DEVELOPMENT OF SPEECH RECOGNITION SYSTEM FOR FORENSIC APPLICATION

YOU KEAN HENG

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

DEVELOPMENT OF SPEECH RECOGNITION SYSTEM FOR FORENSIC APPLICATION

YOU KEAN HENG

This Report Is Submitted In Partial Fulfillment Of Requirements For The Bachelor Degree of Electronic Engineering (Computer Engineering)

Faculty Electronic Engineering and Computer Engineering
Universiti Teknikal Malaysia Melaka

June 2014

Pengakuan

"Saya akui laporan ini adalah hasil kerja saya sendiri kecuali ringkasan dan petikan yang tiap-tiap satunya telah saya jelaskan sumbernya."

Tandatangan : ……

Nama Penulis : YOU KEAN HENG

Tarikh : 5 June 2013

Pengesahan Penyelia

"Saya/kami akui bahawa saya telah membaca karya ini pada pandangan saya/kami karya ini adalah memadai dari skop dan kualiti untuk tujuan penganugerahan Ijazah Sarjana Muda Kejuruteraan Elektronik (Kejuruteraan Komputer)."

Tandatangan : ……

Nama Penyelia : ……

Tarikh : ……

Dedication

This property is dedicated to my beloved family and friends.

Acknowledgement

Throughout this project, I would like to thank, DR. ABDUL MAJID BIN DARSONO my supervisor for guiding me in this project for the difficulties I faced, and helping me understand the project better. My gratitude extends to my families and my friends that has support me in this project.

ABSTRACT

Speech Recognition is the process of recognizing the speech spoken by a particular speaker based on important feature of speech waveform. Forensics is the use of science or technology in the investigation and establishment of evidence in the court of law. This process involves the comparison of recordings of an unknown voice with one or more recording of a known voice. However, there have variability in individual's speech and so the optimum technique for the speech recognition system is yet to be decided through the vast efforts of researchers. Signal processing front end is the process for extracting the data feature in speech recognition system. There have techniques of signal processing front-end such as Linear Prediction Coding (LPC), Mel Frequency Cepstrum Coefficient (MFCC), and other. The state-of-the-art in feature matching techniques used in speech recognition includes Dynamic Time Warping (DTW), Hidden Markov Modelling (HMM), and Vector Quantization (VQ). The objective is designing a speech recognition model using MFCC extraction technique and also with Vector Quantization model. The voice is based on isolated or single word recognition. In this project, there are 8 persons providing 5 speeches of each person for database. Every person utter twice for each speech as database. These files are recorded in Microsoft WAV format. The algorithm has been developed and the speech waveform has been analysed. The accurate and reliable speech recognition system has been developed for forensic application.

Abstrak

Pengecaman pertuturan merupakan proses untuk mengecamkan tuturan orang tersebut berdasarkan ciri-ciri gelombang tuturannya. Forensik adalah penggunaan sains atau teknologi dalam siasatan dan penubuhan keterangan dalam mahkamah undang-undang. Proses ini melibatkan perbandingan rakaman suara yang tidak diketahui dengan satu atau lebih rakaman suara yang dikenali. Walau bagaimanapun, setiap orang mempunyai pelbagai ciri gelombang tuturan. Oleh hal yang demikian, optimum teknik tersebut untuk sistem pengecaman tuturan belum dimutuskan dengan kebanyakan usaha penyelidik. Isyarat pemprosesan akhir dan hadapan merupakan proses untuk mengekstrak ciri data itu dalam sistem pengecaman pertuturan. Teknik-teknik untuk signal pemprosesan akhir dan hadapan adalah *Linear Prediction Coding* (LPC), *Mel Frequency Cepstrum Coefficient (MFCC),* dan sebagainya. Keadaan seni dalam teknik berpadan ciri merupakan *Dynamic Time Warping* (DTW), *Hidden Markov Modelling* (HMM), dan *Vector Quantization* (VQ). Objektif tersebut adalah reka bentuk model pengecaman pertuturan dengan penggunaan teknik *Mel Frequency Cepstrum Coefficient* (MFCC) dan model *Vector Quantization* (VQ). Suara tersebut berdasarkan pengecaman tuturan yang berasingan atau satu perkataan. Setiap 8 orang menyediakan 5 tuturan untuk pangkalan data dalam projek ini. Setiap orang cakap tuturan tersebut dua kali untuk pangkalan data. Fail-fail tersebut dirakamkan dalam format Microsoft WAV. Algoritma tesebut telah diwujudkan dan gelombang tuturan telah dianalisiskan. Sistem Pengecaman pertuturan yang tepat telah diwujudkan untuk penggunaan forensik.

