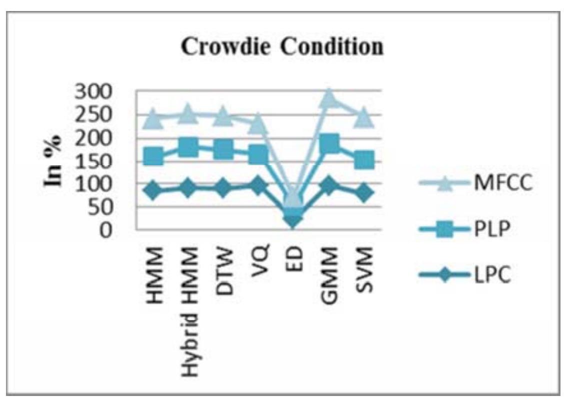

Fig. 8. Crowdie condition behavior for different feature extraction techniques.

Application behavior simulation results give idea that which application is most suitable in different environment and proves to be best for the speech recognition task. Same feature extraction technique as well as classifier is used for the analysis. The three real time application like id exchange, pin distribution etc can be done for remote voice, local voice i.e. from nearby location and the crowdie voice.

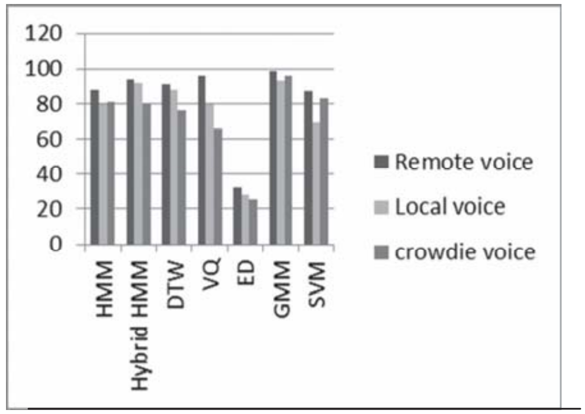

Fig. 8. Behavior of different voice for the features extraction and its classifiers.

VIII. CONCLUSION

The above result shows that the performance of MFCC is quite efficient with GMM for the surrounding condition when n-number of people involved as compared to PLP and LPC highlighted in above representation. The performance of MFCC with VQ shows good percentage of improvement as compared to PLP and LPC for the weather condition. The MFCC proves to be good with GMM for the crowdie condition as compared to PLP and LPC s. So, we can prefer to go with MFCC along with GMM for the surrounding condition and crowdie condition where as for weather surrounding we can go for MFCC with VQ technique. The remote voice detection i.e. one of its application we can go for VQ technique which shows good percentages.

REFERENCES

[1] Campbell J.P. and Jr. “Speaker recognition: A Tutorial” Proceeding of the IEEE. Vol. 85, 1437- 1462 1997.
[2] S. Fruit. “Fifty years of progress in speech and speaker recognition,” Proc. 148th ASA Meeting, 2004.
[3] A. Rosenberg, “Automatic speaker recognition: A review,” Proc. IEEE, vol. 64, pp. 475487, Apr.1976.
[4] G. Doddington, “Speaker recognition-Identifying people by their voices,” Proc. IEEE, vol. 73, pp. 1651-1664, 1985
[5] A.S.B.A Sukor, “Speaker identification system using MFCC procedure & noise reduction method”, M.tech Thesis, Universiy Tun Hussein Onn Malaysia, 2012.
[6] T.Kinnunean, K.Kaprov & PFrank, “Real Time Speaker Identification & Verification”, IEEE Transaction Speech & Audio Processing, pp. 01-13, 2007.
[7] J. Kumar, O. P. Prabhakar. N. K. Sahu, “Comparitive analysis of different feature extraction and classifier technique for speaker identification system: a review”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, pp. 2760-2769, January 2014.
[8] E. M. Mohammed, M. S. Sayed, A. M. Moselhy, and A. A. Abdelnaiem, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification", International Journal of Signal Processing, Image Processing and Pattern Recognition, Vol. 6, No. 3, pp. 55-66, June 2013.
[9] L. R. Rabiner and R. W. Schafer, Digital Processing of Speech Signals. Englewood Cliffs, New Jersey: Prentice-Hall, 1978.
[10] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Englewood Cliffs, New Jersey:Prentice- Hall, 1978.
[11] H. Hermansky, B. A. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," Proc. IEEE Int. Conf. on Acoustic, speech, and Signal Processing," pp. 509-512, Aug.1985.
[12] R. P. Ramachandran, K.R. Farrell, R. Ramachandran, R. J. Mammone, “Speaker Recognition—General Classifier Approaches and Data Fusion Methods”, Pattern Recognition in Information Systems, vol. 35, issue-12, pp. 2801-2821, December, 2002.
[13] Ghahramani Z., “An Introduction to Hidden Markov Models and Bayesian Networks”, International Journal of Pattern Recognition and Artificial Intelligence, vol. 5, issue-1, pp. 9-42, 2001.