Speech Recognition System to Leverage the Accuracy of Training Sample using Optimized Matching Window

Gunjan Thakur
M.Tech
Research Scholar
Department of Electronics and Communication
RIEIT, Ropar

Anudeep Goraya
Associate Professor
Department of Electronics and Communication
RIEIT, Ropar

ABSTRACT
In this voice recognition system is to recognize the voice samples spoken by human and recognize over the system. In this select the most commanly used features of the voice samples with the help of MFCC(Mel frequency cofficient ceptrum) and that feature match with the real time voice sample features using DTW(Dynamic time wrapping) and it is accepted by the system.

Keywords
Dynamic Time Wrapping (DTW),Mel Frequency Cepstral Coefficient (MFCC),Voice recognition.

1. INTRODUCTION
In voice recognition the system has to recognize the every word in the voice sample that is spoken by human in the exact manner and with same tone and pitch of the voice sample and that presents the best output when extract the features of voice samples and make the data set of these voice samples as trained set of voice sample. Voice recognition system use the trained data set when it is dependent and when there is no need of training data set then it is independent and for proposed work use the training data set to get the best results and extract the mostly used features and then these features is match with the real time voice sample using dynamic time wrapping and is also called as the matching algorithm. It is defined as best matching of the samples at different interval and acceleration. Voice recognition is useful in various applications like education purpose, army and medicated area, TV, phone, computer system and for navigation. In this paper section 2 describes the methodology used for proposed work. Section 3 presents the experimental results of proposed work and section 4 shows the conclusion and future scope.

2. METHODOLOGY
In proposed algorithm have two section, first is MFCC (Mel-frequency cepstral coefficient) is used for the features extraction in which extract the most commonly used features of training part and testing part and the second part for features matching of training sample and testing sample using dynamic time wrapping.

2.1 Algorithm
Step 1: Take N numbers of voice samples pass through the high pass filter for amplification of the signal which is compress during the production of voice sample from human then is used to remunerate the signal.

Step2: Filtered sample is divided into frames. Every frame consists of N samples and frames are equally spaced and the voice sample in the frame is as shown below:

$$r(n)= n= 0,1,2\ldots N-1$$

Step3: Every frame is multiplied with the hamming window for continuity occurs at first and last point in the frame. The use of hamming window is to make the peaks in the frequency response is well defined and pointed. The mathematically equation of hamming window equation is given below:

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

The sample in the frame r(n) is multiplied with the hamming window w(n) is given below equation 2.

$$y[n] = r[n]*w[n] \quad (2)$$

where r[n]=voice sample in the frames.

w[n]= hamming window.

y[n]=output of voice sample.

Step4: Take fourier transform of the sample to change the domain of the signal from time domain to frequency and the fourier transform of the sample is given below:

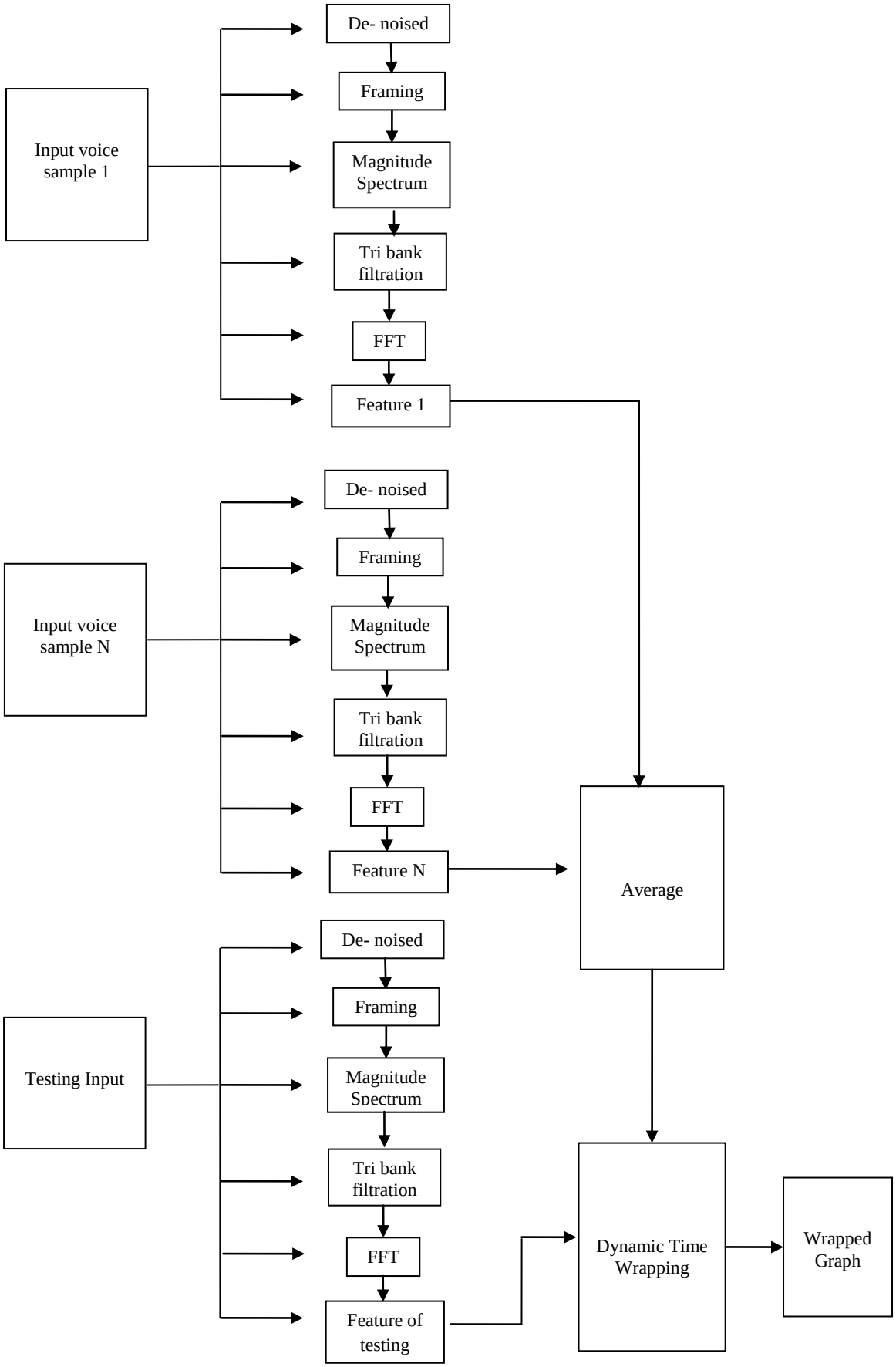

Figure 1: Flow chart of our proposed algorithm.

$$N_k = \sum_{m=0}^{M-1} n_i \, e^{-j\frac{2\pi k_m}{M}} \qquad (3)$$

where N_k= fourier transform of the voice sample.

n_i = voice sample.

A sample in the frames is periodic and continuous in nature when fourier transform is perform in the frames for magnitude response of the signal. If the voice samples does not shows the continuity then it effect on the frequency response and to minimize the frequency response of the voice samples then multiplied with hamming window shown in above step to maintain continuity in the first and last point of the frame.

Step5: Use Mel frequency to convert the frequency domain into Mel frequency scale. In this first take the magnitude response of the voice sample and after that multiplied with triangular band pass filter banks to give the appropriate results understand by human and triangular filter banks are use for smooth results of magnitude spectrum.

$$M(f)=2595*\log_{10}(1+\frac{f}{700}) \qquad (4)$$

where M(f)= mel frequency.

f= linear frequency.

equation 4 shows that the mel frequency is directly proportional to the log of linear frequency.

Step6: For large calculation FFT is used because it computes the calculation as there is large calculation in the DFT to avoid this calculation use of FFT and after that extracts the features of the voice samples and compare with testing voice sample there is large calculation.

Step 7: Dynamic time wrapping is used to match the features of training voice sample with the testing voice sample at different time and speed and then wrapped the voice samples. The graph of testing sample features and training data set features and they are very similar make optimal match.

3. EXPERIMENTAL RESULTS

3.1 Training Phase: In this training phase to remove the noise and background voice use filtration process that filtered the voice samples.