CONTENTS

List of Figures

List of Acronym

ASR	-	Automatic Speech Recognition
MFCC	-	Mel Frequency Cepstrum Coefficients
LPC	-	Linear predictive coding
VQ	-	Vector Quantization
DTW	-	Dynamic Time Warping
HMM	-	Hidden Markov Model
GMM	-	Gaussian Mixture Model
LBG	-	Linde, Buzo and Gray

CHAPTER 1

INTRODUCTION

1.1 Introduction

Speech recognition is in terms of including all of the many different tasks of distinguishing a person from other person based on the spoken speech. Forensics is the use of science or technology in the investigation and establishment of evidence in the court of law [8]. Speech is distinct due to its non-intrusive nature. In the case of forensic, the speech is usually collected without the speaker‟s knowledge and be processed as the biometric after the speech is collected for purposes. Since the dawn of civilization, the invention and widespread use of the telephone, audio-phonic storage, television and radio has given further importance to speech processing [3].

The advancement in digital signal processing technology has affected the use of speech processing in many different application areas such as speech recognition, synthesis, enhancement, and compression. In this thesis, the issue of speech recognition is studied and a speech recognition system is developed for forensic application using Mel Frequency Cepstrum Coefficient (MFCC) and Vector Quantization model (VQ).

1.1.1 Motivation

The motivation for automatic speech recognition (ASR) is a desirable and convenient mode of communication with machines. A robust and efficient speech communication which routing to computer based speech recognition is the modeling of the human system. However, human recognizes speech through a very complex interaction between many levels of processing, for instance, using semantic and syntactic information as well very powerful level pattern classification and processing. Powerful classification algorithms and sophisticated front ends are not enough in many other forms of knowledge, for example, pragmatic, semantic, and linguistic must be created into the recognizer. Automatic speech recognition (ASR) is therefore an engineering compromise between the ideal, for instance, a complete model of the human, and the practical. In addition, the example is the tools that science and technology provided and that cost allowed.

In speech recognition systems, there are two main modules at highest level as shown in Figure 1.1 which are feature extraction and feature matching. Feature extraction is to extract the important amount of data from the speech signal. This useful data is used for representing the speaker. Feature matching relates the crucial procedure to identify the unknown suspect by comparing extracted features from user‟s speech waveform with the ones from a set of known speakers [7]. There are two different phases have to be served in all recognition systems. There are training phase and testing phase respectively. Each of registered speakers has to provide the samples of their speech in the training phase; therefore, the system can train the reference model for that speaker. The threshold of speaker-specific is calculated from training sample in the speaker verification system. Figure 1.2 show that the recognition decision is made in the testing phase.

