# Robust Features for Connected Hindi Digits Recognition

A. N. Mishra[1], Mahesh Chandra[2], Astik Biswas[3], S. N. Sharan[4]
*[123]Department of ECE, BIT, Mesra, Ranchi, India*
*[4]Director GNIT, Greater Noida, India*
*an_mishra53@rediffmail.com[1],shrotriya@bitmesra.ac.in[2],*
*talktosayak@yahoo.co.in[3], snathsharan@yahoo.com[4]*

***Abstract***

*Connected digits recognition is important in many applications such as voice-dialing telephone, automated banking system, automatic data entry, PIN entry, etc. In this paper robust features such as Revised perceptual linear prediction (RPLP), Bark frequency cepstral coefficients (BFCC) and Mel frequency perceptual linear prediction (MF-PLP) are used for speaker-independent connected Hindi digits recognition in clean and noisy environments. The recognition performance of these features is compared with recognition performance of Mel frequency cepstral coefficient (MFCC), ΔMFCC and Perceptual linear prediction (PLP) features. MF-PLP features have shown best recognition efficiency for clean as well as for noisy database. MFCC features are calculated by using feature extraction tool of Hidden Markov model Toolkit (HTK). All other features except MFCC are calculated using Matlab and saved in HTK format. Training and testing for speech recognition is done using HTK.*

***Keywords:*** *Connected Hindi Digits, BFCC, RPLP, HTK, MF-PLP.*

.

## 1. Introduction

Recent research on mobile phone users all over the world and the number of telephone landlines in operation confirm that the voice is the most accessible biometric as no extra acquisition device or transmission system is needed. This fact gives voice an advantage over other biometrics especially when remote users or systems are taken into account [1]. Voice based features can be used to verify the identity of the person and allow access to services such as banking by telephone, database access services, security control for confidential information areas and remote access to computers [2]. Speech recognition is the key technology for effective man-machine interface. There has been a lot of research in the area of speech recognition for different languages like English, Mandarin, Arabic etc but little work has been carried out for Hindi speech recognition. Due to this reason only a small percentage of computer literate Indians are able to take advantage of the new advancement in the computer technology. The connected digits recognition task for Hindi language is difficult due to a large variability's in Hindi dialect.

A speech recognition system has two major components, feature extraction and classification as shown in Figure 1. The recognition performance heavily depends on the performance of the feature extraction block. Thus choice of features and its extraction from the speech signal should be such that it gives high recognition performance with reasonable amount of computation. Previously the performance of isolated spoken English and Hindi

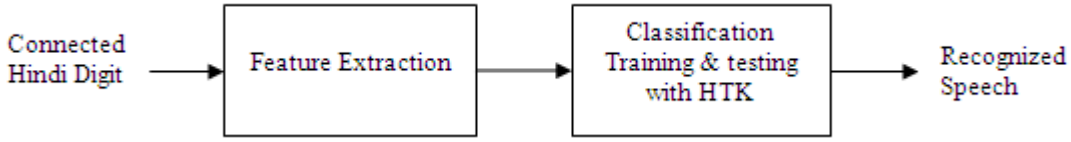


**Figure1. Speech Recognition System**

digits [3, 4] was evaluated using different feature extraction technique. The introduction of hidden Markov models and statistical language modeling techniques has greatly improved the performance of speech recognition systems in clean environments.

Nevertheless, speech recognition accuracy still degrades significantly in noisy environments. Many algorithms have been proposed to address this problem and they have demonstrated significant improvement in performance for stationary noise.

In this research the main work is to show that the recognition performance of connected Hindi digits using HTK 3.4 with Mel-frequency cepstral features and other feature extraction techniques like Perceptual linear prediction (PLP) [5], Revised PLP [6], BFCC [6] and Mel frequency PLP [7] .The experiments were conducted on clean as well as on noisy data for connected Hindi digit recognition.

The details of database preparation are given in next section. The details of feature extraction methods are given in section 3. Brief overview of recognizer model and HTK toolkit are provided in section 4 and section 5 respectively. Experimental results are explained in section 6. Finally the conclusions are drawn in section 7.

