Classification of the Spoken Hindi Partially Reduplicated Words using Artificial Neural Network

Varsha Gupta
Research Scholar
Dehradun Institute of Technology
Dehradun (INDIA)

Anuj Sharma
Asst. Professor
Dehradun Institute of Technology
Dehradun (INDIA)

ABSTRACT
The most ordinary way of information exchange is Speech. It provides an efficient way of man-machine communication using speech interfacing. Speech interfacing involves two process, speech synthesis and speech recognition. Speech recognition allows a computer to identify the words that a person speaks to a microphone or telephone. The two main mechanism, used in speech recognition, are signal processing mechanism at front-end and pattern matching mechanism at back-end. In this paper, a setup for recognition of Spoken Hindi Partially Reduplicated Words (SHPRW), that uses Mel frequency cepstral coefficients at front-end and artificial neural networks at back-end has been developed to perform the experiment.

General Terms
Automatic Speech Recognition, Endpoint Detection, Feature Extraction, Pattern Recognition, Classification, Recognition Rate.

Keywords
Automatic Speech Recognition (ASR), Spoken Hindi Partially Reduplicated Words (SHPRW), Endponit Detection (EPD), Mel Frequency Cepstral Component (MFCC), and Artificial Neural Network (ANN).

1. INTRODUCTION
Speech recognition is known as automatic speech recognition (ASR) or computer speech recognition which means to understand voice of the computer and perform any required task or the ability to match a voice against a provided or acquired vocabulary [1]. The job is to getting a computer to understand spoken language. By the word "understand" we mean to respond properly and convert the input speech into other medium e.g. text. Speech recognition is therefore sometimes referred to as speech-to-text (STT) conversion. A speech recognition system consists of a microphone, for the person who is speaking into; speech recognition software; a computer to take and interpret the speech; a good quality soundcard for input and/or output; a proper and good pronunciation.

The final objective of ASR research is to allow a computer to recognize in real-time, with100% accuracy all words that are intelligibly spoken by any person, independent of vocabulary size, noise, speaker characteristics or accent. When a speaker uttered something, the linguistic content, speaker characteristics (e.g. vocal tract length, shape and gender), speaking rate and acoustic environment, simultaneously affect the acoustics of the net spoken production [2]. Speech signal not only include the meaning of a word but it will also include the emotions which will play a major role in speech recognition [3]. For advance of robust speech systems there is a need to analyze and characterize the emotions present in the speech signal [4]. There is a unique type of words called as spoken words it is of different types such as short words, moderate words and long words [5]. But spoken words have an advantage over short and long words in terms of pre-processing time, misrecognition rate and requirement of large memory space for storing speech templates [6]. Gap between paired word act like a speech code and play significant role in recognition process [7, 8]. Partially Reduplicated words have space between two words, which act as a speech code and can play significant role in recognition process [9].

Now if we talk about the isolated word example- single then in this word no gap is available here which is a major parameter of advance of protection of speech signal, while in connected or paired words example- Kaam – Vaam, we have one gap between these two words. So in case of connected words we have used only Hindi language in the example although gap is their between words which will provide better security comparable to isolated words, but if we use one language comparable to two or more different language then this concept will gives birth to spoken Hindi Partially reduplicated words in which we uses words from same one language in which one word is always from Hindi language and the other is partially repeated and also from same language. If we use noun headed paired word which is again categorized in four types such as copulative, reduplicated, partially duplicated and hybrid type. Since this gives birth to the reduplicated paired word. Reduplication in linguistics is a morphological process in which the root or stem of a word (or part of it) is repeated exactly. When one part has meanings and second part has meaning etc., then reduplication is called a partially reduplication.

An example of spoken Hindi reduplicated words is Chai - Shai in which first word is from Hindi language and second word is a part of the first word and from same language the gap between these two words will act like a speech code and it is also same due to the use of only one language. There are several advantages of partially reduplicated words over other words such as information content is more, security level is high and recognition rate is high.

Organization of this paper is as follows. Section 2 explains the Spoken Hindi Partially Reduplicated Words (SHPRW) database. Section 3 deals with feature extraction technique, i.e. Mel Frequency Cepstral Coefficient (MFCC). In section 4 Artificial Neural Networks (ANN) classifier is described. Section5 discusses the experimental works and results.

