Speech Recognition using MFCC and Neural Networks

[1]Divyesh S. Mistry, [2]Prof.Dr.A.V.Kulkarni

Department of Electronics and Communication,
Pad. Dr. D. Y. Patil Institute of Engineering & Technology, Pimpri, Pune, Maharashtra, India
[1]divyeshsmistry@gmail.com, [2]anju_k64@yahoo.co.in

***Abstract*—The most common mode of communication between humans is speech. As this is the most preferred way, humans would like to use speech to interact with machines also. That is why, automatic speech recognition has gained a lot of popularity. Many approaches for speech recognition exist like Dynamic Time Warping (DTW), Hidden Markov Model (HMM). This paper shows how Neural Network (NN) can be used for speech recognition and also investigates its performance in speech recognition. Learning Vector Quantization Neural Network has been applied. For the feature extraction of speech Mel Frequency Cepstrum Coefficients (MFCC) has been used which gives a set of feature vectors of speech waveform. Earlier research has shown MFCC to be more accurate and effective than other feature extraction techniques in the speech recognition. The work has been done on MATLAB and experimental results show that system is able to recognize words at sufficiently high accuracy.**

***key words* — Speech Recognition ; Mel Frequency Cepstrum Coefficients (MFCC) ; Neural Networks ; Learning Vector Quantization**

I. INTRODUCTION

Speech recognition is the machine on the statement or command of human speech to identify and understand and react accordingly. It is based on the voice as the research object, it allows the machine to automatically identify and understand human spoken language through speech signal processing and pattern recognition. The speech recognition technology is the high-tech that allows the machine to turn the voice signal into the appropriate text or command through the process of identification and understanding. Speech recognition is a cross-disciplinary and involves a wide range. It has a very close relationship with acoustics, phonetics, linguistics, information theory, pattern recognition theory and neurobiology disciplines. With the rapid development of computer hardware and software and information technology, speech recognition technology is gradually becoming a key technology in the computer information processing technology. Products to develop speech recognition technology is also widely used in voice activated telephone exchange query information networks, medical services, banking services, industrial control every aspect of society and people's lives.

In this paper, for speech recognition first Speech Production was carried out followed by the speech classification. The paper is divided into seven sections. Section III describes Structure of a standard speech recognition system, Section IV the feature extraction techniques of speech recognition, section V describes LVQ neural network, section VI the experimental setup and results, and section VII gives the conclusion.

II. SPEECH PRODUCTION

The sound that we know as speech begins with the lungs contracting to expel air, which carries sound of an approximately Gaussian frequency distribution [2]. This air is forced up through the bronchial tract past a set of muscle folds at the top of the trachea called vocal chords, and sets these vibrating. The air then enters the rear of the mouth cavity where it follows one of two paths to the outside. The first path is over and around the tongue, past the teeth and out through the mouth. The second route is through the nasal cavity – and this is the only route possible when the velum is closed. Figures shows a diagram of the speech production apparatus (otherwise known as the human head).

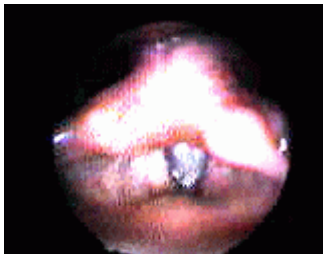


Fig. 1.Closed glottis

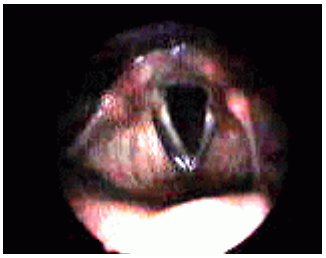


Fig. 2. Open glottis

The actual sound being produced depends on many criteria including the lung power and pressure modulation, the constriction at the glottis, the tension in the vocal chords, the shape of the mouth, and the position of the tongue and teeth.

