# Broad Acoustic Classification of Spoken Hindi Hybrid Paired Words using Artificial Neural Networks

Anand Singh
M.Tech. Student
Department of Electronics & Communication Engineering
D.I.T., Dehardun, affiliated to UKTU, Dehradun, India

Dinesh Kumar Rajoriya
Professor
Department of Electronics & Communication Engineering
D.I.T., Dehradun, affiliated to UKTU, Dehradun, India

Vikash Singh
Assistant Professor
Department of Electronics & Communication Engineering
D.I.T., Dehradun, affiliated to UKTU, Dehradun, India

## ABSTRACT
In this paper, comparison of recognition rate on the basis of domination of vowel and consonant sound in Spoken Hindi Hybrid Paired Words (SHHPW) has been carried out, with 660 utterances as database; Linear Prediction Cepstral Coefficient (LPCC) is used as a feature extraction method and Artificial Neural Networks (ANN) as a classifier. It has been observed that Consonant dominated words provides better recognition rate as compared to vowel dominated words. Average recognition rate of 93.56% has been observed for the group with consonant dominated words on the considered data base.

## General Terms
Speech Recognition, Feature Extraction, Pattern Recognition, Classification, Recognition Rate.

## Keywords
Spoken Hindi Hybrid Paired Words (SHHPW), Linear Prediction Cepstral Coefficients (LPCC), Artificial Neural Networks (ANN).

## 1. INTRODUCTION

The speech is primary mode of communication among human being and also the most earthy and efficient form of interchanging information among human. Speech recognition by machine is one of the most interesting areas for research from last many decades. Basically speech recognition is the process of automatic extracting and finding out the linguistic information conveyed by a speech signal using computers [1]. Speech processing and recognition is intensive field of research due to the broad variety of applications. Speech recognition is involved in our daily life activities like mobile applications, weather forecasting, agriculture, healthcare, speech assisted computer games, telephone assistance systems, biometric recognition etc [2]. Scientists and researchers have been trying to develop software which can easily hear, understand and speak to the users [3]. Processing of speech signal can be categorized into three main parts, Speech recognition, which allows the machine to understand words, phrases and sentences that human speaks, Natural language processing, which allows the machine to understand the need of users, Speech synthesis, with the help of which machines can speak [4,15].

When a speaker speaks, the linguistic content, speaker characteristics (e.g. vocal tract length, shape and gender), speaking rate and acoustic surroundings, simultaneously affect the acoustics of the net spoken productivity [5]. Speech signal not only contain the meaning of a word but it will also contain the emotions which will play a major role in speech recognition [6].

For development of robust speech systems there is a need to analyze and characterize the emotions present in the speech signal [7]. There is a unique category of words called as spoken words it is of different types such as short words, moderate words and long words [8]. But spoken words have an advantage over short and long words in terms of preprocessing time, misrecognition rate and requirement of large memory space for storing speech templates [9]. Gap between paired word act like a speech code and play significant role in recognition process [10, 17]. In Hindi isolated words ex- ek, then in this word there is no gap as in the case of Hindi connected words ex- haathi-godha-palki there is a gap available between every two words. This gap will act as a speech code and play a vital role in speech recognition. In Hindi connected words only Hindi language is used, to strengthen the speech code and to make it unique there is a requirement to use two languages instead of one. This concept will give a new category of words called as Spoken Hindi Hybrid Paired Words (SHHPW). There are several advantages of Spoken Hindi Hybrid Paired Words (SHHPW) over other words such as enhancement in security level, enhancement in information content as well as recognition rate [11].

Organization of this paper is as follows. Section 2 explains the Spoken Hindi Hybrid Paired Words (SHHPW) database. Section 3 deals with feature extraction technique, i.e. Linear Prediction Cepstral Coefficient (LPCC). 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 HYBRID PAIRED WORDS (SHHPW) DATABASE

Database is created for Spoken Hindi Hybrid Paired Words (SHHPW) using 22 speakers. 15 Male and 7 Female speakers are selected from different regions of India, Uttar Pradesh (Five), Uttarakhand (Seven), Madhya Pradesh (One), Jammu & Kashmir (One), Rajasthan (One), Bihar (Two), West Bengal (One), Haryana (Two), Punjab (One), as well as from Nepal (One). Male and Female speakers are selected from different geographical regions in order to accommodate acoustical variations in their utterances [18]. Thirty different words are selected for database and further these Words are categorized into three different groups G1, G2 and G3 on the basis of vowels and consonants using Broad Acoustic classification [12], which is shown in Table1. Words belongs to different groups are shown in Table2, where G1 consist of

vowel dominated words, G2 consist of consonant dominated word and G3 consist of words having equal number of vowel and consonant, each group have ten different words.