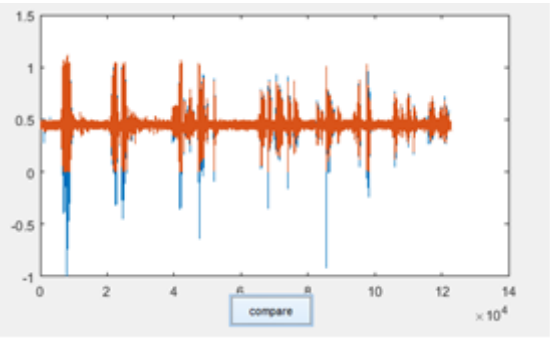

Figure 2: Comparison of filtered and noisy sample.

The noisy version of voice sample compares it with filtered voice sample. The brown part of the sample shows the filtered part and blue is noisy part (see Figure 2).Then framing of the voice sample and use hamming window for multiplication with the voice sample and accommodate periodicity at starting and ending of voice sample in the frame (see Figure 3).

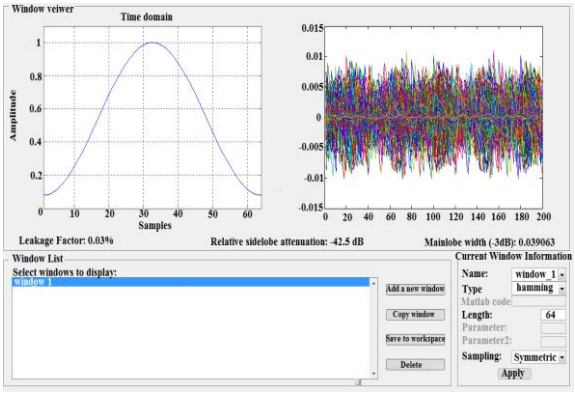

Figure 3: Framing of voice sample

Triangular filter banks are use for providing the continuity in magnitude spectrum and are unity at the central frequency and reduce to zero when two filter banks are adjoining. Fast fourier transform is used for computation of large calculation that occur in discrete fourier transform (see Figure 4) and extract the mostly used features of voice sample. Same process occurs for next voice samples (see Figure 5). Average the features of voice samples make the training set of these voice samples (see Figure 6).

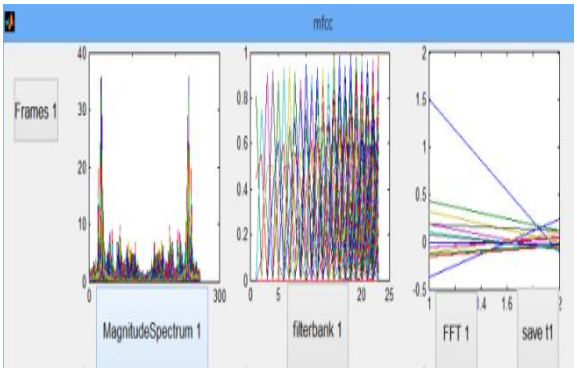

Figure 4: Magnitude spectrum, filter banks and FFT of voice sample 1

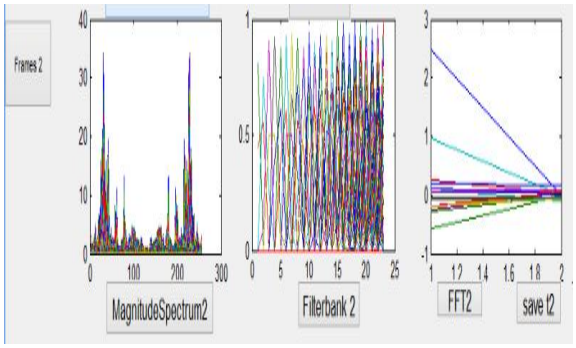

Figure 5: Magnitude spectrum, filter banks and FFT of voice sample 2

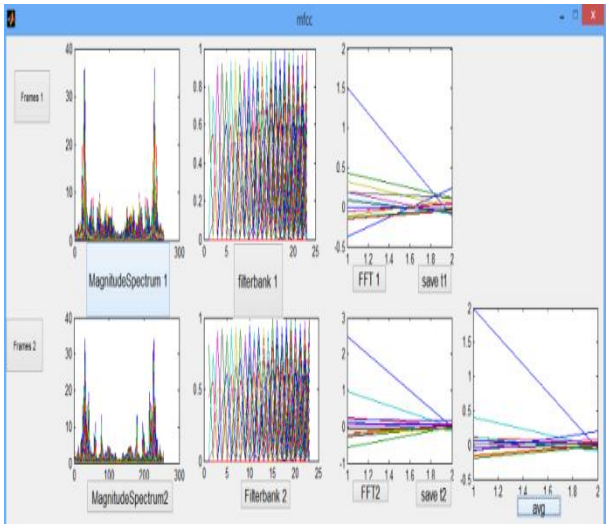

Figure 6: Average of voice samples.