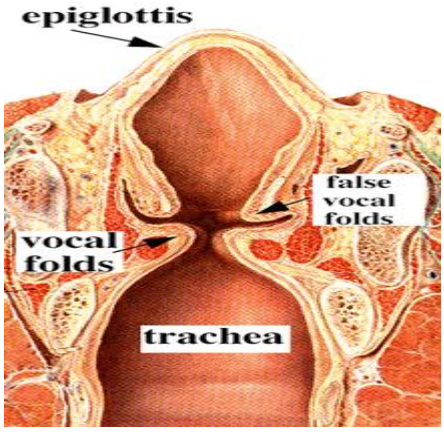


Fig.3 .The Larynx

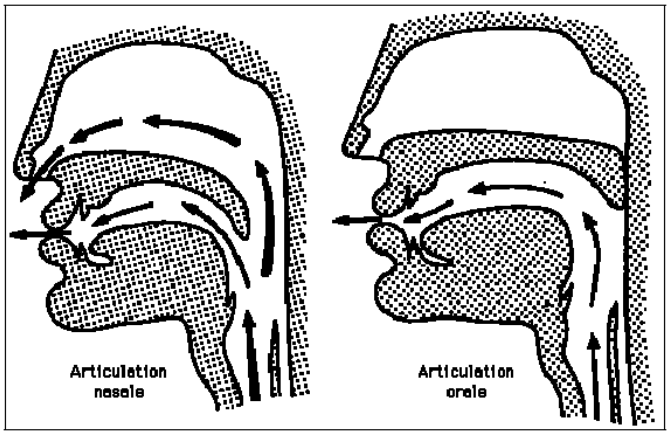


Fig.4 .The oro-nasal process

Speaker Dependent and Speaker Independent System

A speaker dependent system is developed to operate for a single speaker. These systems are usually easier to develop, cheaper to buy and more accurate, but not as flexible as speaker adaptive or speaker independent systems.

A speaker independent system is developed to operate for any speaker of a particular type (e.g. American English). These systems are the most difficult to develop, most expensive and accuracy is lower than speaker dependent systems. However, they are more flexible.

III. STRUCTURE OF A STANDARD SPEECH RECOGNITION SYSTEM

Speech recognition is a multileveled pattern recognition task, In which acoustical Signals are examined and structured into a hierarchy of sub word units (e.g., phonemes),words, phrases, and sentences. Each level may provide additional temporal constraints, e.g., known word pronunciations or legal word sequences, which can compensate for errors or uncertainties at lower levels. This hierarchy of constraints can best be exploited by combining decisions probabilistically at all lower levels, and making discrete decisions Only at the highest level.

The structure of a standard speech recognition system is illustrated In Figure. The Elements are as follows:

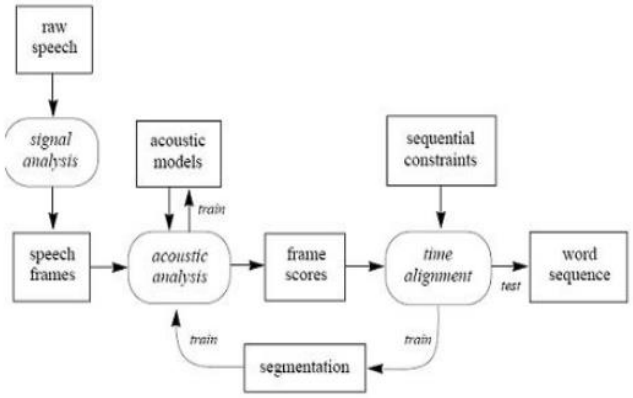


Fig 5 Structure of a standard speech recognition system

- **Raw speech**: Speech is typically sampled at a high frequency, e.g., 16 KHz over a microphone or 8 KHz over a telephone. This yields a sequence of amplitude values over time.

- **Signal analysis:** Raw speech should be initially transformed and compressed, in order to simplify subsequent processing. Many signal analysis techniques are available which can extract useful features and compress the data by a factor often without losing any important information. Among the most popular:

- Fourier analysis (FFT) yields discrete frequencies over time, which can be interpreted visually. Frequencies are often distributed using A Mel scale, which is linear in the low range but logarithmic in the high range, corresponding to physiological characteristics of the human ear.

