Designing a Real Time Speech Recognition System using MATLAB

Neha Sharma
Student ME-EC
Department of Electronics & Communications,
Chandigarh University, Mohali, Chandigarh, India

Shipra Sardana
Assistant Professor
Department of Electronics & Communications,
Chandigarh University, Mohali, Chandigarh, India

ABSTRACT
Real time speech to text conversion system introduces conversion of the uttered words instantly after the utterance. This paper introduces a unique way of interaction of human and a computer through the specific way ofnatural language processing which is basically a speech recognition system. In this paper nine voice samples were recorded through a microphone and the system was trained according to the recoded voice samples. MFCC features of speech sample were calculated and words were distinguished according to the energies associated with each sampled word. This system provides a high accuracy in case of text conversion.

Keywords
Speech Recognition System

1. INTRODUCTION
Speech recognition is an important application of Natural Language Processing (NLP). Speech is the most important part of communication. We express our ideas through a specific language.Computers understand our language (natural language) by speech recognition. Speech or word by word recognition is the process of automatically extracting and determining linguistic information conveyed by a speech wave using computers. Linguistic information, the most important information in a speech wave, is called phonetic information. The term speech recognition means the recognizing the spoken words only. However, the recognition system has no idea what those words mean. It only knows that they are words and what words they are. To be of any use, these words must be passed on to higher-level software for syntactic and semantic analysis. It is a technique of pattern recognition, where acoustic signals are tested and framed into phonetics (number of words, phrases and sentences)[1].To perform such task one needs to record a voice sample and then convert this voice sample into .wav format. Spectrum based parameters are obtained when a word is recognized. Near about twenty four parameters can be obtained in the analysis of spectrum of speech signal.

These parameters are mean, median, standard deviation(STD), root mean square(RMS), maximum peak, minimum peak, slope of the maximum peak, width of maximum peak, signal to noise ratio, peak frequency, peak amplitude, total power, total harmonic distortion(TDH), total harmonic distortion(TDH)+noise, inter modulation distortion(IMD) etc. Various statistical methods are used for the analysis of words which give some specific value of words. Words fluctuate between in its bounded range of occurrence. In the improvement of word recognition process, one of the important tasks is to find the most informative parameters of speech signal. To perform such tasks some of the techniques are used namely, Linear Predicted coding coefficients(LPC) and Mel Frequency Cepstrum Coefficients(MFCC)[2]. By using such techniques new spectrum is obtained that is different from the previous spectrum of spoken words.

2. DESIGN AND DEVELOPMENT OF OUR SPEECH RECOGNITION SYSTEM
The development of our speech recognition system is divided in two stages, first is training stage and second is the testing stage.

Training stage

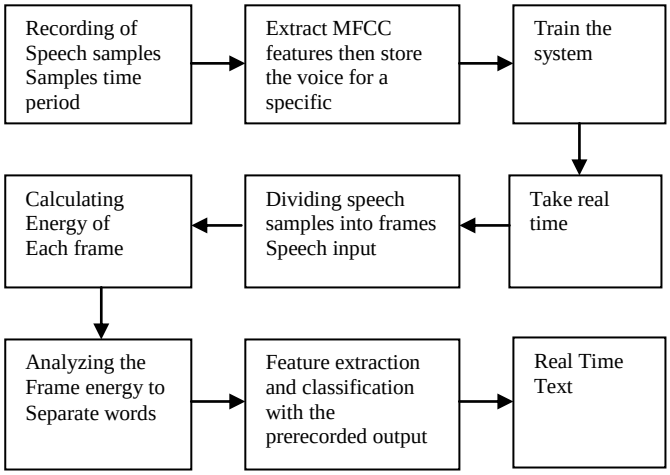

Figure 2.1: Speech recognition process

3. TRAINING STAGE
In this stage a database is created by recording some speech samples by the user. Then the recorded speech samples are stored into .wav format in Matlab. After this stage it is necessary to train the speech recognition system.