## 2. Database Preparation

A database of connected Hindi digits of forty speakers, twenty three females and seventeen males has been prepared by using cool edit software. Thirty six pre-determined sets of connected digits of length seven like 0098765 (shoonya shoonya nou aath saath che paanch), 5432101 (paanch chaar teen do ek shoonya ek) etc. were recorded from all the speakers. Database was prepared at sampling frequency of 16 kHz and 16 bits per sample. Every speaker was asked to utter the same thirty six sets of connected digits from the given list at normal speaking rate one after another with short pauses between sets. After recording, all thirty six sets were manually segmented and stored with a logical name. Speakers were chosen from different Indian states students studying at Birla Institute of Technology, Mesra, Ranchi, India. The age group of 18-26 years was chosen as students of different dialects in this age group were easily available. A distance of 4-6 inch was maintained between microphone and the speaker at the time of database recording. Two different microphones made by Sony and I-ball were used for recording the database in first and second phases respectively. Hindi pronunciations of digits and its corresponding English digits are shown in Table 1.

Artificial noisy database was prepared for all thirty six sets of connected Hindi digits by adding different types of noises from NOISEX-92 database [8] to clean Hindi digits database. To generate noisy speech, babble noise, white noise, pink noise, and F16 noise from this

**Table 1. Hindi Digits, English Digits and their Pronunciations**

| Hindi Digits | Hindi Pronunciations | English Digits | English Pronunciations |
|---|---|---|---|
| ० | Shoonya | 0 | Zero |
| १ | Ek | 1 | One |
| २ | Do | 2 | Two |
| ३ | Teen | 3 | Three |
| ४ | Chaar | 4 | Four |
| ५ | Paanch | 5 | Five |
| ६ | Che | 6 | Six |
| ७ | Saath | 7 | Seven |
| ८ | Aath | 8 | Eight |
| ९ | Nou | 9 | Nine |

database were artificially added to clean speech at different signal-to-noise ratio levels (SNRs) in the range 20dB to 5dB**.** Sampling frequency of different noises is also down sampled to 16 kHz to match with the sampling frequency of the clean speech samples.

## 3. Robust Feature Selection

The raw speech signal is complex and may not be suitable for feeding as input to the speaker recognition system; hence the need for a good front-end arises. The task of this front-end is to extract all relevant acoustic information in a compact form compatible with the acoustic models. All the feature extraction techniques are shown in Figure 2. The details of all the feature extraction techniques used in this paper are given below.

### 3.1. Mel Frequency Cepstral Coefficients (MFCC) and ΔMFCC

The Mel frequency cepstral co-efficient (MFCC) are perhaps the most widely used features in speech recognition today. Stevens and Volkmann [9] developed the Mel scale as a result of a study of the human auditory perception. The Mel scale was used by Mermelstein and Davis [10] to extract features from the speech signal for improving the recognition performance. The Mel scale is logarithmic scale resembling the way that the human ear perceives sound. Mel scale is given by the Equation 1.

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

Where *f* is the frequency. The natural logarithm is taken to transform into the cepstral domain and the discrete cosine transform (DCT) is finally applied to get the 24 MFCCs. The component due to the periodic excitation source may be removed from the signal by simply discarding the higher order coefficients. DCT de-correlates the features and arranges them in descending order of information, they contain about speech signal. Hence 13 coefficients out of 24 coefficients are used as MFCC features in our case.

The performance of a speech recognition system is enhanced by adding time derivatives to the static parameters. The first order derivatives are referred as delta features. Regression analysis is used to compute delta features. Δ MFCC [11] features are calculated by Equation 2.

$$\Delta\text{-Cep}(i) = \alpha \times \sum_{j=1}^{2} j \times (\text{Cep}(i+j) - \text{Cep}(i-j)) \quad (2)$$

Where Δ-Cep denotes delta features, Cep denotes the cepstra, and $\alpha \approx 0.2$ is used to scale these features. The index 'i' varies from 1 to 13 to get thirteen features. Since the regression technique needs past and future speech parameter values, suitable modifications are performed on beginning and end of the data stream.

### 3.2. Perceptual Features (PLP and MF-PLP)

PLP speech analysis method models the speech auditory spectrum of low order all pole model. The detailed procedure for PLP and MF-PLP extraction is given below.

1. Compute power spectrum of windowed speech.
2. Perform grouping to 24 critical bands in bark scale.
3. Perform loudness equalization and cube root compression to simulate the power law of hearing.
4. Perform IFFT.
5. Perform LP analysis by Levinson-Durbin procedure.
6. Convert LP coefficients to cepstral coefficients.

After these steps, all signal components are perceptually equally weighted to form the modified signal. In case of MF-PLP the Mel scale triangular filter bank taken from the MFCC algorithm is applied to the power spectrum instead of the magnitude spectrum.

### 3.3. Hybrid Features (BFCC & RPLP)

In this research two main blocks as shown in Figure 2 were interchanged to develop two hybrid feature extraction techniques. The interest is to see the influence of the spectral processing on different cepstral transformation. Figure 2 shows the steps of parameterization for basic method and beside PLP and MFCC, the way of computing the hybrid techniques has been shown by dashed arrow in Figure 2.