- Perceptual Linear Prediction (PLP) is also physiologically motivated, but yields coefficients that cannot be interpreted visually.

- Linear Predictive Coding (LPC) yields coefficients of a linear equation that approximate the recent history of the raw speech values.

- Cepstral analysis calculates the inverse Fourier transform of the logarithm of the power spectrum of the signal.

- **Speech frames:** The result of signal analysis is a sequence of speech frames, typically at 10 msec intervals, with about 16 coefficients per frame. These frames may be augmented by their own first and/or second derivatives, providing explicit information about speech dynamics; this typically leads to improved performance. The speech frames are used for acoustic analysis.

- **Acoustic models:** In order to analyze the speech frames for their acoustic content, we need a set of acoustic models. There are many kinds of acoustic models, varying in their representation, granularity, context dependence, and other properties.

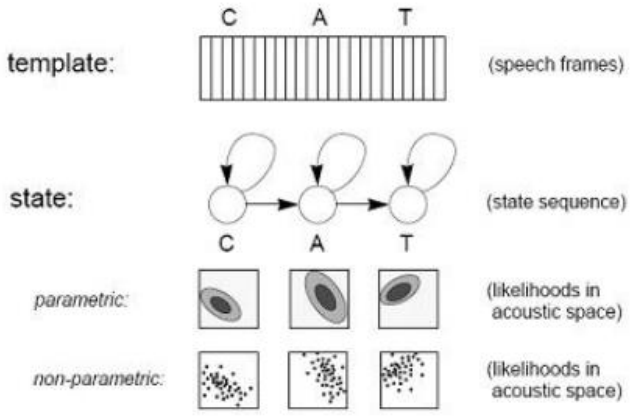


Fig. 6. Acoustic models: template and state representations

Figure shows two popular representations for acoustic models. The simplest is a template, which is just a stored sample of the unit of speech to be modeled, e.g., a recording of a word. An unknown word can be recognized by simply comparing it against all known templates, and finding the closest match.

Templates have two major drawbacks: (1) they cannot model acoustic variabilities, except in a coarse way by assigning multiple templates to each word; and (2) in practice they are limited to whole-word models, because it's hard to record or segment a sample shorter than a word – so templates are useful only in small systems which can afford the luxury of using whole-word models.

A more flexible representation, used in larger systems, is based on trained acoustic models, or states. In this approach, every word is modeled by a sequence of trainable states, and each state indicates the sounds that are likely to be heard in that segment of the word, using a probability distribution over the acoustic space. Probability distributions can be modeled parametrically, by assuming that they have a simple shape (e.g., a Gaussian distribution) and then trying to find the parameters that describe it; or non-parametrically, by representing the distribution directly (e.g., with a histogram over a quantization of the acoustic space, or, as we shall see, with a neural network).

IV. FEATURE EXTRACTION

Various feature extraction techniques exist like Principal Component Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Component Analysis (ICA), Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) etc. In this paper we have used MFCC for feature extraction as previous research has shown this technique to be the better than other techniques.

A. MFCC

The Mel-frequency Cepstral Coefficients (MFCCs) introduced by Davis and Mermelstein is perhaps the most popular and common feature for SR systems. For speech recognition purposes and research, MFCC is widely used for speech parameterization and is accepted as the baseline. This may be attributed because MFCCs models the human auditory perception with regard to frequencies which in return can represent sound better. [4,5]. They are derived from a mel-frequency cepstrum (inimize-of-spectrum) where the frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands used in the normal cepstrum [6]. The block diagram of MFCC as given in [5] is shown in
Fig.6

