RESEARCH ARTICLE

ISSN: 2321-7758

AUTOMATIC SPEECH RECOGNITION SYSTEM USING ANN

S. N. DANDARE[1], NOOPUR D. DABHADE[2]

[1]Associate Professor, Babasaheb Naik College of Engineering,
Pusad (M.S.), India

[2]M.E (Digital Electronics) Project Student, Babasaheb Naik College of Engineering,
Pusad (M.S.), India

Article Received:12/05/2015 | Article Revised on:19/05/2015 | Article Accepted on:25/05/2015

ISSN:2321-7758
International Journal of Engineering Research-Online
A peer-reviewed, open access international scientific journal
www.ijoer.in

ABSTRACT

Speech is the powerful tool of information exchange. Speech Recognition (SR) is the process of determining which speech is present based on individual's utterance. Automatic Speech Recognition (ASR) system can be divided into two different parts, namely feature extraction and feature recognition. The speech samples are taken through microphone and recorded using the algorithm developed using the MATLAB software. The word detection is employed using energy and zero crossing rate of the signal.

The MFCC algorithm is employed that makes use of Mel-frequency filter bank along with several other signal processing operations including pre-processing, framing, windowing, FFT, Mel frequency Warping, Log and DCT. These obtained Mel Frequency Cepstrum Coefficients (MFCC) are used as a set of feature vector for feature recognition process. In feature recognition the supervised learning is employed and target vectors for speech recognition are created. Learning Vector Quantization (LVQ) Neural Network has been applied for classification purpose. The experimental results show that system is able to recognize words at sufficiently high accuracy. The developed ASR system is tested for 100 samples with 10 samples of each word. For which the recognition accuracy is obtained up to 93%. Further the accuracy can be increased by increasing the number of samples.

Keywords — Artificial Neural Networks, LVQ Neural Network, Mel Frequency Cepstral coefficients, Speech Recognition.

I. INTRODUCTION

Speech Recognition is also known as Automatic Speech Recognition (ASR) or computer speech recognition which is the process of converting a speech signal to a sequence of words by means of an algorithm implemented as a computer program. It has the potential of being an important mode of interaction between humans and computers. Generally, machine recognition of spoken words is carried out by matching the given speech signal against the sequence of words which best matches the given speech sample. The main goal of speech recognition area is to develop techniques and systems for speech input to machine.

The real time automatic speech recognition system faces big challenge of increasing accuracy and recognition speed. The performance of speech recognition system degrades due to noise. It also gets affected by varying speech data due to dependence on gender of speaker, environmental conditions and speaking styles. The recognition accuracy depends on the method of feature extraction and training so recognition accuracy is one of the core issues of speech recognition research. The objective of this system is to develop speech recognition system with less recognition time and high recognition accuracy. Also to improve system performance in presence of noise with the use of noise robust feature extraction and training technique. This system can be used to provide security to different systems. One can access the system if the spoken word is recognized. Also this system can be easily extended for voice dialling applications, voice command recognition systems, voice interactive systems and appliance control system through voice.