After that the spectral difference and the frame error rate calculated and this error is less than that of previous work and its value is 5.0045 and 18.4443 respectively (see Figure 7).

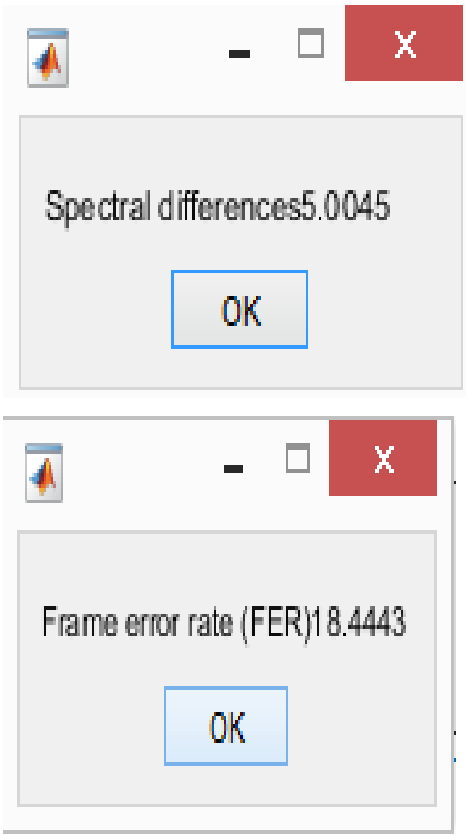

Figure 7: Calculated spectral difference and frame error rate.

3.2 Testing Phase

The testing phase consist of real time voice sample and then filtered the voice sample and same process is occur as describe in training phase of voice sample(see Figure 8) and extract the most commonly used features. Compare it with the training data set of voice sample.

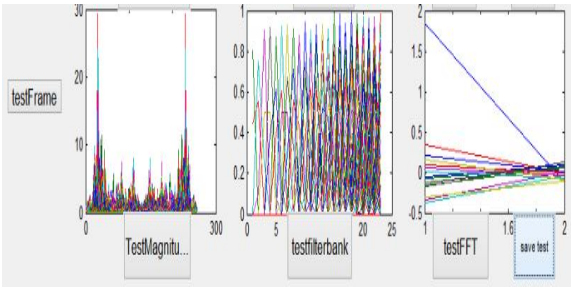

Figure 8: Magnitude spectrum, filter banks and FFT of testing sample

3.3 Classification

Now calculate the dynamic time warping and need to show the matching results of the training sample and testing sample (see Figure 9). There are resemblances between the training samples and the testing sample which show the linearity and the graph of warped signals. Dynamic time wrapping gives the perfect matched result at different time and speed of voice sample.

