Analysis of Voice Recognition Algorithms using MATLAB

Atheer Tahseen Hussein
Department of Electrical, Electronic and Systems Engineering
University Kebangsaan Malaysia
Malaysia, 43600 Bangi, Selangor, Malaysia.

***Abstract*— Voice recognition has become one of the most important tools of the modern generation and is widely used in various fields for various purposes. The past decade has seen dramatic progress in voice recognition technology, to the extent that systems and high-performance algorithms have become accessible. Voice recognition system performance is commonly specified in terms of speed and accuracy, recognition accuracy is the most important and straightforward measure of voice recognition performance. This research were proposed to review several voice algorithms in terms of detection accuracy and processing overhead and to identify the optimal voice recognition algorithm that can give the best trade-offs between processing cost (speed, power) and accuracy. Also, to implement and verify the chosen voice recognition algorithm using MATLAB. Ten words were spoken in an isolated way by male and female speakers (four speakers) using MATLAB as a simulation environment, these word were used as a reference signal to trained the algorithm, for evaluating phase, all algorithms dictates to subject them to similar test criteria. From the simulation results, the Wiener Filter algorithm outperform the other four algorithms in terms of all measure of performance, and power requirement with the moderate complexity of the algorithm and its prospective implementation as a hardware. Wiener filter algorithm scored accuracy of 100%, 5%, and 50% for test cases i,ii,and iii respectively, with recognition speed range of (695-867) msec and estimated power range of (750-885) µW.**

Keywords— ***Voice recognition, spectrum normalization, cross-correlation, auto- correlation , wiener filter , hidden markov model, Matlab*.**

I. INTRODUCTION

Voice recognition is a popular theme in today's life. Voice recognition's programs are available which make our life far better. Voice recognition is a technology that the system can be controlled by people with their language. Rather than typing controlling the buttons for the system or the computer keyboard, using language to control system is more suitable. Additionally, it may reduce the price of the business production at the exact same time. The efficiency of the daily life enhances, but also makes people's life more diversified [1]. Voice recognition technology is the process of identifying, understanding and converting voice signals into text or commands. There are different types of authentication mechanisms available today like alphanumeric passwords, graphical passwords etc. Along with these, biometric authentication mechanisms like fingerprint recognition system, voice recognition system, iris recognition system etc. add more security for data. One of the important areas of research is voice recognition technology. Research in voice recognition involves studies in physiology, psychology, linguistics, computer science, signal processing, and many other fields. Voice recognition technology consists of two different technologies such as speaker recognition and speech recognition [2]. Speech recognition is a technique that enables a device to recognize and understand spoken words, by digitizing the sound and matching its pattern against the stored patterns. Speech is the most natural form of human communication and speech processing has been one of the most exciting areas of the signal processing. Speech recognition technology has made it possible for computer to follow human voice commands and understand human languages [1]. Speech recognition algorithms can be in general divided into speaker dependent and speaker independent. Speaker dependent system focuses on developing a system to recognize unique voiceprint of individuals. Speaker independent system involves identifying the word uttered by the speaker [3].

A. Types of speech recognition

There are three types of ASR depends on speakers, size of vocabulary, and bandwidth are the different basis on which researchers have worked[4]; these types are classified as folowing:

- Isolated word.
- Connected word.
- Continuous word.
- Spontaneous word.

This research concentrates on the isolated word speech recognition. Isolated word speech recognition system requires the user to pause after each utterance. The system composes of two phases; training phase and recognition phase. During the training phase, a training vector is generated for each word spoken by the user. The training vectors extract the spectral features for separating different classes of words. Each training vector can serve as a template for a single word or a word class. These training vectors (patterns) are stored in a database for subsequent use in the recognition phase. During the recognition phase, the user speaks any word for which the system was trained. A test pattern is generated for that word and the corresponding text string is displayed as the output using a pattern comparison technique [5].

II. LITERATURE REVIEW

Many algorithms have been proposed to implement speech recognition. These methods are autocorrelation, cross-correlation, spectrum normalization, Wiener Filter, and hidden Markov model. This research investigates several approaches for implementing the speech recognition system of isolated words "Digit" using the built-in functionality in Matlab and related products to develop the algorithms associated with each approach. These algorithms are as follows: spectrum normalization, cross –correlation, auto- correction, FIR Wiener filter, and hidden Markov model. This part of the thesis focuses on the theoretical framework for these algorithms.

