English Digits Speech Recognition System Based on Hidden Markov Models

Ahmad A. M. Abushariah[(1)], Teddy S. Gunawan[(2)], Othman O. Khalifa[(3)]
Electrical and Computer Engineering Department, Faculty of Engineering, International Islamic University Malaysia, Gombak, 53100 Kuala Lumpur, Malaysia.
ahmad2010@hotmail.com[(1)], tsgunawan@iiu.edu.my[(2)], khalifa@iiu.edu.my[(3)]

Mohammad A. M. Abushariah[(4)]
Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia.
shariah@perdana.um.edu.my[(4)]

***Abstract*—This paper aims to design and implement English digits speech recognition system using Matlab (GUI). This work was based on the Hidden Markov Model (HMM), which provides a highly reliable way for recognizing speech. The system is able to recognize the speech waveform by translating the speech waveform into a set of feature vectors using Mel Frequency Cepstral Coefficients (MFCC) technique This paper focuses on all English digits from (Zero through Nine), which is based on isolated words structure. Two modules were developed, namely the isolated words speech recognition and the continuous speech recognition. Both modules were tested in both clean and noisy environments and showed a successful recognition rates. In clean environment and isolated words speech recognition module, the multi-speaker mode achieved 99.5% whereas the speaker-independent mode achieved 79.5%. In clean environment and continuous speech recognition module, the multi-speaker mode achieved 72.5% whereas the speaker-independent mode achieved 56.25%. However in noisy environment and isolated words speech recognition module, the multi-speaker mode achieved 88% whereas the speaker-independent mode achieved 67%. In noisy environment and continuous speech recognition module, the multi-speaker mode achieved 82.5% whereas the speaker-independent mode achieved 76.67%. These recognition rates are relatively successful if compared to similar systems.**

Keywords: English digits; Features extraction; Hidden Markov Models; Mel Frequency Cepstral Coefficients.

I. INTRODUCTION

The field of Automatic Speech Recognition (ASR) is about 60 years old. There have been many interesting advances and developments since the invention of the first speech recognizer at Bell Labs in the early 1950's. The development of ASR increased gradually until the invention of Hidden Markov Models (HMM) in early 1970's. Researchers' contribution were to make use of ASR technology to what can be seen nowadays of various advancements in fields like multi-modal, multi-lingual/cross-lingual ASR using statistical techniques such as HMM, SVM, neural network, etc [1].

Speech recognition or more commonly known as automatic speech recognition (ASR) was defined as the process of interpreting human speech in a computer [2]. However, ASR was defined more technically as the building of system for mapping acoustic signals to a string of words [3]. In general, all ASR systems aim to automatically extract the string of spoken words from input speech signals as illustrated in Figure 1.

The main objective of this paper is to design and implement an English digits speech recognition system based on Hidden Markov Model (HMM) using MATLAB, which is capable of recognizing and responding to digits speech inputs. This English digits speech recognizer would be applicable and useful for various digits-based applications, such as banking systems, phone dialing systems and various other systems. In this research, we utilized statistical modeling method based on the Hidden Markov Models to recognize English language digits.

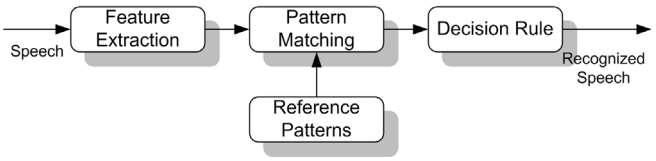

Figure 1. Speech Recognition System Concept

The following section, Section 2, describes the English digits speech recognition system in more depth. Experimental setup is presented in Section 3. Section 4 presents the experimental results and discussion. Finally, the conclusion is presented in Section 5.

II. ENGLISH DIGITS SPEECH RECOGNITION SYSTEM

A. Scopes

There are ten distinct English digits namely zero, one, two, three, four, five, six, seven, eight and nine. Speakers could say any of these ten digits in isolation which is presented under the isolated words speech recognition module or in a sequence or multiple digits which is presented under the continuous speech recognition module. Although training data were recorded in a studio environment which is clean data, however, the system was tested with data recorded in the same environment as well as in various noisy environments to test the performance and accuracy of the system in different environments.