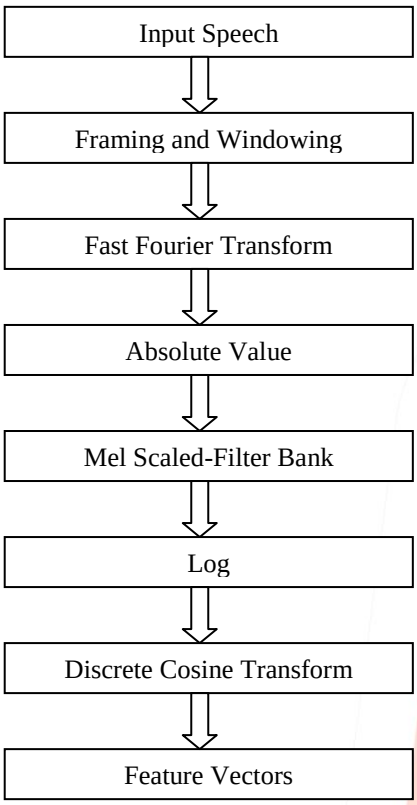


Fig. 7. Block Diagram Of MFCC

- **Pre–emphasis :** Pre-emphasis of the speech signal at
higher frequencies has become a standard preprocessing step in many speech processing applications such as linear prediction (LP) analysis-synthesis [1,2] and speech recognition [3-7].pre-emphasis serves a useful purpose because, at the analysis stage, it reduces the dynamic range of the speech spectrum and this helps in estimating the LP parameters more accurately.

- **Framing :** Framing is used to cut the long-time speech to the short-time speech signal in order to get relative stable frequency characteristics. Features get periodically extracted. The time for which the signal is considered for processing is called a window and the data acquired in a window is called as a frame. Typically features are extracted once every 10ms, which is called as frame rate.

- **Windowing:** Windowing is mainly to reduce the aliasing effect, when cut the long signal to a short-time signal in frequency domain. There are different types of windows,
There are: Rectangular window, Bartlett window,
Hamming window
Out of these, the most widely used window is Hamming
window.

- **Fast Fourier Transform :** To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is 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:

$$Y(w) = FFT\ [h(t) * X(t)] = H(w) * X(w)$$

……(1)

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

- **Mel Filter Bank Processing** : The speech signal consists of tones with different frequencies. For each tone with an actual Frequency, f, measured in Hz, a subjective pitch is measured on the 'Mel' scale. The mel-frequency scale is a linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz.
  We can use the following formula to compute the mels for a
  given frequency f in Hz: mel(f) = 2595*log10(1+f/700) [7].

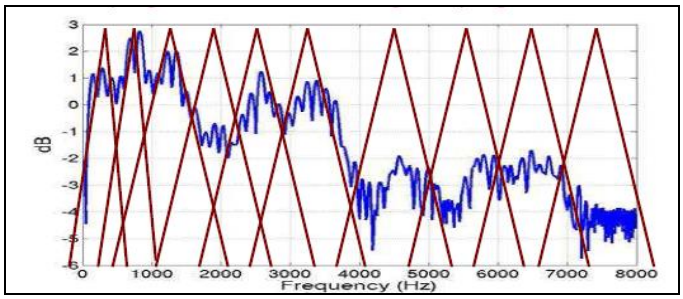


Fig.8. Filter Banks

- **Discrete Cosine Transform:** This is the process to convert 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 vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

V. LVQ NEURAL NETWORK

Neural network is the data-driven, nonlinear and nonparametric model. Neural network has been an important tool in the complex signal processing and classification in the last decades. Classification has been one of the most active application fields in neural network researches. There are usually two kinds of neural network models in seafloor classification. One is the supervised learning neural network which needs geologic grab samples [9]; the other is the unsupervised learning neural network which does not need any grab samples [6].

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 (Fig. 8)[9]. 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. However, the LVQ algorithm is a competitive approach under the supervised learning. By means of the supervised and unsupervised learning, LVQ neural network can distinguish the target vectors from the input vectors, and then divide targets into different types. The third output layer of LVQ neural network can change the transferred information from competitive layer into the defined target classes which we need.