A. Spectrum normalization(SN)

Comparing spectrums in different measurement standards could be difficult when comparing the differences between different speech signals. Hence, the use of normalization could maintain the measurement standard. In other words, the normalization reduces the error when comparing the spectrums. Before analyzing the spectrum differences for different words, the first step is to normalize the spectrum by the linear normalization. After normalization, the values of the spectrum are obtained or normalized spectrum is set to the interval [0, 1]. Normalization changes the value range of the spectrum, but does not change the shape or the information of the spectrum itself. After the normalization of the absolute values of FFT, the next step in programming the speech recognition is to observe spectrums of the recorded speech signals. Then, the algorithms are compared based on the differences between the test or target signal and the training signals or reference signals [6].

B. Cross – correlation(CC)

Assuming the recorded speech signals for the same word are the same, the spectrums of two recorded speech signals are also similar. When doing the cross-correlation of the two similar spectrums and plotting the cross-correlation, the cross-correlation should be symmetrical according to the definition of the cross-correlation. After calculating the cross-correlation of two recorded frequency spectrums for the speech recognition comparison, find the position of the maximum value of the cross-correlation and use the values right to the maximum value position to minus the values left to the maximum value position. Then, take the absolute value of this difference and find the mean square error of this absolute value. The cross-correlation symmetry of two signals indicates the matching level of both signals. Moreover, the more symmetric is the cross-correlation, the smaller is the value of the mean square error. By comparing the mean square errors of processing trained words and the target word, the system decides the training word that has a better match with the test signal based on the minimum mean square error of the cross-correlation differences at *different lags*[7]**.**

C. Autocorrelation (AC)

Autocorrelation can be treated as computing the cross-correlation for the signal and itself instead of the correlation with different signals. Autocorrelation algorithm aims to measure how the signal is self-correlated. Thus, training signal autocorrelations are compared to find the minimum difference between autocorrelations [8].

D. Wiener filter(WF)

FIR Wiener filter is used to estimate the desired signal from the observation process to get the estimated signal. The desired or test and training signals are assumed to be correlated and are jointly wide-sense stationary. The purpose of Wiener filter is to choose the suitable filter order and to find the filter coefficients with which the system can get the best estimation. In other words, the system can minimize the mean-square error with proper coefficients. The recorded signals are wide-sense stationary processes. Then, the reference or training signals can be used as input signals, and recorded test signals can be used as desired signals [9].

E. Hidden markov model(HMM)

A hidden Markov model (HMM) is a triple (π, A, B) where each probability in the state transition matrix and the emission matrix is time independent, which means that the matrices do not change in time as the system evolves. In practice, this model is one of the most unrealistic assumptions of Markov models about real processes. HMMs, described by a vector and two matrices (A, B), which are of great value in describing real systems. Although, the model is usually an approximation, the results are amenable to analysis. The commonly solved problems are [10]:

- Matching the most likely system to a sequence of observations–evaluation, and solved using the forward algorithm.
- Determining the hidden sequence most likely to have generated a sequence of observations – decoding, solved using the Viterbi algorithm;
- Determining the model parameters that are most likely to have generated a sequence of observations – learning, and solved using the forward-backward algorithm.

