Jurnal Teknologi, 40(D) Jun. 2004: 133–144

DESIGN OF EDUCATIONAL SOFTWARE FOR AUTOMATIC SPEECH RECOGNITION (ASR) TECHNIQUES

HONG KAI SZE[1] & SHEIKH HUSSAIN SHAIKH SALLEH[2]

Abstract. Speech recognition has been an important subject for research, and it has come to a stage where it has been actively applied in a lot of industrial and consumer applications, overseas. However, speech recognition research is still in its infancy stage in Malaysia. The main reason is that speech recognition systems are highly complex and teaching students in this subject matter with the underlying technologies is a challenging task. Currently, some instructors use slide show presentations and white board in giving such courses. At the end of the course, students are not able to figure out the real output of the algorithms given. In this case, students are not exposed to the real technical systems and would easily get bored. This research is mainly on the improvement over the limitations and problems of the traditional teaching method in speech recognition by developing a set of interactive and practical education software to guide and assist students in studying, and performing experiments for speech recognition.

Keywords: Speech recognition, pattern recognition, object oriented programming, graphical user interface, human computer interaction

Abstrak. Pengecaman suara merupakan satu subjek penting dalam penyelidikan, dan perkembangannya telah tiba di satu tahap di mana ia telah diaplikasikan di dalam banyak aplikasi industri dan pengguna di luar negeri. Walau bagaimanapun, penyelidikan pengecaman suara masih berada di tahap awalnya di Malaysia. Sebab utama adalah pengecaman suara sungguh kompleks dan pengajaran subjek ini terutamanya teknologi di sebaliknya merupakan satu tugas yang mencabar. Masa kini, sesetengah pengajar menggunakan persembahan tayangan slide dan papan putih dalam pemberian kursus begini. Di akhir kursus, pelajar tidak dapat meninjau *output* daripada algoritma yang diberi ataupun menguji sistem ini dalam masa nyata. Dalam kes ini, pelajar tidak dapat didedahkan kepada sistem teknikal yang sebenarnya dan mudah berasa bosan. Penyelidikan ini terutamanya memberi perhatian kepada kemajuan terhadap had dan masalah yang dihadapi oleh kaedah pengajaran tradisional di dalam pengecaman suara dengan membangunkan satu set perisian pendidikan yang interaksi dan praktikal untuk membimbing dan membantu pelajar dalam pembelajaran, menjalankan pengujian dan membangunkan aplikasi pengecaman suara.

Kata kunci: Pengecaman suara, pengecaman corak, pengaturcaraan berdasarkan objek, perantara muka pengguna bergrafik, interaksi manusia-komputer

1.0 INTRODUCTION

Recent advances in speech technology and computing power have created a surge of interest in the practical application of speech recognition. In Malaysia, speech

[1&2] Dept. of Microelectronics and Computer Engineering, Faulty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia. E-mail: hussain@fke.utm.my

recognition is still in its infancy stage, however, there are very few speech recognition systems built for Malay speeches. Most of the researches were based on the experiments and there are very few end user products for Malay speeches that is available in the market.

In order to develop the speech recognition research for Malay speeches, a set of educational toolkits is needed. This educational toolkits will help the researchers and students to understand the basic theories of speech recognition, as well as performing experiments, in accordance with certain conditions they may wish to specify.

The second main reason to have good educational software is to develop the fundamental structure of the speech recognition research at university level. Currently, there are a few speech recognition related researches successfully done either by researchers or students in Malaysia. However, these researches are normally done separately in their own formats without any standard formats agreeable by others. This make the source codes for the research works not reusable by other researchers without restudying everything or understanding the source codes and rewriting them into their own applications. This creates an obstacle as the research team grows larger and their softwares become more complex. It is undoubtedly clear that there is a need for some of the standard format of programming coding to be introduced into current researches, which will be discussed later in this paper.

Besides that, by developing such teaching tools, it could help students and other interested people in getting an appropriate starting knowledge and experience in speech recognition topics. There are some related tools available on the Internet. However, they are either too expensive or not suitable for such research work. To tackle this problem, a new educational software for automatic speech recognition techniques has been designed and developed.

