International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 8 (2017) pp. 1633-1638
© Research India Publications. http://www.ripublication.com

Speaker Identification and Verification Using Different Model for Text-Dependent

[1]Shrikant Upadhyay, [2]Sudhir Kumar Sharma & [3]Aditi Upadhyay

Department of Electronics & Communication Engineering,
[1]Cambridge Institute of Technology, Ranchi, Jharkhand, India.
[2, 3] Jaipur National University, Jaipur, Rajasthan, India.

Abstract

Human voice is one of the medium through which everybody communicate and share their experiences, thought, feeling etc. and make it meaningful by understanding their meaning and actions. Issue is not that person is speaking in front of about 10-20m and getting clear voice to the next person sitting in front. The person stays at far location and tries to access their various need applications like password, PIN number, account access etc. with the help of voice then it will be quite difficult to do so. Speaker is the medium through which we can transfer your voice in such occasion so, speaker identification and verification is very important. Text-dependent sample is quite easy to determine and verified by the person sitting to the remote location as all information has prior saved to the database of unknown user and doesn't require any additional information. Different proposed model have been analyzed in terms of efficiency, loss and error using different feature extraction techniques in this paper. We put your effort and try to find out that how the model behaves using different feature extraction combinations.

INTRODUCTION

Within the past decade, technological growth advances remote collaborative data processing over huge computer network, hands-free sound capture, telebanking have increased the demand for improved method of information security. For personal information including credit history, bank details, medical records, the ability to verify the identity of individuals attempting to access such data is challenging and critical. To date, low-cost strategy such as magnetic cards, personal identification numbers (pins), passwords have been widely used. More advanced security measures have also been developed (e.g., retinal scanners, face recognition, as well as automatic finger print detectors). The use of these procedures has been limited by both cost and ease of use. In recent years, speaker recognition (recognizing a person from his/her voices by a machine) and verification algorithms have also received considerable attention. In particular, speech provides a convenient and natural form of input conveys a significant amount of speaker dependent information, and it is inexpensive to collect and analyze. Voice is a phenomenon that is highly dependent on the speaker who produced it. Many physical aspects of speech such as the timber, tone or intensity vary a lot from a speaker to another. The same happens with other linguistic aspects as the single intonation and range of vocabulary or expressions a speaker normally uses. All these properties make voice a very powerful biometric key to be used in security systems since the physical characteristics of speech are easy to measure and compare in comparison to other biometric keys. In addition, the speech signal is quite well-known and has been deeply studied for many years so many powerful algorithms can be found to deal with this kind of signal.

Text-dependent task involves some form of pre-determined or prompted password, in order to obtain the required text. It can be used for applications such as voice mode password or signature verification. In such applications, there is a need to change the password frequently and it can be done easily by changing the pre-determined text. In text-dependent speaker verification, during enrolment phase a limited number of utterances of the fixed text are collected. Therefore, approaches based on template matching are used for pattern comparison instead of approaches based on statistics or artificial neural networks, which needs a large amount of training data. Research in speech processing and communication, for the most part, was motivated by people's desire to build mechanicals models to emulate human verbal communication. Research interest in speech processing today has done well beyond the notion of mimicking human vocal apparatus [1].

Speech is the primary communication medium between people. This communication process has a complex structure consisting not only of the transmission of voice but also includes transmission of many speaker specific information. Therefore, from the last five decades people have come forward to investigate various aspects of speech such as mechanical realization of speech signal, human machine interaction and speech & speaker recognition. Speech is a biometric that combines physiological and behavioral characteristics. It is particularly useful for remote-access transactions over telecommunication networks. Presently this task is the most challenging one to researchers. People can reliably identify familiar voices and about 2-3 seconds is enough to identify a voice, although performance decreases for unfamiliar voices [2]. Even if duration of the utterance was increased, but played backward (which distorts timing and articulatory cues), the accuracy decreases drastically. Widely varying performance on this background task suggested that cues to voice recognition vary from voice to voice, so that voice patterns may consist of a set of acoustic cues from which listeners select a subset to use in identifying individual voices.