F. Related works.

Isolated speech recognition by mel-scale frequency cepstral coefficients (MFCCs) and dynamic time warping (DTW) have been proposed. In these studies, several features were extracted from a speech signal of spoken words. DTW is used to measure the similarity between two sequences that may vary in time or speed. The experimental results were analyzed using Matlab and were proven to be efficient. DTW is the best nonlinear feature matching technique in speech identification, having minimal error rates and fast computing speed. DTW will be of utmost importance for speech recognition in voice-based automatic teller machines [11]. The data of one successful system was based on an HMM pattern recognition procedure. The successful system used three states and one Gaussian mixture. This system had a success rate of 87.5% with an average running time of 7.9 sec compared with DTW, which has an average running time of 22.1 s and a success rate of 25%; thus, the hidden Markov system is evidently the better system [12]. A new approach is developed for real-time isolated-word speech recognition systems for human-computer interaction. The system was a speaker-dependent system. The main motive behind the development of this system was to recognize a list of words uttered by a speaker through a microphone. The features used were MFCCs, which allowed good discrimination of speech signal. The dynamic programming algorithm used in the system measured the similarity between the stored templates and the test templates for speech recognition and specified the optimal distance. The recognition accuracy obtained under the system was 90.1%. A simple list of four words, comprising four spelled-out numbers was made (i.e., one, two, three and four), and stored these words under certain command names. When a particular word was spoken into the microphone, the system recognized the word and displayed the respective command name under which the word was stored. With a few modifications, this system could be used in many areas for specific functions (e.g., to control a robot using simple commands) and in many applications [13]. Speech recognition is the process of converting an acoustic waveform into text that is similar to the information being conveyed by the speaker. In the reported study, the implementation of an automatic speech recognition system (ASR) for isolated and connected words (i.e., for the words of Hindi language) were discussed. The HTK based on HMM, as a statistical approach, is used to develop the system. Initially, the system was trained for 100 distinct Hindi words. By presenting the detailed architecture of the ASR system, which was developed using various HTK library modules and tools, the process has been described by which the HTK tool works, which was used in various phases of the ASR system. The reported recognition results show that the system has overall accuracies of 95% and 90% for isolated and connected words, respectively [14]. A speech recognition system can be easily distributed by way of speaking and by noise. When the reference signal is a word that is identical to the target signal, the speech recognition algorithm based on an FIR Wiener filter may be applied, in which case, the reference signal for modeling the target will have fewer errors. When reference and target signals are recorded by one person, in the same accent, both systems would work well for distinguishing different words, regardless of the identity of the person. Meanwhile, when the target and reference signals are recorded by different persons, both systems would perform unsatisfactorily. Therefore, to improve system performance in terms of recognition, the noise immunity of the system must be improved and the distinctive features of speech uttered by various people must be characterized reported [1]. Speech recognition is a technology that enables people to interact with machines. A speech recognition system design capable of 100% accuracy has not been achieved. One speaker-dependent speech recognition system capable of recognizing isolated spoken words with high accuracy have been the system was verified using Matlab, achieving an overall accuracy above 90%. This work emphasized the memory efficiency offered by speech detection for separating words from silence. As supported by experimental results, improved system performance was achieved using DTW while maintaining consideration for the overall design process [15]. The performance of the system was highly dependent on the speech detection block. The detection of speech signals from the silence periods resulted in improved results, improving the template generated by the MFCC processors according to the uttered words. In addition, system performance was improved using multiple templates per word. Endpoint detection, framing, normalization, MFCCs, and DTW algorithm have been used to process speech samples. The speech samples were spoken English numbers (zero to nine); these words were spoken in an isolated way by different male and female speakers. A speech recognition algorithm for English numbers using MFCC vectors provided an estimate of the vocal tract filter. Meanwhile, DTW was used to detect the nearest recorded voice that had appropriate constraint. The system was then applied to recognize the isolated spoken English numbers (i.e., “one,” “two,” “three,” “four,” “five,” “six,” “seven,” “eight,” and “nine”). The algorithm was tested on recorded speech samples. The results indicated that the algorithm managed to achieve an accuracy of almost 90.5% in recognizing the English numbers among the recorded words [16]. Speech recognition using HMM, which works well for n users, is widely used in pattern recognition applications. A recorded signal (test data) was compared with its corresponding original signal (trained data) using HMM algorithms simulated in Matlab. The examination of improved techniques for modeling audio signals can result in additional depth and helps realize improved speech recognition. For the training set, 100% recognition was achieved. The recognizer performed well overall and worked efficiently in a noisy environment. However, the performance of the system could be improved if additional training samples were available; the system could then be compared with neural network algorithms. Disturbances exist in a real environment, which might influence the performance of the speech recognizer. The speech recognizer was realized only in Matlab. The speech recognizer achieved good results (75–100%) under a simulation in a car with the engine running. The recognition rate varied from 100% in a noise-free environment to 75% in a noisy environment. As indicated by a comparison of the recognition rates under white Gaussian noise and that under the noise recorded in a car, the recognizer is more robust against the noise in the car than the white Gaussian noise because the noise in the car mainly affects low frequencies (0–2000 Hz), whereas the white Gaussian noise affects all frequencies [17]. The Wiener filter is an adaptive approach in speech enhancement that depends on the adoption of a filter transfer function based on speech signal statistics (mean and variance) from sample to sample. To accommodate the varying nature of speech signal, the adaptive Wiener filter is accomplished in the time domain rather than in the frequency domain. Results show that this approach provides better SNR improvement than the conventional Wiener filter approach or than a spectral subtraction in frequency domain. Furthermore, this approach is superior to the spectral subtraction approach in treating musical noise and can eschew the drawbacks of the Wiener filter in the frequency domain (Abd El-Fatta

III. METHODOLOGY

In principle, the fundamental concept of a speech recognition system is designed to perform two operations. These operations are the modeling of the signal under investigation and matching of patterns using the extracted features obtained from the first step). Signal modeling aims to find a set of parameters from the conversion process of the speech signal. The task is to find a parameter set from "previously processed" memory with matching properties of the parameter set obtained from the target speech signal "to certain extent" that represents the pattern matching step.

