Speech-To-Text Converter System for Teaching and Learning Application

LAM YING YING

This Report Is Submitted In partial Fulfillment of Requirements For The Bachelor Degree in Electronic Engineering (Industrial Electronic)

Fakulti Kejuruteraan Elektronik dan Kejuruteraan Komputer
Universiti Teknikal Malaysia Melaka

JUNE 2016

UNIVERSTI TEKNIKAL MALAYSIA MELAKA
FAKULTI KEJURUTERAAN ELEKTRONIK DAN KEJURUTERAAN KOMPUTER

BORANG PENGESAHAN STATUS LAPORAN
PROJEK SARJANA MUDA II

Tajuk Projek : Speech-To-Text Converter System for Teaching and Learning Application

Sesi Pengajian :

1	**5**	**/**	**1**	**6**

Saya LAM YING YING ……………………………………………………………………………………………

mengaku membenarkan Laporan Projek Sarjana Muda ini disimpan di Perpustakaan dengan syarat-syarat kegunaan seperti berikut:

1. Laporan adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan dibenarkan membuat salinan untuk tujuan pengajian sahaja.
3. Perpustakaan dibenarkan membuat salinan laporan ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. Sila tandakan (√) :

□ **SULIT*** *(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

□ **TERHAD**** **(Mengandungi maklumat terhad yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

□ **TIDAK TERHAD**

Disahkan oleh:

____
(TANDATANGAN PENULIS)

____
(COP DAN TANDATANGAN PENYELIA)

Tarikh: ………………………..

Tarikh: ………………………..

“I hereby declare that the work in this project is my own except for summaries and quotations which have been duly acknowledge.”

Signature : ....

Author : LAM YING YING

Date : 14th JUNE 2016

"I acknowledge that I have read this report and in my opinion this report is sufficient in term of scope and quality for the award of Bachelor of Electronic Engineering (Computer Engineering) With Honours."

Signature : ....

Supervisor"s Name : DR. ABD MAJID BIN DARSONO

Date : 15th JUNE 2016

Specially dedicate to my beloved parents and also to my siblings and friends who give encouragement and support for me to complete this project. Thank you to my supervisor, Dr. Abd Majid bin Darsono who gave me a lot of guidance and advices throughout this project until successfully. Thank you very much to all of you.

ACKNOWLEDGEMENT

Firstly, I would like to express my deepest appreciation to the Universiti Teknikal Malaysia Melaka (UTeM) for letting me fulfill my degree holder in Bachelor of Electronic Engineering. I would also like to thank the Faculty of Electronic and Computer Engineering (FKEKK) for giving me the opportunity to write an honours thesis. Besides, I would like to express my sincere gratitude to my advisor Dr. Abd Majid bin Darsono for the continuous support of my final year project study and related research, for his patience, motivation, and immense knowledge. His guidance helped me all the time in research and writing of this thesis.

ABSTRACT

Speech-To-Text (STT) is the recognition of spoken language and converts it into written text by using computer program. It is known as speech recognition (SR) or automatic speech recognition (ASR) as well. Undeniable, speech is our primary means in communication between people. Therefore, the conversion of our speech directly into written word can make our life easier and efficient. In our real life, education has become essential and important for everyone. Undoubtedly, the needs of the documentations, reports and even teaching materials in education are a must regardless students, teachers, or even lecturers. The documents preparation is a time consuming process as we need to collect, analyze and type the information. Hence, as the need of documentation in education, it caused to the problem of long preparation time for documents and teaching materials. Therefore, the main objective of this system is to help in handle of documentations like notes or report with the conversion of user‟s speech into text and then documents. This application system was developed by using Microsoft Visual Studio (VS) with C# programming language. The audio processing was carried out within the system library in order to match user‟s speech with the language model and word databases. Then, by the matching of the speech and the word in database, the text word displayed out in graphical user interface (GUI) when it met the confidence level set programmatically.