**3.3.1. Bark Frequency Cepstral Coefficient:** BFCC is the process where PLP processing of the spectra and cosine transform are combined to get the cepstral coefficients. Instead of using Mel filter bank, Bark filter bank has been applied and equal loudness pre-emphasis with intensity to loudness power law has been applied to the MFCC like features. Only first thirteen cepstral features of each windowed frame of speech utterances were taken.

**3.3.2. Revised Perceptual Linear Prediction:** In second approach instead of using bark filter bank, Mel filter bank has been applied to compute RPLP. The signal is pre-emphasized before segmentation and FFT spectrum is processed by Mel scale filter bank. The resulting

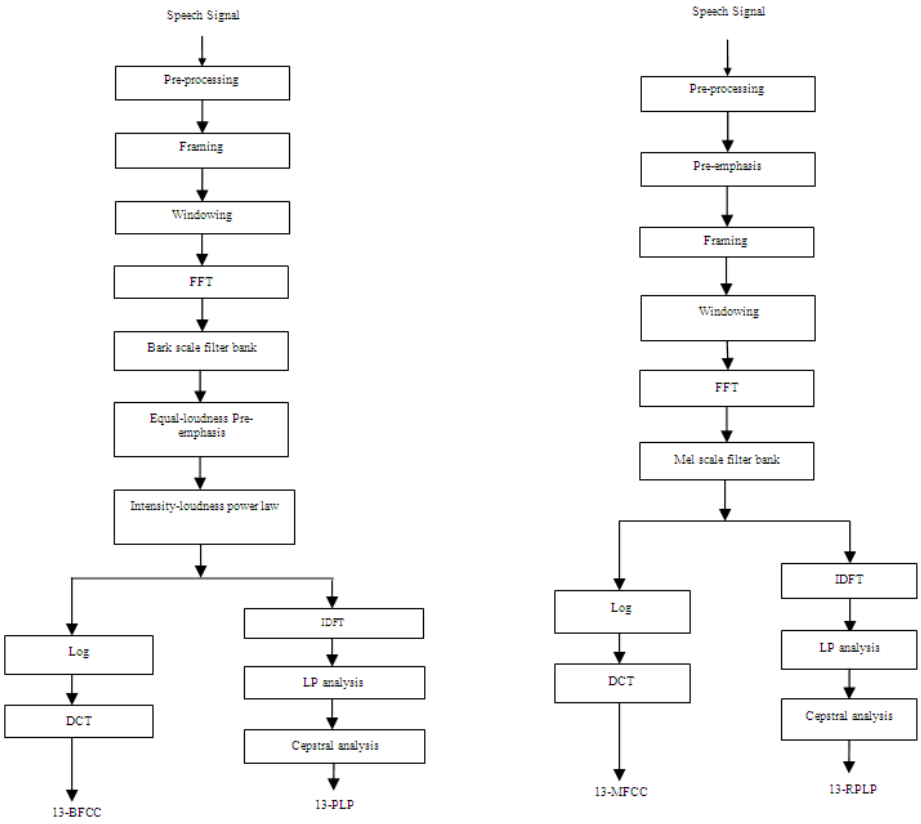


**Figure 2. BFCC and RPLP Features Extraction Methods**

spectrum is converted to the cepstral coefficients using LP analysis with prediction order of 13 followed by cepstral analysis.

## 4. Recognizer Model

To model the sequence of feature vectors, the recognizer uses a set of continuous density phone models. Each model is a three state first-order left-to-right continuous density hidden markov model with Gaussian mixture observation densities. The phone contexts to be modeled are automatically selected based on their frequencies in the training data. The models may be triphone models, left-context phone models, right-context phone models, or context-independent phone models. The covariance matrices of all the Gaussian components are diagonal.

Since phone duration is not accurately modeled with a three state Markov chain, a separate duration density is associated with each phone model. Duration is thus modeled with a gamma distribution per state. As proposed by Rabiner et al. [12], the HMM and duration parameters are estimated separately and combined in the recognition process during the Viterbi search. The main advantage of continuous density modeling over discrete or semi-continuous (or tied-mixture) observation density is the number of parameters used to model an HMM observation distribution, which can easily be adapted to the amount of available training data associated to this state. So as a consequence, high precision modeling can be

achieved for highly frequented states without the explicit need of smoothing techniques for the densities of less frequented states. Discrete and semi-continuous modeling use a fixed number of parameters to represent a given observation density and therefore cannot achieve high precision without the use of smoothing techniques. This problem can be alleviated by tying some states of the Markov models in order to have more training data to estimate each state distribution [13, 14]. This kind of tying requires careful design, some a priori assumptions, and results in a more complex training procedure. However, these techniques are of interest, particularly in situations where the training data is limited and cannot be increased easily.