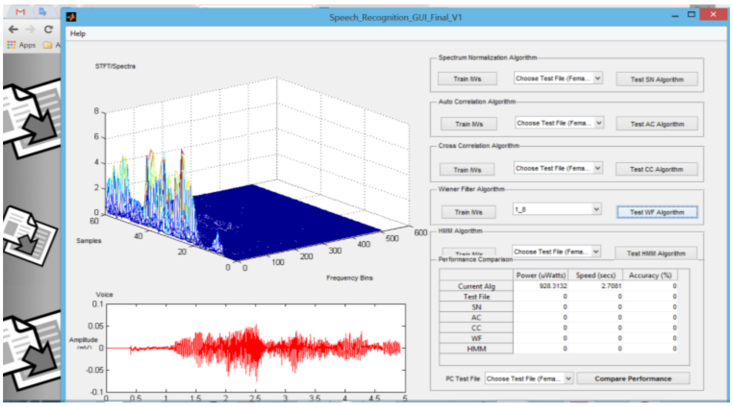
The five speech recognition algorithms were trained using pre-recorded voice signals of words in '.wav' format. In this research, 10 words were selected as a reference signal recorded by male voice and female voices. Therefore, we were able to obtain 20 reference signals. Table I lists these signals. Evaluating speech recognition algorithms dictates to subject them to similar test criteria. A performance criterion was set to obtain the results of the recognition. The performance index consists of three components, and the speed of recognition was calculated in "milliseconds". The accuracy rate was calculated as a percentage (%). Finally, consumed power was hypothetically estimated once the algorithm implemented in the hardware and calculated as "OC/µWatt".

Table 1. Pre-recorded word

No.	File Name '.wav' Format		
	Word	Female voice	Male voice
1	Mariam	1_1	1_11
2	Sunday	1_2	1_12
3	Monday	1_3	1_13
4	Tuesday	1_4	1_14
5	School	1_5	1_15
6	Holiday	1_6	1_16
7	Football	1_7	1_17
8	Happy	1_8	1_18
9	Beautiful	1_9	1_19
10	Crazy	1_10	1_20

IV. RESULT AND DISCUSSION

Starting the performance evaluation by running h algorithms in Matlab, running the algorithm codes shall display a GUI(graphical user interface) panel to train, test and display the tested voice signal and the short time Fourier transform (STFT) plots within the panel, from this panel, the operator can chose one test signal for all five algorithm and press the performance comparison button and display results; power in µW, speed in msec, and accuracy as percentage with the plots as shown in Figure 1.



Wiener Filter Algorithm: 1_8

	Power (uWatts)	Speed (secs)	Accuracy (%)
Current Alg	928.3132	2.7081	0
Test File	0	0	0
SN	0	0	0
AC	0	0	0
CC	0	0	0
WF	0	0	0
HMM	0	0	0

PC Test File: Choose Test File (Fema...

Fig 1. GUI panel

The simulation results of the testing data set used to evaluate the performance of the developed speech recognition algorithm. All five developed algorithms were subjected to similar training and testing data sets in each case. These five algorithms were tested with three different case scenarios:

- The first test uses a data set that consists of isolated words uttered "spoken" by male and female voice and consequently recorded, the first set of words is listed in Table II.

Table 2. Isolated word used in the evaluation of the speech recognition algorithms

No.	Isolated Words
1	Mariam
2	Sunday
3	Monday
4	Tuesday
5	School
6	Holiday
7	Football
8	Happy
9	Beautiful
10	Crazy

- The second test uses a data set that consists of the 10 words as in the previous case and this set that consists of 20 ".wav" files is used for training. The same 10 words were uttered again and recorded by different speakers (female and male). The set that includes different speakers was used as a test set to check the capabilities of the different five algorithms to recognize the correct uttered. The set that includes different speakers was used as a test set to check the capabilities of the different five algorithms to recognize the correct uttered word.
- The third test uses a data set that consists of the 10 words as in the case of the first test. The 20 recorded files are used as training data. Two words, namely, "beautiful" and "mariam" were selected for the test data. These two words were recorded 10 times each by the same speaker (male voice) which resulted in 20 recorded files that were used as test data to determine the recognition rate as percentage for all five algorithms of these specific two words.