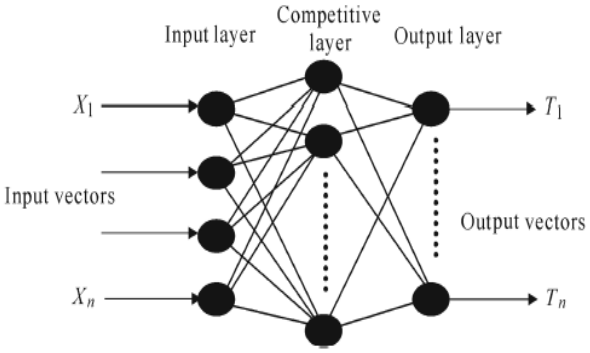


Fig 9 Schematic depiction of LVQ neural network

The core of LVQ neural network is based on the nearest-neighbor 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[9].

$d(X, Wc) = \min \{ d(X, Wi) \}, (i = 1, 2, \cdots , n),$ ……(2)
where,
X = The input vector,
Wi= The reference vector,
d(X,Wi)= The distance between X and Wi,
Wc= The winner subclass.

The following equations define the basic LVQ algorithm process:

When i = c,
if X and Wc belong to the same class

$$Wc(t + 1) = Wc(t) + \_(t)[X(t) - Wc(t)], \quad \ldots\ldots(3)$$

if X and Wc belong to the different classes

$$Wc(t + 1) = Wc(t) - \_(t)[X(t) - Wc(t)], \quad \ldots\ldots(4)$$

When i 6= c,

$$Wi(t + 1) = Wi(t), \quad \ldots\ldots(5)$$

where 0 < _(t) < 1, and learning rate _(t) is usually made to decrease monotonicity with time. It plays a very important role in network convergence.
By the iterative learning, the input vector X will be assigned to the class which the reference vector W belongs to. The class of each input vectors can be obtained through the competitive learning process.

VI. IMPLEMENTATION & RESULT

The Experiment has been conducted using MATLAB R2010a with Neural Network toolbox. In this study we took four words. These word were recording in MATLAB R2010a.The sampling frequency for all recording was 8000HZ.

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. As already mentioned NN consists of neurons. These neurons can use any differentiable transfer function *f* to generate their output. The transfer function used in our case for hidden layers is tan-sigmoid, and for the output layer is linear. Figure show below speech analysis results and Mean square error.

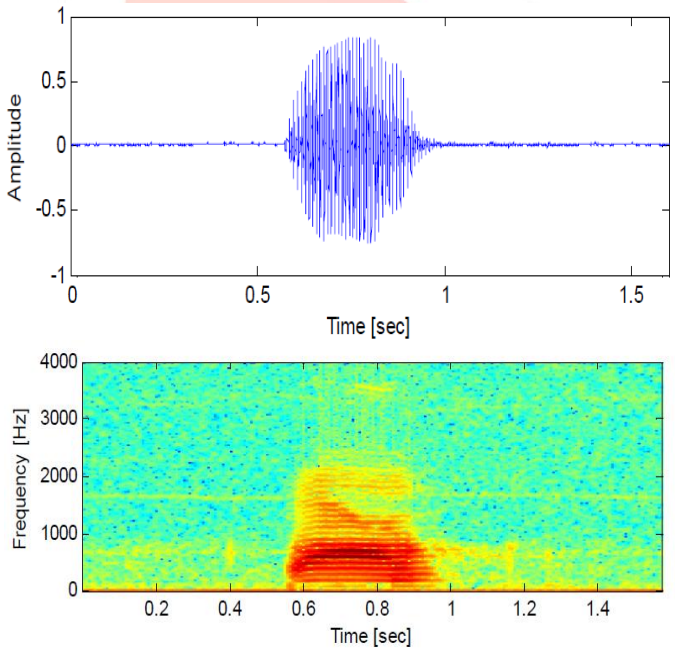
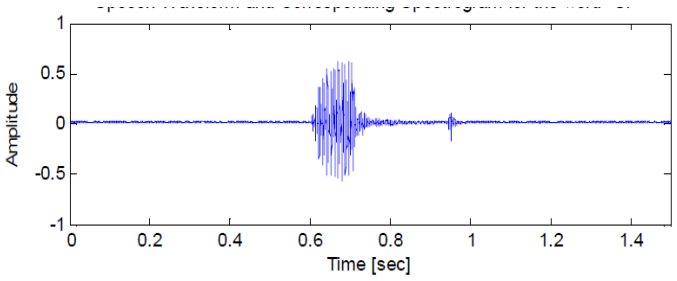


Fig. 10. Speech waveform (top plot) and associated spectrogram (bottom plot) of the word "down".