2.0 OVERVIEW OF SPEECH RECOGNIZER

All speech recognizers include an initial signal processing front end that converts a speech waveform into features useful for further processing. This front end is required to extract the important features from the speech waveform. The first stage of processing the speech waveform is by reducing the data rate. This research only concentrates on three types of speech feature extraction techniques which will be described later. These techniques are Linear Predictive Coding (LPC) [1], LPC-based Cepstrum [2], and Mel-Frequency Cepstrum Coefficients (MFCC) [3].

Another important process of speech processing is a lossy data compression method based on the principle of block coding called Vector Quantization (VQ) [4]. The outputs of the Vector Quantization are a series of vectors characteristic of time-varying spectral characteristic of the speech signal. VQ further reduces the raw spectral representation of speech to small, finite number of 'unique' spectral vectors. The key advantage of the VQ representation is the reduced storage and computation time for spectral analysis information.

Another key question in speech recognition is how speech patterns are compared to determine their similarity. Depending on the requirements of the recognition system, pattern comparison can be done in a wide variety of ways. People never speak words at exactly the same uniform rate. Sometimes words are spoken quickly and at other times, slowly, so a method of time alignment is required in order to compare the test pattern with the reference word patterns. Dynamic Time Warping (DTW) technique is a time-alignment algorithm that can handle this problem [5]. Another well-known and widely used statistical method of characterising the spectral properties of the frames of a pattern is the hidden Markov model (HMM) approach [6]. The use of HMM for speech recognition has become increasingly popular in the past few years. This research focuses only on these two pattern comparison techniques.

Speech input, or human voice, is first converted to a digital form. This can be accomplished using a computer microphone, and then using a sound card (analog to digital converter) to convert the analog signal to its digital form. Thus, we will have digital data representing each level of signal at every discrete time step.

For speech recognizer training purposes, the digitized speech samples are then processed, using the feature extraction technique, to produce speech features. Speech features can go through VQ stage to produce a sequence of codebook indexes. As can be seen from Figure 1, speech features or codebook indexes will be stored in the template library.

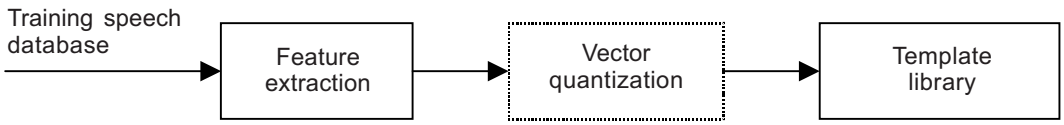

Figure 1 Overview of speech recognizer training process

For recognition purposes, the same process will be carried out to produce speech features or codebook indexes. Then, these features or indexes will be compared with those stored in the template library. The best-matched template will be determined so that a suitable action can be done. The whole process is illustrated in Figure 2. However, VQ is an optional step for DTW matching technique but is essential for discrete density HMM.

