

Performance Analysis of Hindi Voice for Speaker Recognition and Verification Using Different Feature Extraction

**Shrikant Upadhyay¹, Sudhir Kumar Sharma²,
Pawan Kumar³ & Aditi Upadhyay⁴**

^{1,3}Cambridge Institute of Technology/ Department of Electronics & Communication Engineering, Ranchi, Jharkhand India.

^{2,4}Jaipur National University/ Department of Electronics & Communication Engineering, Jaipur, Rajasthan, India.

Abstract

Hindi is the most common language in India and being spoken by about 80% and understanding rate is about 93% by Indians. Different region has many common languages and they pronounce according to their culture environment. Pronunciations may be different but the meaning of the word being spoken is same. So, we try through this paper with the help of different Hindi voice database taken from different part of India and focus to analyze the performance of voice using different feature extraction techniques. This will to identify the proper feature extraction method according to their voices and analyze the performance of these feature extraction techniques. The efficiency of these feature extraction techniques will also help to analyze recognize and verification of the speaker performance.

Keywords: Hindi database, Feature extraction technique, speaker recognition & verification.

INTRODUCTION

The natural and fundamental way of communication is speech for humans. Every human voice has various attributes to communicate the information such as emotion, gender, attitude, health and identity. The aim of Speaker Recognition (SR) by machine is the task of recognizing a person automatically based on the information received from his/her speech signal (Kumar et al. 2011; Shaughnessy 1986; Chauhan et al. 2013). No two individual's sound are identical because their vocal tract shapes, larynx sizes and other parts of their voice production organs are different and each voice has its own characteristic manner of speaking namely rhythm, intonation style, pronunciation, vocabulary etc., (Gbadamosi 2013; Kinnunen and Li 2010). There is a variation between speech recognition (what is being said) and speaker recognition (who is speaking). sr is categorized into speaker identification and speaker verification (Sumithra et al. 2011) [1].

Applications like Speech and Speaker Recognition efficient feature extraction techniques from speech signal where most of the voiced part contains Speech or Speaker specific attributes [2]. From security point of view speech can also be used to protect your asset and privacy. It can help in other commercial application like account access from remote location or help any physical challenges users also. The broad area of speech is quite broad and it has many dimensions but we have only focus on the speech and the language is taken in consideration in terms of Hindi.

SPEAKER IDENTIFICATION & VERIFICATION

Speaker recognition is the general term used to include different ways of discriminating people based on their voices. The main categories are: speaker identification system and speaker verification system.

A. *Speaker Identification*

The objective of speaker identification is to classify an unlabeled utterance belonging to one of the N reference speakers [3]. It can be closed set identification or open set identification shown in Figure1. The objective of speaker identification is to decide the identity of speaker based on the speaker's voice, from set of N speakers i.e., one-to-many matching.

B. *Speaker Verification*

The objective of speaker verification is to accept or reject the identity claim of speaker [4]. If the match between test and reference is above threshold level, the claim is accepted shown in Figure1. Speaker verification is an open set problem. Speaker recognition systems can be further classified as text-dependent and text-independent systems.

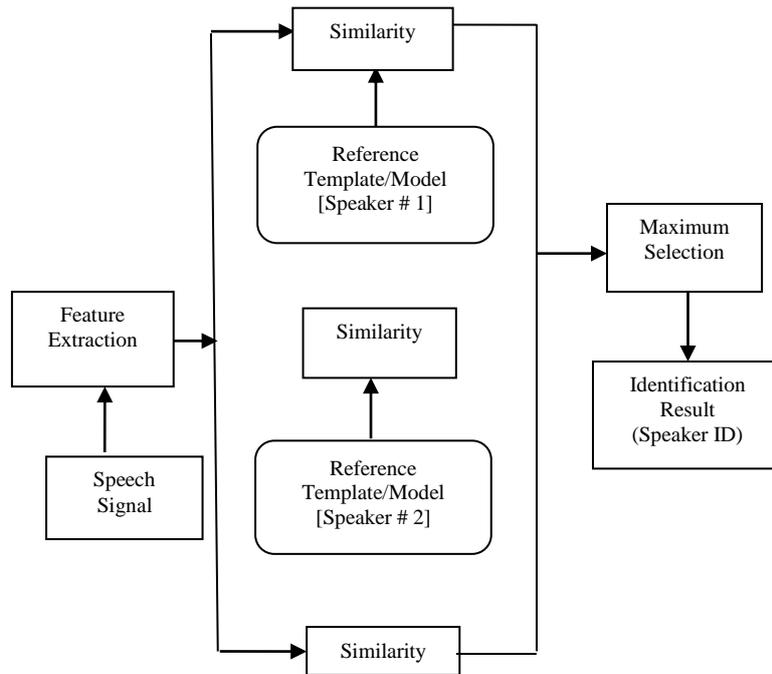


Figure 1. Structure of speaker verification system.

