International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

Speech Recognition by Linear Prediction

Shipra Soni

Abstract—Speech recognition is fundamentally pattern classification task. It is divided mainly into two components. The first component is speech signal processing and the second component is speech pattern recognition technique. The speech processing stage includes speech end point detection, pre-emphasis, frame blocking, windowing, calculating the Linear Predictive Coding (LPC) coefficients and finally generating the codebook by vector quantization. The second part includes pattern recognition system using Neural Network (NN). We use feed forward back propagation neural network for classification, both for speaker dependent and speaker independent system. Speech signals are recorded by an audio wave recorder in the normal room environment. The research work has been done using 50 different hybrid paired word (HPW) by 10 male and 10 female speakers. The performance of 95.7455% recognition rate for speaker dependent and 70.305% recognition rate for speaker independent was established.

Index Terms—Speech Recognition, pattern classification, feed forward back propagation neural network (FFBPNN), CRR, Spoken hybrid paired word (SHPW), k-fold cross validation technique.

1 Introduction

Humans can hear are one of the most important sources of information i.e. called signals. Humans access much information not only from voices but also non-verbal sounds, such as cars, airplanes, boat, ships, etc this arrangement of sounds in our daily lives, called "environmental sounds"[1]. The Speech is the most primary mode of communication among human being. The communication among human computer interaction is called human computer interface. Speech has potential of being important mode of interaction with computer. A speech signal carries information in both the time and frequency domain [2]. The human speech production process begins when the taller/speaker formulates a message or word that he/she wants to transmit to the listener via speech. Conversion of the message into a language code is the next step of the process. After choosing the language code, the speaker must execute neuromuscular commands to cause the vocal cords to vibrate appropriately and also shape the vocal tract. It results into is creation of proper sequence of speech sounds by the speaker, thereby producing an acoustic signal as the final result i.e. utterances. It produces meaningful words, phrases, sentences and non meaningful sound such as whispers, humming, whistling, etc [3].

Research in the area of speech and natural language processing has been on-going for over forty years with foundations in a number of overlapping disciplines [4]; however, there is still some improvement with mainstream speech recognition systems [5]. Spoken language is quite pervasive which leads to frustrations when spoken language systems do not meet a user's expectation. In theory, these system have the ability to save both time and money, all while executing on a consistent basis, something humans are not easily able to do.

• *Shipra soni, M.Tech from Indian Institute of Technology Rookee, Roorkee, Uttrakhand, 247667, India.PH-+91 8175087833. E-mail: shiprasoni298@gmail.com.*

Speech signal not only transfer the linguistic information but also a lot of information about the speaker himself: gender, age, social and regional origin, health and emotional state and, with a rather strong reliability, his identity. Beside intra-speaker variability (emotion, health, age), it is also commonly admitted that the speaker uniqueness results from a complex combination of physiological and cultural aspects [6].The achievement of speech recognition systems reduces significantly when they are operated in noisy conditions.

Speaker recognition systems divided into two phases. The first one is training phase while the second one is testing phase. In the training phase, each speaker has to provide samples of their speech so that the system can train a reference model for that speaker. In the testing phase, the input speech is matched with stored reference model(s) and recognition decision is made.

2 Consequences of Spoken Hybrid Paired Word

A spoken hybrid pair word is a word which etymologically has one part derived from one language and another part derived from a different language. There is a high misrecognition rate in short words. Long words require large preprocessing time, computation time and large memory space for storing speech data. Hybrid paired words are free from above problems, as they have moderate word length that requires less memory space.

3 Problem Description

Many events and incidences in our daily life relate to the perception that spoken hybrid paired word recognition system have been a barely addressed problem. The system/machine only authenticates a user at the login procedure and do not monitor the user during the entire time of computer usage. As one of the components of information system, and one of the elements in authentication system, human voice is working as authentic bioinformatics fingerprint taking a vital place in the security world. This means that people can use this property of spoken word, as speaker dependent recognition system, for

International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

minimizing threatening related problem in login-ID/password procedures.

The objective of this work is to make the community aware of the existence and the properties of the spoken PWS in speech recognition area. This work present significant contributions and uniqueness of the effort of the author to prove suitability and existence of spoken PW in speech recognition set-ups. Gap between words play significant role when paired word (moderate word) is articulated. Gap between phonetically uttered two words is low level signal and generates helpful information for recognition of utterance. The objective is to utilize this gap conceptually in two ways; first as dependent speaker recognition system (extremely protected unit, secured voice enabled system) and independent speaker recognition system(public voice enabled system).