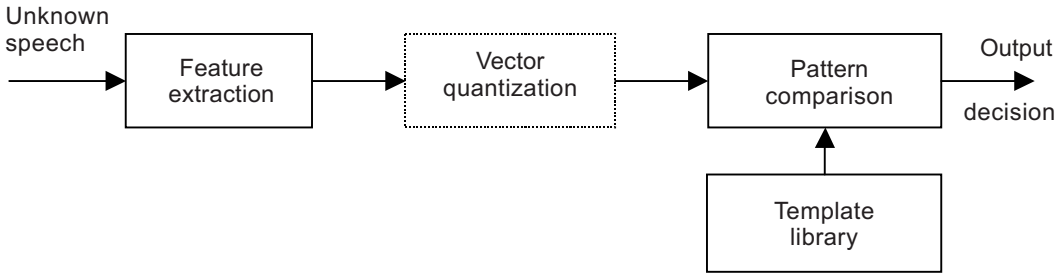

Figure 2 Overview of speech recognition

3.0 EXISTING SOFTWARE RELATED TO SPEECH RECOGNITION EDUCATION

In this section, literature study on the existing software related to speech recognition education will be discussed. The study is basically based on the recent published papers and materials related to speech recognition education software. There are a lot of speech analysis tools that provide most of the standard signal processing functions together with a visual presentation of the input and output data. All of them have some advantages and disadvantages in speech recognition education. These softwares, are WaveEdit [7], VISPER [8], Multi-Speech, Model 3700 [9], WINDSK [10], new collaborative active learning tool for signal processing [11], an Internet-based signal processing laboratory [12], SAPPHIRE [13], SPANNET [14], and software tool for introducing speech coding fundamental in a DSP Course [15, 16]. Another set of software tools to enhance the local and distance learning for speech and signal processing classes [17] are POST [18] and ARES [19]. Reviews on the above existing softwares are essential for building a good education software. All these softwares are developed for educational and research purposes. They have their own strengths and weaknesses in handling some specific problems. Thus, this project is targeting some specific topics in speech recognition that are important for education purposes.

4.0 GRAPHICAL USER INTERFACE AND SOFTWARE DESIGN

In this section, the graphical user interface of the education software will be described in details. The discussion will also include the design of various supporting modules. These supporting modules are lab sheets, help files, demonstration movies, and designation of setup files in a Setup CDROM.

4.1 Main Module

Graphical User Interface (GUI) is the interaction bridge between the computer users and the software. In this research and development, Microsoft Visual C++ 6.0 was utilised to create the attractive and user-friendly GUI. Figure 3 shows the GUI design of the main module.

4.2 Speech Studio

Figure 4 shows the GUI design of the Speech Studio, where its main functions are to perform speech recording and endpoints detection. The module was designed based on the multi-document interface, which means many documents can be opened at the same time. This is important because most of the time speech recognition systems have to deal with many speech files. Besides its main function for speech recording, the Speech Studio was also created for speech endpoint detection.