## 5. Hidden Markov Model Toolkit

The choice of the HTK [17] as the recognition engine for the simulations was to get a standard benchmark for the performance of various feature extraction methods as well as for easy migration and reproduction of the simulations by other researchers. The HTK consists of three major blocks, feature extraction, HMM classifier [15] for training and the reorganization block. The feature extraction block supports various feature extraction techniques such as Linear Prediction Coefficients (LPC), Reflection Coefficients (RC), Mel Frequency Cepstral Coefficients (MFCC) and more. This can also estimate the dynamics of the features in time i.e. the Derivative and Acceleration. The simulation parameters used for feature extraction are given in Table 2. The HMM training block supports both continuous and discrete density modeling and uses the Baum-Welch algorithm to create the HMMs. The HMM was a simple Left-Right with no skip model, which was trained for each digit. The training tools in HTK (HInit and HRest) were used with their default settings. The speech data is parameterized using HCode and transcription label files are created using HLEd. HTK has an incremental build philosophy at the core of which is the HMM editor HHEd and the embedded re-estimation tool HERest. Starting with a very simple prototype system, the HMMs are repeatedly edited and re-estimated until the required level of model complexity and performance is reached.

The recognizer block calculates the likelihood of a given feature set to be generated from each HMM or HMM network using the Viterbi algorithm [10]. The simulations were performed for connected digit recognition for which testing between the 10 trained HMMs were performed using HVite tool.

**Table 2. Feature Extraction Parameters**

| HTK Parameter | Value |
|---|---|
| WINDOWSIZE | 25ms |
| TARGETRATE | 10ms |
| USEHAMMING | TRUE |
| PREEMCOEF | 0.97 |
| NUMCEPS | 13 |
| CEPLIFTER | 24 |

## 6. Experimental Results

The database of 40 speakers is divided into a training database with 35 persons and a testing database with 5 persons. In the testing database, there are three females and two males. The speech utterances were recorded in a quiet room.

Using NOISEX-92 database four different types of noise (babble, white, pink and F16) have been added to clean database to create an artificial noisy database at 5dB, 10 dB and 20dB SNR levels. Training and testing without language model was done in the following manner. To build digit recognition system at first acoustic models of digits [16] has been prepared. The proper dictionary was needed to develop good recognition system. At first using feature extraction block of HTK toolkit MFCC features have been extracted from the raw speech unit. The basic architecture of the recognizer must be set of observation sequences used to get the initial estimates of parameters. It uses Viterbi alignment for segmenting training observations and then pools the vectors in each segment to re-compute the parameters by counting the number of times each state is visited. During alignment process it starts to estimate the transition probabilities. HCompv tools forms the first stage of flat start training scheme and initializes the parameters such that global data variances and means are equal to the component variances and means. The next step was to re-estimates the optimal values of parameters of a single HMM like transition probabilities, mean and variance vectors of each observation function using the Baum-Welch algorithm. Re-estimation is done several times till measures do not change and a convergence is reached.

### 6.1 Effect of Using Different Features for Clean Database

Since the HTK doesn't support the features like PLP, RPLP, BFCC, MF-PLP etc so these features have been first extracted from the clean database using Matlab programs. The features extracted from clean database have been converted to HTK format using VoiceBox toolbox for Matlab. Figure 3 presents the results from series of recognition experiments to determine the effect of different features. In all the experiments, 5-states HMM with 9-Gaussian mixtures were used.