The problem of speech recognition is handled as pattern recognition problem. Speech recognition as pattern classification consists of two components: feature analysis i.e. parameters extraction and pattern classification. The objectives associated with proposed approach of SHPW recognition can be summarized as follows:

- To collect and study data for Hindi language utterances
- Number of speakers
- Nature of the utterance
- Vocabulary size
- Differences between speakers
- Language complexity
- Environment conditions

4 Architecture of Hindi Language and its Characteristics

Hindi is mostly written in a script called Nagari or Devanagari which is phonetic in nature. Hindi sounds are broadly classified as the vowels and consonants. Hindi shares major linguistic characteristics with other indo Aryan languages. It has ten vowels. The length of vowels is phonemic. All vowels can be nasalized and nasalization is phonemic. The Hindi syllable contains a vowel as its nucleus, followed or preceded by consonants. Words usually have two or three syllable. Apart from consonants and vowels, there are some other characters used in Hindi language are: anuswar (ं), visarga (ः), chanderbindu (ँ). Anuswar indicates the nasal consonant sounds. Anuswar sound depends upon the character following it. Depending upon the varg of following character, sound wise it represents the nasal consonants of that vargs [7].

5 Detail Study of Database Generation

Speech database is required to be either generated or be used as available database for speech driven applications/machines. For developing Hindi Paired Word Recognition (HPWR) system, it is necessary to collect speech templates from a large number of speakers. A vocabulary of 50 words templates is designed 7the general purpose database environment. The recording of hybrid paired word is done in the format of 16 bit PCM, mono, audio sampling rate 10kHz.A set of database has been generated and recorded file format '*.wav' has been considered. Recording is done in a laboratory environment with closed talking microphone. Utterances samples from individuals with different gender and age groups have been recorded on computer system and converted into digitized speech signal.

To start with the process, generalized database is generated. The database consists of paired word of mixed types. While selecting the paired words to be included in this bunch of database, it is important to select appropriate type of paired words. At the outset, Hindi paired words are gathered, taking into consideration following important factors:

- PWs should be easy to speak and comprehensible.
- They should be prominent in daily life use as for common people.
- They should be easy to memorize.

Initially it is felt slightly difficult to select the paired word that should be included as part of the database. After searching numerous paired words in Hindi language, the database is generated with diverse words. Salient points about database generation process are as follows:

- Paired word with different phonemes (vowels or consonants), for the first word in paired structure are chosen.
- In the paired structure, first phoneme of second word is generally different from first phoneme of first word, as far as possible.
- In some words, a part of the first word (group of consonant and vowel) is repetitive, in the second word.
- Some words are expressive type or echo words.
- It is tried to cover up all phoneme either in first word or second word of PW for complete analysis of phonemes.

Table 1 shows the spoken hybrid paired word database, their meanings and interpretations.

6 Process of Recognition of Spoken Hybrid Paired Word

The goal of speech recognition is for a machine to be able to "hear," understand," and "act upon" spoken information. Speech recognition systems can be divided into a number of classes based on their ability to recognize that words and list of words they have. This process comes under automatic speech recognition (ASR) system.ASR is considered as pattern recognition (PR) problem. The PR method is divided into two basic tasks, such as description and classification. In the description process, features are extracted from the utterances and in classification process, feature vector are categorized into suitable classes through classifier [8].