B. System's Architecture and Algorithms

It has been noticed that the success of any automatic speech recognition system requires a combination of various techniques and algorithms, each of which performs a specific task for achieving the main goal of the system. Therefore, a combination of related algorithms improves the accuracy or the recognition rate of such applications. Figure 2 shows the

978-1-4244-6235-3/10/$26.00 ©2010 IEEE

architecture of the HMM based English digits speech recognition system.

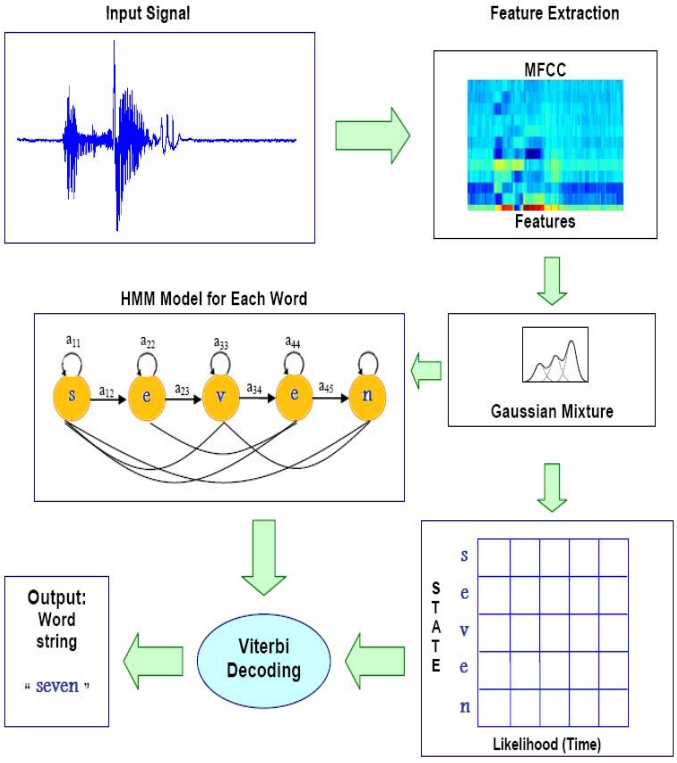

Figure 2. Architecture of the HMM Based English Digits Speech Recognition System

Figure 2 had just shown the main steps to perform the HMM based speech recognition system as follows:

1. Receiving and digitizing the input speech signal.
2. Extracting features for all input speech signals using MFCC algorithm, where its computational steps are shown in Fig. 3 and Fig. 4, then converting and storing each signal's features into a feature vector.
3. Classifying the feature vectors into the phonetic based categories at each frame using HMM algorithm.
4. Finally, performing a Viterbi search which is an algorithm to compute the optimal (most likely) state sequence in HMM given a sequence of observed outputs.

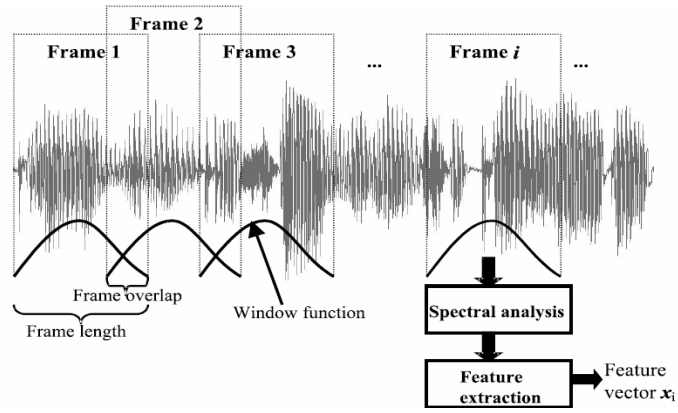

Figure 3. Features Extraction Concept [4]

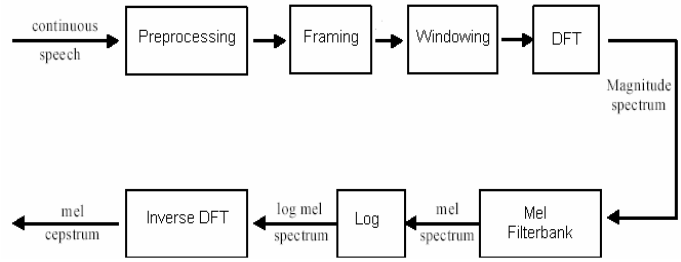

Figure 4. MFCC Computational Process [4]

Hidden Markov Model (HMM) is one of the most powerful and dominating statistical approaches, which has been applied for many years. The basic theory of HMM was published in a series of classic papers by Baum and his colleagues in the late 1960s and early 1970s which was then implemented for speech recognition applications by Baker at Carnegie Mellon University (CMU) and by Jelinek and his colleagues at IBM in the 1970s [5].