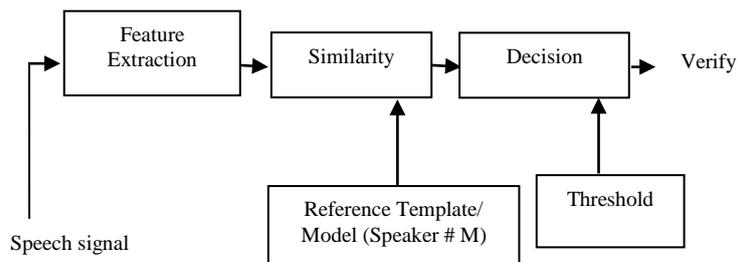


Figure 2. Structure of speaker identification system.

FEATURE EXTRACTION TECHNIQUES

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:

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

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

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

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

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.

B. Linear Predictive Coding Analysis (LPC)

It is a frame based analysis of the speech signal performed to provide observation vectors [5]. The relation between speech sample S (n) and excitation X(n) for auto regressive model (system assume all pole mode) is explained mathematically as:

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

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:

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

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 [6].

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

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 $\partial E_n / \partial \alpha_i = 0$ & $i = 0, 1, 2, \dots, p$ thus getting the equation:

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

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

Thus, equation (13) rewritten as:

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

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 [6].

Speech segment is windowed in autocorrelation method as discuss below:

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

Where, w (m) is finite window length

Then, we have

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

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

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 [7]

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

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

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

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

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

$$k_i = \text{PACOR coefficients}$$

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 [8]:

$$C_0 = \ln G \quad (21)$$

$$C_m = \alpha_m + \sum_{k=1}^{m-1} \binom{k}{m} C_k a_{m-k} \text{ for } 1 \leq m \leq p \quad (22)$$

$$C_m = \sum_{k=1}^{m-1} \binom{k}{m} C_k a_{m-k} \text{ for } m > p \quad (23)$$

C. Linear Prediction Cepstral 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 (2) below:

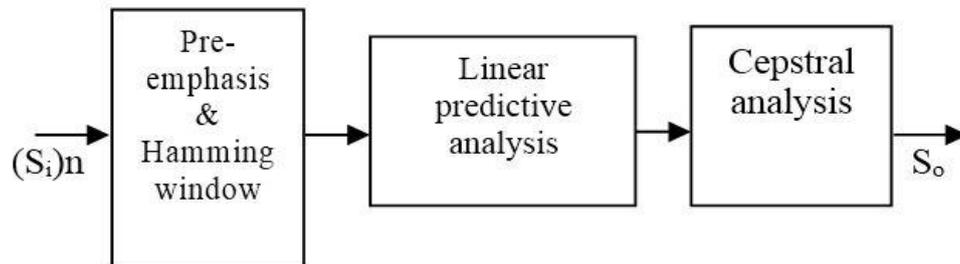


Figure 3. Steps involved in LPCC algorithm processing.

A digital all-pole filter is used to model the vocal tract and has a transfer function represented in z-domain as:

$$V(z) = \frac{G}{1 - \sum_{k=1}^p a_k z^k} \quad (24)$$

where, $V(z)$ is the vocal tract transfer function, G is the gain of the filter, a_k is the set of auto regression coefficients known as linear prediction coefficients (LPC) and p is the order of all-pole filter. One of the efficient method for estimating the LPC coefficients and filter gain is autocorrelation [8]. The inverse FFT transform of the logarithm of the speech magnitude spectrum and it is defined as:

$$\hat{s}[n] = \frac{1}{2\pi} \int_{-\pi}^{+\pi} \ln[s(w)] e^{jwn} dw \quad (25)$$

RELATED WORK

This paper completely based on analyzing the best possible solution in signal processing domain by detecting and identifying the performance of above mention three feature extractions techniques while training and testing the real time voice Hindi database sample and try to judges the efficient one which suites to be the best and suitable for real time applications. Acoustic condition is one the major issues/challenges that must be taken into consideration. Voice samples taken from the three different region which includes Punjabi, Marathi and Oriya. These sample where taken using Cool and Gold Wave software and microphone having sampling rate 8000Hz, channel type mono with resolution 16bit.