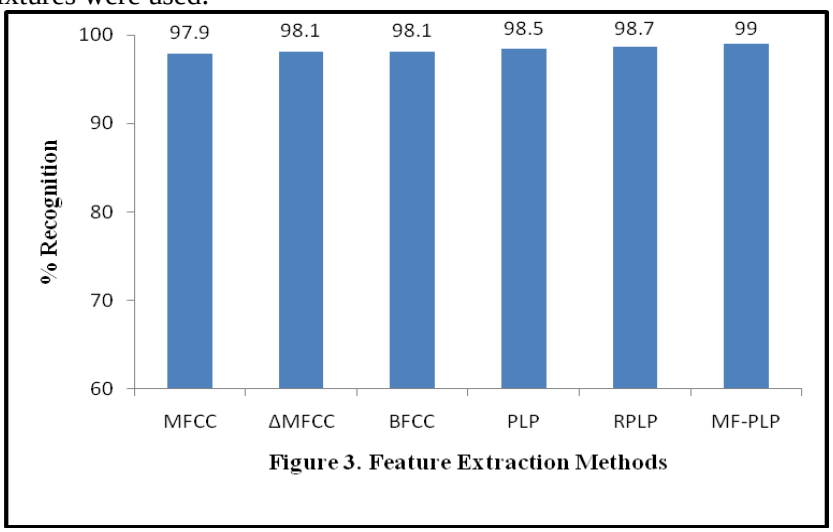


**Figure 3. Feature Extraction Methods**

As observed from results that inserting the Mel filter in the perceptual feature extraction results in the best recognition efficiency. Evaluating temporal coefficients also increases the recognition efficiency. So we observe that ΔMFCC gives better efficiency than MFCC and

MF-PLP gives best recognition efficiency among all. Bark-frequency scaling, equal loudness pre- emphasis, intensity-loudness power law and DCT seem to have reasonable influence on the recognition accuracy. Testing was done with thirty sets of connected Hindi digits of length seven of five speakers. Hence a total of 5×30×7=1050 digits are tested. The recognition efficiency has been calculated on the basis of total number of times a correct digit has been recognized in a connected digit sets.

### 6.2 Effect of using different features for noisy database

The features extracted from noisy database have been converted to HTK features using VoiceBox toolbox for Matlab. Table 3 presents the results from the series of recognition experiments to determine the effect of different noises on different features. In all the experiments, we use the constant characteristics, which are 5-states HMM with 9-Gaussian mixtures. Experiments are performed only for 5dB, 10dB and 20dB SNR levels. It has been observed that system doesn't work properly for SNR level below 5 dB and it gives almost same result as clean database for SNR level more than 20 dB. While on average most of the times MF-PLP features along with Pre-emphasis, Mel–frequency scaling and LP based analysis have shown best recognition result. From results it has also been observed that at low SNR most of the times modified features perform better than conventional features. For all kinds of feature extraction techniques there was considerable decrease in the recognition efficiency for SNR below 10dB. The impact of the babble and white noise addition shows the importance for the feature extraction methods to have this kind of robustness in HMM-based recognition. The recognition efficiency has been calculated on the basis of total number of times a correct digit has been recognized in a connected digit string. Recognition efficiency in presence of babble, white, pink and F16 noise are shown in Figure 4, Figure 5, Figure 6, and Figure 7 respectively.

**Table 3. Recognition Rate for Noisy Connected Hindi Digits**

| NOISE | SNR LEVEL / FEATURES | % RECOGITION RATE | | | | | |
|---|---|---|---|---|---|---|---|
| | | MFCC | ΔMFCC | BFCC | PLP | RPLP | MF-PLP |
| Clean | Infinite | 97.9 | 98.1 | 98.1 | 98.5 | 98.7 | 99.0 |
| BABBLE | 5 dB | 70.48 | 73.71 | 73.33 | 77.14 | 78.57 | 79.52 |
| | 10 dB | 85.24 | 87.14 | 88.10 | 89.05 | 90.00 | 90.48 |
| | 20 dB | 96.67 | 97.62 | 97.62 | 98.10 | 98.57 | 98.57 |
| WHITE | 5 dB | 64.76 | 70.00 | 73.33 | 76.67 | 77.62 | 77.62 |
| | 10 dB | 87.14 | 90.48 | 93.33 | 92.38 | 94.29 | 94.76 |
| | 20 dB | 96.19 | 97.14 | 96.67 | 97.62 | 98.10 | 98.57 |
| PINK | 5 dB | 61.90 | 65.24 | 66.67 | 70.48 | 72.38 | 73.81 |
| | 10 dB | 90.00 | 91.43 | 91.43 | 92.86 | 93.81 | 94.76 |
| | 20 dB | 96.67 | 97.62 | 97.62 | 98.10 | 98.57 | 98.57 |
| F16 | 5 dB | 40.00 | 43.81 | 51.90 | 49.05 | 54.29 | 53.33 |
| | 10 dB | 85.24 | 87.14 | 88.10 | 89.52 | 90.00 | 91.43 |
| | 20 dB | 94.76 | 95.24 | 96.19 | 96.67 | 97.14 | 98.10 |