An HMMs are specified by a set of states *Q*, a set of transition probabilities *A*, a set of observation likelihoods *B*, a defined start state and end state(s), and a set of observation symbols *O*, which is not drawn from the same alphabet as the state set *Q* [6].

C. System's MATLAB Graphical User Interfaces (GUIs)

The user can initiate the system by typing 'main' in the MATLAB command window in order to enter the main interface of the system. As a result, the main interface of the English digits speech recognition system automatically appears as shown in Figure 5.

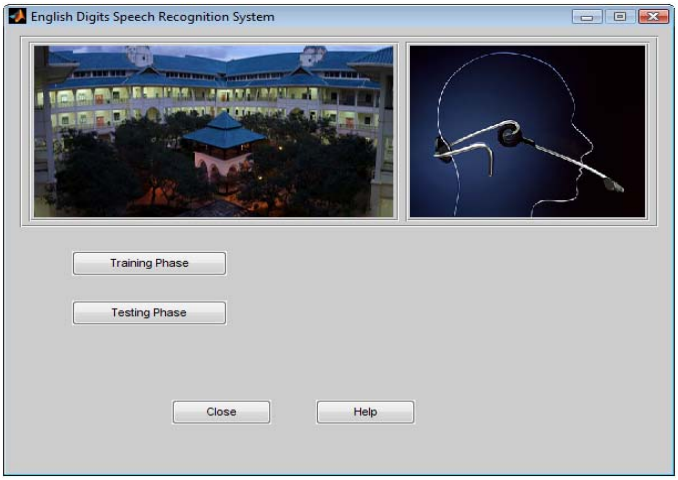

Figure 5. Main GUI of the English Digits Speech Recognition System

From the main GUI of the system as illustrated in Figure 5, the user has to either click on 'Training Phase' button, 'Testing Phase' button, 'Help' button in order to train, test, or view the user manual, or 'Close' button to exit the system. However, first, users are advised to click 'Training Phase' button to start training the HMM. Figure 6 shows the "Training Phase" interface as a result of the user's action.

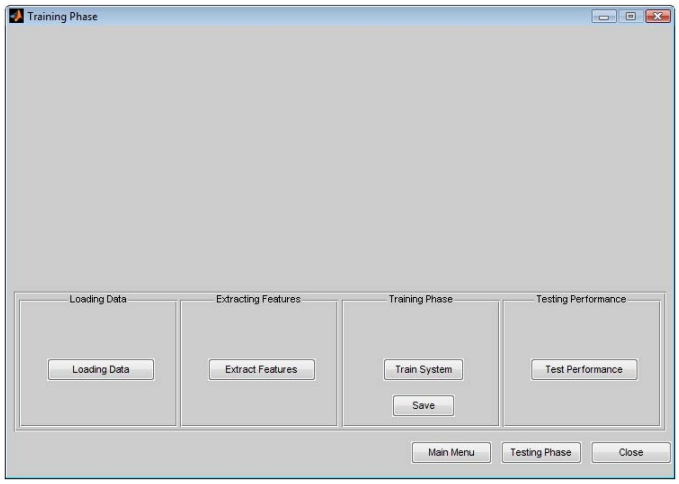

Figure 6. GUI of the System's Training Phase

The 'Training Phase' GUI requires users to follow the following steps:

1. Click 'Loading Data' button, in order to load all recorded speech samples, which were recorded in the noise free Visual-Audio Studio. These speech samples are clean samples and they are 630 speech samples. They represent speech samples of 21 speakers, for the ten words, whereby each speaker recorded each word three times. Therefore, a total of 21 Speakers * 10 Words * 3 Repetitions = 630 Speech Samples, whereby each digit has been recorded for 63 times by all 21 speakers.
2. Click 'Extract Features' button, in order to apply the MFCC algorithm on 43 out of 63 loaded speech samples per digit. These 43 speech samples per digit are used for training the HMM in the next step, whereas the remaining 20 speech samples per digit will be used to test the performance of the system.
3. Click 'Train System' button, in order to train the HMM using the 43 feature vectors per digit produced from the previous step. Now, each digit has its own HMM.
4. Click 'Save' button, which saves the HMM models for all ten digits as in one 'HMM.mat' database. This database will be used for testing the performance of the system in the next step and can be uploaded when the user does not want to make the training all over again every time he/she wants to test the system in the 'Testing Phase'.
5. Click 'Test Performance' button, which tests the remaining 20 speech samples of the loaded speech data against the HMM models of the 43 trained speech samples. Figure 7 shows the performance testing accuracy of the system which is 99.5%, whereby this is the multi-speakers mode recognition rate in noise free studio.

The testing options interface contains two major mechanisms for testing the system; which are 1) Isolated Words Recognition and 2) Continuous Speech Recognition. However, both are meant to recognize English digits, except that Isolated Words Recognition GUI permits users to test one and only one digit at a time, whereas the Continuous Speech Recognition GUI permits users to test a string that contains more than one digit at a time.

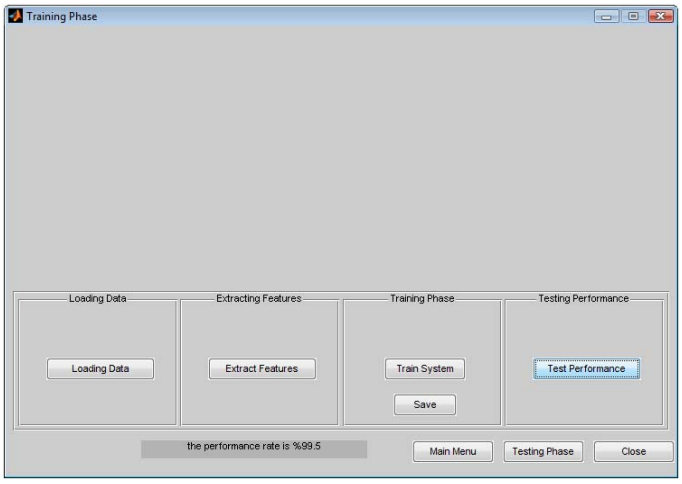

Figure 7. Testing the System's Performance for Multi-Speakers Mode

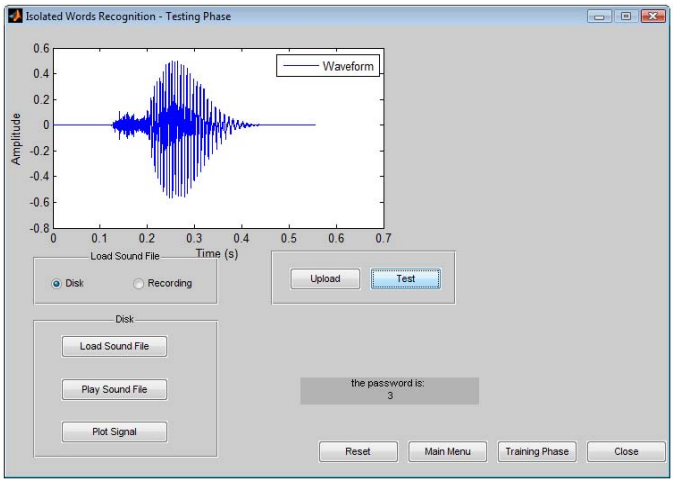

Figure 8. Example of Loading and Testing the English Digit 'Three' in the GUI of the Isolated Words Recognition – Testing Phase

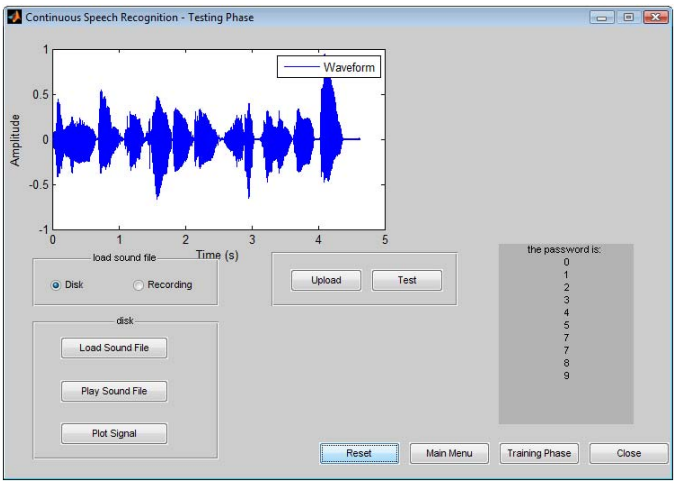

Figure 9. Example of Loading and Testing the Digits Sequence 'zero, one, two, three, four, five, six, seven, eight, nine' in the GUI of the Continuous Speech Recognition – Testing Phase