Conclusion and future scope of the work has been derived in section 6.

2. SPOKEN HINDI PARTIALLY REDUPLICATED WORDS (SHPRW) DATABASE

Database is created for Spoken Hindi Partially Reduplicated Words (SHPRW) using 16 speakers. 8 Male and 8 Female speakers are selected from different regions of India and of different age groups also. Male and Female speakers are selected from different geographical regions and from different age groups in order to accommodate acoustical variations in their utterances [11].

Recording is done using stereo headset with microphone H250 with noise cancelling feature at a sampling rate 11025 Hz using MATLAB 7.9.0. Total 640 utterances have been recorded by 16 speakers.

Table1. Words from database and their Broad Acoustic Classification, where V= vowel, C= consonant and WN= word no

S. No.	WORDS IN DATABASE	BROAD ACOUSTIC CLASSIFICATION
1.	PYAS – VYAS	CCVC – CCVC
2.	HASNA – VASNA	CCCV – CCCV
3.	PANI – VANI	CVCV-CVCV
4.	KAM – VAM	CVC – CVC
5.	CHAI – SHAI	CVC – CVC
6.	PAN – VAN	CVC – CVC
7.	NAM – VAM	CVC – CVC
8.	KHUN – VUN	CVC – CVC
9.	ASAN – VASAN	VCVC – CVCVC
10.	PHAL – VAL	CC – CC
11.	PYAR – VYAR	CCVC – CCVC
12.	GANA – VANA	CVCV – CVCV
13.	KHANA – VANA	CVCV – CVCV
14.	GHAR – VAR	CC – CC
15.	GADI – VADI	CVCV – CVCV
16.	RAJAI – VAJAI	CCVV – CCVV
17.	PARDA – VARDA	CCCV – CCCV
18.	TAR – VAR	CVC – CVC
19.	DHONA – VONA	CVCV – CVCV
20.	NAHANA – VAHANA	CCVCV – CCVCV
21.	BAAT – SHAAT	CVC – CVC
22.	CINEMA – VINEMA	CVCVCV-CVCVCV
23.	HOTEL – VOTEL	CVCC – CVCC
24.	RAHNA – VAHNA	CCCV – CCCV
25.	PHUL – VUL	CVC – CVC
26.	TALA – VALA	CVCV – CVCV
27.	SABJI – VABJI	CCCV – CCCV
28.	KAPDA – VAPDA	CCCV – CCCV
29.	MAKAN – VAKAN	CCVC – CCVC
30.	DHUP – VUP	CVC – CVC
31.	ROTI – VOTI	CVCV – CVCV
32.	BAL – VAL	CVC – CVC
33.	KAAN – VAAN	CVC – CVC
34.	ALAG – THALAG	VCC – CCC
35.	HATH – VATH	CVC – CVC
36.	AAR – PAAR	VC – CVC
37.	AAS – PAAS	VC – CVC
38.	THAPPAD – VAPPAD	CCCC – CCCC
39.	KAGAZ – VAGAZ	CVCC – CVCC
40.	GOL – MOL	CVC – CVC

3. FEATURE EXTRACTION

The next step is an important one, namely to extract relevant from the speech blocks while removing redundant and unwanted information. The aim is to sufficiently represent the characteristics of the speech signal with reduced redundancy. Features can be defined as the smallest unit which distinguishes maximally closed classes. It is essential that good features must have a wide changes from class to class and should be insensitive to the irrelevant variations and should have a low correlation with other features. The good feature quality is that it gives maximum information about the class within a much smaller dimension. Further, these features play an important role in deciding the overall recognition system.

3.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficient (MFCC) [12] is commonly used feature extraction front-end process in speech recognition systems. This technique is so-called FFT-based, which means that the feature vectors are extracted from the frequency spectra of the windowed speech frames. Figure 1 show the steps included in the MFCC algorithm.