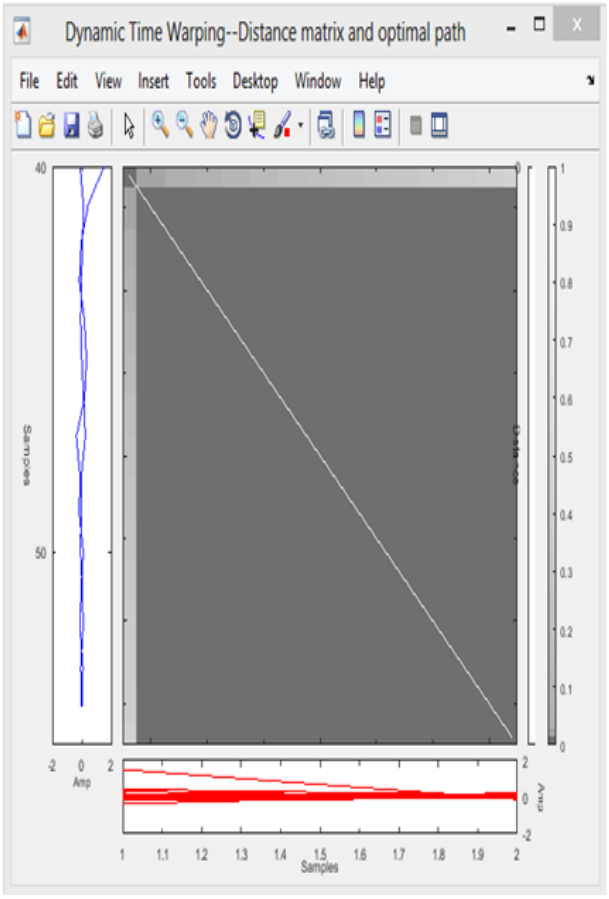

Figure 9: Matching voice samples

The almost similarity between the testing and training samples as seen that both of the signals color in blue and red go beyond each other (see Figure 10). The voice samples use dynamic time wrapping for optimal match and voice sample features overlaps each other and word error calculated for voice sample is 0.40969(see Figure 11).

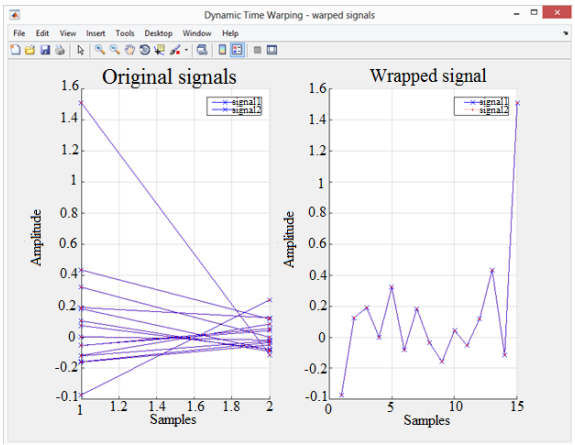

Figure 10: Wrapped voice sample

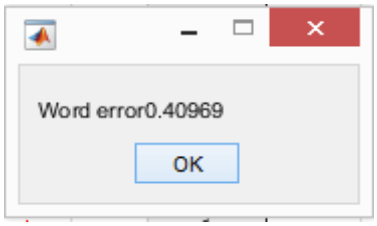

Figure 11: Calculate word error

In the proposed work frame error rate in noisy and noiseless environment is 9.4443 and 4.3212 respectively which is comparatively low than the previous work. The word error results is also low in noisy and noiseless environment is 2.3838 and .40969 are better than the previous work (see table 1).

Table 1. Calculated Errors

S.No	Previous Results	Proposed work
1	Frame error (noisy) 11.7	Frame error (noisy) 9.4443
2	Frame error (noiseless) 12	Frame error (noiseless) 4.3212
3	Word error(noisy) 30	Word error(noisy) 2.3838
4	Word error (noiseless) 28.2	Word error (noiseless) .40969

4. CONCLUSION AND FUTURE SCOPE

The main motive of the proposed work is to remove the noise from the voice sample. In past work there is occurrence of noise in the voice sample. In this extract the most commonly used features and to give optimal match using feature of training data set and real time voice sample. For matching features use the dynamic time wrapping and calculate the frame error rate word error rate and spectral difference. Take sample at different environment like living room, meeting room and class room with noisy and noiseless background and the results shows that the noisy environment has more errors than noiseless (see Figure 12).

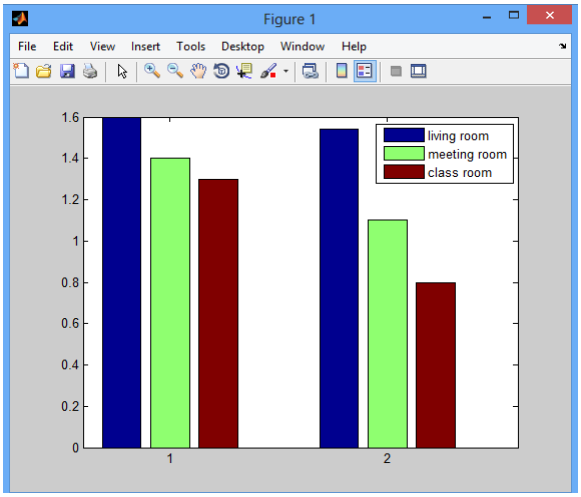

Figure 12: Comparison of noisy and de-noised environment

For future wide the area of research, work on sentence recognition which is efficient as a biometric for sensitive data security, for matching algorithm technique use adaptive learning technique.