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 660 utterances have been recorded by 22 speakers. 660 utterances are divided into three groups where each group contained 220 utterances.

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

| WN | WORDS FROM THE DATABASE | BROAD ACOUSTIC CLASSIFICATION |
|---|---|---|
| 1. | TERI-INAYAT | CVCV-VCVVCVC |
| 2. | TERI-IBADAT | CVCV-VCVVCVC |
| 3. | BADIYA-LATEEFAY | CVCVCVV-CVCVVCVC |
| 4. | PURANA-ZAMANA | CVCVVCVV-CVCVVCVV |
| 5. | PURANI-AARZOO | CVCVVCV-VVCCVV |
| 6. | BADA-EHSAAN | CVCVV-VCCVVC |
| 7. | BADIYA-SAWAAL | CVCVCVV-CVCVVC |
| 8. | NAYA-AASHIYANA | CVCVV-VVCCVCVVCVV |
| 9. | RAIL-GARI | CVVC-CVVCV |
| 10. | KAALA-SAYA | CVVCVV-CVVCVV |
| 11. | SACHI-MOHABBAT | CVCCV-CVCVCCVC |
| 12. | ACHHE-LAFZ | VCCCV-CVCC |
| 13. | GALAT-TARIKH | CVCVC-CVVCVCC |
| 14. | EK-SHAKS | VC-CCVCC |
| 15. | ANOKHA-SHAKS | VCVCCVV-CCVCC |
| 16. | TEZ-BARISH | CVC-CVVCVCC |
| 17. | KATHIN-IMTIHAAN | CVCCVC-CCCVCVVC |
| 18. | BUTTER-ROTI | CVCCVC-CVCV |
| 19. | GALAT-HARKAT | CVCVC-CVCCVC |
| 20. | NAYI-KHABREIN | CVCV-CCVCCVVC |
| 21. | KHAS-AADMI | CCVVC-VVCCV |
| 22. | ACHI-TALEEM | VCCV-CVVCVVC |
| 23. | ACHI-SHAYARI | VCCV-CCVVCVCV |
| 24. | BEWAFA-LADKI | CVCVCVV-CVCCV |
| 25. | LAMBA-INTEZAAR | CVCCVV-VCCVCVVC |
| 26. | JHOOTHI-UMEED | CCVVCCV-VCVVC |
| 27. | PURANA-ZAKM | CVCVVCVV-CVCC |
| 28. | ANZAAN-MUSAFIR | VCCVVC-CVCVVCVC |
| 29. | SARKARI-MULAZIM | CVCCVVCV-CVCVVCVC |
| 30. | DOUBLE-ROTI | CVVCCV-CVCV |

**Table2. Words from database and their classification on the basis of no. of vowels and consonants, where WFD= words from the database, NV= no. of vowels, NC= no. of consonants, GN= group no. and WN= word no.**

| WN | WFD | NV | NC | GN |
|---|---|---|---|---|
| 1. | TERI-INAYAT | 6 | 5 | 1 |
| 2. | TERI-IBADAT | 6 | 5 | 1 |
| 3. | BADIYA-LATEEFEY | 8 | 7 | 1 |
| 4. | PURANA-ZAMANA | 10 | 6 | 1 |
| 5. | PURANI-AARZOO | 8 | 5 | 1 |
| 6. | BADA-EHSAAN | 6 | 5 | 1 |
| 7. | BADIYA-SAWAAL | 7 | 6 | 1 |
| 8. | NAYA-AASHIYANA | 10 | 6 | 1 |
| 9. | RAIL-GARI | 5 | 4 | 1 |
| 10. | KAALA-SAYA | 8 | 4 | 1 |
| 11. | SACHI-MOHABBAT | 5 | 8 | 2 |
| 12. | ACHHE-LAFZ | 3 | 6 | 2 |
| 13. | GALAT-TARIKH | 5 | 7 | 2 |