Users can now start testing the system real-time using their own speech recordings or by loading any of the previously recorded speech samples. This step is included in both Isolated Words Recognition and Continuous Speech Recognition. Figure 8 shows Testing Isolated Word, whereas Figure 9 shows Testing Continuous Speech.

III. Experimental Setup

Speech samples collection is mostly concerned with recording various speech samples of each distinct English digit by different speakers. However, there are four main factors that must be considered when collecting speech samples, which affect the training set vectors that are used to train the HMM [5]. Those factors include who the talkers are; the speaking conditions; the transducers and transmission systems and the speech units.

This system had 34 different speakers out of whom their speech samples were collected. Those 34 speakers include 24 male and 10 female speakers belonging to different ages, genders and races. Table I summarizes the first factor in more details about the profiles of talkers/speakers.

TABLE I. Summary of Talkers' Profile

Age	Gender	Race
Adults ranging from 18 to 46 years	**1.** "24" males **2.** "10" females	**1.** Arabs **2.** Malays

The English digits speech recognition system's speech samples collection was done in both noise free (clean) and noisy environments. The speech samples of 24 speakers were collected in a clean environment using specialized Visual-Audio studio. The Visual-Audio studio is a noise proof studio is shown in Figure 10. Other speech samples of 10 other speakers were collected in noisy environments, which include faculty class rooms and students' rooms in the university's hostels. This is important in order to examine the English digits speech recognition system in different environments with different levels of noise.

Figure 10. Visual-Audio Studio

The transducers and transmission systems. The speech samples collected from the Visual-Audio studio were recorded and collected using specialized equipments including microphones, speakers, headsets, PC with some special software, and some hardware that control the entire recording process, which includes noise filters, stereo controller and many other options as shown in Figure 11.

Figure 11. Visual-Audio Studio's Equipments

The English digits speech recognition system's main speech units are specific isolated words and their combination, which includes continuous speech recognition. In other words, the purpose of the system is to recognize words that belong to isolated word recognition category of applications and to some extent continuous speech recognition. A set of ten distinct English digits were recorded, which are (Zero, One, Two, Three, Four, Five, Six, Seven, Eight and Nine) each of which was recorded separately. Therefore, this system is mainly an isolated word recognizer and serves to some extent as a continuous word recognizer.

Each distinct digit was recorded by every speaker and their recordings were saved for further processing. There are two types of sound files recorded by the speakers, namely **1)** Sound Files in Clean Environment and **2)** Sound Files in Noisy Environments. There are also two modes, namely **1)** Multi-Speaker Mode (i.e., the same set of speakers were used in both the training and the testing phases) and, **2)** Speaker-Independent Mode (i.e., speakers used for training are different from those used for testing). Another factor to consider is that, there are two types of sound files depending on the type of testing namely, **1)** Isolated Words Recognition and, **2)** Continuous Speech Recognition.

IV. Experimental Results and Discussion

The database consists of 10 distinct words (digits) of 34 male and female speakers. It also contains 1,380 sound files used for training and testing the Isolated Words Recognition module and 28 sound files for testing the Continuous Speech Recognition module in clean and noisy environments for both multi-speaker and speaker-independent modes. Recognition rate of the trained HMM is defined as follows:

$$RR = \frac{N_{Correct}}{N_{Total}} \times 100\% \quad (1)$$

where RR is the recognition rate, $N_{Correct}$ is the number of correct recognition of testing speech samples per digit, and N_{Total} is the total number of testing speech samples per digit.

