Comparative Analysis of LPCC, MFCC and BFCC for the Recognition of Hindi Words using Artificial Neural Networks

Taabish Gulzar
M Tech student
Department of Electronics and Communication Engineering,
D.I.T, Dehradun, India

Anand Singh
Assistant Professor
Department of Electronics and Communication Engineering,
D.I.T, Dehradun, India

Sandeep Sharma, Ph.D
Professor and HOD
Department of Electronics and Communication Engineering,,
D.I.T, Dehradun, India

ABSTRACT
Most important way of communication among humans is language and primary medium used for the said is speech. The speech recognizers make use of a parametric form of a signal to obtain the most important distinguishable features of speech signal for recognition purpose. In this paper, Linear Prediction Cepstral Coefficient (LPCC), Mel Frequency Cepstral Coefficient (MFCC) and Bark frequency Cepstral coefficient (BFCC) feature extraction techniques for recognition of Hindi Isolated, Paired and Hybrid words have been studied and the corresponding recognition rates are compared. Artifical Neural Network is used as back end processor. The experimental results show that the better recognition rate is obtained for MFCC as compared to LPCC and BFCC for all the three types of words.

General terms
Speech Recognition, Recognition rate, Bit-wise, Zero crossing rate.

Keywords
Hindi Hybrid words, Spoken Paired words, Feature Extraction, Artifical Neural Networks.

1. INTRODUCTION
The need for most high-flying speech recognition systems and spoken language systems to be robust with respect to their acoustical environment had a dramatic impact in current era. Listeners outperform Automatic speech recognition (ASR) systems in each and every speech recognition task. Modern high-tech automatic speech recognition systems perform very well in environments, where the speech signals are reasonably clean. Currently there has been a growing body of research in extending various speech recognition tasks. A complex relationship is observed between physical speech signal and the corresponding words and is very hard to understand [1]. The Very known applications of the said systems include physical access entry and where remote identity verification is necessary [2]. However, the emergence of classy technologies in different areas of ASR systems makes the safe operation of these systems certain. However, some areas of ASR systems opposes the same status in terms of possessing proficient techniques or refined methods for solving many difficulties within the domain. In most of the cases recognition by machines degrades dramatically with slight adjustment in speech signals or speaking environment, thus complex algorithms are used to represent this unpredictability [3]. The complicated speech processing task has been divided into three relatively simpler classes (a) Speech recognition: that permits the machines to understand the words, sentences, phrases spoken by different speakers, (b) Natural language processing: this lets the system to understand the needs of different speakers (c) Speech synthesis: here the machines respond to the needs of users [4], [5].

Every time a speaker speaks, the linguistic content, speaker characteristics i.e. (length of the vocal tract, emotions, gender and origin), speaking rate and acoustic environment all together influence the acoustics of the overall spoken productivity [6]. Speech signal not only contains the meaning of an utterance but also the emotions within it, which plays a vital role in speech recognition [7]. For developing a robust ASR system, it is essential to analyse the emotions contained within the speech signal [8]. The detection of emotions has been the area under discussion of several recent studies. Currently, several general classes of emotion are defined and correlated with measurable characteristics of speech [9], [10]. As far as misrecognition rate, processing time, memory allocation is concerned Spoken paired words have advantage over short and long words as they have medium word length and make use of less computation time and memory allocation [11]. In case of paired words there is always a gap known as speech code in between the two words between two words and it is this speech code that plays a significant role in recognition process [12]. In order to fortify this speech two different languages are used. Thus, a new class of words known as Spoken Hindi Hybrid Paired words came into existence. The concept of Hindi Hybrid words lies in the fact that one word is necessary from Hindi origin and the other may be from Urdu, English or from any other language.

The rest of the paper is outlined as: section 2 briefly describes the database creation details, section 3 explains the feature extraction methods i.e. Linear prediction cepstral coefficient (LPCC) and Mel frequency cepstral coefficient (MFCC). In section 4, classification model i.e. Artifical neural network is described. Section 5 deals with the experimental results and finally conclusion is presented in section 6.