II. LITERATURE REVIEW

Many researchers have given their contribution in this area and which is explained as here.H. Hattori described a text-independent speaker recognition method using predictive Neural Network and compared the distortion-based methods, hidden Markov model (HMM)-based methods, and a discriminative neural-network-based method through text-independent speaker recognition experiments on 24 female speakers. The recognition accuracy was claim 100% [1]. Wouhaybi R. H and Adnan. Al Alaoui. also compared different Neural Networks for speaker Recognition for which different algorithms were tested and results were compared. [2]. H.P. Combrinck and E.C. Botha reported on superior performance of MFCC especially under adverse conditions. Also concluded that it represents a good trade-off between computational efficiency and perceptual considerations [3].M. A. M. Abu Sarah, R. N. Anion, R. Zainuddin, and O. O. Khalifa developed an isolated word automatic speech recognition (IWASR) system based on vector quantization (VQ). Their experimental results showed that the recognition rate has been improved with the increase of codebook size and showed that the codebook size of 81 feature vectors had a recognition rate exceeded 85% [4].Mahdi Shaneh and Azizollah Taheri suggested Voice command recognition system based on MFCC and VQ algorithms. The training is done initially with one repetition for each command and once in each in testing sessions and got 15% error rate. Secondly the training samples are increased then got zero error rates[5].A.Anusuya, S.K.Katti in the paper" Speech Recognition by Machine: A Review" summarized and compared some of the well known methods used in various stages of speech recognition system and identify research topic and applications which are at the forefront of this exciting and challenging field[6].Lindasalwa Muda, Mumtaj Begamand I. Elamvazuthi provided voice recognition algorithms using MFCC and DTW technique also concluded that with these techniques the particular speaker was correctly recognized. It is found that DTW is best for non linear time (speech) sequence alignment [7].Ahmad A. M. Abushariah, Teddy S. Gunawan, Othman O. Khalifa explained English Digits Speech Recognition System Based on Hidden Markov Models. They developed two modules, which were tested in both clean and noisy environments in which the recognition accuracy of multi-speaker model obtained was 99.5% whereas for speaker independent 79.5% accuracy was obtained[8].Ibrahim Patel and Dr Y. Shrinivasa Rao explained speech recognition using Hidden Markov Model with MFCC Sub band technique and concluded that with these methods quality metrics of speech recognition with respect to computational time, learning accuracy get improved[9].Geetika Munjal explained that the performance of ANN depend on many factors including the quality of input pattern fed in the neural network, the quantity of input pattern, the scale of neural network including the number of hidden layers and number of hidden units in each layer. Also compared MLP, LVQ and SOM in audio pattern recognition [10].Shumaila Iqbal, Tahira Mahboob and Malik Sikandar Hayat Khiyal presented Voice Recognition using HMM with MFCC for secure ATM which provided recognition accuracy to be 86.67% [11].Sharada C. Sajjan, Vijaya C gave Comparison of DTW and HMM technique for isolated word recognition and comparison of LPC and MFCC method as feature extraction. They showed that recognition accuracy is 69% for LPC & DTW, 86% For LPC & HMM, 77% for MFCC & DTW and 90% for MFCC &HMM[12].Sanjib Das provided review of Speech Recognition Techniques and through this review it is found that MFCC is used widely for feature extraction of speech, and GHM and HMM are best among all modeling techniques[13]. Geeta Nijhawan, M.K. Soni provided a Comparative Study of two different neural models i.e. back propagation (BP) neural network and radial basis function (RBF) network's performance as applied to the speaker recognition. They proved that the RBF neural network is more efficient and accurate than BP neural network in speaker recognition, and thus more suitable for practical application [14].Ms. Vrinda, Mr. Chander Shekhar presented Speech Recognition system for English language and concluded that the accuracy of the system depends upon the training time. Also provided that Time is directly proportional to accuracy if training time increases then accuracy will increase automatically[15].Vaibhavi Trivedi, Chetan Singadiya presented a detailed technical overview of MFCC and VQ algorithm. It was clearly mentioned that MFCC handles the features extraction process, which then produces outputs of speech feature vectors that are then considered as the training set used in the LBG VQ algorithm to train the VQ codebook. This research also studies the possibility of using this combination in telephony speech recognition systems [16].Lilia Lazli1, Mounir Boukadoumv presented the test of two types of hybrid models in the framework of speech recognition and medical diagnosis. The first hybrid model is the MNN structure based on the KM clustering: the work use involves LVQ and RBF neural network. The second hybrid model explains a discriminate training algorithm for hybrid HMM/MLP system based on the KM clustering. [17].Nidhi Srivastava explained speech recognition using Artificial Neural Network and MFCC technique. Training was done using different training functions of which the trainscg performed the best and was used. The result showed high accuracy when simulation was performed [18].Siddhant C. Joshi, Dr. A.N.Cheeran presented MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition. They showed that Back-propagation training for a multi-layer feed-forward network results in neural network architecture whose weights are modified in such a way that it acts as an independent word speech recognizer. Here 1000 epoch of 50 samples for training the back-propagation network and 20 samples for testing are used. The recognition accuracy of the back-propagation network implemented for pattern recognition obtained is 80 % [19].