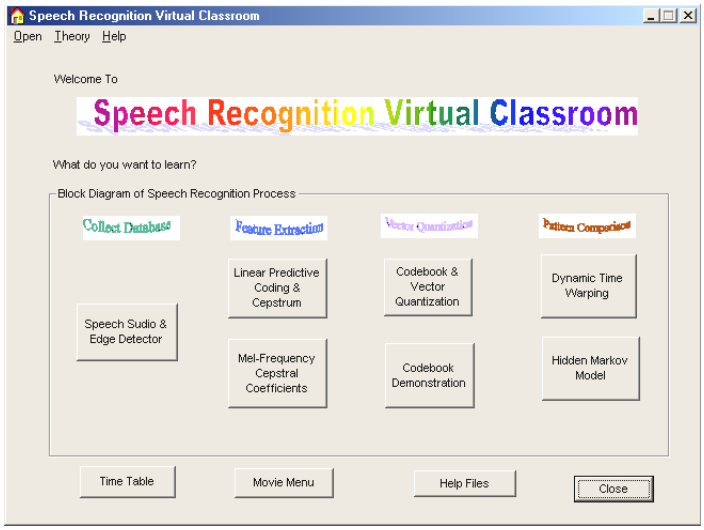

Figure 3 Main module

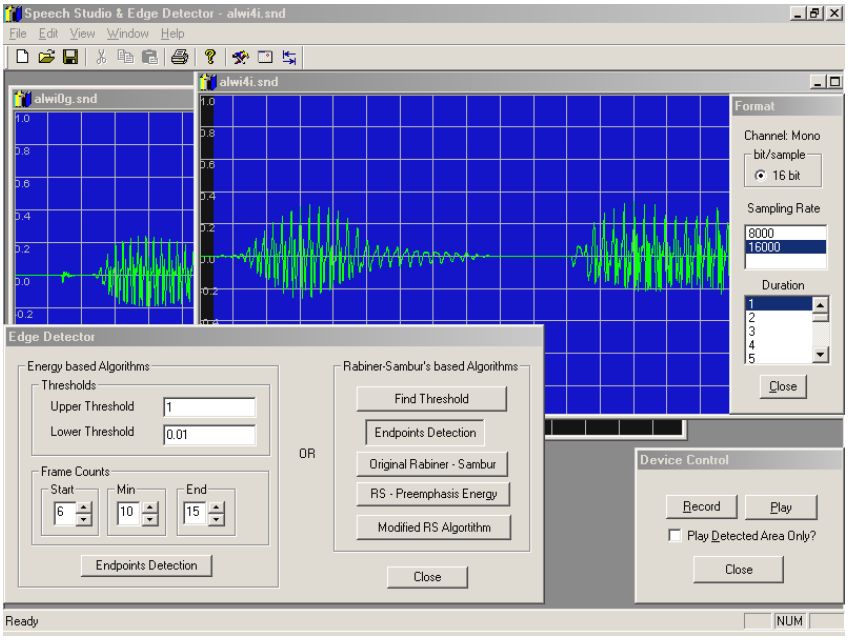

Figure 4 Speech studio

4.3 LPCC Classroom

As can be seen from Figure 5, the document view of this module is divided into 8 main windows. The user may view the output of every blocking frame by clicking the arrows on the toolbar and the waveforms will be redrawn accordingly. The LPC and CEP orders can be changed using the 'LPC Parameters' dialog box. The LPC Cepstrum Classroom can perform batch processing on many speech files. The LPCC Classroom utilizes the Microsoft Agent component to create attractive and effective education course. The 'peedy' (the green bird) was chosen to become the vitual lecturer. Once activated, the 'peedy' will travel around and give explanation to every LPCC steps. The 'peedy' will speak through the computer speakers. Many animations were used, such as Surprise, Announce, Explain, Blink, Read, Write, Show, Hide, and Wave.

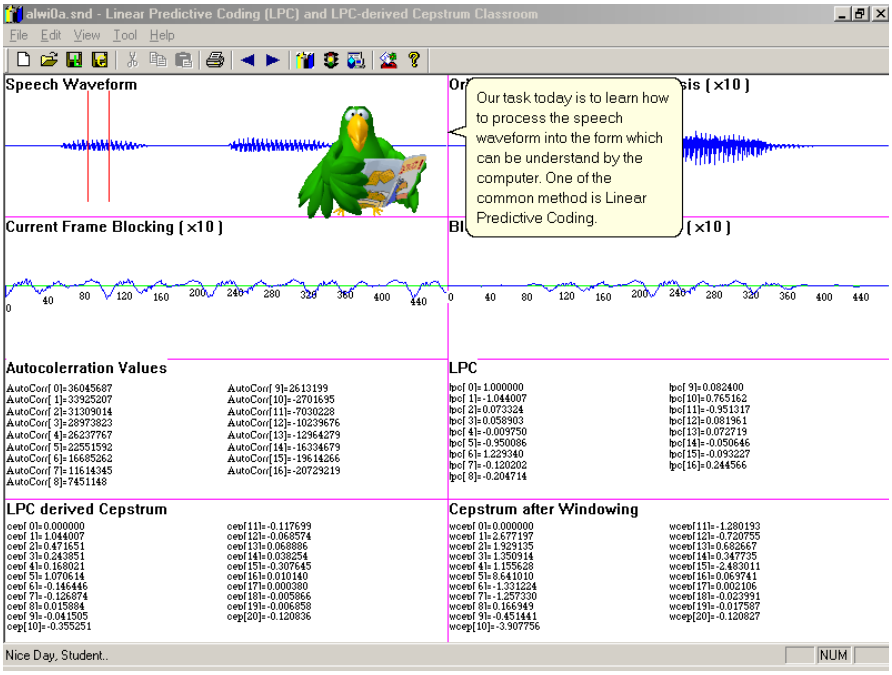

Figure 5 LPCC classroom

4.4 VQ Classroom

Figure 6 shows the VQ Classroom. Before a codebook can be built, the input type, LPC, Cepstrum or MFCC, has to be chosen. Then, a list of training files has to be loaded into the list box. After selecting the desired number of stage, a codebook can be built by clicking 'Build a New Codebook' button. VQ can be performed by clicking the 'Build Vq' button and these 'vq' files can be viewed in the notepad.