2. DATABASE DEVELOPMENT DETAILS
The first and the foremost step towards the ASR systems is generation of database. In this paper three types of words

namely Isolated words, Spoken paired words and spoken Hindi Hybrid paired words using five different emotions namely normal, happy, anger, surprise, sad are being analysed. Five words are taken from each class of words for analysis, thus making a total of fifteen words. Five male and five female speakers are taken for recording session from different areas. Each of the word is uttered by the individual speaker five times thus making a total of (15×5×10) 750 utterances which are enough for obtaining better results. Recordings are done in room environment making use of a stereo headset with microphone H250 with frequency response of 20 Hz-20 KHz , channel selected as mono and with a feature of noise cancelling at a sampling rate of 16000 Hz using MATLAB 7.14.0.739 and are stored in a .wav format. . A clear description of the database development is shown in tables 1

Table 1. Database details

S.No	Isolated words	Paired words	Hybrid words
1	EK	HUM-TUM (HT)	KALA-SAYA (KS)
2	DO	YANHA-WANHA (YW)	KHAS-AADMI (KA)
3	TEEN	DIN-RAAT (DR)	BADIYA-ITEM (BI)
4	CHAR	JINA-MARNA (JM)	BADA-EHSAAN (BE)
5	PAANCH	KHANA-PINA (KP)	PURANI-JEANS (PJ)

3. FEATURE EXTRACTION

In order to achieve higher accuracy in speech recognition, selecting appropriate features from a speech signal is the central concern. Feature extraction process is grounded on the basis of discarding the irrelevant information from the speech signal and only keeping the useful content. As the raw speech signal is always complex thus feeding the said as an input to the classifier may not be suitable, hence the requirement for a high-quality front-end arises. The primary aim of feature extraction is to find a set of properties of an utterance that have acoustic correlations to the speech-signal, that is parameters that in some way can be computed or estimated through processing of the signal waveform. Such parameters are termed as features. There are many properties of features which includes high discrimination between sub-word classes, low speaker variability, invariableness to degradations in the speech signal due to noise and channel [13]. The feature extraction methods considered in this paper are Linear Prediction cepstral coefficients (LPCC), Mel frequency cepstral coefficients (MFCC) and Bark Frequency cepstral coefficients (BFCC). A concise explanation of each of the feature extraction method is given below.

3.1 Linear Prediction Cepstral Coefficients

For estimating the basic parameters of a speech signal, LPCC has become one of the predominant techniques. . The basic theme behind this method is that one speech sample at the current time can be predicted as a linear combination of past speech samples. Algorithm for LPCC is shown in Figure 1.

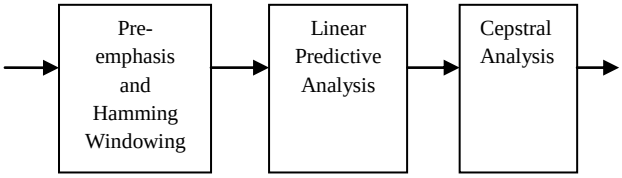

Figure 1. Steps for LPCC.

The input signal is first pre-emphasized using a first order high pass filter. Since the energy contained within a speech signal is distributed more in the lower frequencies than in the higher frequencies. In order to boost up the energies contained within high frequencies, Pre-emphasis of the signal is done. The transfer function for this filter in z-domain is expressed as:

$$H_p(z) = 1 - az^{-1} \quad (1)$$

Where, a (filter coefficient) is a constant with a typical value of 0.97. The pre- emphasized signal is blocked into frames. In order to reduce the signal discontinuities at the edges of eachframe, windowing of the signal is performed. Most commonly window used is hamming window because of its smoothness in low-pass and very low side lobe [14] and is described in equation below.

$$w(n) = 0.54 - 0.46\cos\left(2\pi\frac{\mathrm{n}}{\mathrm{N}}\right); \quad 0 \leq \mathrm{n} \leq \mathrm{N} \quad (2)$$