After exhaustive study of above research work, it is learn that there is a scope for improvement in ASR. For effective results, ASR can employ an approach that is closer to human perception.

III. METHODOLOGY

Speech recognition is a complex and challenging task. The steps involved in the speech recognition are speech pre-processing, feature extraction and speech classification. The details of methodology adopted are explained as follows.

3.1. Speech

In this the speech of a person is recorded. Many types of software are available by which the speech of a person can be recorded and converted to ".wave" form. Database consists of two groups of speech samples, recorded under most similar setting condition such as the same length of recording time, and the level of sound amplitude. In training, MATLAB programs named "train" extracts features of all the words and are stored in a file named "train.m". In testing phase when a MATLAB program named "test" is executed it postulates to the user to choose any speech sample from the test group that are pre-recorded in the database.

3.2. Speech pre-processing

To enhance the accuracy and efficiency of the extraction processes, speech signals are normally pre-processed before features are extracted. There are two steps in Pre-processing.

- Pre-emphasization**:** The digitized speech waveform has a high dynamic range and suffers from additive noise. In order to reduce this range and spectrally flatten the speech signal, pre-emphasis is applied. First order high pass FIR filter is used to pre-emphasize the higher frequency components.
- Voice Activation Detection (VAD): VAD facilitates speech processing, and it is used to deactivate some processes during non-speech section of an audio sample. The speech sample is divided into non-overlapping blocks of 20ms. It differentiates the voice with silence and the voice without silence. End

point detection is done using energy and zero crossing rate of the speech signal.

3.3 Feature Extraction

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a speech classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Several feature extraction algorithms can be used to do this task, such as - Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC), and Human Factor Cepstral Coefficient (HFCC).

The MFCC algorithm is chosen to extract the features for the following reasons:

- MFCC is the most important features, which are required among various kinds of speech applications.
- It gives high accuracy results for clean speech.
- MFCC can be regarded as the "standard" features in speaker as well as speech recognition.

3.3.1 MFCC PROCESS

MFCC is based on the human peripheral auditory system. The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency *t* measured in Hz, a subjective pitch is measured on a scale called the 'Mel Scale' .The Mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1kHz.As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 Mels. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech suppresses undesired distortions or enhances some image features important for further processing. The detailed process of MFCC extraction is as explained below:

- Pre-processing: The continuous time signal (speech) is sampled at sampling frequency at 8000Hz. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This pre-emphasis is done by using a filter.
- Framing: The speech signal is split into several frames such that each frame can be analysed in the short time instead of analysing the entire signal at once. The frame size is of the range 0-20 ms. Then overlapping is applied to frames. Overlapping is done because on each individual frame, hamming window is applied.
- Windowing**:** Windowing is performed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. Hamming window gets rid of some of the information at the beginning and end of each frame. Overlapping reincorporates this information back into our extracted features.
- FFT**:** FFT is executed to obtain the magnitude frequency response of each frame and to prepare the signal for the next stage i.e. Mel Frequency Warping.
- Mel-frequency warping: The mel filter bank consists of overlapping triangular filters with the cut-off frequencies determined by the centre frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale. To compute the mels for a given frequency f in Hz, the approximate formula is used is Mel (f) = Sk = 2595*log10 (1 + f/700)

  The subjective spectrum is simulated with the use of a filter bank, one filter for each desired mel-frequency component. The filter bank has a triangular band pass frequency response, and the spacing as well as the bandwidth is determined by a constant mel-frequency interval.
- Log**:** The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition.
- DCT: In this final step, we convert the log mel spectrum back to time. The result is called the Mel Frequency Cepstrum Coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the discrete cosine transform

(DCT). By doing DCT, the contribution of the pitch is removed.