The results of the test were tabulated in Tables 3, Tables 4, Figure 2, and Figure 3. The measure of performance for the three algorithms was determined from the first test result case I, Table 4 shows that the estimated required power was minimum for the SN algorithm and maximum for the HMM algorithm and this can be attributed to the computational requirements for pre-processing, feature extraction and classification. HMM highly demands power, whereas the SN usually requires less computer operation. Thus, the basic method used in DSP and FPGA applications to estimate and measure processing power and perform a rough estimation of required operational power is based on operation count or count of floating point operations required to perform the task that is usually in the form of vector and matrix multiplications and inversions.

During simulations, the same approach was used for the calculation of power based on floating point operations, which used to be a function in the older versions of Matlab. When the Matlab data handling structure was changed, this flop count function was also removed. Three different parameters were used while each algorithm ran to form an estimate of the required processing and operational power. These parameters are the number of created arrays by each algorithm, the amount of CPU usage by each algorithm and the amount of memory access by each algorithm. Based on these three measurements, a relative measure of required power for each algorithm was formulated. Thus, the power measurements should be considered as relative power requirements for each algorithm.

The speed measurements are based on Matlab runtime and CPU time measurement. They accurately represent the time required for each algorithm to execute by pressing the start button until the elapsed time value is shown in the command window. Accuracy of calculation is based on a simple comparison of the initial test file and the file recognized. Thus, the results are either 100 for correct recognition or 0 for unsuccessful or wrong recognition of the initial test file.

Table 3. Performance comparison (power, speed, and accuracy) for test case i.

No.		Metric parameters		
	algorithms	Speed (msec)	Power (µW)	Accuracy (%)
1	SN	386 - 516	387 – 662	100
2	CC	765 - 985	776 -1103	100
3	AC	492 - 556	552 - 740	100
4	WF	695 - 867	750 - 885	100
5	HMM	10255 - 11085	10425 - 11244	100

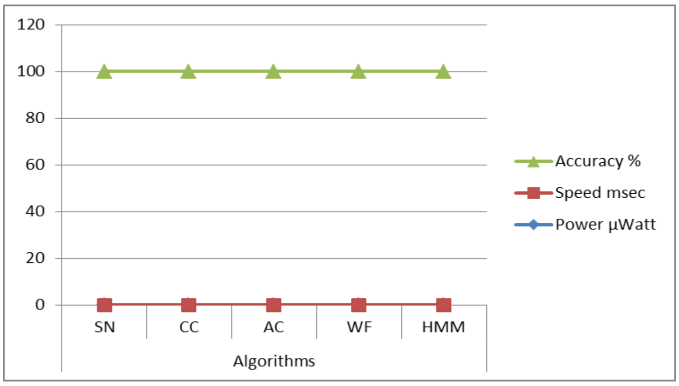


Fig 2. Performance comparison (power, speed, and accuracy) for test case I.

Table 4. Recognition rate results of the five algorithms based on three test cases (I, II, and III).

No.		Algorithm recognition rate %		
	algorithms	Case I	Case II	Case III
1	SN	100	5	20
2	CC	100	5	5
3	AC	100	10	0
4	WF	100	5	50
5	HMM	100	0	10

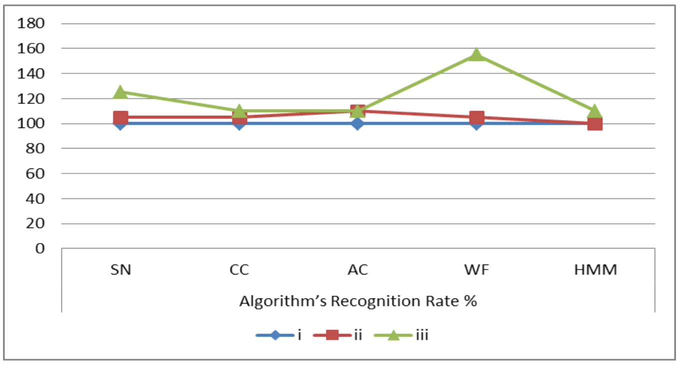


Fig 3. Recognition rate results of the five algorithms based on three test cases (I, II, and III).