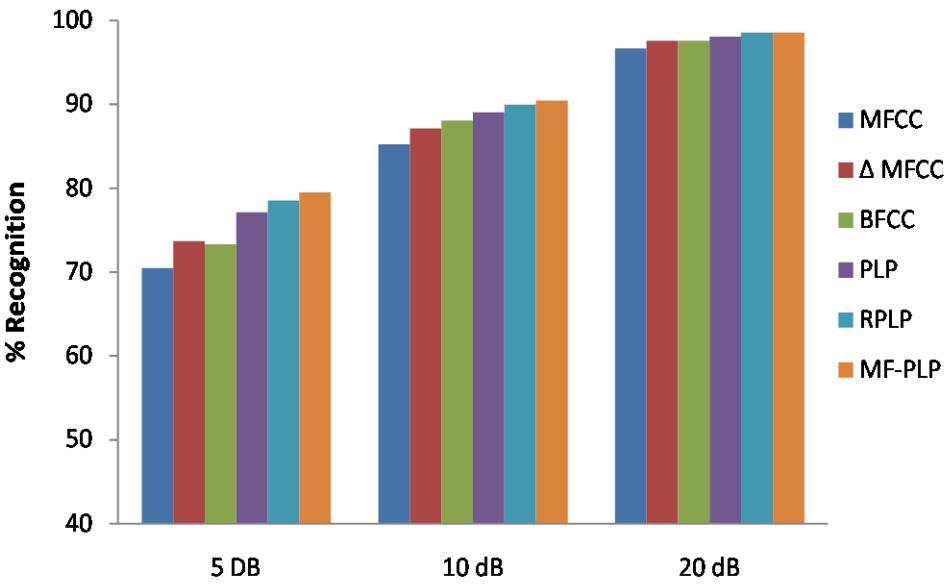


**Figure 4. Comparative performance of feature extraction techniques with Babble Noise**

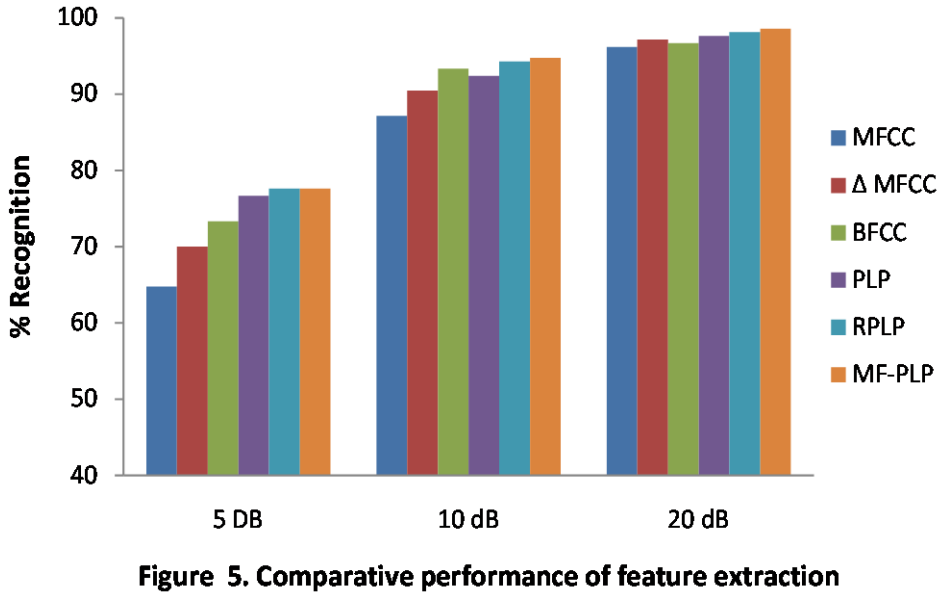


**Figure 5. Comparative performance of feature extraction techniques with White Noise**

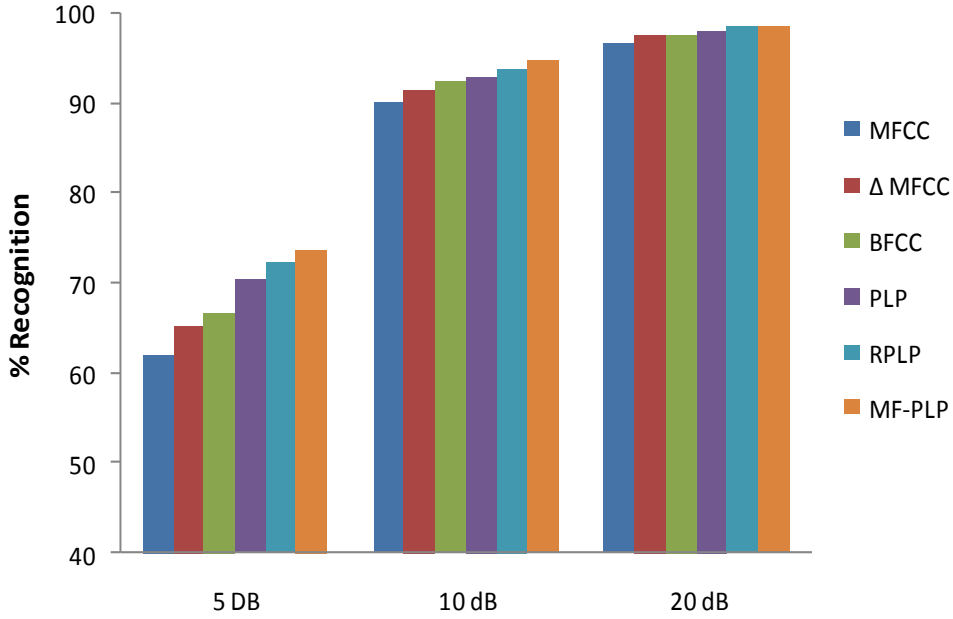


**Figure 6. Comparative performance of feature extraction techniques with Pink Noise**

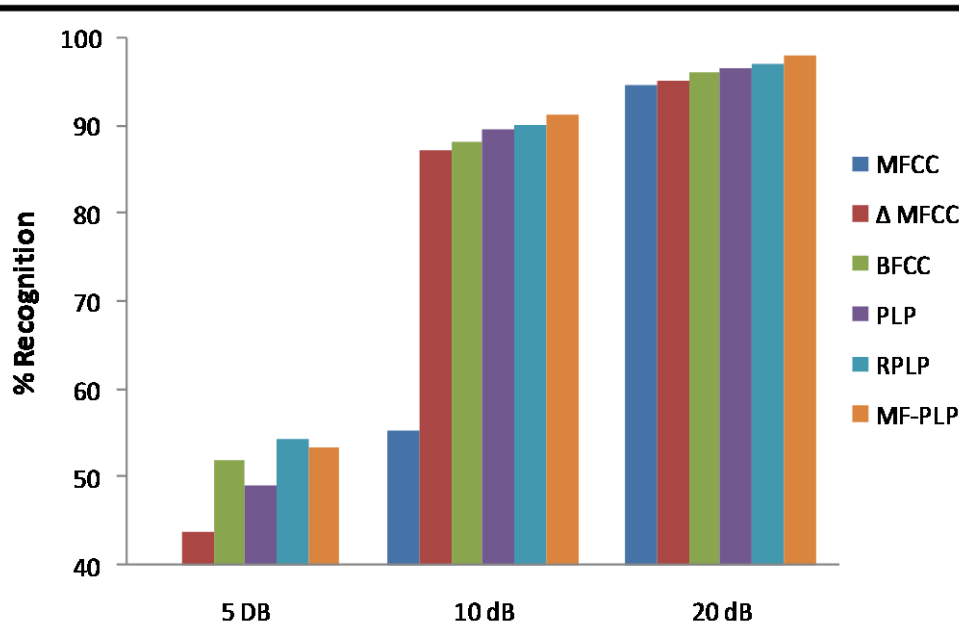


**Figure 7. Comparative performance of feature extraction techniques with F16 Noise**

## 7. Conclusion

This paper has illustrated the speaker-independent connected Hindi digits recognition using HMM. For training and testing HTK is used. The choice of the HTK as the recognition engine for the simulations was to get a standard benchmark for the performance of various feature extraction methods and for easy migration and reproduction of the simulations by other researchers.

From all the experiments, it was concluded that MF-PLP has shown best recognition performance compared to other feature extraction techniques because it incorporates Mel filter into a perceptual linear feature extraction method. PLP features have also shown improvement in recognition performance as compared to MFCC. PLP features performed better because the signal was pre-emphasized by a simulated equal-loudness curve to match the frequency magnitude response of the ear as well as all signal components were perceptually equally weighted. RPLP features have also shown good results for clean as well as noisy data. This is due to the fact that it takes advantage of pre-emphasis filter, Mel scale filter bank along with linear prediction and cepstral analysis.

Future work will be directed towards investigation of low SNR Hindi digits recognition (0dB & 5dB), by taking more contexts into consideration during the feature extraction and optimizing the primary time-frequency analysis.

## References

[1] K Jain, P. Flynn, and A.A. Ross, "Handbook of Biometrics", *Springer,* 2008.

[2] A.O. Afolabi, A. Williams, and O. Dotun, "Development of a text dependent speaker identification security system", *Research Journal of Applied Sciences*, 2 (6), pp. 677-684, 2007.

[3] K. Samudravijaya, Barot & Maria, "A Comparison of Public-Domain Software Tools for Speech Recognition", *In WSLP*, pp.125-131, 2003.