Where N is the length of the windowing function. Linear predictive analysis is grounded on the hypothesis, that the shape of the vocal tract decides the character of the sound being produced. A digital all-pole filter is used to model the vocal tract and has a transfer function represented in z domain as

$$V(z) = \frac{\mathrm{G}}{1 - \sum_{\mathrm{k}=1}^{\mathrm{p}} a_k z^{-k}} \quad (3)$$

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 Linear Prediction Coefficients (LPC), p is the order of the all-pole filter. One of the efficient method for estimating the LPC coefficients and the filter gain is Autocorrelation method [15]. Last stage of this algorithm is cepstral analysis which refers to the process of finding out the cepstrum of speech sequence. Basically there are two types of cepstral approaches: FFT cepstrum and LPC cepstrum. In the former case the real cepstrum is defined as the inverse FFT transform of the logarithm of the speech magnitude spectrum defined by following equation

$$\hat{s}[n] = \frac{1}{2\pi}\int_{-\pi}^{+\pi} \ln\left[S(\omega)\right] e^{j\omega n} d\omega \quad (4)$$

Where$S(\omega)$ and $\hat{s}[n]$ shows the Fourier spectrum of a signal and cepstrum respectively [16]. However, one more method for estimating these cepstral coefficients is from the LPC via a set of recursive procedure and the coefficients thus obtained are known as linear prectioncepstral coefficients (LPCC).

3.2 Mel Frequency Cepstral Coefficients

MFCC is the most widely used feature extraction technique. These coefficients represent audio based on perception and are derived from the mel frequency cepstrum. This method is considered to be the best available approximation of human ear. The block diagram of the structure of an MFCC processor is shown in Figure 2.

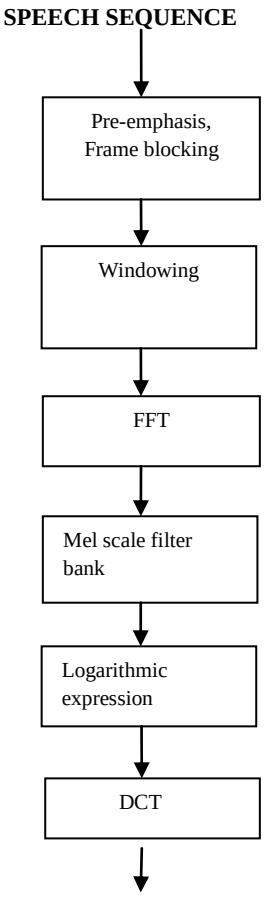

Figure 2. Basic layout of MFCC

Like LPCC, the input signal is first passed through the 1st order digital all- pole filter for pre- emphasis so as to spectrally flatten the signal and then this resultant signal is passed for windowing where it is divided into frames using hamming windows. After windowing first FFT and then Mel scale filter banks are applied so as to obtain the Mel-spectrum. FFT is basically used for the conversion of the speech signal from time domain to frequency domain. Mel scale filter bank consists of a series of triangular bandpass filter banks which are arranged in such a way so that the lower boundary of one filter is located at the centre frequency of the previous filter and the upper boundary of the same filter is situated at the centre frequency of the next filter [15]. The Mel scale is logarithmic scale that resembles the way in which human ear perceives sound. Mel scale filter bank maps the powers of the spectrum obtained above onto the Mel scale by using triangular overlapping windows. The Mel scale is represented by the following formula:

$$Mel_f = 2595 \ln\left(1 + \frac{f}{700}\right)(5) \quad (5)$$

Where Mel_f is Mel frequency in mels and f is linear frequency in hertz. After the signal is passed through the filter banks, log energy at the output of each filter bank is calculated. The natural logarithm is taken to transform into cepstral domain. Finally, DCT is applied to each Mel spectrum (filter output) to convert the values back to real values in time domain. This transformation decorrelates the features and first few coefficients are joined together as a feature vector of a particular speech frame. Since DCT accumulates most of the information contained in the signal to its lower order coefficients by discarding the higher order coefficients, so a considerable reduction in computational cost is accomplished. Hence 12 coefficients out of 24 coefficients are used as MFCC features in our paper.