3.1 Feature Extraction
Feature extraction converts the speech waveform into some parametric information for further analysis and processing. This is often referred as the signal-processing front end. The speech signal is a slowly time varying signal.When examined over a sufficiently short period of time, its characteristics are fairly stationary. However, over long period of time (on the order of 1/5 seconds or more) the signal characteristic change to reflect the different speech sounds being spoken. Therefore, short-time spectral analysis is the most common way to characterize the speech signal. For that use MFCC features are used.

3.1.1 Mel Frequency Cepstral Coefficients (MFCCS)
We used MFCC features for this system.The word 'Mel' in the MFCCs represents the melody of a speech signal. MFCC features are based on the human ear perception which means

human's ear's critical bandwidth frequencies filters the spaced linearity between the high frequency and low frequency of the speech signal and capture the useful information of that particular signal. The human perception for the frequency contents of the speech signals follows a nonlinear scale. That's why pitch is measured on a scale which is actually a Mel scale.TheMel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [3].

MFCCs are calculated as follows:

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f / 700) \quad (1)$$

3.2 CREATING THE DATABASE

To recognize the uttered word of the speaker, a database is created to resemble the pronounced word. To create such database, we first recorded some numerals from one to nine and achieved following plots:

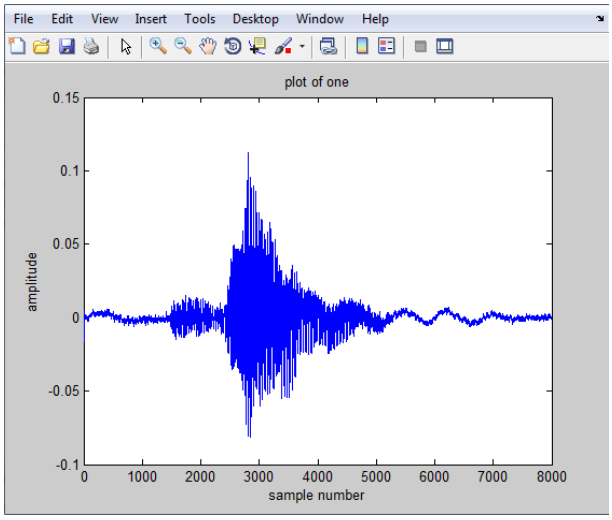

Figure 3.2.1: spectrum of 'one'

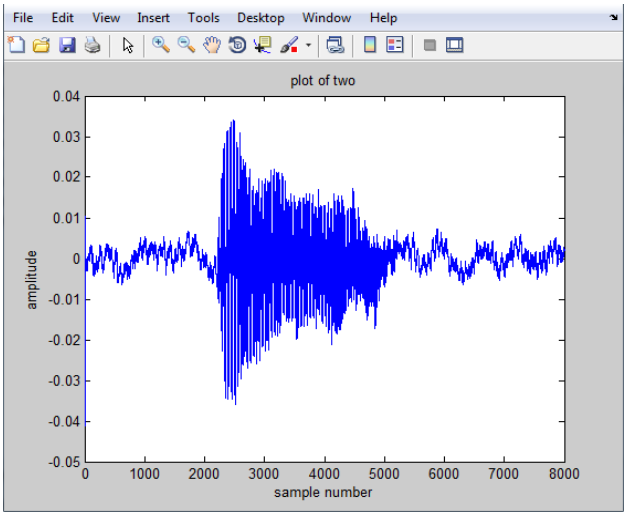

Figure 3.2.2: spectrum of 'two'

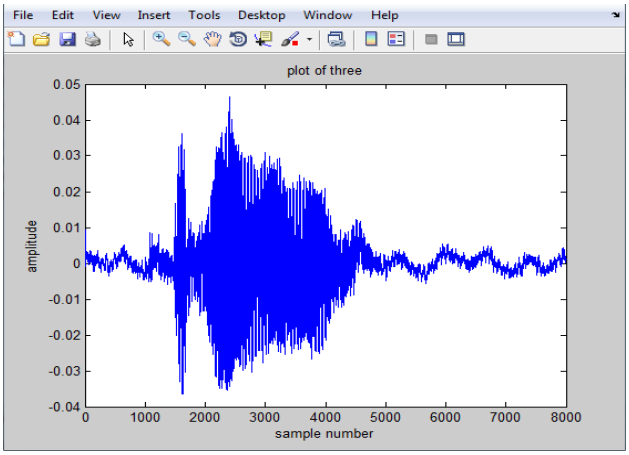

Figure: 3.2.3: Spectrum of 'three'