As shown in Figure 1, MFCC consists of seven computational steps [13]. Each step has its own function and mathematical approaches as discussed briefly in the following:

Step 1: Pre–emphasis

In this step the signal is passed through a filter which emphasizes higher frequencies. This process will enhance the energy of signal at higher frequency.

$$Y[n] = X[n] - 0.95\,X[n-1] \qquad (1)$$

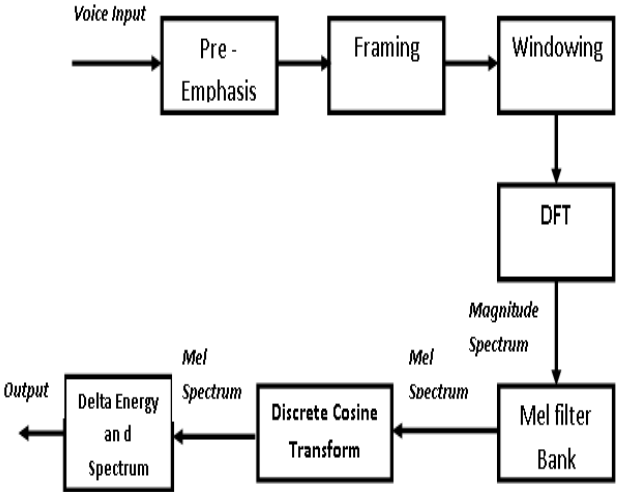


Fig 1: Illustrates the MFCC extraction procedure

Let's consider a = 0.95, which make 95% of any one sample is supposed to originate from previous sample.

Step 2: Framing

In this process the speech samples are segmented obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. Adjoining frames

are being separated by M (M<N). Typical values used are M = 100 and N= 256.

Step 3: Hamming windowing

Hamming window is used as window shape which is the next block in feature extraction processing chain and it integrates all the closest frequency lines. The Hamming window equation is written as: If the window is defined as W (n), $0 \leq n \leq N-1$ where

N = number of samples in each frame

Y(n) = Output signal

X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) \times W(n) \quad (2)$$

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \quad 0 < n < N-1 \quad (3)$$

Step 4: Fast Fourier Transform

This process is used to convert each frame of N samples from time domain into frequency domain. The Fourier Transform is used to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports the equation below:

$$Y = FFT[h(t) * x(t)] = H(\omega) * X(\omega) \quad (4)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively**.**

Step 5: Mel Filter Bank Processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. The bank of filters according to Mel scale as shown in figure 4 is then performed.

Figure shows a set of triangular filters that are used to calculate a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [14, 15]. Then, each filter output is the sum of its filtered spectral components. After that the

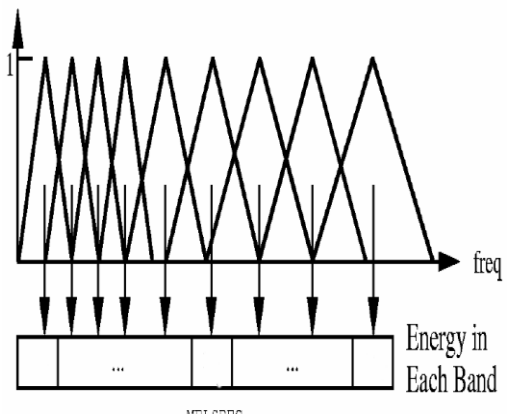


Fig 2: Mel Scale Filter Bank

Following equation is used to compute the Mel for given frequency f in HZ:

$$F_{mel} = [2595 * log10[1 + f]700] \quad (5)$$

Step 6: Discrete Cosine Transform

This process converts the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic feature vectors. Thus, each input utterance is transformed into a sequence of acoustic vector.

4. SPEECH CLASSIFICATION

Pattern recognition can also be seen as a classification process. Its ultimate aim is to optimally extract patterns based on certain conditions and is to separate one class from the others. Basically pattern classification is the process of comparing unknown pattern (i.e. called test pattern) with class reference pattern and measuring similarly between them. Basically in short, a pattern is a cluster of data points in an n-dimensional feature space, and classification is the procedure for discriminating that cluster from other data sources in the feature space.