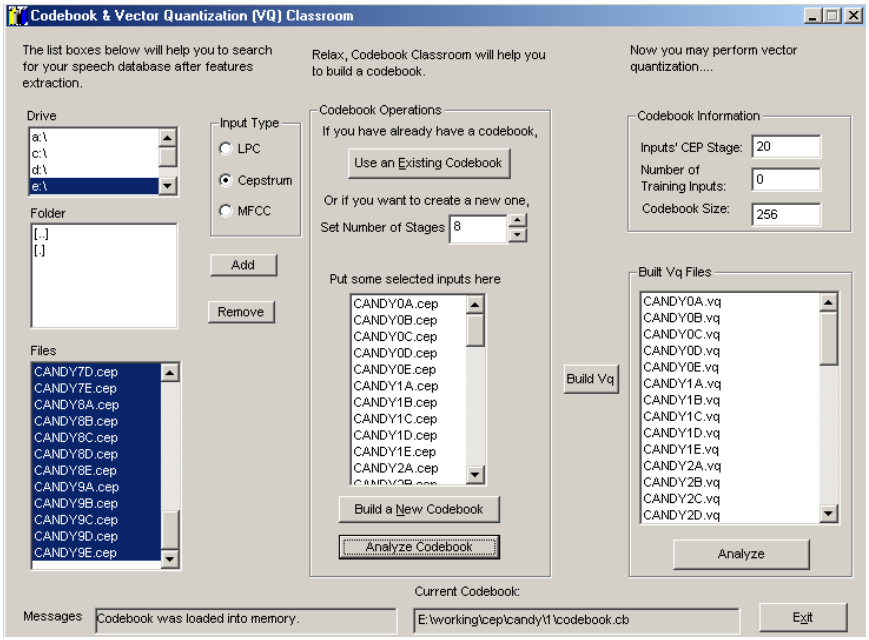

Figure 6 VQ classroom

Besides, the existing codebook can also be used. Figure 8 shows the codebook building and VQ processes. This module was built to demonstrate the VQ process. The splitting of codeword can be controlled by the user. The position of every codeword and input can be viewed by the user. The inputs of this demonstration can be randomized automatically.

4.5 DTW Classroom and HMM Classroom

The DTW Classroom, which is used for pattern matching, requires training templates and unknown inputs to be specified. This can be accomplished by putting all the related template files (the database) into the DTW template library. This module compares four most commonly feature files - '.lpc', '.cep', '.mfc', and '.vq'. The outlook of the DTW template library is shown in Figure 7.

HMM Classroom demonstrates the model training and speech recognition. Figure 8 shows the GUI design of the HMM Classroom. As can be seen from this figure, the HMM Classroom can perform pattern comparison and shows the best model in the document view.

4.6 Lab Sheets

Besides creating interesting and useful GUI in the education software, lab sheets are also prepared so that the students may learn speech recognition through following the