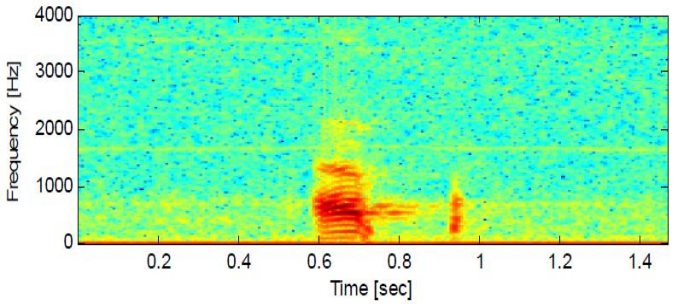


Fig. 11. Speech waveform (top plot) and associated spectrogram (bottom plot) of the word “up”.

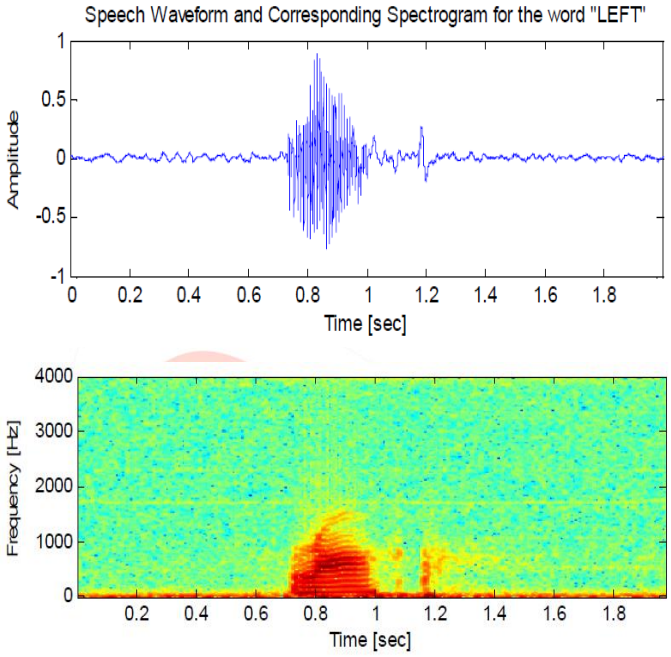


Fig. 12. Speech waveform (top plot) and associated spectrogram (bottom plot) of the word “left”.

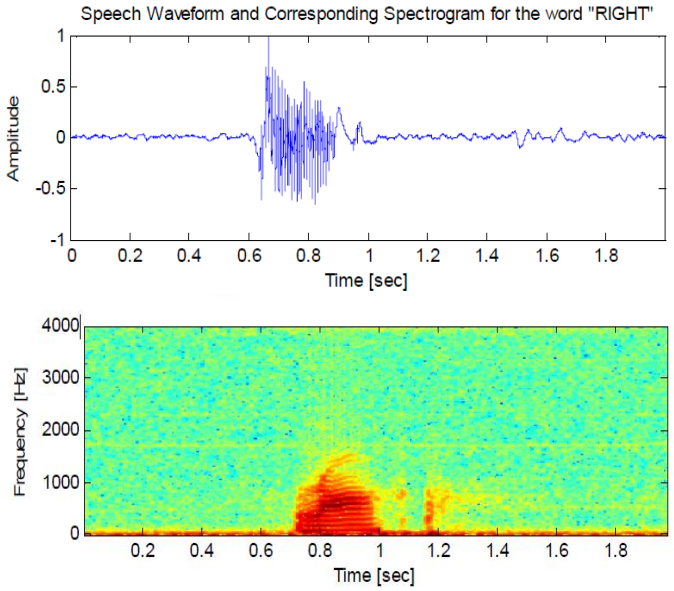


Fig. 13 : Speech waveform (top plot) and associated spectrogram (bottom plot) of the word “right”.