IDENTIFICATION & VERIFICATION

Speaker recognition by a machine involves three stages. They are: (1) Extraction of features to represent the speaker information present in the speech signal. (2) Modelling of speaker features. (3) Decision logic to implement the identification or verification task. The primary task in a speaker recognition system is to extract features capable of representing the speaker information present in the speech signal. It is known to us that human beings use high-level features such as style of speech, speech dialect and verbal mannerisms (for example, a particular kind of a laugh or use of particular idioms and words) to recognize speakers. Intuitively, it is clear that these features constitute important speaker information. Difficulty arises due to limitations of the existing feature extraction techniques [3]. Current speaker recognition systems follow segmental features such as the shape of the vocal tract to represent the speaker-specific information. These features show significant variations across speakers, but they also show considerable variations from time to time for a single speaker. In addition to this, the characteristics of the recording equipment and transmission channel are also reflected in these features [3].

Once a proper set of feature vector is obtained, then in the next phase task in speaker recognition is to develop a model (prototype) for each speaker. The development of speaker modelling is called the training phase. The block diagram of training phase is shown in Figure 1.

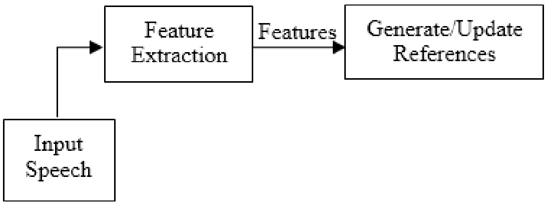

Figure 1. Training in speaker recognition system

Feature vectors representing the voice characteristics of the speaker are extracted and used for building the reference models. The performance of a speaker recognition system depends primarily on the effectiveness of the models in capturing the speaker-specific information, and hence this phase plays a major role in determining the performance of a speaker recognition system. The final stage in the development of a speaker recognition system is the decision logic stage, where a decision to either accept or reject the claim of a speaker is taken based on the result of matching techniques used. The block diagram of decision logic and testing phase process is shown in the Figure 2.

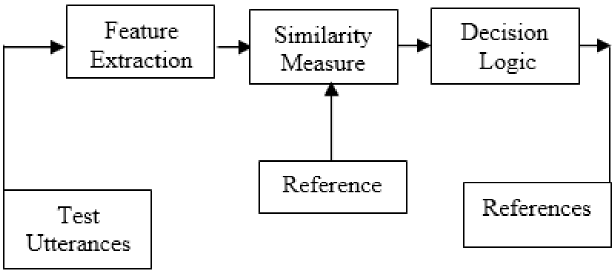

Figure 2. Testing in speaker recognition system

In speaker identification a speech utterance from an unknown speaker is analyzed and compared with models of all known speakers. The unknown speaker is identified as the speaker whose model best matches the input utterance. Figure 3 shows the basic structure of speaker identification system. Speaker identification can be a closed set identification or an open set identification. In a closed set identification, it is assumed that the test utterance belongs to one of N enrolled speakers (N decisions). In the case of open set identification, there is an additional decision to be made to determine whether the test utterance was uttered by one of the N enrolled speakers or not, that is, there are N+1 decision levels.

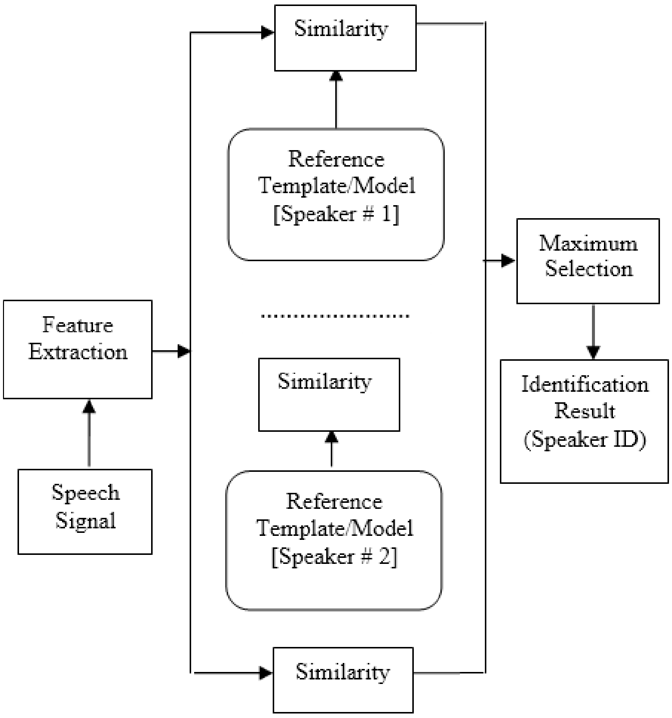
Similarity
Reference Template/Model [Speaker # 1]
....
Similarity
Reference Template/Model [Speaker # 2]
Similarity
Feature Extraction
Speech Signal
Maximum Selection
Identification Result (Speaker ID)