ABSTRAK

Speech-To-Text (STT) ialah pengesanan bahasa pertuturan dan menukarkannya kepada teks yang bertulis dengan menggunakan komputer program. STT juga dikenali sebagai *speech recognition* (SR) atau *automatic specch recognition* (ASR). Tidak dapat dinafikan bahawa percakapan adalah saluran utama dalam komunikasi antara manusia. Hal ini bermakna bahawa apa yang kita tuturkan adalah apa yang kita maksudkan. Oleh itu, penukaran ucapan kita terus ke dalam perkataan yang ditulis boleh menjadikan kehidupan kita lebih mudah dan cekap. Dalam kehidupan kita, pendidikan telah menjadi sesuatu yang penting dan diperlukan untuk semua orang. Tidak dapat dinafikan, keperluannya juga tinggi dari segi dokumentasi, laporan dan juga bahan-bahan pengajaran dalam pendidikan tanpa mengira pelajar, guru, atau pensyarah. Penyediaan dokumen biasanya akan mengambil masa kerana melibatkan proses mengumpul, menganalisis dan menaip maklumat. Oleh itu, hal ini menyebabkan penggunaan masa penyediaan yang panjang untuk dokumen dan bahan-bahan pengajaran. Oleh itu, objektif utama sistem ini adalah untuk membantu dalam mengendalikan dokumen seperti nota atau melaporkan dengan penukaran ucapan pengguna ke dalam dokumen. Pemprosesan audio telah dijalankan dalam perpustakaan sistem untuk penyepadanan ucapan pengguna dengan model bahasa dan pangkalan data perkataan. Sistem aplikasi ini telah dibangunkan dengan menggunakan *Microsoft Visual Studio* (VS) dengan penggunaan C# sebagai bahasa pengaturcaraan. Kemudian, dengan pemadanan ucapan dan perkataan di dalam pangkalan data, perkataan teks akan dipaparkan di dalam antara muka pengguna grafik (GUI) apabila memenuhi tahap keyakinan yang ditetapkan dalam pengaturcaraan.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Speech-To-Text (STT) Converter System for Teaching & Learning Application is an application which can convert user‟s speech into a sequence of words by a computer program. The aim of this study is to help in handle of documentations like notes or report. The output text which successfully display after the conversion of user‟s speech can directly produce documents for user.

1.1 Introduction

In our real life, education has become essential and important for everyone. Undoubtedly, the needs of the documentations, reports and even teaching materials in education are a must regardless students, teachers, or even lecturers. Undeniable, the documents preparation is a time consuming process as we need to collect, analyze and type the information. The typing process actually is considered as a double work because we need to regenerate the sentences based on what we found, learnt and analyzed.

Speech-To-Text (STT) Converter System for Teaching & Learning Application not only able to recognize user‟s speech, but also help in time saving at which documentation preparation can be done instantly for teaching and learning process. As we are given the ability of speaking, we can make our works easily by the use of STT application. User can give their command by a simple mouse clicking to generate their document either in Microsoft Word or PDF version. User can choose to directly print out the document generated at the moment as well.

Besides that, STT application can be considered as long lasting as long as there is a need of the documentations. Every document can be generated easily and conveniently with this application. Consequently, user can still able to conduct other work simultaneously when using STT application. This shows that STT application not only able to sustain for a long period of time and also is a time-efficient application to use. This application do not need any external hardware or devices, it included only software application. Thus, it is considered as cost-effective as well.

1.2 Problem Statement

Education is currently conducted by using different methods, such as e-learning [1], games [2], Computer Assisted Language Learning (CALL) [3], and so forth. The needed of documentation in education caused to the problem of long preparation time for documents and teaching materials. For teachers and lecturers, they need to sacrifice their time in produce teaching plans and discussion report. This might delay their other work as well.

Even though there is the implementation of speech technology in CALL [3] and education support for deaf students [4], however it is not an application which could directly assist lecturer during normal discussion process with their students. The implementation in CALL is purposely assist teacher to provide their guidance to deaf students academically. This is because sign language takes time to deliver the messages out and some students would miss some messages delivered by sign language as well.

On the other hand, some students might not have enough time to jot down the important information that delivered by lecturer. Therefore, an application which could assist mainly in documentation is needed so that it can display what lecturers delivered without using whiteboard in writing. This can assist lecturer during class when give some questions, quizzes or exercised to students. Lecturer can able to know what the questions are given even those questions given at the spot.

1.3 Objective

The main objectives of this project are:

- To investigate the speech conversion techniques and its process.
- To develop a C# application which turns speech into text for teaching & learning purpose
- To analyse the performance of speech conversion in teaching & learning application

1.4 Scope of Project

The scope of this project is to design and develop an application which is able to convert speech into a sequence of words by a computer program. This application is developed by using Microsoft Visual Studio (VS) C#. The speech conversion process behind the application is first investigated in order to understand the working principle of speech-to-text system. Then, the speech to text conversion is done by loading the speech recognition library in VS C# into application for the further development purpose. A GUI will be designed by the use of VS to display for the text converted and some other features like export the text into documentation form. The application is designed as a simple interface which suitable for education purpose and easy to use for the teaching and learning process. At the end of this project, the performance and quality of application will be examined and analyse for further improvement.