4.1 ARTIFICIAL NEURAL NETWORK

The human brain is known to be wired differently than a conventional computer; in fact it operates under a radically different computational paradigm [16]. While conventional computers use a very fast & complex central processor with explicit program instructions and locally addressable memory, by contrast the human brain uses a massively parallel collection of slow & simple processing elements (neurons), densely connected by weights (synapses) whose strengths are modified with experience, directly supporting the integration of multiple const raints, and providing a distributed form of associative memory. ANN is adaptive in nature where learning by examples replaces programming in solving problems [17]. It can process information in parallel, at a very high speed, and in a distributed manner. If the signal flows from inputs $X_1, \ldots \ldots X_n$is considered and neuron's outputs is defined

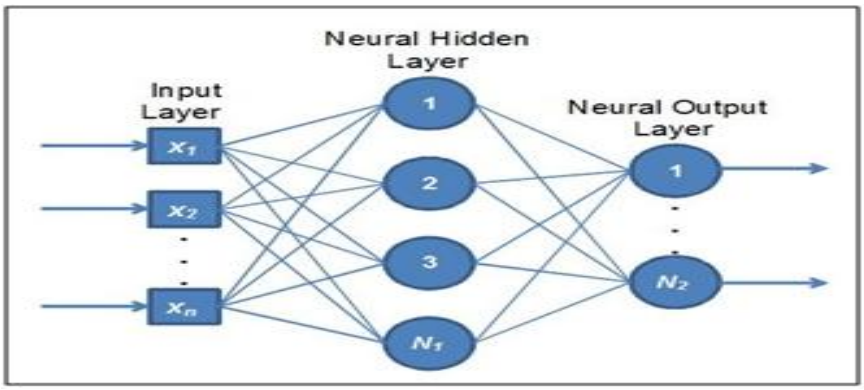


Fig 3: Multi layer ANN

As (Y), then the output signal (Y) of neuron is defined by the following equation

$$Y = f(net) = f\left(\sum_{j=1}^{n} w_j x_j\right) \qquad (6)$$

Where $w_j$is the weight vector and functions $f(net)$ is referred as an activation (transfer) function. A scalar product of the weights and input vectors is defined by variable net

$$net = w^T = w_1 x_1 + \cdots + w_n x_n \qquad (7)$$

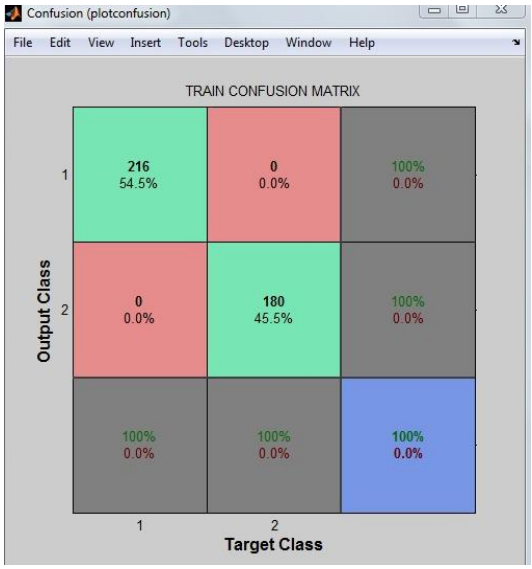
Where T is the transpose of a matrix. Figure 3 explains the structure of Multilayered ANN.

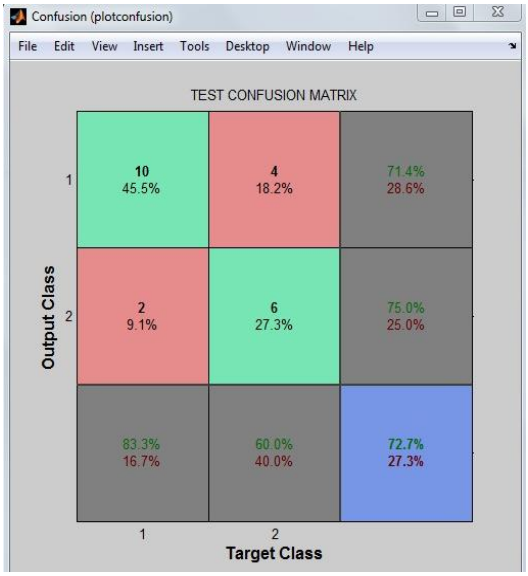
The basic architecture of a neural network is shown in above figure which is divided into three types of layers-

input, hidden and output. The signal will flow stringently in a feed forward direction from input to output. Non linear separable classes are recognized by the extra layers. In this research work Multi Layer Perceptron network is used which consists, input, one or more hidden and output layers [18]. Two layer feed forward network is used with sigmoid output neurons. Various algorithms are used in ANN for classification purpose. For Data Division Function- Random Data Division Function (dividerand), for Training Function- Scaled Conjugate Gradient training function (trainscg) and for Performance Function- Mean Squared Error Performance (mse) are used [20].