Figure 3. Speaker verification system

FEATURE EXTRACTION METHODS

Feature extraction involved in signal modeling that performs temporal and spectral analysis. The need of feature extraction arises because the raw speech signal contains information to convey message to the observer or receiver and has a high dimensionality. Feature extraction algorithm derives a characteristics feature vector with lower physical or spatial properties.

A. Mel-Scale Cepstrum Co-efficient (MFCC)

MFCC technique is basically used to generate the fingerprints of the audio files.

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:

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

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

$$X_k = \sum_{i=0}^{N-1} x_i\ e^{\frac{2\pi ik}{N-1}} \tag{2}$$

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

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

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

The result of the conversion is known as MFCC and the set of co-efficient is called acoustic vectors.

B. Linear Predictive Coding Analysis (LPC)

It is a frame based analysis of the speech signal performed to provide observation vectors [3]. The relation between speech sample S (n) and excitation X(n) for auto regressive model (system assume all pole mode) is explained mathematically as:

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

The system function is defined as:

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

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

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

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

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

The prediction error e (n) is defined as:

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

The transfer function of prediction error sequence is:

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

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):

$$H(z) = G/A(z) \tag{11}$$

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 [4].

$$\begin{aligned} E(n) &= \textstyle\sum_m (e_n(m))^2 \\ &= \textstyle\sum_m \{s_n\ (m) - \sum_{k-1}^{p} \alpha_k\, s_n\ (m-k)\} \end{aligned} \tag{12}$$

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:

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

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

Thus, equation (13) rewritten as:

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

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 [4].

Speech segment is windowed in autocorrelation method as discuss below:

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

Where, w (m) is finite window length

Then, we have

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

$$\varphi_n\, (i,k) = R_n\ (i-k) \tag{18}$$

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 [5]

$$\textstyle\sum_{k-1}^{p} \alpha_k\ R_n\, (|i-k|) = R_i\ (i) \text{ for } 1 \leq i \leq p \tag{19}$$

Thus using Durbin's recursive procedure the resulting equation is solved as:

$$E^{(i)} = (1 - k_i^2)\, E^{(i-1)} \tag{20}$$

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

A very essential LPC parameter set which is derived directly from LPC coefficients is LPC cepstral coefficients C_m. The recursion used for this discussed as [6]:

$$C_0 = \ln G \tag{21}$$

$$C_m = \alpha_m + \sum_{k=1}^{m-1}\left(\frac{k}{m}\right) C_k a_{m-k} \text{ for } 1 \le m \le p \tag{22}$$

$$C_m = \sum_{k=1}^{m-1}\left(\frac{k}{m}\right) C_k a_{m-k} \text{ for } m > p \tag{23}$$

C. Linear Prediction Cepestral Coefficients (LPCC)

The basic parameters for estimating a speech signal, LPCC play a very dominant role. This method is that where one speech sample at the current time can be predicated as a linear combination of past speech sequence or sample. LPCC algorithm in term of block diagram is shown in Figure 4. below:

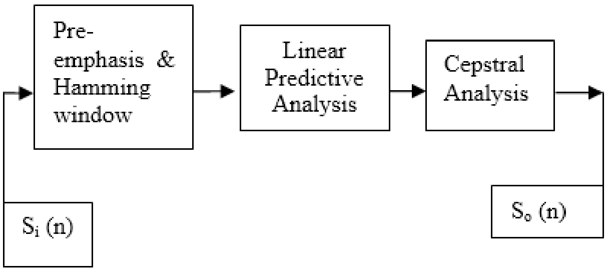

Figure 4. LPCC algorithm process

RELATED WORK

This paper focus on the Hindi database text-dependent sample spoken from SHUNAY (0) to NAU (9). This database is prepared using GoldWave software using version 5.55. This sample is tested using MATLAB tool for their efficiency and error rate. The proposed model is tested using various combinations of feature extraction techniques like LPC+LPCC, LPCC+MFCC and MFCC+LPC. The considered parameters for analysis are mention in Table 1.