International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

TABLE 1
HYBRID PAIRED WORDS, MEANING AND INTERPRETATION

Hindi word	English	Meaning	Interpretation
आधी रात	Aadhi raat	Half night	Midnight
आधा हिस्सा	Aadha hissa	Half part	Half
आज प्रातः	Aaj pratah	Today morning	Today morning
आजाद देश	Aazad desh	Free country	Independent country
अजनबी व्यक्ति	Ajnabi vyakti	Unknown person	Unidentified people
अनोखा व्यापार	Anokha vyapaar	Strange business	Different kind of business
अन्तिम इन्तकाम	Antim intkaam	Last revenge	Last revenge
बड़ा ड्रामा	Bada drama	Big drama	Big drama
बड़ी स्क्रीन	Badi screen	Big screen	Big screen
बाग बगीचा	Baag bagicha	Garden	Garden
बहती हवा	Bahti hawa	Flowing wind	Flowing wind
बहुत ज्यादा	Bahut jyada	Too much	Too much
बैंड बाजा	Band baaja	Musical instruments	Musical instruments
बेऔलाद स्त्री	Beaulad shtri	Childless woman	Barren woman
भेजा फ्राई	Bheja fry	Brain fry	Exhaust brain
ब्रेड मक्खन	Bread makkhan	Bread butter	Bread butter
ब्रेड पकौड़ा	Bread pakora	Bread sandwich	Fried sandwich
बुलंद दरवाजा	Buland darwaza	Strong door	Exalted door
चालाक मनुष्य	Chalak manushya	Cunning man	Cunning person
चतुर बीवी	Chatur biwi	Clever wife	Clever wife
सिनेमा घर	Cinema ghar	Cinema home	Studio
दवा दारू	Dava daru	Medicine	Related with medicine
डेली उठना	Daily uthna	Daily rising	Daily wake up
दो लफ्ज	Do lafz	Two words	Two word
डबल धमाका	Double dhamaka	Double offer	Double offer
डबल रोटी	Double roti	Double bread	Bread
दूर दराज	Dur daraaz	Too far	Too far
गोडेज र्दुगा	Goddess durga	Goddess durga	Goddess durga
जाली नोट	Jaali note	Lattice note	Lattice rupee
पह्ली आरजू	Pahli aarzu	First wish	First wish
काला साया	Kaala saya	Black shadow	Black shadow
कम्बख्त रात्री	Kambakht ratri	Damn night	Damn night
खुश मिजाज	Khush mizaj	Jolly nature	Jolly
मधुर लफ्ज	Madhur lafz	Sweet word	Sweet word
पनीर पराठा	paneer paratha	Paneer paratha	Paneer paratha
मुर्ख इंसान	Murkh insaan	Stupid human	Stupid person
नाक नक्श	Naak nakhsh	Features	Feature
नीला आसमान	Neela asmaan	Blue sky	Blue heaven
नुक्कड़ नाटक	Nukkad natak	Street drama	Street play
पहला मनुष्य	Pahla manushya	First man	First person
र्पदा प्रथा	Parda pratha	Curtain custom	Veil custom
प्रातः काल	Pratah kaal	Early morning	Early morning
रेल गाड़ी	Rail gaadi	Track vehicles	Train
साजो सामान	Saajo samaan	Goods	Types of goods
सातवॉ अजूबा	Satva ajubaa	Seventh wonder	Seventh wonder
सही वक्त	Sahi waqt	Exact time	Proper time
साहसी मुसाफिर	Sahasi musafir	Brave traveler	Brave traveler
सुहानी रात	Suhani raat	Pleasant night	A pleasant night
तन्दूरी चिकेन	Tandoori chicken	Roasted chicken	Roasted chicken
टेढ़ा रास्ता	Tedha rasta	Skewed way	Bridle way

International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

6.1 Front-end Processing

Extracted feature vector are the outcome of the front end processing unit. Each step is to be described in order to understand the overall process of frontend processing. Recorded HPW signals are required to be preprocessed first to apply further for the successive steps of features extraction. The preprocessing of an audio signal requires four steps to be applied on to the raw utterances of recorded spoken hybrid paired word signal.

Pre Emphasis

A pre-emphasis filter was applied to the digitized speech signal to reducing the high spectral dynamic range. It is used to spectrally flatten the signal. It has the effect to make it less susceptible to finite precision effect later in the signal processing. The transfer function of the first order FIR filter is

$$H(z) = 1 - \bar{a}z^{-1}\ , \qquad 0.9 \le a \le 1.0$$

Thus, the output of pre-emphasis network $\bar{s}(n)$ is

$$\bar{s}(n) = s(n) - \bar{a}s(n-1)$$

The values of a is 0.95[3].

Endpoint Detection

The aim of the endpoint detection is to remove the background noise from the spoken hybrid paired word utterances. Background noise is evaluated at the starting and end point of the spoken word signal. As in most of the cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is divided together as silence/unvoiced and is defined by voiced part.