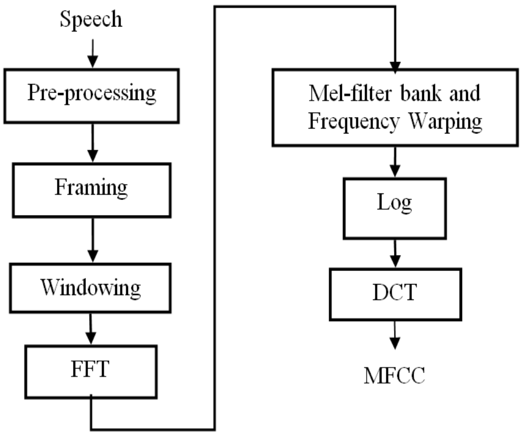

Figure 3.3.1: Block Diagram of MFCC Process

3.4. Speech Classification

The "training" process builds a library of feature vectors. The classifier uses the spectral feature set (feature vector) of the input (unknown) word and attempts to find the "best" match from the library of known words. Simple classifiers, use a simple "correlation" metric to decide. While advanced recognizers employ classifiers that utilize *Hidden Markov Models* (HMM), *Artificial Neural Networks* (ANN), and many others. The most sophisticated ASR systems search a huge database of all "possible" words to find the best (most probable) match to an unknown input.

There are many methods used for feature recognition, normally used methods nowadays are listed as below:

- Hidden Markov models
- Dynamic Time Warping (DTW)-based speech recognition
- Artificial Neural Network (ANN).

3.4.1LVQ Neural Network

LVQ neural network is the integrated network structure of supervised and unsupervised learning and its learning rate is much faster than Back Propagation (BP) neural network's. LVQ neural network fundamentally is composed of input layer, competitive layer and output layer. The foregoing first layer and second layer constitute a competitive-learning neural network. As a traditional competitive-learning neural network, such as Kohonen's Self-Organizing Map (SOM) neural network, it can automatically learn the classification of input vectors according to the nearest-neighbour method by calculating the Euclidean distance.

The LVQ performs very well if suitable initialization of weights is done. Training an LVQ is accomplished by presenting input vectors and adjusting the location of hidden units based on their proximity to the input vector. The nearest hidden units based is moved a distance proportional to the learning rate. The hidden layer weights are trained in this manner for an arbitrary number of iterations, usually with learning rate decreasing as the training progresses. The objective is to place the hidden units so as to cover the decision regions of the training set. The core of LVQ neural network is based on the nearest-neighbour method by calculating the Euclidean distance. Distances between each input vectors and competitive layer neural nodes can be calculated, and the output node which is of minimum distance is designated as a winning node.

IV. EXPERIMENTATION

The Experiment has been conducted by using MATLAB R2009b with Neural Network toolbox. Thus, different speech samples are recorded using MATLAB R2009b developed algorithm. The sampling frequency for all recording was 8000 Hz with 8 bits per sample.Word detection is done using energy and the time length in seconds. The output of this component is a wave file and then MFCC coefficients are calculated for all the input 'wave' files. The supervised learning based LVQ Neural Network is employed to create target vectors i.e. desired output vectors for inputs. Here Neural Network toolbox of MATLAB was used to create, train and simulate the networks and mean square error was used to evaluate its performance.

V. RESULT

The sample words like 'computer', 'ok', 'windows', 'paint', 'down', 'left', 'right', 'up', 'circle', 'rectangle' were tested with 10 samples of each word. Total 100 samples were used for testing purpose. Recognition accuracy is obtained to be 93%as shown in Fig. 5.1. It is desirable to have the values as close to the desired values and get a practical using this implementation.The recognition accuracy can be increased further by increasing the number of samples.