| | | | | |
|---|---|---|---|---|
| 14. | EK-SHAKS | 2 | 5 | 2 |
| 15. | ANOKHA-SHAKS | 5 | 7 | 2 |
| 16. | TEZ-BARISH | 4 | 6 | 2 |
| 17. | KATHIN-IMTIHAAN | 5 | 9 | 2 |
| 18. | BUTTER-ROTI | 4 | 6 | 2 |
| 19. | GALAT-HARKAT | 4 | 7 | 2 |
| 20. | NAYI-KHABREIN | 5 | 7 | 2 |
| 21. | KHAS-AADMI | 5 | 5 | 3 |
| 22. | ACHI-TALEEM | 6 | 6 | 3 |
| 23. | ACHI-SHAYARI | 6 | 6 | 3 |
| 24. | BEWAFA-LADKI | 6 | 6 | 3 |
| 25. | LAMBA-INTEZAAR | 7 | 7 | 3 |
| 26. | JHOOTHI-UMEED | 6 | 6 | 3 |
| 27. | PURANA-ZAKM | 6 | 6 | 3 |
| 28. | ANZAAN-MUSAFIR | 7 | 7 | 3 |
| 29. | SARKARI-MULAZIM | 8 | 8 | 3 |
| 30. | DOUBLE-ROTI | 5 | 5 | 3 |

## 3. FEATURE EXTRACTION

Feature extraction is the process of retaining important information of the speech signal while removing redundant and unwanted information. There are several properties of features such as high discrimination between sub-word classes, low speaker variability, invariableness to degradations in the speech signal due to noise and channel [23]. The goal with feature extraction is to attained and to comb out the speech signal into various acoustically recognizable components as well as to obtain a set of features with low rates of change in order to keep computations viable. Feature extraction can be subdivided into three basic operations-spectral analysis, parametric transformation and statistical modeling [24]. The complete succession of steps is summarized in the Figure 1.

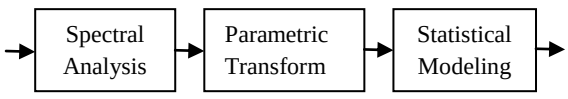


Figure 1.Steps of Feature Extraction Process

### 3.1 Linear Prediction Cepstral Coefficients

LPCC has been ordinarily used in several speech recognition applications. The opinion behind LPCC is to model the human vocal tract by a digital all pole filter. The following Figure 2. represents the process of LPCC algorithm.

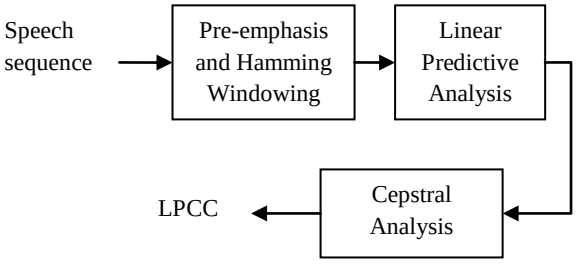


Figure 2. Steps involved in LPCC algorithm

The very first step of the algorithm is pre-emphasis. The concept of pre-emphasis is to spectrally flatten out the speech signal and equalize the inherent spectral inclination in speech. It is implemented by a first order Finite Impulse Response digital filter. The transfer function of the pre-emphasis digital filter can be shown by the following equation

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

Where $a$ is a constant and it is having a typical value of0.97. The configuration of the windowing function is an essential parameter. Rectangular window is not prescribed due to several spectral deformations to the speech frames. Instead of rectangular window other windowing function should be considered. Hamming window is mainly preferred for windowing purpose. The shape of the vocal tract in human speech production leads the behavior of the sound being developed.

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

The vocal tract is modeled by a digital all pole filter and its transfer function is expressed by the following equation in $z$ domain

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

Where $V(z)$ is the vocal tract transfer function, $G$ is the gain of the filter and $a_k$ is a set of auto regression coefficients called as Linear Prediction Coefficients (LPC). The upper limit of the summation p stands for the order of the all pole filter. The set of the Linear Prediction Coefficients finds the features of the vocal tract transfer function.