Table 1: Parameters used for analysis

Parameters	Values
Total number of speakers	24
Single channel	16KHz
Sampling rate	16 bit (For less quantization error)
Language	Hindi (Phonetics is rich)

PROPOSED MODEL

The proposed model designed based on the aims to achieve better performance by exploiting differences in the nature of the errors occurred at the level of different modeling structures known as recognizer output voting error rate. This method treats the output of multiple ASR systems as independent knowledge sources, and then combines these outputs to produce a composite output with a lower word error rate than an individual ASR output. It proceeds in two stages as shown in Figure 5.

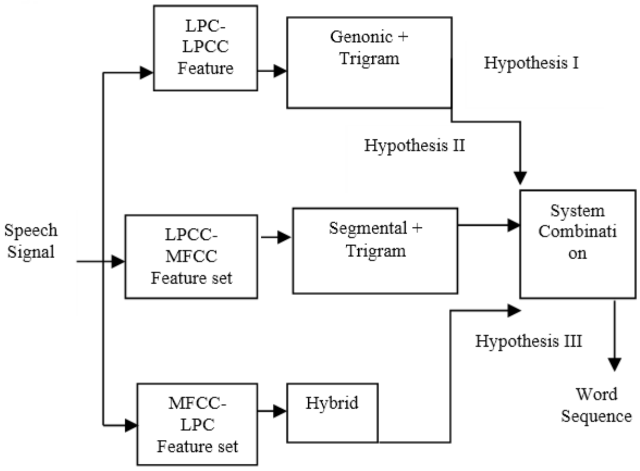

Figure 5. Proposed model (ROVER based)

The second proposed model is based on combined combination known as CNC as shown in Figure 6. below approach to speech recognition, the decoder outputs the string of word hypotheses corresponding to the path with highest posterior probability given the acoustic and a language model, and represented by either N-best lists or word lattice.

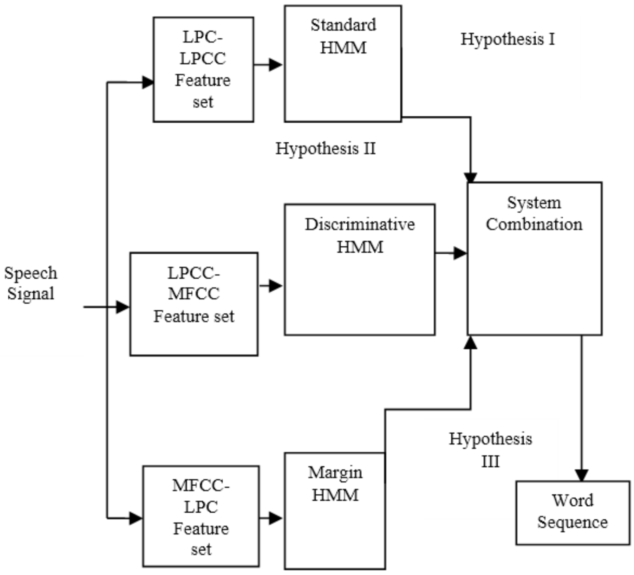

Figure 6. Proposed model (CNC based)

These two models have been used to calculate the efficiency and error rate of the speech signal spoken from SHUNYA (0) to NAU (9). And with the help of this model we exactly know that which model performs better also that which combinations of feature extraction techniques will better for both the model. Earlier no such models have been develop using such combination. So, it is quite useful for speaker recognition.

RESULTS

The simulation has been done using MATLAB tool also taking help with COOL and GoldWave software to analyze the database.