Short Time Energy (STE) and Zeros Crossing Rate (ZCR) are two methods which are widely accepted for silence removal. But there are some limitations regarding setting thresholds as an ad hoc basis. In fact STE in voiced sample is greater than silence/unvoiced sample. However, it is not specific as to how much greater to be needed for proper classification, and it varies from case to case. On the other hand if the ZCR of a portion speech exceeds 50 then this part will be labeled as unvoiced or background noise whereas any segment showing ZCR at about 12 is treated voiced one [9].

Frame Blocking

Speech data contains information that represents speaker identity. Selection of proper frame size and overlap for analysis is crucial in order to extract relevant features that represent speaker identity. Need to analyze the signal over many short segments is called frames. After end point detection speech signal is blocked into the frames of N samples, with adjacent frames being separated by M samples [3]. If we denote the lth frame of speech by x_l (n) and there are L frames within the entire speech signal, then

$$x_l(n) = \bar{s}(Ml + n),$$

$$n = 0,1, \dots \dots \dots, N-1$$

$$l = 0,1, \dots \dots \dots, L-1$$

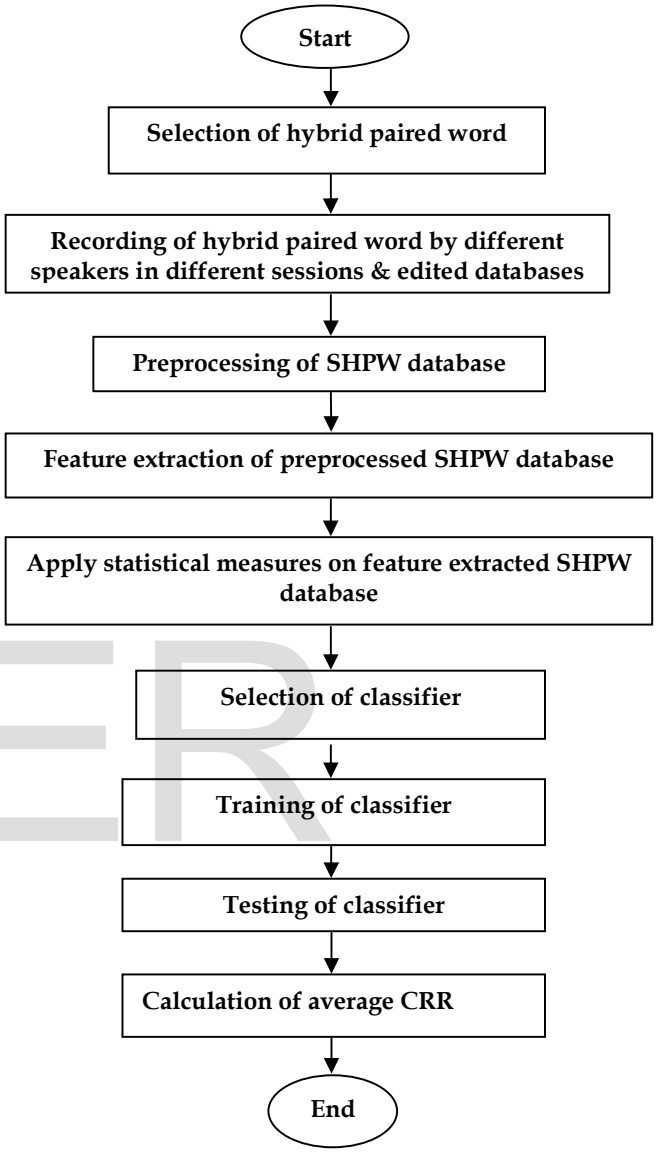

Fig 1: A flow chart of SHPW recognition

Windowing

Windowing is performed on the framed signal to smooth the abrupt frequencies at the end points of the frames. The rectangular window (i.e., no window) can cause problems, when we do Fourier analysis; it abruptly cuts off the signal at its boundaries. A Hamming window will be used to smooth the abrupt and undesirable frequencies in the speech frames. It is used to gradually tamper the window frame to zero at its beginning and end boundaries [8].Then the resulting windowed signal is defined as:

$$x_l(n) = w(n)x_l(n),$$

International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

$$0 \le n \le N-1$$

Where

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

$$0 \le n \le N-1$$

Feature Extraction

Goal of feature extraction is to transform the input waveform into a sequence of acoustic feature vectors, each vector representing the information in a small time window of the signal. Feature extraction transforms high-dimensional input signal into lower dimensional vectors. Linear predictive analysis of speech is demonstrated. The methods used are either the autocorrelation method or the covariance method. The autocorrelation method assumes that the signal is identically zero outside the analysis interval (0<=m<=N-1). Then, it tries to minimize the prediction error wherever it is nonzero, that is in the interval 0<=m<=N-1+p, where p is the order of the model used. The error is likely to be large at the beginning and at the end of this interval. This is the reason why the speech segment analyzed is usually tapered by the application of a Hamming window, for example. For the choice of the window length it has been shown that it should be on the order of several pitch periods to ensure reliable result. Thus, summing the total energy over this interval is mathematically equivalent to summing over all time. Each frame of windowed signal is auto correlated to give:

$$r_l(m) = \sum_{n=0}^{N-1-m} \bar{x}_l(n)\bar{x}_l(n+m),$$

$$m = 0,1,\ldots\ldots\ldots..p$$

The objective of Linear Predictive Coding (LPC) is to predict the next output of the system based on previous outputs and inputs [3].

6.2 Statistical Measures

For increasing the computing efficiency of classifier, statistical measures apply on extracted diverse feature vectors. Few statistical measures are as follows:

If N numbers are given, each number denoted by Xi, where i = 1... N, the average is the [sum] of the Xi's divided by N:

$$\text{Average} = \frac{1}{N}\sum_{i=1}^{N} Xi$$

The most commonly used estimator for σ is an adjusted version, the sample standard deviation, denoted by s and defined as follows:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items and $\bar{x}$ is the mean value of these observations.

Variance of size N is given by

$$\sigma^2 = \frac{1}{N}\sum_{\mathbf{i=1}}^{\mathbf{N}}(\mathbf{x_i} - \boldsymbol{\mu})^2$$

Where

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

is the mean.

6.3 Back-end Processing

It includes three essential sub processing steps namely cross validation followed by training phase and testing phase of classifier, which are essential for recognition of the system. Recognition rate of the system is dependent on classifier type. The first type of neural nets used for speech classification is a Multilayer Feed forward Network using Back Propagation algorithm for training. This type of NN is the most popular NN and is used worldwide in many different types of applications. Our network consists of an input layer, hidden layer and an output layer. NN classifiers like FFBPNN may lead to very good performances because they allow to take into account speech features information and to build complex decision regions. The mean square error of the NN is achieved at the final of the training of the ANN classifier by means of Levenberg-Marquardt Back propagation [10].

In the ANN training phase, over-fitting problem can be solved by k-cross validation technique. In the k fold cross validation technique, the original examples are partitioned into k subsets, usually of same sizes.k-1 subsets are used for training the FFBPNN and one for testing. The training is done k number of times. With this arrangement, each subset is used exactly once as the validation set. The k result can then be combined in some other way to validation while the rest are used for training.

7 Experimental Result and Analysis for Database

With the help of the process of signal preprocessing and classification, several possibilities exist in terms of selection of various parameters and various techniques. Various possibilities are explored conceptually and experimentally in order to reach up to a final conclusion. In Hybrid Paired Word Recog-

International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

nition (HPWR) system, the experimentation started with recording from different speakers with different gender, age, and different laboratory environment. Ten male and ten female speakers spoke 50 hybrid paired words for five times each. For improving performance of the system, k-fold cross-validation technique is used on extracted feature vector database. The experiment has been conducted for speaker dependent and speaker independent SHPW recognition for each speaker separately, female as well as male. The group consists of five utterances of each paired word, i.e. 50 paired words and hence make 250 templates. The 80% of extracted feature vector data set, i.e. four groups out of five have been used for training of FFBPNN model, while one group (20% of total dataset) has been used for testing. This procedure has been repeated five times to determine the suitable result. In the experiment the value of k is 5. CRR of the speaker-dependent HPWR system is 95.7455% and for speaker-independent HPWR, it is 70.305%.