1.5 Report Structure

This thesis is a combination of five chapters that contain the introduction, literature review, methodology, result and discussion and the last chapter is the conclusion and recommendation of the project.

The introduction of this project was discussed in Chapter 1. In Chapter 1, the introduction and objective of the project was discussed. Besides that, the initiative behind of the project and overall overview of the project also discussed within this chapter.

The literature review of the Speech-To-Text or Speech Recognition system and its implementation in different fields was discussed in the Chapter 2.

Chapter 3 explained about the project methodologies of this project. This chapter had shown the process of conversion of the speech into text in Speech Recognition System of Microsoft Visual Studio. The project flow of the application working principle will be demonstrated in detail.

Chapter 4 discussed about the project outcome of the project. This chapter had shown the results that produced from the method that used in Chapter III. The results were mainly consisted of the accuracy of the speech detected by five speakers with different education background, gender and race. The speech that spoke by the five speakers was based on the same paragraph of words.

The final chapter which is Chapter 5 explained about the conclusion and recommendation of this project.

CHAPTER II

LITERATURE REVIEW

This chapter will discuss in the literature review of the Speech-To-Text or Speech Recognition system and its implementation in different fields.

2.1 Speech-To-Text (STT) and Speech Recognition (SR)

Speech-To-Text (STT) is the recognition of spoken language and converts it into written text by using computer program. It is known as speech recognition (SR) or automatic speech recognition (ASR) as well. Undeniable, speech is our primary means in communication between people. It means that what we speak is what we meant. Therefore, the conversion of our speech directly into written word can make our life easier and efficient. This is why the concept of STT technologies implements until now.

However, STT is just one part of the SR. SR system is not only involve the conversion into text, it is even powerful which can use to recognize speaker‟s voice and giving commands. SR application has included in many fields, including telephone application[5], hands-free operation[6], application for the physically handicapped[7], dictation[8], translation[9], home automation[10], education for children[3], education for deaf students[4], and so forth. The application of SR shows that it is potentially and useful in our daily life.

Undoubtedly, SR is difficult to perform no matter in system or application. This is because the signal of speech is continuous, so that it is hardly to define where the word boundaries are as it has no pauses between words. Besides that, natural speech can be change with different speech‟s accent, slang, pronunciation of words, and phonemes in different contexts. Therefore, the speech spectrum will change under these circumstances. The signal processing of the words are initially affected no matter how accurate of the SR system is.

Moreover, the large vocabularies of SR system are always confusable as there are some words that sound like the same. These makes the words matching process complicated and lead to misdetection of words. The environmental factors would affect the detection of speech as well. The recorded speech is variable over room acoustics, microphone characteristics and background noise. All of these factors will change the characteristics of speech and thus decrease the accuracy of SR system.

Although SR technologies cannot understand all the speech 100% successfully, however it allows a computer to recognize in real-time and recognize all words in its databases up to 90%. Commercially available SR systems normally can only support a short period of speaking process and may successfully capture continuous speech with a large vocabulary with a high accuracy if speakers are speak in normal speed.

2.2 History of Speech Recognition (SR)

The first speech recognition system was built to recognize digit from a single speaker at normal speech rate. This system was a circuitry which built at Bell Labs in 1952. The accuracy of system is very high which varied around 97 to 99 percent. It measured a simple function of the spectral energy over time in two wide frequency bands, which is below 900 Hz and above 900 Hz[11]. The system worked with the concept that the speech detected was filtered into the low and high frequency component with cutoff frequency at 900 Hz. Although the system idea was intentionally good, but the technology was not able to support for the advance system idea as it used analog electrical components that difficult to modify[12].

In 1956, Olson and Belar tried to recognize 10 distinct words of a single speaker. The ten words used as word memory in the system are: are, see, a, I, can, you, read, it, so, and sir. The spectral display was established by speaking and the connection between spectral memory and the word memory can be made. Olson and Belar found that the typing accuracy is 98 percent for several hundred series of words[13]. In 1959, Fry and

Denes tried to build a phoneme recognizer at University College in England in order to recognize four vowels and nine consonants. They had increased the overall phoneme recognition accuracy for words consisting of two or more phonemes[14].

Forgie, J. W., and Forgie, C. D. at MIT Lincoln Laboratories developed a system which was able to recognize 10 vowels embedded in a /b/ - vowel - /t/ format in a speaker-independent manner in 1959[15]. At the same time, Suzuki and Nakata at Radio Research Lab in Japan built the first Japanese system which was a hardware of vowel recognizer[16]. In general, speech recognition in the 1950s was more on detecting phoneme, some selected words and carried out within a small number of people.