V. CONCLUSSION

This research study presents the challenges in implementing a speech recognition system using the various techniques namely;spectrumnormalization,cross-rrelation,autocorrelation, wiener filter, and hidden markov model. To develop algorithms for isolated word recognition only in the relatively quiet environment. From the simulation result, the Wiener filter algorithm produced the best trade-off over the other four techniques in terms of all measure of performance as a whole that are; speed, accuracy, and power. Wiener filter algorithm scored accuracy of 100%, 5%, and 50% for test cases i,ii,and iii respectively, with recognition speed range of (695-867) msec and estimated power range of (750-885) µW. The performance of speech recognition algorithm can be easily disturbed by the noise and the pronunciation of the word. To improve the developved algorithms particularly for hidden markov model algorithm requires solution to many factors for example, noise immunity, robustness, ability to handle varying speaking accents and styles, humman interaction design (HCI) issues that are suitable for speech- based interaction.

ACKNOWLEDGMENT

The author would like to thank the Ministry of Electricity of Iraq (MOE) or their continued support and encouragement for the completion of this research, as well as thanks to the Ministry of Higher Education and Scientific Research (MOHER) for their support to me and to facilitate a lot of the difficulties that I faced during the study period.

REFERENCES

[1] Deepak, M. Vikas, "Speech Recognition using FIR Wiener Filter", International Journal of Application or Innovation in Engineering & management (IJAIEM),pp.204-20,2013.

[2] Paul, Teenu Therese, and Shiju George. "Voice recognition based secure android model for inputting smear test result." *International Journal of Engineering Sciences & Emerging Technologies, ISSN*: 2231-6604,2013.

[3] Mohan, Bhadragiri Jagan, and N. R. Babu. "Speech recognition using MFCC and DTW." *Advances in Electrical Engineering (ICAEE), 2014 International Conference on*. IEEE, 2014.

[4] Ghai, Wiqas, and Navdeep Singh. "Literature review on automatic speech recognition." *International Journal of Computer Applications* 41.8 (2012),pp, 42-50,2012.

[5] Amin, Talal Bin, and Iftekhar Mahmood. "Speech recognition using dynamic time warping." *Advances in Space Technologies, 2008. ICAST 2008. 2nd International Conference on*. IEEE, 2008.

[6] Buera, Luis, et al. "Cepstral vector normalization based on stereo data for robust speech recognition." *Audio, Speech, and Language Processing, IEEE Transactions on* 15.3 (2007),pp, 1098-1113,2007.

[7] Chen, Jingdong, Jacob Benesty, and Yiteng Huang. "Robust time delay estimation exploiting redundancy among multiple microphones." *Speech and Audio Processing, IEEE Transactions on* 11.6 (2003),pp,549-557,2003.

[8] Varshney, N., & Singh, S, " Embedded Speech Recognition System", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering ,pp,9218-9227,2014.

[9] Patel, V. D, " Voice recognition system in noisy environment",. California State University, Sacramento,2011.

[10] Ye, J., " Speech recognition using time domain features from phase space reconstructions" ,Marquette University Milwaukee, Wisconsin, 2004.

[11] Dhingra, Shivanker Dev, Geeta Nijhawan, and Poonam Pandit. "Isolated speech recognition using MFCC and DTW." *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering* 2.8 (2013),pp. 4085-4092,2013.

[12] Price, Jamel, and Sophomore Student. "Design of an Automatic Speech Recognition System Using MATLAB." *Dept. of Engineering and Aviation Sciences, University of Maryland Eastern Shore Princess Ann* ,2005.

[13] Makhijani, Rajesh, and Ravindra Gupta. "Isolated word speech recognition system using dynamic time warping." *International Journal of Engineering Sciences & Emerging Technologies (IJESET)* 6.3 (2013): pp.352-367,2013.

[14] Choudhary, Annu, Mr RS Chauhan, and Mr Gautam Gupta. "Automatic Speech Recognition System for Isolated & Connected Words of Hindi Language By Using Hidden Markov Model Toolkit (HTK)", 2013.

[15] Amin, Talal Bin, and Iftekhar Mahmood. "Speech recognition using dynamic time warping." *Advances in Space Technologies, 2008. ICAST 2008. 2nd International Conference on*. IEEE, 2008.

[16] Limkar, Maruti, Rama Rao, and Vidya Sagvekar. "Isolated Digit Recognition Using MFCC AND DTW." *International Journal On Advanced Electrical And Electronics Engineering,(IJAEEE), ISSN (Print)* (2012),pp 2278-8948,2012.

[17] Srinivasan, A. "Speech recognition using Hidden Markov model." *Applied Mathematical Sciences* 5.79 (2011),pp.3943-3948,2011.

[18] Abd El-Fattah, M. A., et al. "Speech enhancement using an adaptive wiener filtering approach." *progress in electromagnetics research M* 4 (2008),pp. 167-184,2008.