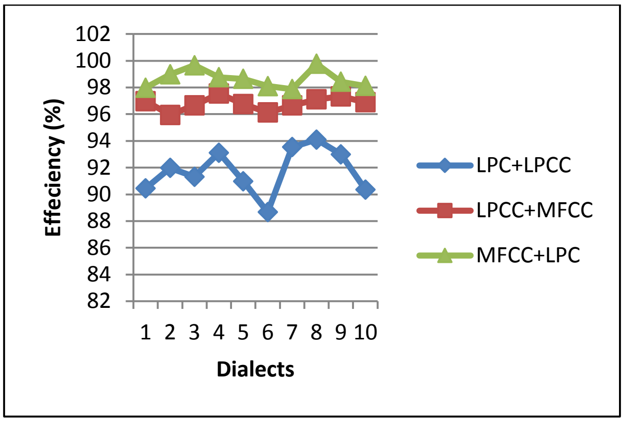

Figure 7. Efficiency for ROVER model

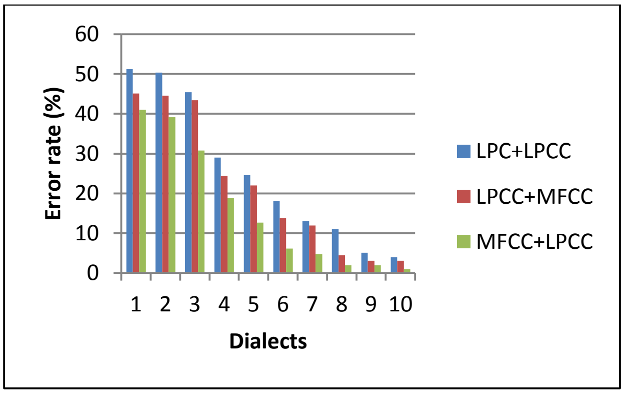

Figure 8. Error rate for ROVER model

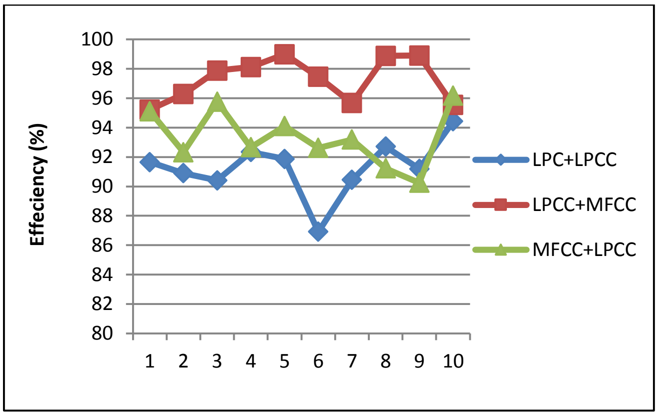

Figure 9. Efficiency for CNC model

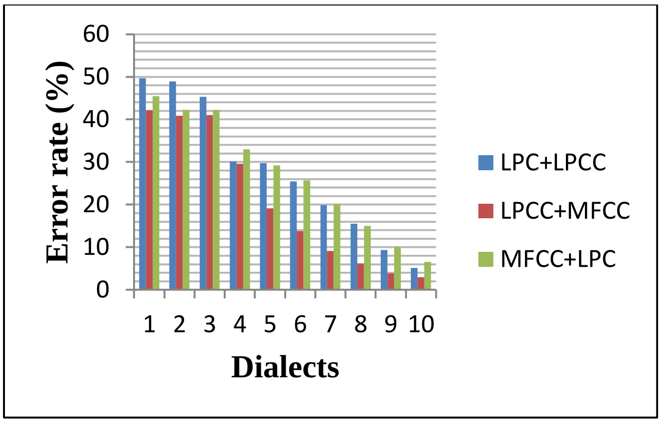

Figure 10. Error rate for CNC model

Calculations:

GMMs (Gaussian Mixture Models) are trained separately on each speaker's enrolment data using the Expectation Maximization (EM) algorithm [255]. The update equations that guarantee a monotonic increase in the model's likelihood value are:

Mixture weight: $\overline{p_\iota} = \frac{1}{T} \sum_{t=1}^{T} p(i \mid \overrightarrow{x_t}, \lambda)$ (24)

Means: $\overrightarrow{\overline{\mu_\iota}} = \frac{\sum_{t-1}^{T} p(i|\overrightarrow{x_t}, \lambda) \overrightarrow{x_{t,}}}{\sum_{t-1}^{T} p(i|\overrightarrow{x_t}, \lambda)}$ (25)

Variance: $\overline{\sigma_\iota}^2 = \frac{\sum_{t-1}^{T} p(i|\overrightarrow{x_t}, \lambda) \overrightarrow{x_t}^2}{\sum_{t-1}^{T} p(i|\overrightarrow{x_t}, \lambda)} - \overrightarrow{\overline{\mu_\iota}}^2$ (26)

Recognition Rate $= \frac{Successfully\ detected\ words}{Number\ of\ words\ in\ test\ set}$ (27)

Experiments were performed with different types of speech, that is, for slow speech (less than 12 phones per second), for normal speech (12 to 18 phones per second) and for fast speech (more than 18 phones per second). Maximum accuracy is achieved for normal speech as shown in Figure 11.