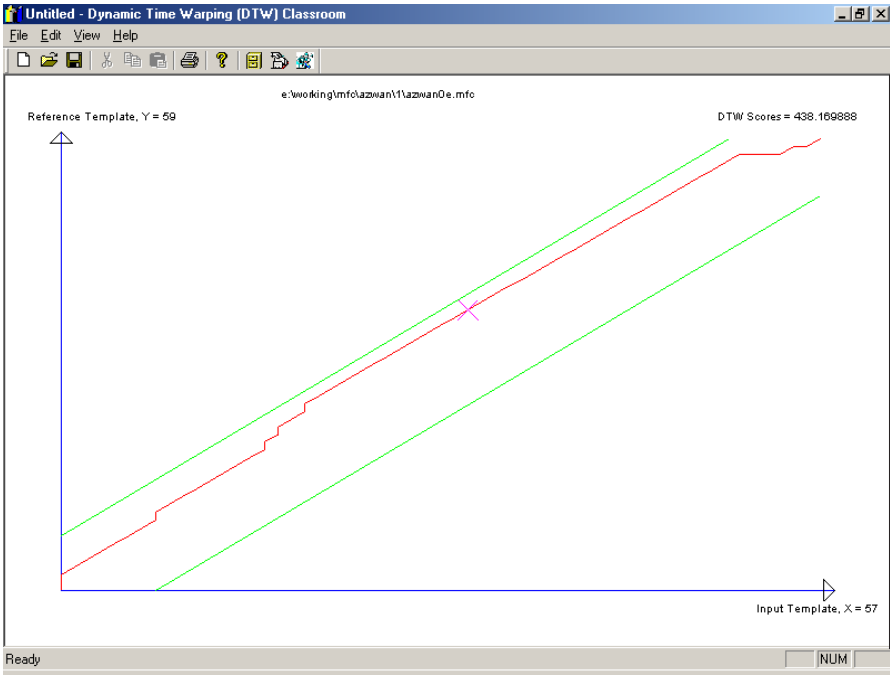

Figure 7 DTW classroom

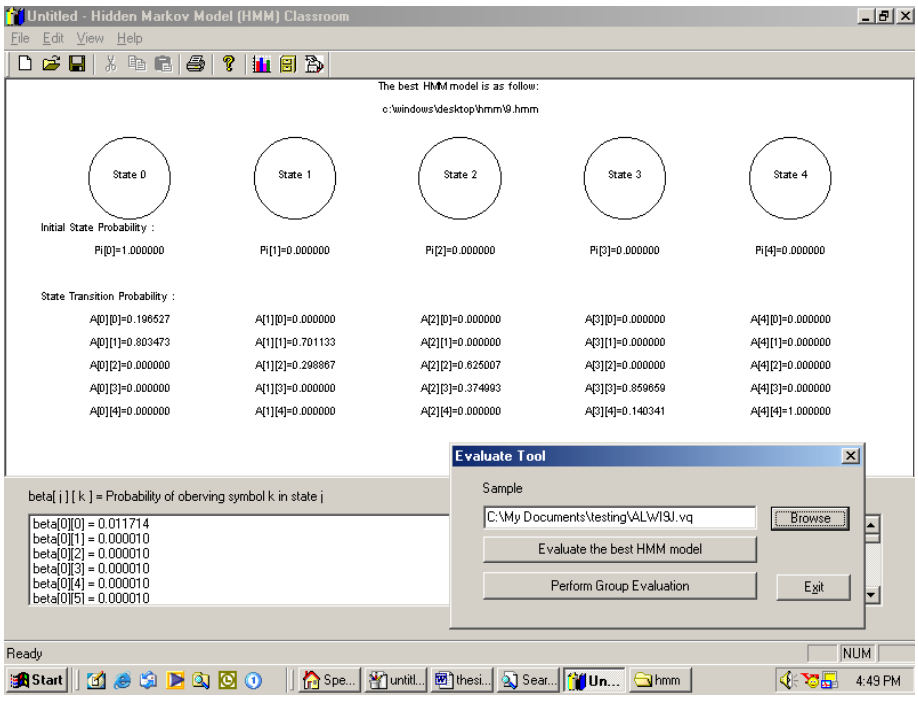

Figure 8 HMM classroom

instructions in the lab sheets. There are a total of 6 lab sheets that cover most of the focused topics in this software.

4.7 Help Files

The main purpose of creating help files for the education software is to expose students to the background theory of speech recognition as well as helping them to use the functions provided by the software. The background theory was initially created in the Microsoft Word 2000 format. The same process as shown in the previous section was performed to convert the background theory into the HTML Help Files (extension .chm).

4.8 Movies

Besides developing help files to help students in every aspect, many animation movies were created to visually guide students in performing various tasks. Camtasia Studio Version 1.0.1 is used to perform screen recording. 23 animation movies (with .avi extension) were carefully created to provide visual guidance to the tasks.

4.9 Set Up Files

Since this education software is intended to be distributed to many students, an installation setup CDROM is needed. In order to build an installation CD for Speech Recognition Virtual Classroom, Installshield Version 7.0 was selected as the development platform. Following the assistance of the Installshield, the setup CDROM was made. The setup process is very easy and standard with other programs.

5.0 APPLICATIONS AND DEMONSTRATIONS OF THE EDUCATION SOFTWARE

In this section, two related applications will be described. The first one is the speech-control robotic manipulator arm. A paper describing this application [5] was published in the proceedings of The Seventh International Conference on Control, Automation, Robotics, and Vision, December 2002, Singapore. Previous works were done by C. S. Lim *et al.* [1] and K. S. Hong *et al.* [3]. Now, this application is integrated as part of this research project. The second application is an intelligent teaching toolkit that becomes the extension to the LPC Cepstrum classroom. This application utilizes the Microsoft Agent to effectively demonstrate the theory of LPC based Cepstrum. A related paper [125] was published in the proceedings of Multimedia University (MMU) International Conference on Information and Communication Technologies Conference 2002, Petaling Jaya.