5. ACKNOWLEDGMENT

My sincere thank and gratitude to my esteemed and honorable supervisor Anudeep Goraya Associate Professor. She has not only been very kind to me but has always stood as an epitome of knowledge and encouragement. My special thanks to Dr. Ramesh Chand Kashyap, Head of the Department of Electronics and communication Engineering, for his support and providing all those facilities required for the completion of this thesis. Also thankful to the faculty and staff members of Electronic and Communication Engineering Department for helping me out in one way or the other

6. REFERENCES

[1] Yogesh Kumar Sen, R. K. Chaurasiya. IEEE International Conference on voice rcoginition-june2014,24:58-95.

[2] Daubechies, I. The wavelet transform, time-frequency localization andsignal analysis. IEEETransformation and Information Theory.2014,36: 961-1005.

[3] Hasan Serhan Yavuz, Hakan Çevikalp .A wavelet Tour of Signal ProcessingIEEE International Conference on signal processing june 2014,34:19-445.

[4] Tiecheng Yu. The Development State of the Voice Identification. The Development communication world.2005,2:56-59.

[5] Dian RetnoAnggraini .The development of a voice recognition system based on Principal Component Analysis(PCA) and unsupervised learning algorithm.2012,4:35-58.

[6] Jiqing Han, Lei Zhang, Tieran Zheng. Voice Signals Processing[M].Beijing: Tsinghua University Press 2004,3:67-94.

[7] Remzi Serdar Kurcan, "Isolated word recognition from in-ear microphone data using hidden markov models (hmm)", Master's Thesis, 2006.

[8] Nikolai Shokhirev ,"Hidden Markov Models ", 2010.

[9] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proceedings of the IEEE Journal, Feb 1989, Vol 77, Issue: 2.

[10] Suma Swamy, Manasa S, Mani Sharma, Nithya A.S, Roopa K.S and K.V Ramakrishnan, "An Improved Speech Recognition System", LNICST Springer Journal, 2013.

[11] Lindasalwa Muda, Mumtaj Begam and Elamvazuthi.,"Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and DTW Techniques ",Journal of Computing, Volume 2, Issue 3, March 2010.

[12] Mahdi Shaneh and Azizollah Taheri ,"Voice Command Recognition System based on MFCC and VQAlgorithms", World Academy of Science, Engineering and Technology Journal , 2009.

[13] Remzi Serdar Kurcan, "Isolated word recognition from in-ear microphone data using hidden markovmodels (hmm)", Master's Thesis, 2006.

[14] Nikolai Shokhirev ,"Hidden Markov Models ", 2010.

[15] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in Speech Recognition", Proceedings of the IEEE Journal, Feb 1989, Vol 77, Issue: 2.

[16] Suma Swamy, Manasa S, Mani Sharma, Nithya A.S, Roopa K.S and K.V Ramakrishnan, "An Improved Speech Recognition System", LNICST Springer Journal, 2013.

[17] M. L. Shire and B. Y. Chen, "Data-driven RASTA filters in reverberation,"in *Proc. ICASSP'00*, 2000, vol. 3, pp. 1627–1630.

[18] T. Takiguchi and Y. Ariki, "Robust feature extraction using kernel PCA," in *Proc. ICASSP'06*, 2006, pp. 509–512.

[19] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech,Signal Process.*, vol. ASSP-34, no. 1, pp. 52–59, Feb. 1986.

[20] O. Ichikawa, T. Fukuda, R. Tachibana, and M. Nishimura, "Dynamic features in the linear domain for robust automatic speech recognition in a reverberant environment," in *Proc. Interspeech'09*, 2009, pp. 44–47.

[21] M. Nakayama *et al.*, "CENSREC-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments," in *Proc. Interspeech'08*, 2008, pp. 968–971.

[22] T. Nishiura *et al.*, "Evaluation framework for distant-talking speech recognition under reverberant environments—Newest part of the CENSREC series-," in *Proc. LREC '08*, 2008.