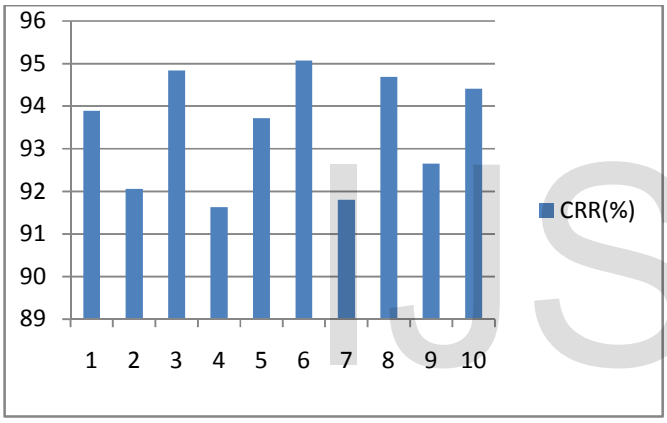

Fig 2: CRR (%) of speaker dependent SHPW for female

X axis represents no of female speaker & Y axis represents CRR (%).

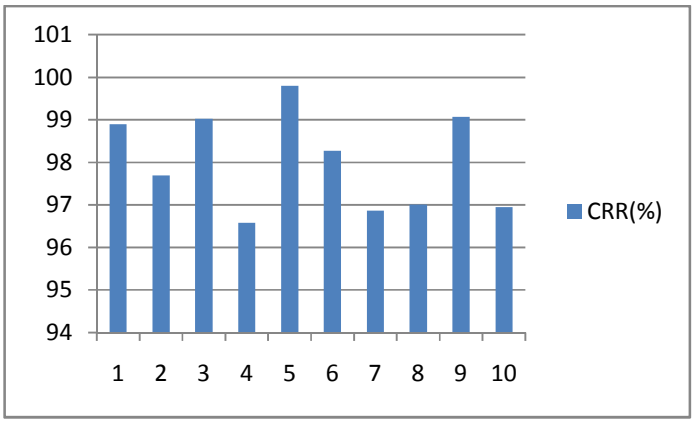

Fig 3: CRR (%) of speaker dependent SHPW for male

X axis represents no ofmale speaker & Y axis represents CRR (%).

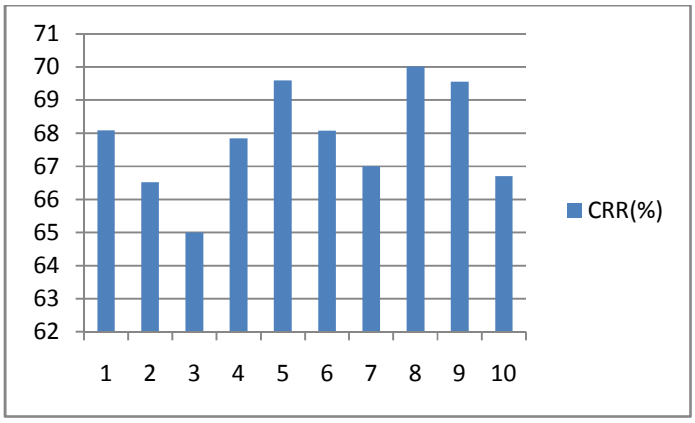

Fig 4: CRR (%) of speaker independent SHPW for female

X axis represents no of female speaker & Y axis represents CRR (%).

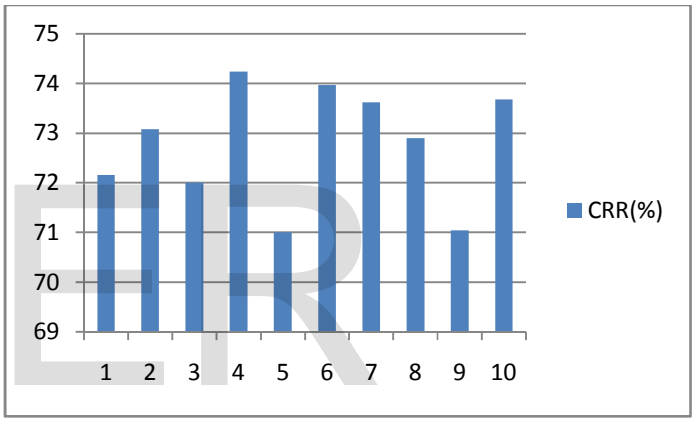

Fig 5: CRR (%) of speaker independent SHPW for male

X axis represents no of male speaker & Y axis represents CRR (%).