5.1 Linear Predictive Coding Classroom with Talking Head [20]

This teaching toolkit is an attempt to integrate Microsoft Agent in the speech education software. The aim of this project is to build an intelligent teaching toolkit that is able to guide the users step by step through the guidance given by the intelligent agent, without referring to the manual or help files. Further more, the Microsoft Agent used in this toolkit will describe the theory behind the speech processing process and help the users to understand the whole process.

For the first time, Microsoft Agent is used as assistance in teaching the algorithm of Linear Predictive Coding. Microsoft Agent is a set of programmable software services that supports the presentation of interactive animated characters within the Microsoft Windows interface. Developers can use the characters as interactive assistants to introduce, guide, entertain, and enhance their web pages or applications, besides the conventional use of windows, menus, and controls. A demo to the public had been held in the lab to see the effectiveness of the toolkit. As a result, the teaching toolkit attracted many students to know more about speech processing.

6.0 CONCLUSION

In this paper, the main contribution and target of the research is the designation and development of the new education software for speech recognition. The new education software solves the problem of lacking speech recognition teaching toolkits at University of Technology, Malaysia. According to Noam Amir [21], undergraduate instructors are often extremely busy people, it is therefore, unrealistic to expect them to spend very large amount of time developing educational software. The software can also be used as a teaching aid in other universities for the speech recognition related subjects. Future research can be done easily because the software was built utilising Object Oriented Programming (OOP) concepts that ease the adding of teaching modules. This software is essential for initialising large scaled Malay-based speech recognition research because the new researchers can join the research and learn from the existing modules. Thus, they can develop the additional modules to expand the current education software. This is a very good way for the students to learn and develop their motivation, curiosity, and creativity.

The experiments and analysis were performed using the Malay Speech Database. A set of Malay digit database was collected, added to the original database and well organised so that it is ready to be used for testing purposes. To date, there is still no proper Malay language based speech recognition engine that is available today. Previous works for Malay-based speech recognition system were done in the laboratory. This user-friendly software will be available in the Installation CD or it can be downloaded from the Internet. This research developed a Multi Speaker and Speaker Independent Malay digit speech recogniser experiments that can later be used by other programmers to enhance their programs with Malay speech recognition capabilities.

In order to improve the current existing system, a few algorithms have been implemented using the current available theories. One of them is Mel-Frequency Cepstral Coefficients (MFCC), one of the new feature extraction techniques. The experiments were performed comparing the recognition accuracy. Using the same database, the former algorithm - Linear Predictive Coding Cepstrum (LPCC) achieved the accuracy of 100%(SS), 98.08%(MS), and 72.56%(SI) while the MFCC managed to get 99.92%(SS), 97.76%(MS), and 83.06%(SI). From the view of speech end points detection algorithms, Rabiner-Sambur's algorithm was implemented and found to be more accurate than the energy-based end points detection algorithm. Some modifications [22] have been added to the original algorithm because most Malay digits have more than one syllabus and these syllabuses are far from each other.

The animation talking head has also been added to education software. The aim of this project is to build an Intelligent Teaching Toolkit in speech processing that is able to guide the learner step by step through the instructions given by the Intelligent Agent. Related work was published in [20]. Help files, lab sheets, and movies have been designed for this new developed software. The software was also integrated to the application such as speech control robotic manipulator arm as published in [23].

ACKNOWLEDGEMENTS

The authors would like to thank the Ministry of Science and Technology and Universiti Teknologi Malaysia for providing the financial support and facilities. This research was supported using IRPA grant of Vot. No. 74073.

REFERENCES

[1] O'Shaughnessy. D. 1988. Linear Predictive Coding. *IEEE Potentials.* (7): 1.

[2] Hwang, I. C. *et al.* 1998. An LPC Cepstrum Processor for Speech Recognition. Proceedings of the 1998 *IEEE* International Symposium on Circuits and Systems. 31 May-3 Jun. 4. 233-236.