PARAMETERS & VALUES

TABLE I: PARAMETERS CONSIDERED FOR ANALYSIS

<i>Feature Extraction</i>	<i>Frequency</i>	<i>Language</i>	<i>Condition</i>
MFCC	50KHz	Punjabi	Acoustic
LPC	100KHz	Marathi	Acoustic
LPCC	150KHz	Oriya	Acoustic

TABLE II: VARIATION IN EXTRACTION TECHNIQUE

<i>Feature Extraction</i>	<i>Variations in %</i>	<i>Language</i>	<i>Condition</i>
MFCC	19.66	Punjabi	Acoustic
LPC	14.23	Marathi	Acoustic
LPCC	18.51	Oriya	Acoustic

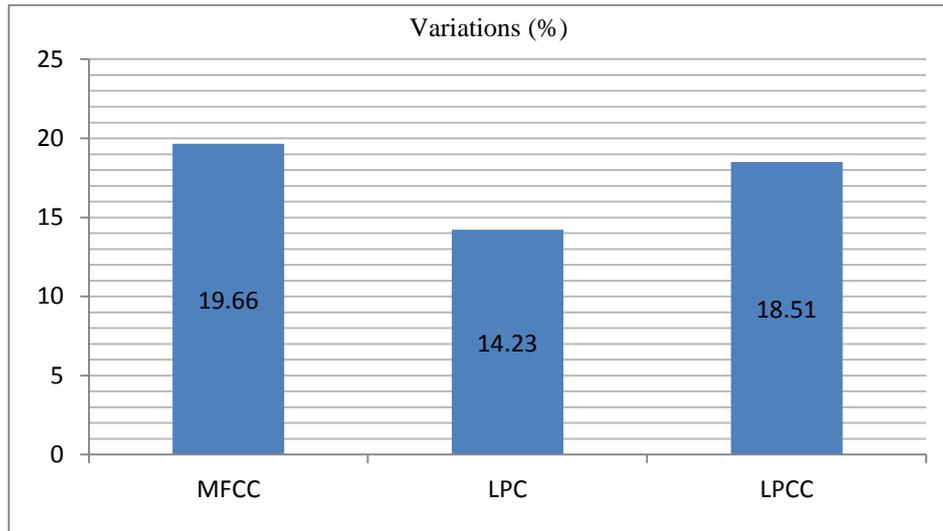


Figure 4. Variation of feature extraction techniques for different above mention languages

DATA COLLECTION FOR HINDI VOICE

The data has been collected using the cool software and the snap short is shown in Figure 5. Here each sample is taken and each frame is collected by just copying the each frame using .wav file in separate folder and using MATLAB each frame is analyzed. The error rate and efficiency is calculated for the data collected. This calculation will help to judge the efficiency of the above mention technique accuracy and helps for optimization. We try to take the sample in noise free environment but it cannot to completely eliminated so, try to keep as minimum as possible. Acoustic condition may vary due to external or internal changes in the environment. So, that the result may not be much affected.

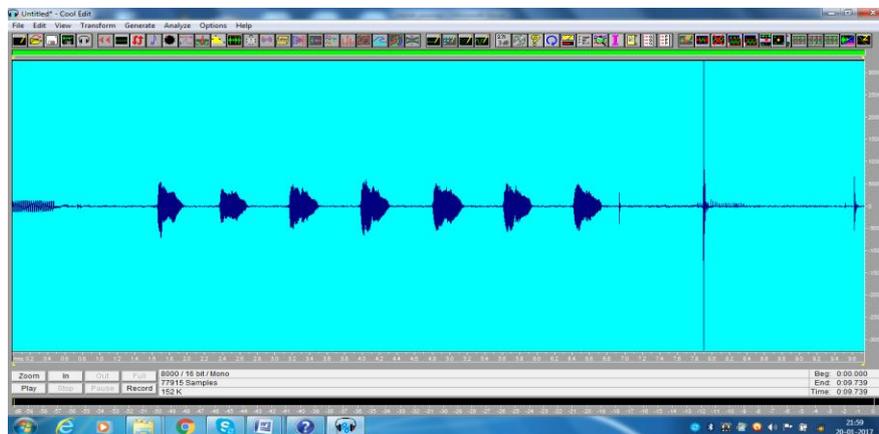


Figure 5. Snapshot taken for voice recording using cool software

A. Dialects

The Hindi digit (sample) where collected for SHUNAY to DUS for different age group with different category as shown in table III.

TABLE III: Dialects

Hindi Digits	Age	Category
SHUNYA	19	Male
EK	26	Male
DO	18	Male
TEEN	20	Female
CHAAR	22	Female
PAANCH	23	Male
CHHEH	25	Male
SAAT	21	Male
AATH	24	Female
NAO	27	Male
DUS	28	Female

B. Parameters

Total number of speaker = 24

Language = Hindi Language (Phonetics is rich)

Single channel = 16 KHz

Sampling rate = 16 bit (For less Quantization error)

C. Vector Quantization Calculation

Vector quantization is important as it helps to get the exact reference values as it is difficult to extract the each feature of voice sample so, to remove the computational burden we take a reference value for each sample. Table IV shows how the reference

value is calculated for the twenty four speaker considered for different Hindi sample. The Hindi sample is spoken to the Punjabi, Marathi and Oriya people of different age group and different categories.