TABLE II. OVERALL RECOGNITION RATE (%) OF THE ENGLISH DIGITS SPEECH RECOGNITION SYSTEM

Environment	Isolated Words Recognition		Continuous Speech Recognition	
	Multi-Speaker Mode	Speaker-Independent Mode	Multi-Speaker Mode	Speaker-Independent Mode
Clean	99.5	79.5	92.5	76.67
Noisy	88	67	72.5	56.25

Table II shows the overall recognition rates (%) for the English Digits Speech Recognition System. Experimental results of the combination of MFCC and HMM algorithms in the English digits speech recognition system are acceptable, but could be improved further to obtain higher accuracy rates. Table III shows a comparison of recognition rates (%) for current speech recognition researches and systems together with feature extraction and classification techniques used.

TABLE III. COMPARISON OF RECOGNITION RATES (%) FOR CURRENT SPEECH RECOGNITION RESEARCHES AND SYSTEM

Reference	Features Extraction Techniques	Features Classification Techniques	Recognition Rate (%)
[4]	MFCC	VQ	**88.88%**
[7]	MFCC	-	**33% to 45%**
	PLP	-	**30% to 40%**
[8]	LPC	VQ and HMM	**62% to 96%**
[9]	MFCC **(Clean)**	HMM **(Clean)**	**86%**
	MFCC **(Noisy)**	HMM **(Noisy)**	**28% to 78%**
[10]	MFCC	HMM	**92%**
[11]	MFCC	VQ	**57% to 100%**
[12]	MFCC	VQ	**70% to 85%**

From the above results, it is clearly found that the English digits speech recognition system performed well in both clean and noisy environment with both multi-speaker and speaker-independent modes. It is noticed that the recognition results on the clean environment are much higher than the recognition results of the noisy environment.

V. CONCLUSIONS

The recognition rates of multi-speaker mode performed better than the speaker-independent mode in both environments. Although it is believed that the recognition rates achieved in this research are comparable with other systems and researches of the same domain, however, more improvements need to be made specially increasing the training and testing speech data. The more training speech data used in a system, the better and higher the system's performance can be obtained.

The accuracy of the identification process can be influenced by certain factors such as different level of surrounding noise during the recording session, the quality of the microphone used to input the speech signals, and many other factors. Even though it is difficult to avoid some of these factors, steps should be taken to minimize the effect.

REFERENCES

[1] Garfinkel (1998). Retrieved on 10th February 2009, www.dragon-medical-transcription.com/historyspeechrecognition.html

[2] M., Forsberg, "Why is Speech Recognition Difficult?". Department of Computing Science, Chalmers University of Technology, Gothenburg, Sweden, 2003.

[3] D., Jurafsky and J.H., Martin, "Speech and Language Processing an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition". Prentice Hall, Upper Saddle River, NJ, USA, 2000.

[4] M. A. M. Abu Shariah, R. N. Ainon, 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] L., Rabiner, and B. H., Juang. *Fundamentals of Speech Recognition*. Prentice Hall, NJ, USA, 1993.

[6] D. S., Jurafsky, and J. H., Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Prentice Hall, Upper Saddle River, NJ, 2nd edition, 2008.

[7] B., Milner, "A Comparison of Front-End Configurations for Robust Speech Recognition". *ICASSP'02*, pp. 797–800, 2002.

[8] S.K., Podder, "Segment-based Stochastic Modelings for Speech Recognition". PhD Thesis. Department of Electrical and Electronic Engineering, Ehime University, Matsuyama 790-77, Japan, 1997.

[9] S.M., Ahadi, H., Sheikhzadeh, R.L., Brennan, and G.H., Freeman, "An Efficient Front-End for Automatic Speech Recognition*". IEEE International Conference on Electronics, Circuits and Systems (ICECS2003)*, Sharjah, United Arab Emirates, 2003.

[10] M., Jackson, "Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language". Master Thesis, Faculty of Computing and Information Technology, Makerere University, 2005.

[11] M.R., Hasan, M., Jamil, and M.G., Saifur Rahman, "Speaker Identification Using Mel Frequency Cepstral Coefficients". *3rd International Conference on Electrical and Computer Engineering,* Dhaka, Bangladesh, pp. 565-568, 2004.

[12] M.Z., Bhotto and M.R., Amin, "Bangali Text Dependent Speaker Identification Using MelFrequency Cepstrum Coefficient and Vector Quantization". *3rd International Conference on Electrical and Computer Engineering*, Dhaka, Bangladesh, pp. 569-572, 2004.