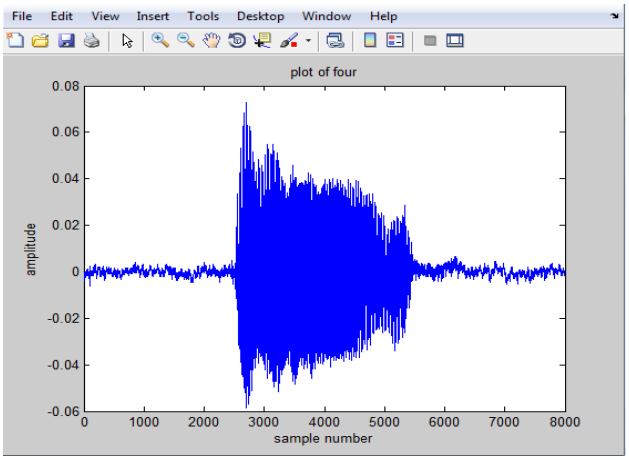

Figure 3.2.4: Spectrum of 'four'

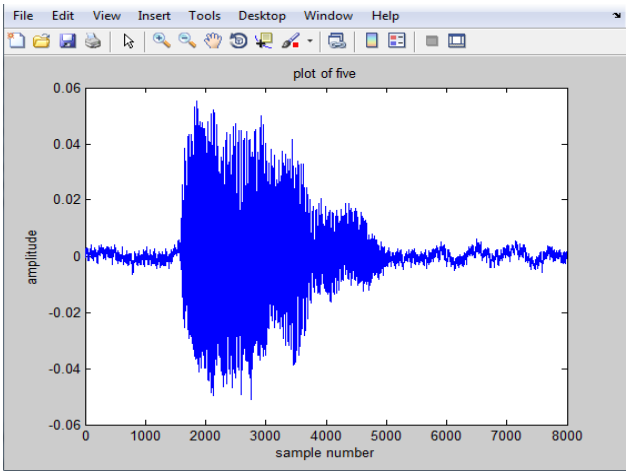

Figure 3.2.5: Spectrum of 'five'

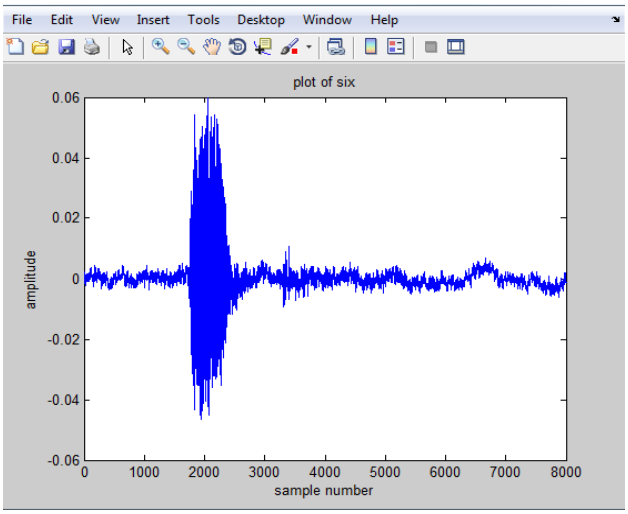

Figure3.2.6: spectrum of 'six'