3.3 Bark frequency cepstral coefficients

BFCC is another method for extracting the features from the speech signal. Figure 3 shows the block diagram of BFCC algorithm. This method is similar to MFCC. Implementation of bark scale filters in place of mel filters is quite obvious. Mathematically bark scale filter is represented by following formula:

$$f_{bark} = 6ln\left[\frac{f}{600} + [(f/600)^2 + 1]^{0.5}\right] \quad (6)$$

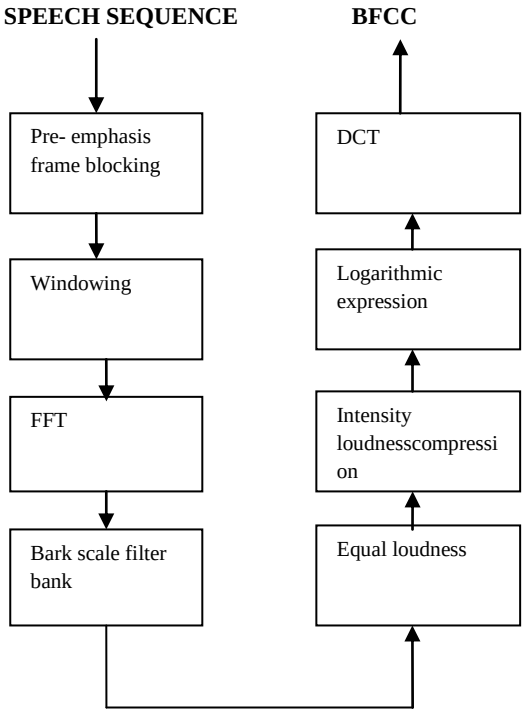

Figure 3.shows the block diagram of BFCC process

Where, f_{bark} is the corresponding Bark frequency in Barks and f is the linear frequency in hertz. The filter outputs are weighted according to the equal-loudness curve, that approximates the sensitivity of human hearing. The signal is then compressed under the intensity-loudness power law where the amplitude of the signal is compressed by the cubic-root to match the non linear relationship between intensity of sound and perceived loudness. Finally signal is first compressed through the logarithmic function and finally DCT is used to decorrelate the features as in case of MFCC

4. ASR BACK-END PROCESSING

Pattern recognition is one of the most important problems in speech recognition task. The task of pattern classification is to allocate an input pattern characterized by a feature vector to one of many prespecified classes [17], [18]. In this paper, Artifical Neural Network (ANN) technique has been used due to its better appropriateness for the said problem.

4.1 Artifical neural network

Perception information is processed much faster by human brain than modern computers, like visual and auditory information. An artificial neural network is an information processing paradigm that is enthused by the manner biological

nervous systems, especially the brain, process information [19]. It comprises millions of interconnected processing elements called neurons that are being used to abstract and model some functionality of human nervous system. Normally ANN is trained in order to achieve configuration for a specific application.

Mathematically, output of any neuron can be represented by scalar product of the input vectors and corresponding weights and is defined by the following equation:

$$Y = f(net) = f\left(\sum_{j=1}^{n} w_j x_j\right) \quad (7)$$

Where Y, is the output signal of the neuron, x_j are input signals w_j are the corresponding weights and $f(net)$ is the activation function responsible for activating the neuron.

In this paper feed forward network (with no loop back) and Multi Layer Perceptron network is used i.e. an additional layer is present in between the input and output neuron layers known as hidden layer [20]. Figure 4 shows the interconnection between neuron of various layers.