The operation of finding the Cepstrum of a speech sequence is called as cepstral analysis. Cepstrum is outlined as the Inverse Fourier Transform of the logarithm of a signal's spectrum. A signal's spectrum is defined by the following equation

$$\hat{s}[n] = \int_{-\pi}^{\pi} \ln[S(\omega)]\, e^{j\omega n}\, d\omega \quad (4)$$

Where $\hat{s}[n]$ is the cepstrum and $S(\omega)$ is the Fourier spectrum of a signal [13].

## 4. SPEECH CLASSIFICATION

Pattern training is the step where pattern representative created for particular class with the help of one or more test patterns belonging to the same class. Pattern classification is

the process of comparing unknown pattern (i.e. called test pattern) with class reference pattern and measuring similarity between them[21,22].

## 4.1 Artificial Neural Networks

Human brain is capable of face recognition, speech recognition and in other controlling activities (e.g. movement of body parts) [16]. It is capable of doing all these activities because of its effective use of its massive parallelism, the highly parallel computation structure and very fast information processing mechanism. Neurons are responsible for all these activities of human brain. The human brain is accumulation of more than 10 billion inter-connected neurons.

ANN are adaptive in nature where learning by examples replaces programming in solving problems [19]. ANN can process information in parallel, at a very high speed, and in a distributed manner. If the signal flows from inputs $x_1, \dots, x_n$ is considered and its neuron's outputs is defined as $(Y)$. The output signal $(Y)$ of neuron is defined by the following equation

$$Y = f(net) = f\left(\sum_{j=1}^{n} w_j x_j\right) \quad (5)$$

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

$$net = w^T x = w_1 x_1 + \cdots + w_n x_n \quad (6)$$

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

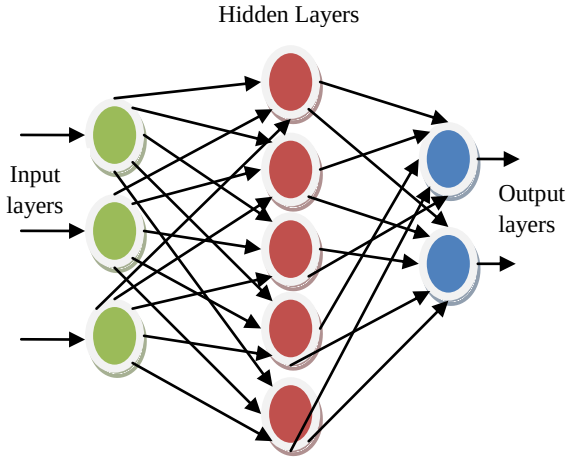


Figure 3. Structure of Multilayered ANN

The basic architecture of a general neural network will be 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 [14].

### *4.1.1 Network and algorithm specifications*

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 RESULTS

Group G1, G2 and G3 consists of 220 utterances of 10 different, vowels, consonants and equal number of vowel and consonant dominated Spoken Hindi Hybrid Paired Words respectively, recorded by 22 different speakers as mentioned in section 2.

Features of utterances belonging to different groups were extracted by using LPCC (of order 18) then these feature vectors were given as input to ANN classifier, where whole data was divided into three parts , 70% for training, 15% for validation and 15% for testing.

Figure 4. Shows the Word wise Recognition Rate of G1, in which highest and lowest recognition rates are 92.40% and 90.11% for W1 and W3 respectively.

Figure 5. Shows the Word wise Recognition Rate of G2, in which highest and lowest recognition rates are 99.14% and 90.68% for W8 and W3 respectively.

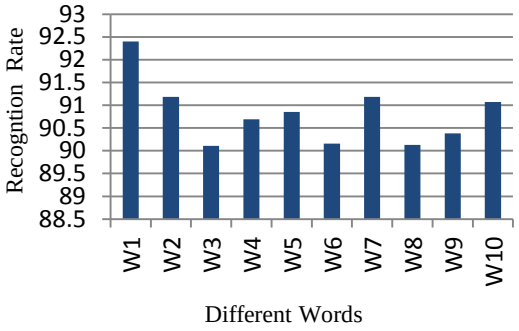


Figure 4. Word wise Recognition Rate (%) of Group 1

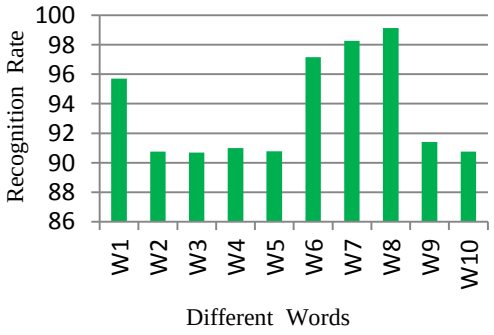


Figure 5. Word wise Recognition Rate (%) of Group 2

Figure 6. Shows the Word wise Recognition Rate of G3, in which highest and lowest recognition rates are 93.30% and 90.04% for W1 and W6 respectively.