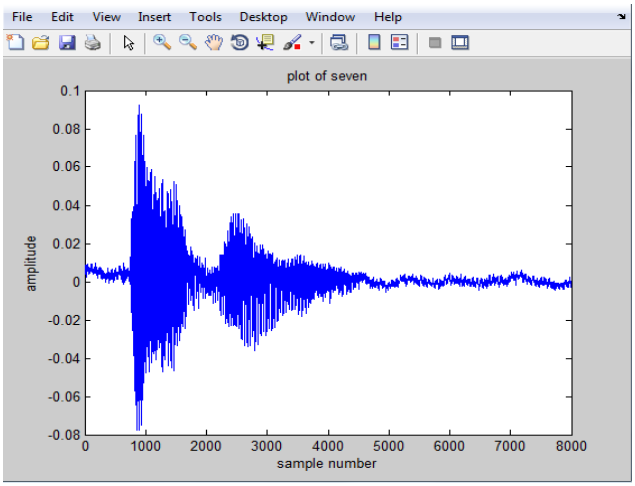

Figure3.2.7: spectrum of 'seven'

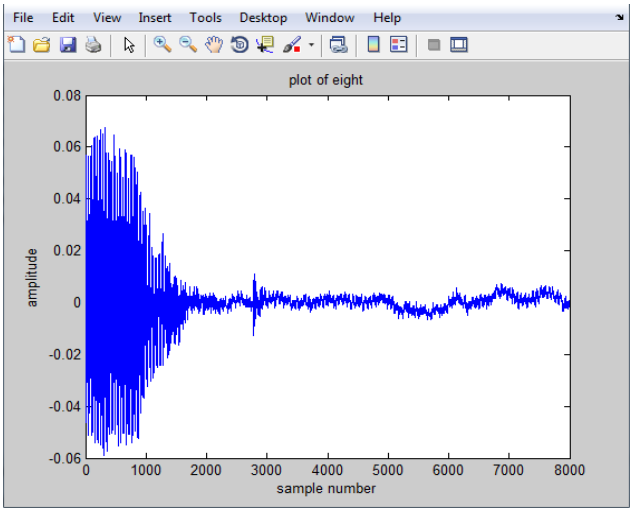

Figure3.2.8: Spectrum of 'Eight'

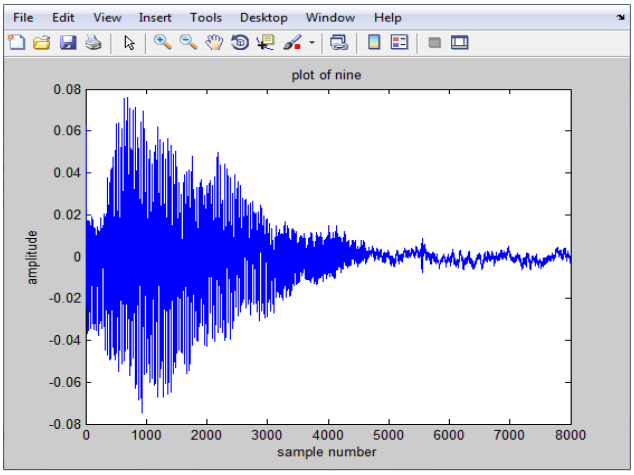

Figure3.2.9: spectrum of 'nine'

3.3 TRANING OF THE VOICE SAMPLES

Speech recognition system is trained before use. This training of the speech samples is a necessary part of the speech recognition system. We have trained our speech samples at sampling frequency 8khz.

The duration of the training can be varied from 20s.After the training of the speech samples the system will separate the frames of speech signal with high energy and the speech signal with low energy. As the figure:3.3 show the training sequence of speech samples. The plot of training of voice samples:

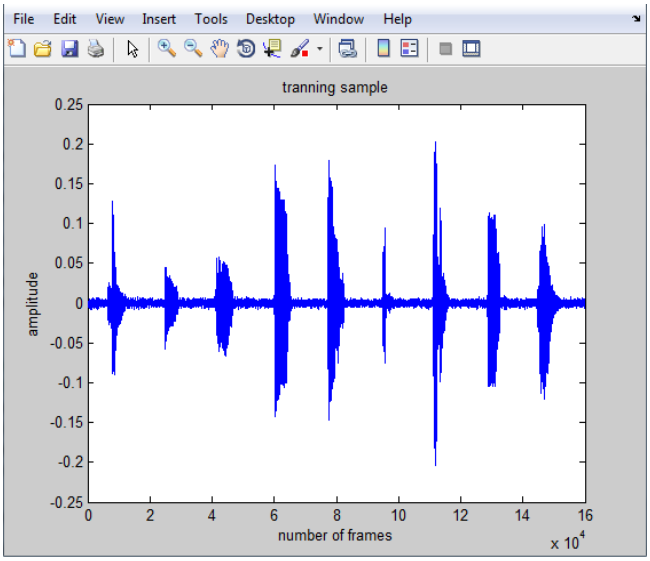

Figure 3.3: Training of the speech samples

4. EXPERIMENTAL TESTING

Our speech recognition system was a speaker dependent system. So it was dependent on the user's voice only. In the training of this system we created a database of nine words. After the training of this system, a real time speech input was given to it through a good quality microphone. The system divided the real time speech sample into small segments of frames or continuous groups of samples. After that the energy of each frame segment was calculated using simple energy formula:

$$E_x = \int_{-\infty}^{\infty} x^2 \, dx \qquad (2)$$

Energy calculated was then analyzed by a speech detection algorithm to separate the words.

4.1 SPEECH DETECTION ALGORITM

The speech detection algorithm was developed by processing the prerecorded speech samples frame by frame within a simple loop. We divided the whole frame into the segment of 160 samples and each of the samples was detected by the system. For the detection of each frame we used a combination of signal energy and a zero crossing rate. This calculation became very simple with the MATLAB mathematical and logical operators.

4.2 ACOUSTICAL MODEL

It is very important to create an acoustical model for the detection of each uttered word. So we created an acoustical model. It is known that different sounds are produced by human vocal cord and different sounds can have different frequencies. To predict the different frequencies it power spectral density measure can be a better way. So we find out the frequencies by power spectral densities measure.

Speech can be termed as short term stationary so MFCC features were again extracted and words pronounced by the user were detected.

5. RESULTS

Real time results were obtained in the lab. The user was speaking through the microphone and the text representation was obtained on the computer screen as shown in the figure 5.1. Implementation results of speech to text conversion system are as follows:

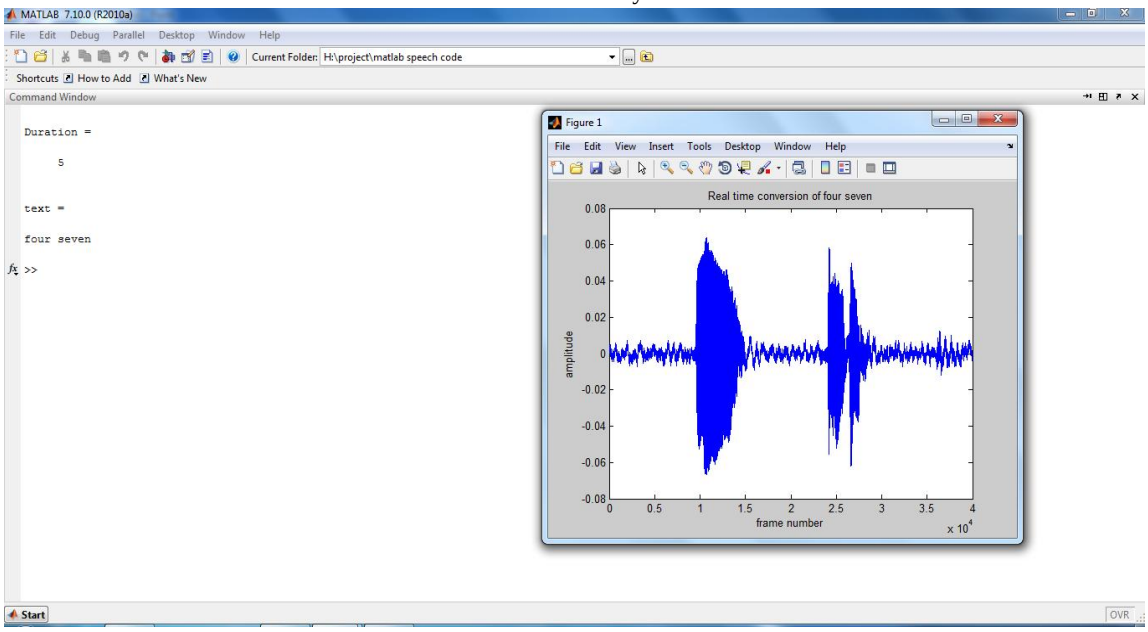

Figure 5.1: STT conversion of four and seven

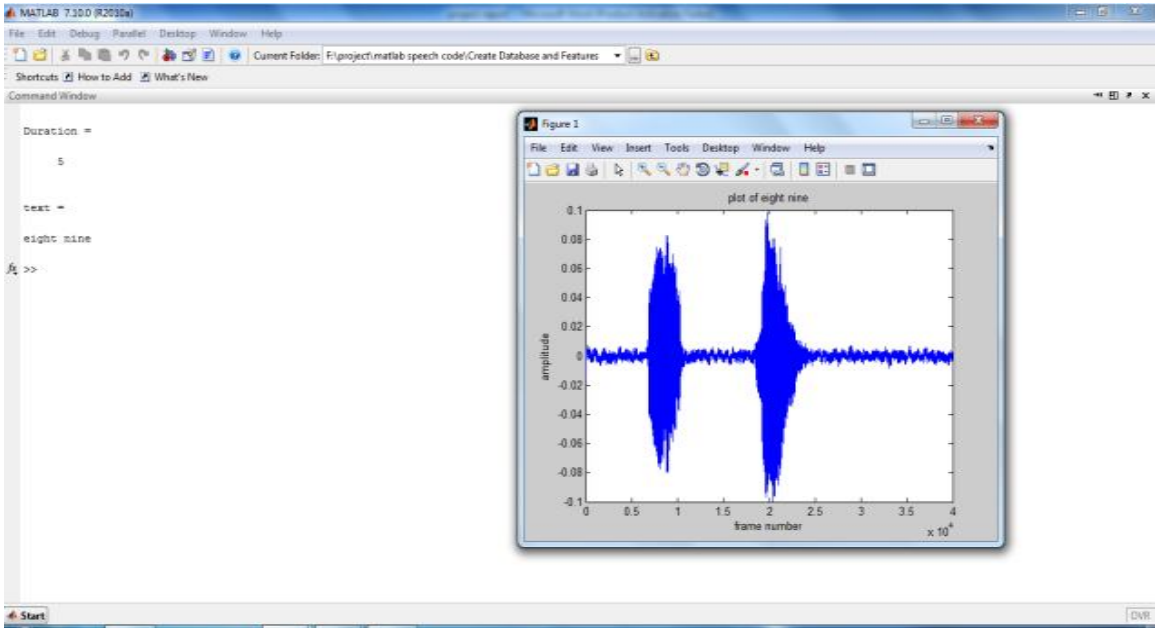

Figure 5.2: STT Conversion of Eight and Nine

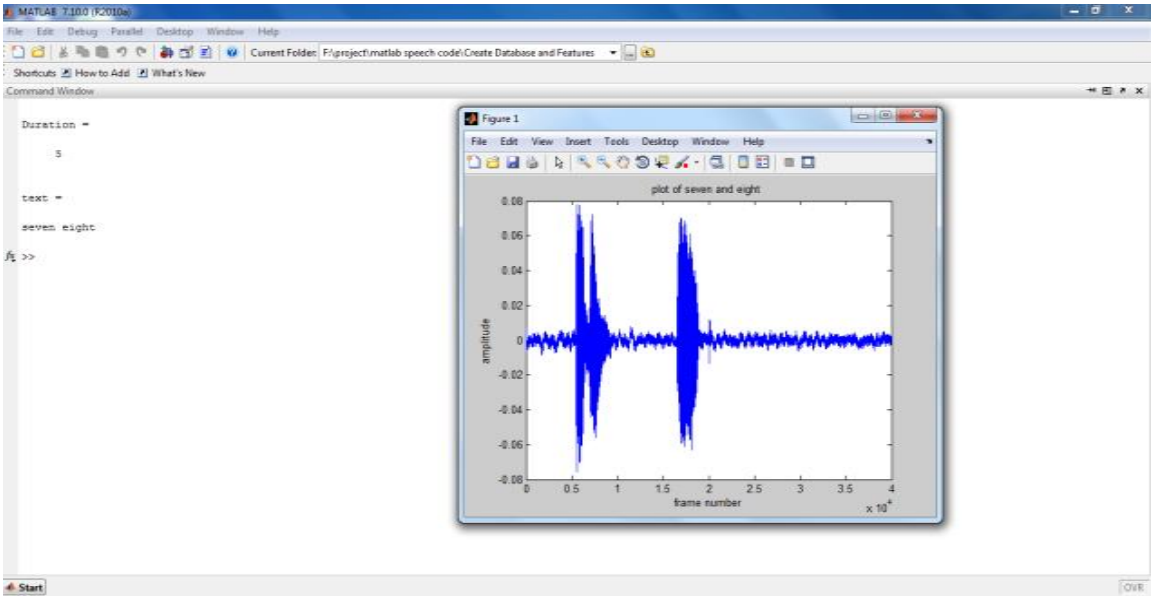

Figure 5.3: STT Conversion of Seven and Eight

6. CONCLUSION

In this project nine words were collected and analyzed. Words were distinguished by energies associated with them. The system was able to separate the words according to their energies. Final output comes out in the form of text. By using this code the system can be trained for more words and paragraphs. Every word parameter has their bounded values in whichthat parameter varies. Each word has some specific range of these parameters. Some words are same but they still have some same parameters which tell us about the word. e.g. in speech seven and one. Words are similar and have some parameters which are same. Here the word seven contains one at the end. So it sounds same sometimes, and the system gives output 'one' when seven is pronounced sometimes. Such type of ambiguities can be removed with large number of samples taken for one particular word.

This system is also very sensitive to noise. In future we can work for this task. Also this system is very sensitive to word pronunciation during training. Words that we have recorded to create a database and the words during training should be pronounced similarly. System is sensitive to tone of pronunciation.

7. REFERENCES

[1] J. D. Tardelli, C. M. Walter, "Speech waveform analysis and recognition process based on non-Euclidean error minimization and matrix array processing techniques". IEEE ICASSP, pp. 1237-1240, 1986.