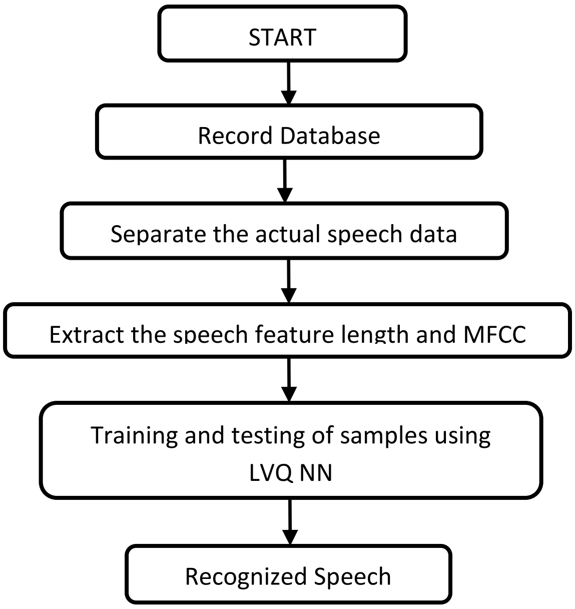
START
Record Database
Separate the actual speech data
Extract the speech feature length and MFCC
Training and testing of samples using LVQ NN
Recognized Speech

Figure 4.1: Flow diagram

Speech recognition is carried out using our own created speech database of six word category. This system had two different speakers whose speech samples were collected for training. The speakers include one female & one male speakers belonging to different ages & genders. Here total 42 samples were used for testing purpose that includes 7 samples of each word. As shown in table Fig 5.2, the 90.47% successful detection accuracy is achieved for male speaker and 95.23% recognition accuracy for female speaker.

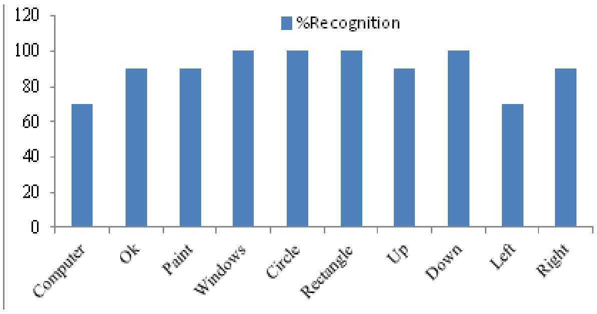

Figure 5.1: Graph of % Recognition

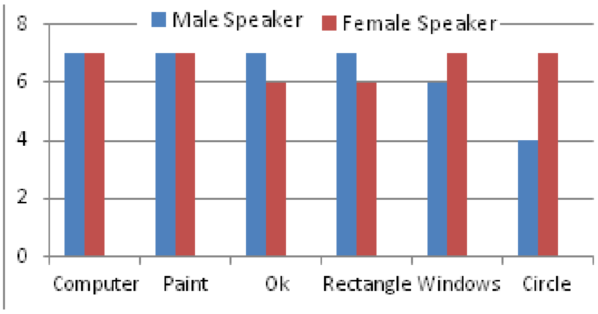

Figure 5.2: Graph of % Recognition for different Speakers

VI. CONCLUSION

Speech recognition is of great significance to many of the applications. In this paper for speech recognition, MFCC and Learning Vector Quantization (LVQ) Neural Network have been used. The exhaustive experiment has been carried out using MATLAB R2009b Neural Network toolbox and it successfully recognizes speech sample. For feature extraction MFCC algorithm and LVQ Neural Network for training is used and it is tested upon own created database. The performance of the proposed method is satisfactorily in terms of recognition accuracy. It shows that the proposed method has high recognition accuracy which illustrates the robustness of the proposed method against the speaking variations like speaking style, gender, speaking environment etc. Extensive training and testing experiments are carried out in order to demonstrate the effectiveness of the proposed method for speech recognition. Speech recognition using ANN gives good result due to resemblance between architecture of ANN and varying speech data. The recognition accuracy increases due to combination of MFCC and LVQ-ANN in noisy environments. The performance is evaluated by finding word wise accuracy for ten different word samples. Here the overall accuracy obtained for above mentioned ten words is 93% also recognition accuracy for six words for male speaker is found to be 90.47% and for female speaker 95.23%.

REFERENCES

[1] H. Hattori "Text independent speaker recognition using Neural Network" IEEE Sanfrancisco, USA, vol 2, pp 153-156, 1992.

[2] Wouhaybi R. H and Adnan. Al Alaoui, "Comparison of Neural Networks for speaker Recognition" IEEE, vol 1, pp. 125-128, 1999.