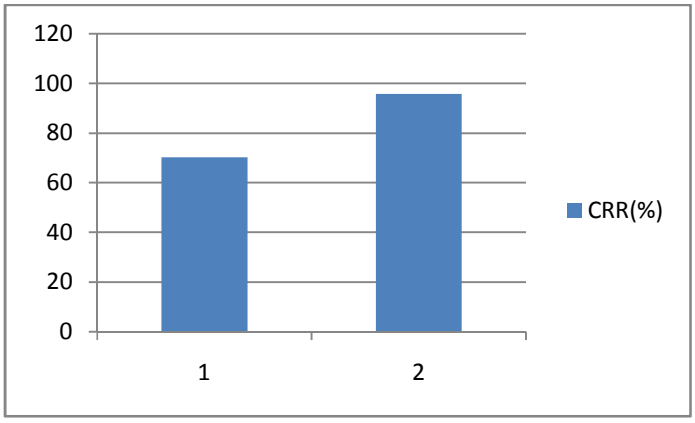

Fig 6: Average CRR (%) for speaker independent and speaker dependent

At X axis No.1 & No.2 represents speaker independent &speaker dependent respectively and Y axis represents CRR(%).

International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015
ISSN 2229-5518

8 CONCLUSION AND FUTURE WORK

The major contribution of this paper is to develop a SHPW recognition approach for voice enabled system in Indian scenario. Speaker independent and speaker dependent SHPW recognition have been evaluated for the maximum performance. In this paper author used different types of languages except Hindi, like English, Sanskrit and Urdu. To derive the SHPW features, lpc method has been used. To reduce the dimensionality of the extracted features, statistical measures as supportive phenomenon have been explored. Besides reduced features extraction, different components of pre-processing unit are explored. With these front-end processing methods, Levenberg-Marquardt Back propagation as a classifier and k-fold cross validation is used. Highest recognition rate of 95.7455% for dependent system and 70.305% for speaker independent system with SHPW database are achieved.

Speaker dependent and speaker independent voice enabled systems are inevitable for futuristic technological requirement as per the society demand. The most important future scope of the work therefore, is to further enhance the CRR of the proposed approach in the case of speaker independent system because the CRR is minimum. More attention is required to develop different PW databases in different Indian languages and other languages for expansion of voice enabled system in multidirectional usages. Different order of LPC and different windowing technique based feature extraction can be explored. In addition to these, acoustic feature of SHPW and waveform image features of SHPW can be considered for improvement of the CRR.

REFERENCES

[1] Uribe, A. O., Meana, P. M. H. and Miyatake, N. M., "Environmental sounds recognition system using the speech recognition system techniques,"*2nd International Conference on Electrical and Electronics Engineering (ICEEE) and XI Conference on Electrical Engineering,* pp. 13-16,2000.

[2] Gaikwad, S. K., Gawali, B. W. and Yannawar, P,"A Review on Speech Recognition Technique," *International Journal of Computer Applications IJCAvol. 10,* pp. 24-28, 2010.

[3] Rabiner, L. and Juang, B. H. *Fundamentals of speech recognition vol. 103: Prentice hall,*1993.(Book style)

[4] Jurafsky, D., Martin, J. H., Kehler, A., et al., *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition vol. 2: Prentice Hall New Jersey*, 2000.

[5] Deng, L. and Huang, X.,"*Challenges in adopting speech recognition,*" Communications of the ACM, vol. 47, pp.69-75, 2004.

[6] Garvin, P. L. and Ladefoged, P.,"*Speaker identification and message identification in speech recognition,*" Phonetica, vol. 9, pp.193-199, 1963.

[7] Kumar K. and Aggarwal R. "*Hindi Speech Recognition System Using Htk,*" Int. J. of Computing and Business Research, ISSN Online: 2229-6166, 2011.

[8] Rajoriya D. K., Anand, R. S.and Maheshwari, R.P.,"Hindi paired word recognition using probabilistic neural network,"*International Journal of Computational Intelligence Studies*, vol. 1, pp.291-308, 2010.

[9] Saha, G., Chakroborty, S. and Senapati, S.,"A new silence removal and endpoint detection algorithm for speech and speaker recognition applications", In Proceedings of NCC 2005, January 2005.

[10] Al Azzawi, K. Y. and Daqrouq, K.," Feed Forward Back Propagation Neural Network Method for Arabic Vowel Recognition based on Wavelet Linear Prediction Coding", International Journal, vol. 1,1963.