However, speech recognition is more to the use of digital processing and computer-based started from 1960. In 1960s, Martin and his colleagues developed a speech- analysis system that used analog-threshold logic in order to recognize consonants by a number of talkers. He developed a time-normalization methods to detect the speech starts and ends[17]. At the same time, Vintsyuk proposed to use dynamic programming methods for time aligning a pair of speech utterances including the algorithms which connected with word recognition[18]. It is known as dynamic time warp (DTW) as well. DTW is an approach to match the time sequence of short-term spectral which estimated to store pattern that represent a words that are being modeled.

During 1970s, there was an active period for speech recognition. There was a big development in terms of speech recognition in large vocabularies and continuous speech. In 1970s, the Advanced Research Projects Agency (ARPA) funded on a large speech-understanding project. The main purpose of this project was to carry out 1000 word recognition from a number of speakers, connected speech, and constrained grammar with less than a 10% semantic error. In 1980s, speech recognition activities were focusing on the system capability to recognize a fluently spoken string of connected word. One of the key technologies is the hidden Markov model (HMM) approach[12].

Starting from 1990s, innovations took place in the field of pattern recognition. The circumstance had made speech recognition inapplicable as speech signal could not be detected accurately. Therefore, the error minimization concept was starting implemented. In 2000s, ARPA program was performed to develop speech-to-text technologies with the aim of achieving more accuracy output. It includes the detection of sentence boundaries, filters and disfluencies in order to perform better job in detection, extraction, summarization and translation of important information[16].

2.3 Working principle of Speech Recognition (SR) system

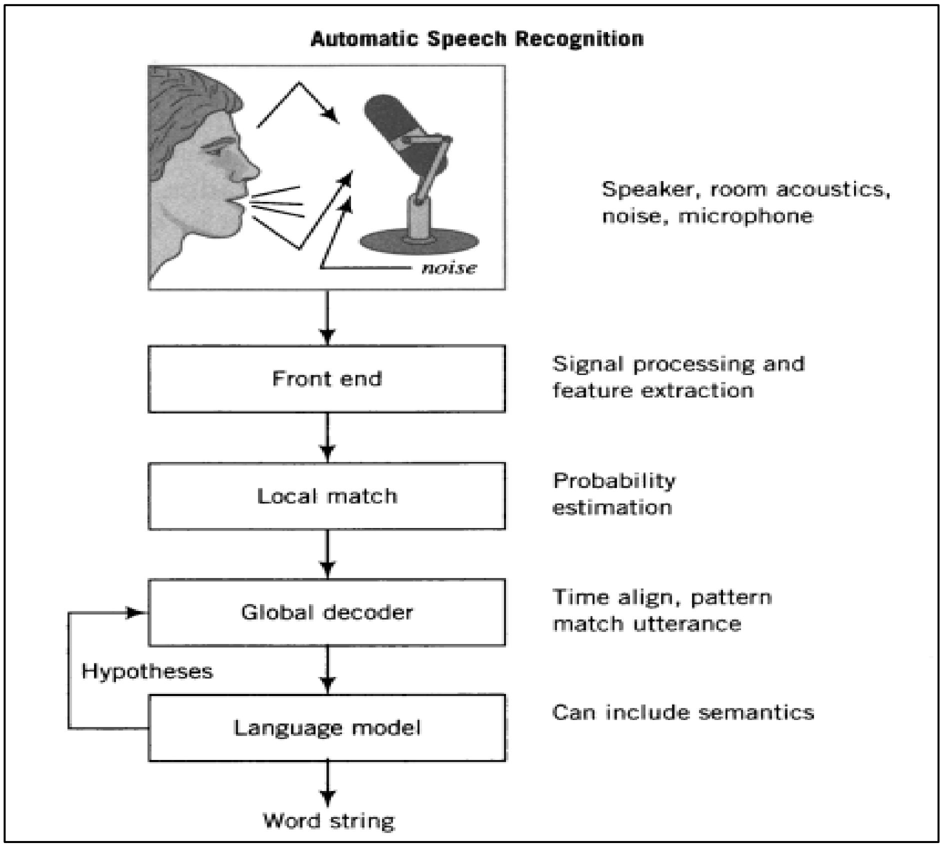
Automatic Speech Recognition
noise
Speaker, room acoustics, noise, microphone
Front end
Signal processing and feature extraction
Local match
Probability estimation
Global decoder
Time align, pattern match utterance
Hypotheses
Language model
Can include semantics
Word string

Figure 2.1: The five subsystems of a speech-recognition system. [12]

The block diagram in Figure 2.1 was defined according to Ben and Nelson in their book[12]. SR system can be considered having five different subsystems which are input acquisition, front end, local match, global decoder and language model before the word string defined.