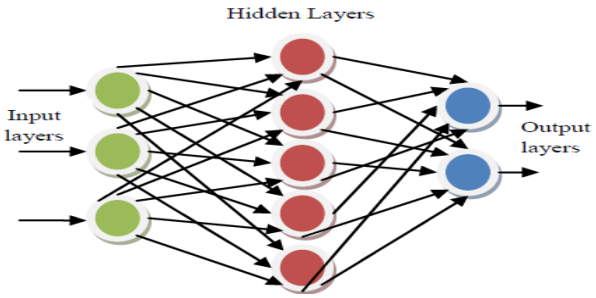

Figure 4. Structure of Multilayered ANN

5. EXPERIMENTAL RESULTS

Here four steps are used in the development and training of the feed-forward neural network [21], namely: (i) collect the training data, (ii) develop the network object, (iii) train the neural network, and (iv) Simulate the network response to test data. Database consists of three different types of Hindi words namely Isolated words, Paired words and Spoken Hindi Hybrid Paired words. Experimentation has been started with recording of the same signal from different speakers with different gender, age groups, origin. Ten speakers, five male and five female speaking each of the word, each five times in different emotions (normal, happy, anger, surprise and sad) have been recorded making a total of 750 utterances. Because of the noise present at the start and end in the recorded signal, some delay is caused and hence the length of the feature vector is increased. Bit- wise Endpoint detection algorithm is employed, so as to minimize the length of feature vectors and the resultant signal is applied as an input to the feature extractor [5]. LPCC and MFCC have been employed as feature extraction methods and the corresponding outputs are given as input to the classifier. Out of the whole data, 80% of the extracted feature vectors were used for training and the remaining 20% for testing purpose. Table 2, Table 3 and Table 4 shows the word wise recognition rates for Isolated, Paired and Hybrid words.

Figure 5, Figure 6 and Figure 7 shows the word wise recognition for Isolated, Spoken paired and Hindi Hybrid words using LPCC, MFCC and BFCC as feature extraction techniques respectively. Figure 8 shows the status of average recognition rate of 95.82%, 99.78%, 95.68% and 97.02%, 99.88%, 95.56% and 96.62%, 99.82%, 97.62% for isolated, paired and hybrid words using LPCC, MFCC and BFCC respectively.

Table 2. Word wise recognition rate for Isolated words.

Words	Extraction Methods	RECOGNITION RATE
EK	LPCC	96.0
	MFCC	99.8
	BFCC	96.5
DO	LPCC	96.0
	MFCC	100
	BFCC	98.3
TEEN	LPCC	90.1
	MFCC	99.8
	BFCC	95.7
CHAR	LPCC	99.0
	MFCC	99.8
	BFCC	93.3
PAANCH	LPCC	98.0
	MFCC	99.8
	BFCC	94.6

Table 3. Word wise recognition rate for Paired words.

Words	Extraction Methods	RECOGNITION RATE
HUM-TUM	LPCC	97.0
	MFCC	100
	BFCC	96.2
YANHA-WANHA	LPCC	93.1
	MFCC	99.9
	BFCC	93.3
JINA-MARNA	LPCC	99.0
	MFCC	99.9
	BFCC	98.1
KHANA-PINA	LPCC	98.0
	MFCC	99.7
	BFCC	99.2
DIN-RAAT	LPCC	98.0
	MFCC	99.9
	BFCC	91.0

Table 4. Word wise recognition rate for Hybrid words.

Words	Extraction Methods	RECOGNITION RATE
KALA-SAAYA	LPCC	98.0
	MFCC	99.9
	BFCC	96.9
KHAAS-AADMI	LPCC	92.1
	MFCC	99.9
	BFCC	97.2
BADIYA-ITEM	LPCC	96.0
	MFCC	99.8
	BFCC	98.7
BADA-EHSAAN	LPCC	99.0
	MFCC	99.6
	BFCC	97.9
PURANI-JEANS	LPCC	98.0
	MFCC	99.8
	BFCC	97.4