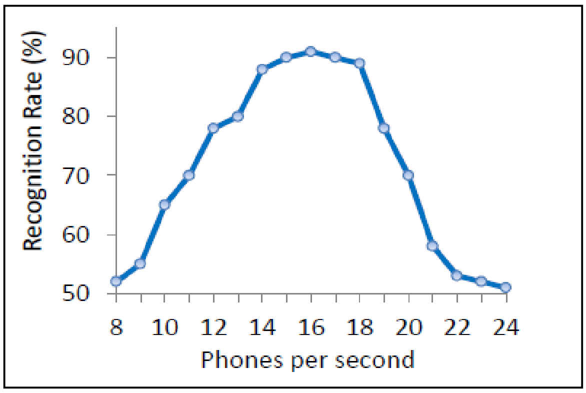

Figure11. Speech rate versus accuracy

CONCLUSION & FUTURE SCOPE

Efficiency rate of MFCC+LPC for ROVER model is 98.22% compared to LPCC+MFCC and LPC+LPCC i.e. 93.99% and 92.10. Error rate of MFCC+LPC is quite low i.e. 1.22% compared to LPCC+MFCC and LPC+LPCC i.e. 3.53% and 4.42%.

On the other hand efficiency rate of LPCC+MFCC for CNC model is 98.99% compared to LPC+LPCC and MFCC+LPCC i.e. 94.62 and 96.23. Error rate of LPCC+MFCC here is 1.29% compared to LPC+LPCC and MFCC+LPCC i.e. 2.54 and 3.39.

So, with the help of above calculation and analysis we can choose ROVER model with the feature extraction combination of MFCC+LPC for getting good and efficient speech recognition. For CNC based model we can have good and efficient combination of LPCC+MFCC which give good efficiency and less error rate compared to other two combinations and may be useful for accurate speaker recognition.

The combination may be used for different types of model taking the acoustic condition in future.

ACKNOWLEDGMENT

We are extremely thankful to the Prof. (Dr.) Sudhir Kumar Sharma (HOD, Deputy Director, Jaipur National University), and Prof. (Dr.) Pawan Kumar (HOD, Cambridge Institute of Technology) for their technical support and guidance.

REFERENCES

[1] D. G. Childers, R. V. Cox, R. Demori, B. H. Juang, J. J. Mariani, P. Price, S. Sagayama, M. M. Sondhi and R. Weischedel, "The past, present and future of speech processing,' *IEEE Processing Magazine,* pp. 24-48, May 1998.

[2] D. Lancker, J. Kreiman and K. Emmorey, "Familiar voice recognition: Pattern and parameters-recognition of backward voices," *Phonetics*, vol. 13, no. 1, pp. 19-38, 1985.

[3] 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, pp. 55-66, June 2013.

[4] L.R. Rabiner and R.W. Schafer, *Digital processing of speech signals*. Englewood cliffs, New Jersey: Prentice-Hall, 1978.

[5] L.R. Rabiner and B.H. Juang, *Fundamental of speech recognition*, Englewood cliffs, New Jersey: Prentice-Hall, 1978.

[6] Han Y, Wang G and Y *Yang, "Speech emotion recognition based on mfcc", Journal of Chong Qing. University of Posts and Telecommunication* (Natural Science Ediion) vol.69, pp. 34-39, 2008.

AUTHORS PROFILE:

Mr. Shrikant Upadhyay, Research Scholar at Jaipur National University, Jaipur. He received his M.Tech degree from Dehradun Institute of Technology (University) in 2011. His current research area includes speech processing, image processing, neural network and the security challenges for speaker identification and verification in signal processing domain.

Dr. Sudhir Kumar Sharma is Professor & Head at the Department of Electronics & Communication Engineering, School of Engineering and Technology, Jaipur National University, Jaipur, Rajasthan, India. Professor Sharma received his Ph.D. in Electronics from Delhi University in 2000. Professor Sharma has an extensive teaching experience of 19 years. He has been keenly carrying out research activities since last 20 years prominently in the field of Optical Communication Signal Processing.

Mrs. Aditi Upadhyay, Research Scholar. She received M.Tech degree from Dehradun Institute of Technology (University) in 2012. Her current research area includes speech processing, image processing using different state of HMM. Image enhancement by using different acoustic model.