5. EXPERIMENTAL RESULT

After extracting features of speech by using MFCC, performance analysis is done by using Artificial NeuralNetwork classifier. Forty words (40) are selected which has been uttered by (16) male and female speakers.





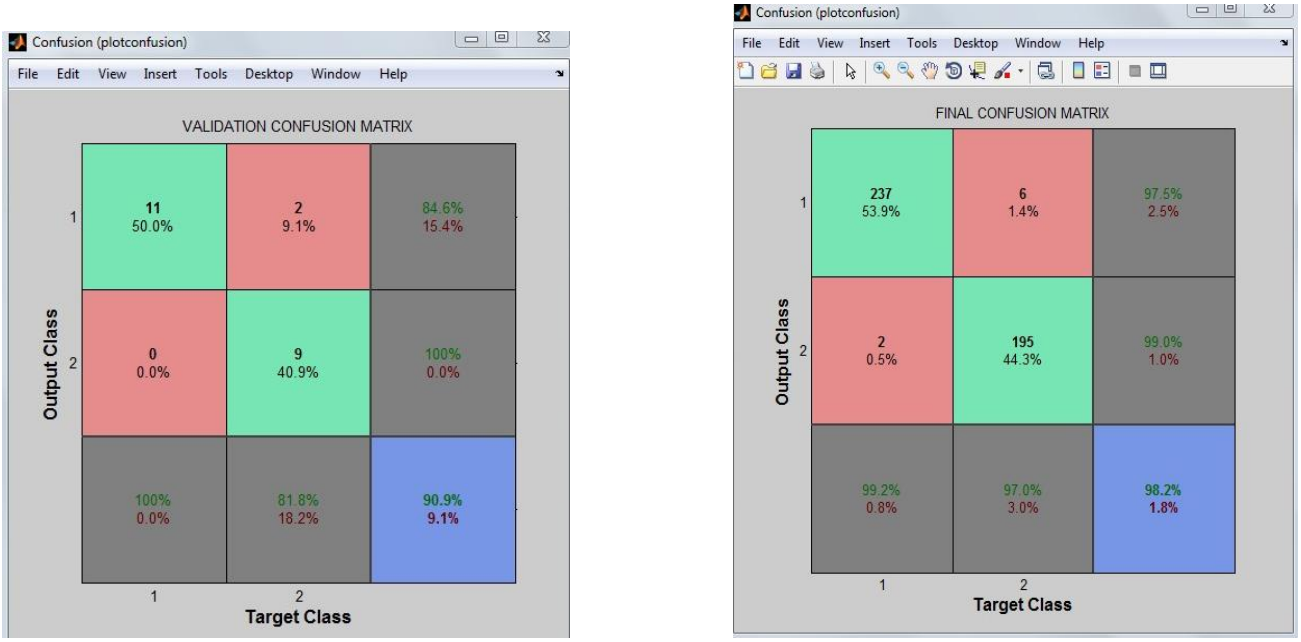


Fig 4: Confusion matrix for second feature extraction method (MFCC)

6. CONCLUSION AND FUTURE SCOPE

Finally recognition rate of SHPRW was determined with the help of feature extraction method (MFCC), along with it ANN was used for classification purpose. MFCC provides high recognition rate. Highest average recognition rate is 98.2%. Recognition rate may be further increased by considering other feature extraction technique and classifier with increased database.

7. REFRENCES

[1] Preeti saini, Parneet kaur, 'Automatic Speech Recognition: A Review', International Journal of Engineering Trends and Technology- Volume4Issue2-2013.