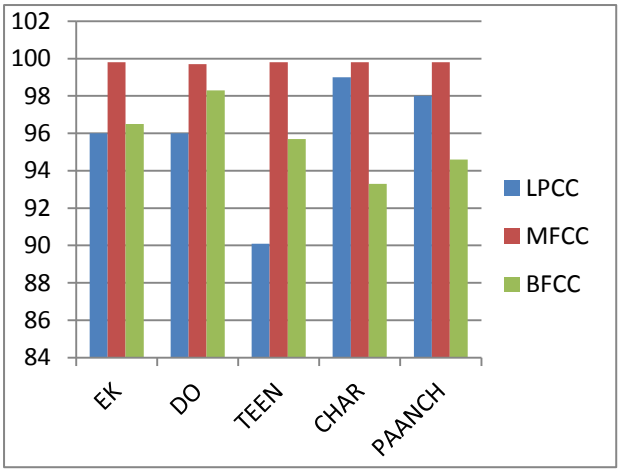

Figure 5. Word- wise recognition for isolated words

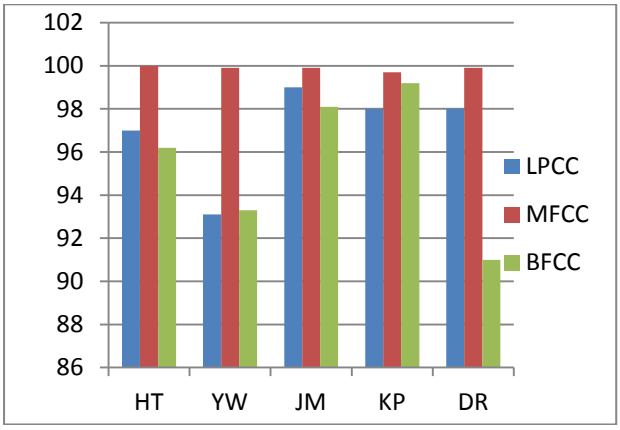

Figure 6. Word- wise recognition for paired words

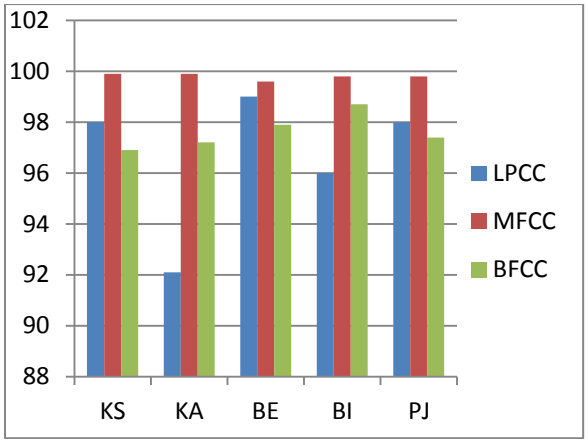

Figure 7. Word- wise recognition for Hybrid words

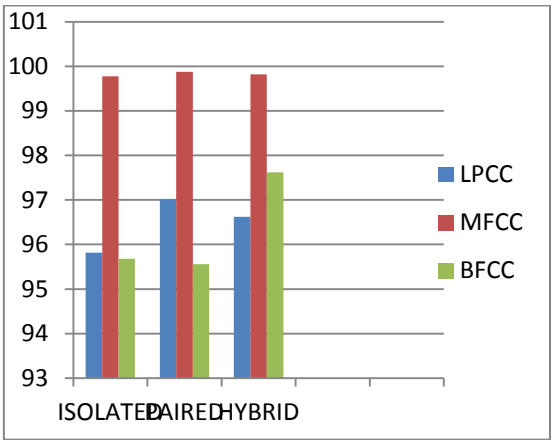

Figure 8. Average recognition rate for Isolated, Paired Hybrid words respectively.

5. CONCLUSION

Most high-flying and prime means of communication among humans is speech. A brief detail regarding the implementation of different algorithm has been provided. This paper surveys a comparative analysis of LPCC, MFCC and BFCC as feature extraction techniques and classifier as ANN. Hindi Isolated, Paired and Hybrid words are used for database purpose. Experimental results demonstrate that MFCC shows better recognition rate with 99.78% for Isolated, 99.88% for paired and 99.82% for Hybrid words. LPCC an BFCC shows the recognition rates of 95.82%, 97.02%, 96.62% and 95.68%, 95.56%, 97.62% for Isolated , Paired and Hybrid words respectively. Results from the set of experiments show that MFCC perform better than the conventional LPCC and BFCC methods.