Fig. 14 shows the MSE in the training phase. Mean Squared Error (MSE) is the average squared difference between outputs and targets. Lower values are better.

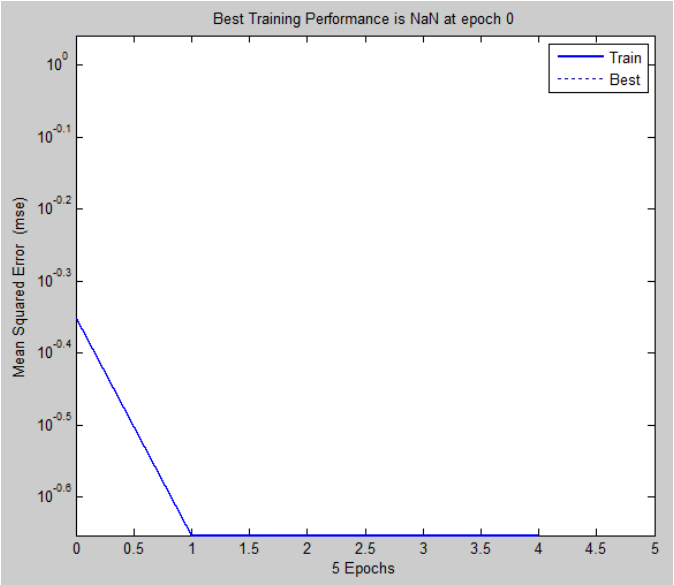


Fig.14.Mean Squared Error (MSE)

VII. CONCLUSION

In this paper we have used MFCC and Neural Network for speech recognition. The whole experiment has been implemented on MATLAB R2010a using Neural Network toolbox and it successfully recognizes speech. The simulation shows high accuracy in result. Further , improvement can be made in this method which will yield more accurate and precise result.

ACKNOWLEDGMENT

I am really thankful to my guide without which the accomplishment of the task would have never been possible. I am also thankful to all other helpful people for providing me relevant information and necessary clarifications.

REFERENCES

[1] David Dean "Synchronous HMMs for Audio-Visual Speech Processing" PhD thesis, Queensland University of Technology, July 2008.
[2] R. P. Lippmann, "Speech recognition by machines and humans" Speech Commun., vol. 22, pp. 1–15, 1997.
[3] Vimala.C,, Dr.V.Radha "A Review on Speech Recognition Challenges and Approaches" WCSIT Vol. 2, No. 1, pp 1-7, 2012.
[4] Hamdy K. Elminir, Mohamed Abu ElSoud, L. M. Abou El-Maged "Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition", International Journal of Science and Technology, Volume 2 No.10, October 2012.
[5] T.B.Adam, Md Alam "Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks" IJCA, Volume 42– No.12, March 2012.
[6] Daniel Jurafsky, James H. Martin, (2008) "Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition", Pearson Education, Second Edition.
[7] Ozdas, A., Shiavi, R.G., Wilkes, D.M., Silverman, M., Silverman, S. ,"Analysis of Vocal Tract Characteristics for Near-term Suicidal Risk Assessment", Meth. Info. Med., vol. 43, pp 36-38, 2004.
[8] Ren Tianping. Application of speech recognition technology [J]. Henan Science and Technology, 2005.
[9] Yin Peng, Li Tao, Wang Haibing. Intelligent neural network system composed of the principle in speech recognition. Mini-Micro Systems, 2000,21(8):836-839.
[10] Yangshang Guo, Yang Jinlong. The speech recognition technology overview [J]. Computer, 2006.