[3] H. Combrinck and E. Botha, "On the Mel-scaled cepstrum" Department of Electrical and Electronic Engineering, University of Pretoria., and Journal of Computer Science, 3 (8): 608-616, ISSN 1549-3636, 2007.

[4] M. A. M. Abu Sarah, R. N. Anion, R. Zainuddin, and O. O. Khalifa ," Human Computer Interaction Using Isolated-Words Speech Recognition Technology" IEEE Proceedings of The International Conference on Intelligent and Advanced

Systems (ICIAS'07), Kuala Lumpur, Malaysia, pp. 1173 – 1178, 2007.

[5] Mahdi Shaneh, and Azizollah Taheri,"Voice Command Recognition System Based on MFCC and VQ Algorithms", World Academy of Science, Engineering and Technology 57, 2009.

[6] M. A. Anusuya, S. K. Katti "Speech Recognition by Machine: A Review" International journal of computer science and Information Security (IJCSIS), Vol 6, No.3, ISSN: 1947-5500, pp 181-205, 2009.

[7] 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, ISSN :2151-9617, pp 138-143,March 2010.

[8] Ahmad A. M. Abushariah, Teddy S. Gunawan, Othman O. Khalifa "English Digits Speech Recognition System Based on Hidden Markov Models", International Islamic University Malaysia, International Conference on Computer and Communication Engineering (ICCCE 2010), 11-13 May 2010, Kuala Lumpur, Malaysia.

[9] Ibrahim Patel, Dr.Y.Srinivasa Rao, "Speech recognition using Hidden Markov Model with MFCC-Sub band Technique" International Conference on Recent Trends in Information, Telecommunication and Computing, Vol 1, NO. 2, pp 101-110, 2010.

[10] Geetika Munjal,"ANN Paradigms for Audio Pattern Recognition" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (4), ISSN: 0975-9646, pp 1555-1558, 2011.

[11] Shumaila Iqbal, Tahira Mehboob, Malik Sikandar Hayat Khiyal, "Voice Recognition using HMM with MFCC for secure ATM", IJCS Vol.8, Issue 6, ISSN : 1694-0814, pp 297-393, Nov 2011.

[12] Sharada C. Sajjan, Vijaya C,"Comparison of DTW and HMM for isolated word Recognition", Proceedings of International Conference on Pattern Recognition, IEEE, ISBN : 978-1-4673-1073-6, pp 466-470, 21-23 March 2012.

[13] Sanjib Das," Speech Recognition Technique: A Review" IJERA, Vol 2, Issue 3, ISSN: 2248-9622, pp. 2071-2087, May-Jun 2012.

[14] Geeta Nijhawan, M.K. Soni "A Comparative Study of Two Different Neural Models For Speaker Recognition Systems" International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-1, Issue-1, ISSN: 2278-3075 , pp 67-72,June 2012

[15] Ms. Vrinda, Mr. Chander Shekhar , "Speech recognition for English language", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 1, ISSN No. 2319-5940, pp 919-922,January 2013.

[16] Vaibhavi Trivedi, Chetan Singadiya "Automatic Speech Recognition using Different Techniques" International Journal of Science and Research (IJSR), Volume 2 Issue 5, ISSN: 2319-7064, pp 144-150, May 2013

[17] Lilia Lazli, Mounir Boukadoum, "Hidden Neural Network for Complex Pattern Recognition: A Comparison Study with Multi- Neural Network Based Approach" International Journal of Life Science and Medical Research Vol. 3 Iss. 6, pp 234-235,Dec. 2013,

[18] Nidhi Srivastava ,"Speech Recognition using Artificial Neural Network" International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 3, Issue 3, ISSN: 2319-5967, pp 406-412, May 2014 .

[19] Siddhant C. Joshi, Dr. A.N.Cheeran, "MATLAB Based Back-Propagation Neural Network for Automatic Speech Recognition "International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Vol. 3, Issue 7,ISSN No. 2320-376 July 2014.p.p 10498-10504.

[20] L. R. Rabiner and B. H. jaung," Fundamentals 5of Speech Recognition" Prentice-Hall, Englewood Cliff, New Jersy, 1993.