2.3.1 Input acquisition

When speech recognition starts to work, it needs an input signal, which is voice input from the user. This part seems like unimportant, however, it is the main source of an SR system. If speech signals are not received properly at this stage, overall spectral slope of the transduced speech would be affected. Therefore, there are some factors needed to take into consideration when recording the sound inputs. One of the factors is related to the position and the choice of microphone. The quality of microphone will affect the overall noise level and influence by room acoustics. Therefore, simple preprocessing may be used to partially overcome the problems. For instance, the preprocessing can be done by flatten the spectral slope[12].

2.3.2 Front End – Feature Extraction

After an input signal received, the speech signal will undergo signal processing or in terms of speech recognition, we have known it as feature extraction. According to Ben and Nelson[12], feature extraction consists of computing representations of the speech signal that are robust to acoustic variation but sensitive to linguistic content. In other words, it means that the same word which spoken should not be vary much in a good speech recognition system. But in fact, the recognition of spoken words always detects different words even it is the same word spoken at different time.

On the other hands, B.Singh, R.Kaur, N.Devgun and R.Kaur said that the feature extraction is the process of retaining useful information of the signal while discarding redundant and unwanted information[19]. Feature extraction involves transformation of signal into digital form. The main purpose of doing feature extraction is to solve the speech signal into various acoustically identifiable components [19]. Feature extraction can be subdivided into three basic operations, which are spectral analysis, parametric transformation and statistical modeling as shown in the Figure 2.2[19].

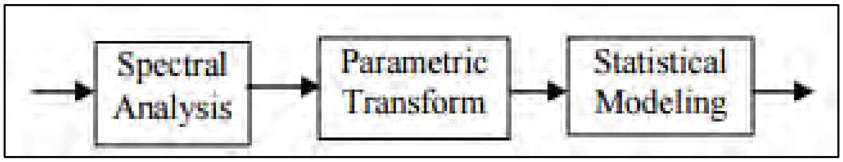

Figure 2.2: Complete sequence of feature extraction process. [19]

2.3.2.1 Spectral Analysis

For spectral analysis, it has six major classes of spectral analysis algorithms:

1. Digital filter bank
2. Fourier Transform (FT)
    i. FT Derived Filter Bank Amplitudes
    ii. FT Derived Cepstral Coefficients
3. Linear Prediction (LP)
    i. Linear Prediction Coding (LPC)
    ii. LP Derived Filter Bank Amplitudes
    iii. LP Derived Cepstral Coefficients.

In feature extraction, LPC is considered as the one of the most powerful speech analysis techniques. It is useful for encoding the quality of speech at a low bit rate. The use of LPC is mainly to minimize the sum of the squared differences between the original speech signal and the estimated speech signal over a finite duration. In fact, the concept of LP is that a specific speech sample at the current time can be approximated as a linear combination of past speech samples. LP is a human speech based model which utilizes a conventional source-filter model. In the model, transfer function of the glottal, vocal tract, and lip radiation are integrated into one all-pole filter that stimulates acoustics of the vocal tract[20].

2.3.2.2 Parameter Transforms

For parameter transforms, signal parameters are generated from signal measurements through two fundamental operations, which are differentiation and concatenation. The use of differentiation operation is to characterize temporal variations in the signal. It involves mathematical theories of differentiation in which a first-order time derivative can be approximated. The concatenation operation is working in the form of a matrix operator because it would be easier to investigate signal model in a matrix measurement.

2.3.2.3 Statistical Modeling

The third step of the feature extraction is Statistical Modeling. This model assumes that signal parameters were generated from some underlying random process. Statistical analysis is to perform on the vectors in order to determine the existence of noise in the spoken word or phrase. Speech recognition system use sophisticated statistical model because it is one of the fundamental functions of a speech recognizer. From Figure 2.3, it is stated that Vector Quantization (VQ) is applied and a training speech sequence is used to generate the codebook.

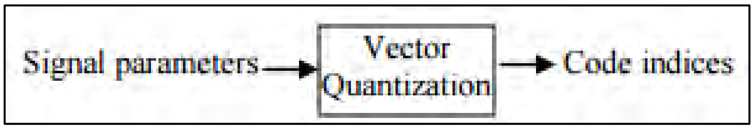

Figure 2.3: Statistical Models in speech recognition system. [20]

2.3.3 Local Match

The local match module in speech recognition system can produce a label for speech segment or in simple word. It would produce some measurement of the similarity between the speech fragment and a reference speech fragment. This process is known as probability estimation at which determine the possible word in the speech signal.