[2] Hisashi Wakita, 1977, Normalization of Vowels by Vocal Tract Length and Its Applications to Vowel Identification, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.25.

[3] Shasidhar G. Koolagudi, Ramu Reddy, Jainath Yadav and K.Sreenivasa Rao, 2011, IITKGP-SEHSC: Hindi speech corpous for emotion analysis, IEEE International Conference on Devices and Communications

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

[5] Dinesh Kumar Rajoriya, R.S. Anand & R.P. Maheshwari, 2011, Spoken Paired Word Pattern Classification Using Whole Word Template, TECHNIA-International Journal of Computing Science and Communication Technologies, Vol.3.

[6] A.K. Jain, R.P.W. Duin, and J. Mao, 2000, Statistical pattern recognition: A Review, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 22.

[7] Dinesh Kumar Rajoriya, R.S. Anand & R.P. Maheshwari, 2011, Enhanced recognition rate of spoken Hindi paired word using probabilistic neural network approach, International Journal of Information and Communication Technology, Inderscience Publishers, Geneva, Switzerland, Vol.3.

[8] Hariharan R., Hakkinan J. and Laurila K., 2001, Robust end of utterance detection for real time speech recognition applications, IEEE International conference on Acoustics, Speech and Signal Processing.

[9] A.K. Jain, R.P.W. Duin, and J. Mao, 2000, Statistical pattern recognition: A Review, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 22.

[10] Anand Singh, Dr. Dinesh Kumar Rajoriya and Vikash Singh, 2012, Database Development and Analysis of Spoken Hindi Hybrid Words Using Endpoint Detection, International Journal of Electronics and Computer Science Engineering, Vol.1.

[11] Rabiner, L.R. and Levinson, S.E., 1981, Isolated and connected word recognition theory and selected applications, IEEE Transactions on Communications, Vol.29.

[12] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Trans. ASSP-28, pp 357-366, 1980.

[13] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, 'Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques', Journal Of Computing, Volume 2, Issue 3, March 2010.

[14] http://www.cse.unsw.edu.au/~waleed/phd/html/node38.html, downloaded on 1st March 2014.

[15] Jamal Price, sophomore student, Design an automatic speech recognition system using maltab, University of Maryland Estern Shore Princess Anne.

[16] V. Tabarabaee, B. Azimisadjadi, S.B. Zahirazami and C. Lucas, 1994, Isolated word recognition using a hybrid neural network, IEEE International conference on Acoustics, Speech and Signal Processing.

[17] Sonia Sunny, David Peter S., K. Poulose Jacob, 2011, Wavelet Packet Decomposition and Artificial Neural Networks based Recognition of Spoken Digits, International journal of machine intelligence, Vol.3.

[18] Petek, B. and Tebelskis, J. (1992). Context- Dependent Hidden Control Neural Network Architecture for Continuous Speech Recognition. In Proceeding IEEE International Conference on Acoustics, Speech and Signal Processing.

[19] Sonia Sunny, David Peter S., K. Poulose Jacob, 2011, Wavelet Packet Decomposition and Artificial Neural Networks based Recognition of Spoken Digits, International journal of machine intelligence, Vol.3.

[20] Demuth, H., Beale, M. and Hagan, M., 2008, Neural network toolbox 6 user's guide, Mathworks Tool Box.

[21] Anand Singh, Dinesh Kumar Rajoriya, Vikash Singh, ' Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks', International Journal of Computer Applications (0975 – 8887) Volume 52– No.12, August 2012.

[22] Mehta, K. and Anand, R.S., 2010, Robust front-end and back-end processing for feature extraction for Hindi Speech recognition, IEEE International Conference on (ICCIC).

[23] Vibha Tiwari, 'MFCC and its applications in speaker recognition', International Journal on Emerging Technologies 1(1): 19-22(2010).

[24] Deeksha Bhatnagar, Vikash Singh, Sandip Vijay, ' Database Enhancement and Analysis of Spoken Hindi Reduplicated Words using Endpoint Detection Algorithm', International Journal of Computer Applications (0975 – 8887) Volume 63– No.9, February 2013.

[25] Varsha Gupta, Mukul Pant, 'An approach to describe methods Of front end processing of Speech signal', International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013.

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