[4] K. Vertanen, "Baseline WJS acoustic models for HTK and Sphinx: Training recipes and recognition experiments", *http://www.inference.phy.cam.ac.uk/kv227/papers*, 2007.

[5] A. Revathi and Y. Venkaramani , "Source and system Features for text independent Speaker Recognition", *Springer-verlog Berling Heidelberg,* pp.21-30 ,2010.

[6] R.Josef and P. Pollak , "Modified Feature Extraction Methods in Robust Speech Recognition", *Radioelektronika, 17th IEEE International Conference,* pp.1-4, (2007).

[7] Andra´s Zolnay , Daniil Kocharov , Ralf Schlüter and Hermann Ney, "Using multiple acoustic feature sets for speech recognition", *Science direct, Speech Communication 49* , pp. 514–525, 2007.

[8] A. Varga, H.J.M. Steeneken & D. Jones, "The noisex-92 study on the effect of additive noise on automatic speech recognition system", *Reports of NATO Research Study Group (RSG.10)*, 1992.

[9] S. S. Stevens and J. Volkman, "The relation of pitch to frequency", *American Journal of Psychology*, vol. 53(3), pp. 329-353, 1940.

[10] P. Mermelstein and S. B. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 357-366, 1980.

[11] Qu Dan, Wang Bingxi, WelXin, "Language identification using vector quantization", *In Proc. IEEE Int. conf. speech processing*, June.2002, pp. 492-495.

[12] L. R. Rabiner, B. H. Juang, S. E. Levinson & M.M. Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", *AT&T Tech. J.*, 64(6), 1985.

[13] S.J. Young, P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *EUROSPEECH-*.

[14] M. Y. Hwang, X. Huang, "Subphonetic Modeling with Markov States - Senone", *ICASSP-92*.

[15] Ma Guangguang, Zhou Wenli, Jing Zheng, You Xiaomei & Ye Weiping, "A Comparison between HTK and SPHINX on Chinese Mandarin", IEEE *International Joint Conference on Artificial Intelligence,* pp. 394-397, 2009.

[16] S. Young, D. Ollason, V. Valtchev & P. Woodland, "HTK Book (for HTK Version 3.4)*", Cambridge: Cambridge University Engineering Department*, 2006.

[17] http://htk.eng.cam.ac.uk.

# Authors

**Mr. A. N. Mishra** has received his B.Tech from Gulbarga University, Gulbarga- India in 2000 and M.Tech from Uttar Pradesh Technical University, Lucknow (UP)-India in 2006. Presently he is pursuing Ph.D. from Birla Institute of Technology, Mesra (Jharkhand)-India. He has published more than 7 research papers in the area of Speech, Signal and Image Processing at National/International level. His areas of interest are Speech, Signal and Image Processing.

**Dr. Mahesh Chandra** received B.Sc. from Agra University, Agra(U.P.)-India in 1990 and A.M.I.E. from I.E.I., Kolkatta( W.B.)-India in winter 1994. He received M.Tech. from J.N.T.U., Hyderabad-India in 2000 and Ph.D. from AMU, Aligarh (U.P.)-India in 2008. He has worked as Reader & HOD in the Department of Electronics & Communication Engg. at S.R.M.S. College of Engineering and Technology, Bareilly (U.P.)-India from Jan 2000 to June 2005. Since July 2005, he is working as Reader in the Electronics & Communication Engg. Department, B.I.T., Mesra, Ranchi (Jharkhand)-India. He is a Life Member of ISTE, New Delhi-India and Member of IEI Kolkata (W.B.)-India. He has published more than 23 research papers in the area of Speech, Signal and Image Processing at National/ International level. He is currently guiding four Ph.D. students in the area of Speech, Signal and Image Processing. His areas of interest are Speech, Signal and Image Processing.

**Mr. Astik Biswas** has received his B.Tech in 2008 from West Bengal University of Technology, Kolkata, India. He has received is ME in 2010 from Birla Institute of Technology, Mesra (Jharkhand)-India in the field of speech Recognition. His areas of interest are Speech and Signal Processing,video pocessing and Digital Electronics.

**Dr. S. N.Sharan** has received his Ph.D. from IIT, Delhi, India. He has published more than 30 research papers in the area of Speech, Signal and Image Processing at National/ International level. He has worked as Prof. at BITS, Pilani, India. Presently he is working as Director GNIT, Gr. Noida, UP, India. He is currently guiding two Ph.D. students in the area of Speech, Signal and Image Processing. His areas of interest are Energy, Signal and Image Processing.