[2] Takao Suzuki, Yasuo Shoji, "A new speech processing scheme for ATM switching systems". IEEE, Digital Communications Laboratories, Oki Electric Industry Co. Ltd., Japan, pp. 1515-1519, 1989.

[3] Siva PrasadNandyala, Dr.T.Kishore Kumar,"Real Time Isolated Word Speech Recognition System for Human Computer Interaction"InternationalJournal of Computer ApplicationsVolume- 12, pp.0975 – 8887, November 2010.

[4] Jeong, S., Hahn, M.: "Speech quality and recognition rate improvement in car noiseenvironments", Electron. Lett.,37, (12),pp. 800–802,2001.

[5] Ma, J., Deng, L.: "Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model", IEEE Trans.Speech Audio Process., 11, (6), pp. 590–602, 2003.

[6] RohitRanchal, , Teresa Taber-Doughty, YirenGuo, Keith Bain,Heather Martin, J. Paul Robinso, and Bradley S. Duerstock, "Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom", IEEE Transactions On Learning Technologies, Vol. 6, No. 4, October-December 2013.

[7] Daryl Ning, "Developing an isolated word recognition system in MATLAB", The Mathworks, Inc. 2009.

[8] Deepak Baby, Tuomas Virtanen, Jort F. Gemmeke, Hugo van hamme, "Coupled dictionaries for Exampler- Based Speech Enhancement and Automatic Speech Recognition", IEEE/ACM Tans. On Audio, speech and Language processing, vol. 23, No. 11, 2015.

[9] Naoki Hirayama, Koichiro Yoshino, KatsutoshiItoyama, Shinsuke Mori, and Hiroshi G. Okuno, "Automatic Speech Recognition for Mixed Dialect Utterances by Mixing Dialect Language Models", IEEE/ACM transactions on Audio, Speech, And Language Processing, vol. 23, no. 2, february 2015

[10] Shaila D. Apte, "speech and audio processing", Wiley India, 2013