[3] Wang. J. C., J. F. Wang., and Y. S. Weng Chip. 2000. Design of Mel Frequency Cepstral Coefficients for Speech Recognition. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

[4] Chang, C. C., and Y. C. Hu. 1998. A Fast LBG Codebook Training Algorithm for Vector Quantization. IEEE Transactions on Consumer Electronics. 1201-1208.

[5] Vuckovic. V. 2001. Dynamic Time-Warping Method for Isolated Speech Sequence Recognition. *5th International Conference on Telecommunications in Modern Satellite, Cable, and Broadcasting Service*. 1: 257 - 260.

[6] Rabiner. L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*. Volume: 77 Issue: 2: 257-286.

[7] Akbar. M. 1997. WaveEdit, An Interactive Speech Processing Environment for Microsoft Windows Platform. *Eurospeech 97 Conference Proceedings*. Greece: Eurospeech. 677-680.

[8] Nouza. J., M. Holada., and D. Hajek. 1997. An Educational and Experimental Workbench for Visual Processing Of Speech Data. *Eurospeech Conference Proceedings*. Greece: Eurospeech. 661-664.

[9] Kay Elemetrics Corp. CSL – Computerized Speech Lab Model 4300b. *CSL Product Brochure*. New York. USA.

[10] Morrow, M. G., and T. B. Welch, WINDSK: 2000. A Windows-based DSP Demonstration and Debugging Program. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

[11] Abut, H., and Y. Ozturk. 1997. Interactive Classroom for DSP/Communication Courses. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

[12] Spanias, A. *et al.* 2000. Development and Evaluation of a Web-based Signal and Speech Processing Laboratory for Distance Learning. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing.*

[13] Hetherington, L., M. McCandless., and SAPPHIRE. 1996. An Extensible Speech Analysis and Recognition Tool Based on Tcl/Tk. *Proceedings of Fourth International of Spoken Language*.

[14] Sehhati, S. Java. 1999. Based Speech Analysis via Internet Spannet. *Proceedings of International Conference on Electronics, Circuits, and Systems*.

[15] Spanias, A., and E. M. Painter. 1996. A Software Tool for Introducing Speech Coding Fundamentals in a DSP Course. *IEEE Transactions On Education.*

[16] Painter, E. M., and A. A Spanias. 1996. MATHLAB Software Tool for the Introduction of Speech Coding Fundamentals in a DSP Course. *Proceedings of 26th Annual Conference Frontiers in Education Conference*.

[17] Patterson. E. K., D. Wu., and J. N. Gowdy. 1998. Multi-Platform CBI Tools Using Linux and Java-Based Solutions for Distance Learning. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*.

[18] Hennebert, J., and D. P. Decacretaz. POST: 1996. Parallel Object-Oriented Speech Toolkit. *Proceedings of Fourth International Conference on Spoken Language.*

[19] Sario.N., A. Paoloni., and B. Saverione. ARES: 1989. An Environment for Speech Analysis and Labelling. *Proceedings of Mediterranean Electrotechnical Conference*.

[20] Tan. T. S., S. H. S. Salleh., and K. S. Hong. 2002. Intelligent Teaching Toolkit in Speech Processing Utilizing Microsoft Agent. Proceedings of Multimedia University (MMU) International Conference on Information and Communication Technologies Conference. Petaling Jaya.

[21] Amir. N. 2000. The Role of Graphical Programming Languages in Teaching DSP. *International Conference on Acoustics, Speech, and Signal Processing.*

[22] Hong. K. S., S. H. S. Salleh., and A. Z. Sha'ameri. 2002. Speaker Independent Speech Recognition for Malay Digits Using Different Feature Extraction Techniques. *Proceedings of Multimedia University (MMU) International Conference on Information and Communication Technologies Conference*. Petaling Jaya.

[23] Salleh. S. H. S., K. S. Hong., and T. S. Tan. 2002. Design and Development of Speech-Control Robotic Manipulator Arm. *Proceedings of The Seventh International Conference on Control, Automation, Robotics and Vision.* Singapore.