In Figure 7. Average recognition rate of 90.81%, 93.56% and 91.20% has been observed for G1, G2 and G3 respectively.

Average recognition rate is highest in case of Group 2(93.56%) and it is lowest in case of Group 1(90.81%).

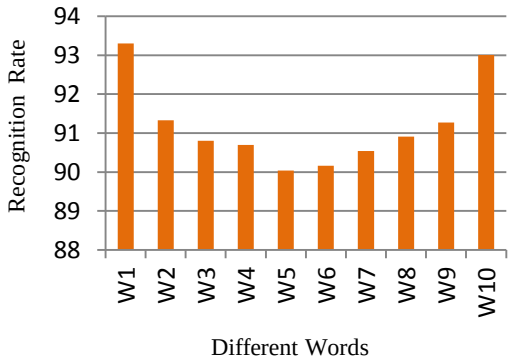


Figure 6. Word wise Recognition Rate (%) of Group 3

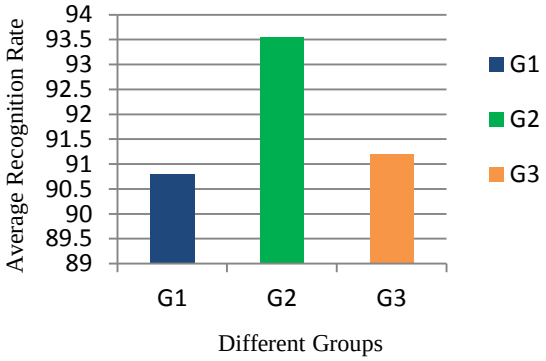


Figure 7. Group wise Average Recognition Rate (%)

## 6. CONCLUSION AND FUTURE SCOPE

Experiment results of this paper suggest that, recognition rate of group with consonant dominated word is higher than that of group with vowel dominated words as well as group with equal number of vowel and consonant words. Recognition rate of consonant dominated group is 2.75% higher than that of vowel dominated group and 2.36% higher than that of group with equal number of vowels and consonants. Thus it was found that recognition rate can be increased by using database with consonant dominated word. ANN classifier provides good recognition rate. Other classifiers such as KNN, SVM and HMM etc. can be used as an extension of this study with increased database to achieve higher recognition rate.

## 7. REFERENCES

[1] Sadaoki Furui, 2005, 50 Years of Progress in Speech and Speaker Recognition Research, ECTI Transformations on Computer and Information Technology, Vol. 1.

[2] L.R. Rabiner and B.H. Juang, B. Yegnanarayana,2009,Fundamentals of speech Recognition, 1st edition, Pearson education in south Asia,

[3] Maxine Eskenazi, 2009, An overview of spoken language technology for education, Speech Communication, Elsevier, Vol. 51.

[4] Biing- Hwang Juang and Sadaoki Furui, 2000 Automatic Recognition and understanding of spoken language- A first step toward natural human-machine communication, Proceedings of the IEEE, Vol.88.

[5] 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.

[6] 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.

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

[8] 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

[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] 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.

[11] 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.

[12] S Vishal Chourasia, Samudravijaya K., Manohar Chandwani, 2005, Phonetically Rich Hindi Sentence Corpus for Creation of Speech Database, In Proceedings of International Conference on Speech Databases and Assessment, Jakarta, Indonesia.

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

[14] 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.

[15] Rajoriya, D.K., Anand, R.S. and Maheshwari R.P., 2010, Hindi paired word recognition using probabilistic neural network, International Journal Computational Intelligence studies, Vol.1.

[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] 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.

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

[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] 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.

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

[23] Rabiner, L., 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.

[24] 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).