6. REFERENCES

[1] Taabish, G., Anand, S., Rajouriya, D.K. and Najma, F. 2014, A Systematic Analysis of Automatic Speech Recognition: An Overview, International Journal of Current Engineering and Technology, Vol.4, No.3

[2] Hasnain, S.K., Maqsood, M., Shazad, M.A. and Bashir, S. 2008, development of speech recognition systems, TECHNOLOGY FORCES (Technol, forces) journal of engineering and science, Vol.2, No.1

[3] Yuan, M. 2004, Speech Recognition on DSP: Algorithm Optimization and Performance Analysis.

[4] Biing,H.J. and Sadaoki,F. 2000, Automatic recognition and understanding of Spoken launguage- A first step towards natural human-machine communication, Proceedings of the IEEE Vol.88.

[5] Taabish, G., Anand, S. And Vijay, S. 2014, An Improved Endpoint Detection Algorithm using Bit Wise Approach for Isolated, Spoken Paired and Hindi Hybrid Paired Words. International journal of computer applications, 0975-8887, Volume 92 – No.15

[6] Hisashi, W. 1977, Normalization of Vowels by Vocal Tract Length and Its Applications to Vowel Identification, IEEE Transactions onAcoustics, Speech and Signal Processing, Vol.25.

[7] Shasidhar G. Koolagudi, Reddy, R. ,Yadav, J. and Rao, K.S., 2011, IITKGP-SEHSC: Hindi speech corpous for emotion analysis, IEEE International Conference on Devices and Communications

[8] Cowie, R. and Cornelius, R.R. 2003, Describing the emotional states that are expressed in speech, Speech Communication, Elsevier, Vol. 40.

[9] Cowie, R., 2000, Emotional states expressed in speech," in Proc. of the ISCA Workshop on Speech and Emotion: AConceptual Framework for Research, pp. 224- 231.

[10] Scherer, K.R. 2000, Emotional effects on voice and speech: Paradigms and approaches to evaluation, In Pmc. Of the ISCA Workshop on Speech and Emotion: A ConceptualFramework for Research

[11] Saon, G. and Padmanabham, M. 2001,Data-driven approach to designing compound words for continuous speech recognition, IEEE Transactions on Speech and Audio Processing,Vol. 9,No.4, pp.327-332.

[12] Singh, A., Rajoriya, D.K. and Singh,V., 2012, Database Development and Analysis of Spoken Hindi Hybrid Words Using Endpoint Detection, International Journal of Electronics and Computer Science Engineering, Vol.1.

[13] Rabiner, L. and Wilpon,J., 1979, Speaker- independent isolated word recognition for a moderate size (54 word) vocabulary, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.27.

[14] Han, Y., Wang, G.Y. and Yang, Y. 2008 , Speech emotion recognition based on mfcc. Journal of Chongqing University of Posts and Telecommunications, 20,

[15] Octavian, C., Abdulla, W. and Zoran, S. 2005, Performance Evaluation of Front-end Processing for Speech Recognition Systems.

[16] Rabiner, L.R., Shafer, R.W. 2009, Digital Processing of Speech Signals, 3rd edition, Pearson education in south Asia.

[17] Schulze, E. , 1982, Hypothesizing of words for isolated and connected word recognition systems based on phonem pre-classification, IEEE International conference on Acoustics, Speech and Signal Processing.

[18] Kellis, S., 2010, Classification of spoken words using surface local field potentials, IEEE International coneference on (EMBC).

[19] Xiaoguo, X. 2013, Joint Speech and Speaker Recognition Using Neural Networks,

[20] Singh, A., Rojouriya, D.K. and Singh, V., 2012, Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks, International Journal of Computer Applications, Vol. 52, No. 12.

[21] Demuth, H., Beale, M. 2002, Neural Network Toolbox For Use with MATLAB, The Math- Works, Inc., Natick, MA