TABLE IV. Reference value for 24 speakers

Speaker Number	Vector Quantization	Reference Value
SPK0	1-10	00
SPK1	11-20	01
SPK2	21-30	02
SPK3	31-40	03
SPK4	41-50	04
SPK5	51-60	05
SPK6	61-70	06
SPK7	71-80	07
SPK8	81-90	08
SPK9	91-100	09
SPK10	101-110	10
SPK11	111-120	11
SPK12	121-130	12
SPK13	131-140	13
SPK14	141-150	14
SPK15	151-160	15
SPK16	161-170	16
SPK17	171-180	17

SPK18	181-190	18
SPK19	191-200	19
SPK20	201-210	20
SPK21	211-220	21
SPK22	221-230	22
SPK23	231-240	23

EEROR RATE AND EFFECIENCY

The error rate and efficiency of each feature extraction will help to analyze the performance. The error rate and efficiency of each technique is calculated using MATLAB version 2015b and it shown in Table V.

TABLE V: Error rate & efficiency

Feature Extraction	Error Rate (%)	Efficiency (%)
MFCC	0.20	98.87
LPC	0.266	96.34
LPCC	0.18	97.76

ANALYSIS RESULTS

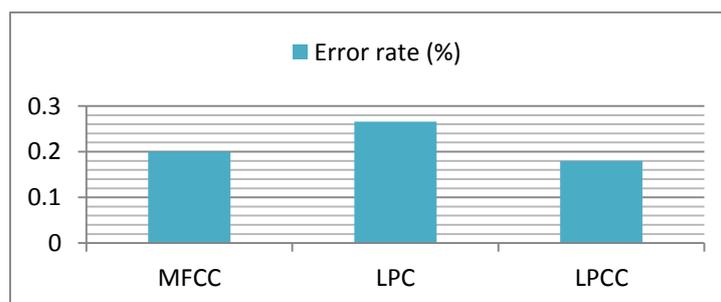


Figure 6. Error rate

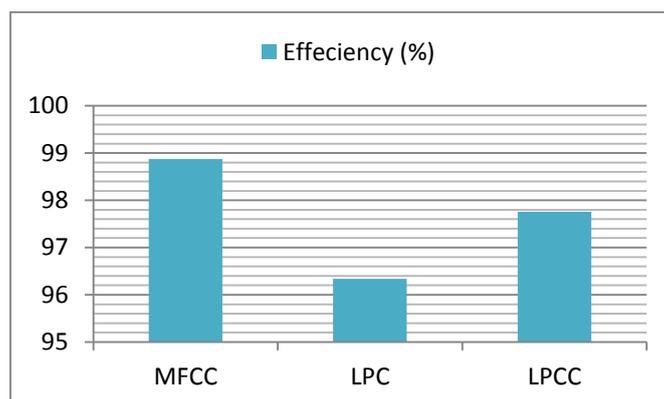


Figure 7. Efficiency rate

CONCLUSION

The analysis result for the above mention feature technique shows that the MFCC shows good performance compared to LPC and LPCC feature extraction technique so, MFCC proves to be efficient. This gives proper idea of feature extraction technique for different speech signal. The error rate of LPC is very high compared to LPCC and MFCC. So, we can go for either LPCC or MFCC for low error rate and may this result be useful for further. This result is useful for acoustic condition also but the more challenging would be the nature of the condition or noise.

ACKNOWLEDGMENT

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

REFERNCES

- [1] R. Visalakshi & P. Dhanalakshmi, "Performance of speaker identification using CSM and TM", PP. 457-465, International Journal of Speech Technology, March 2016.
- [2] V. GUPTA, A. K. Sharma & A. Singh, "Database Development and Comparative Analysis of Isolated, Spoken Hindi Hybrid Paired Words & Spoken Hindi Partially Reduplicated Words Using Endpoint Detection", Conference on Advances in Communication and Control Systems, pp. 485-492, 2013.
- [3] H. Gish and M. Schmidt, "Text-dependent speaker identification," IEEE Signal Processing Magazine, pp. 18-32, October (1994).

- S.Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Trans. Acoust., Speech, Signal Processing, vol. 29, pp. 254-272, April 1984.
- [4] 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.
- 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 Edition) vol.69, pp. 34-39 ,2008.

ABOUT AUTHOR (S):



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, advance digital signal 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.



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 noise model.

Dr. Pawan Kumar is a Professor at the Department of Electronics & Communication Engineering, Cambridge Institute of Technology, Ranchi India. Professor Pawan

received his Ph.D. degree in Electronics from BIT, Mesra. He has an extensive teaching experience of more than 12 years. His research area includes , signal and speech processing.