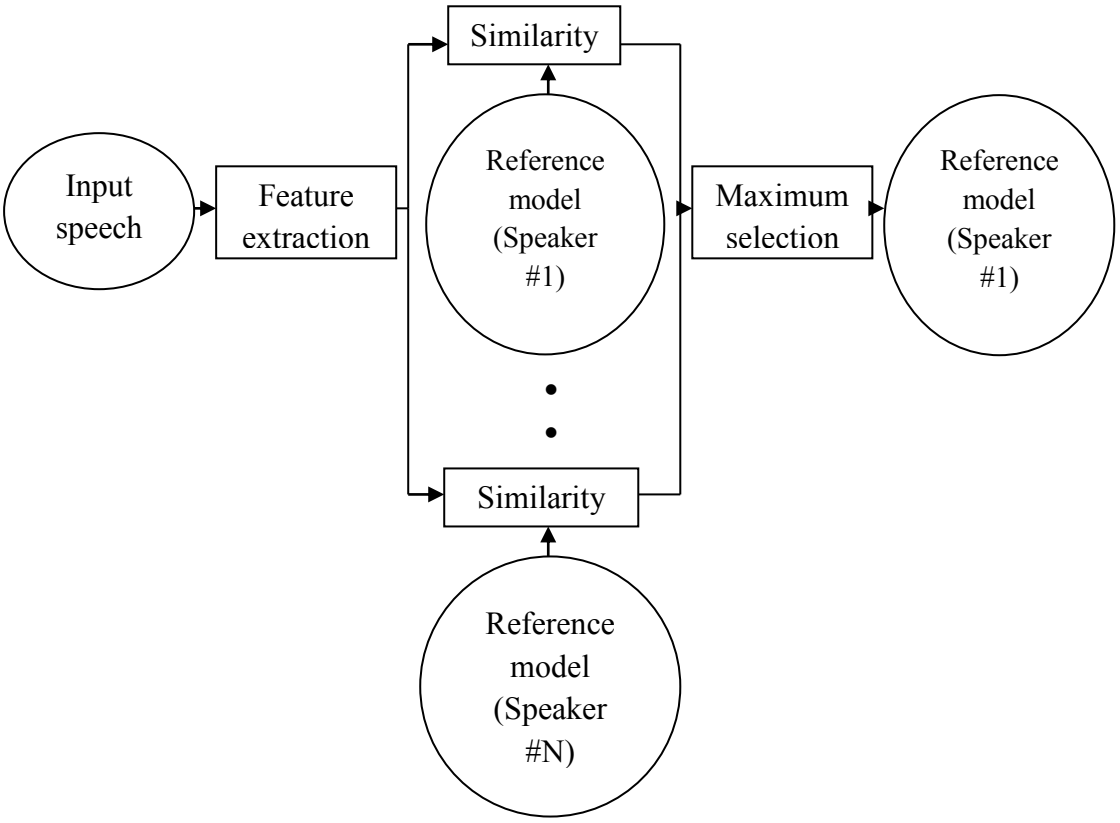

Figure 1.1: Speaker Identification Training

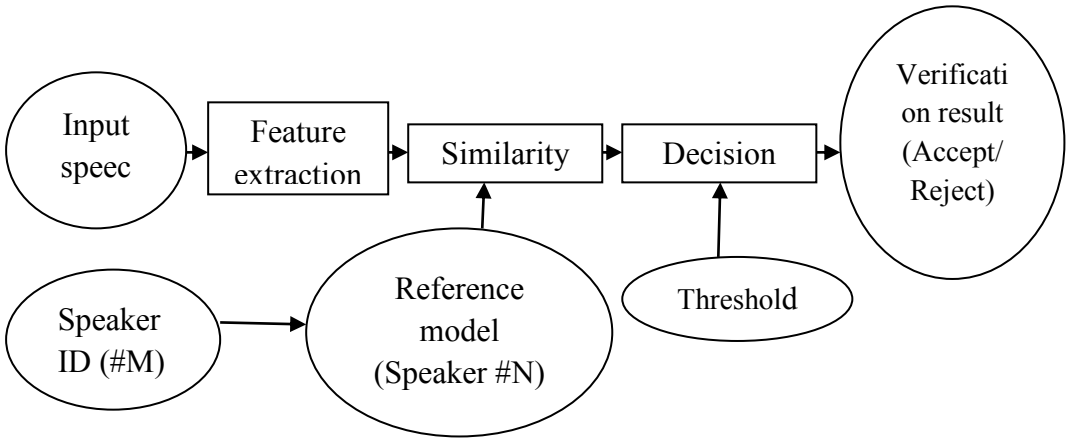

Figure 1.2: Speaker Identification Testing

Speech recognition is a hard task which is still the active research area. Automatic speech recognition works based on the reason that the person‟s speech display characteristic that are distinct to the speaker. Yet, the highly variety of input

speech signals are challenging to the task. The speakers themselves are the principle source of variance. There are different of speech signals in training and testing sessions. The reasons are the facts for instance, user‟s sound with time, the health conditions, and the speaking rate and so on. There are other factors which depending the speaker variability challenging the speech recognition technology. For instance, there are acoustic noise and variation in recording environment. The challenge would be making the system "Robust" meaning that the performance of speech recognition systems are trained with clean speech which does not demote more in the real world if under mismatch conditions between the training and testing environment.

1.2 Problem Statement

Speech Recognition System is the system of automatically recognizing a certain word spoken by a particular speaker based on individual information included speech waves. However, there is variability in individual‟s speech. Phonemes are the basic principle for describing how to take linguistic meaning to the speech. The phonemes are grouped based on the properties of either frequency characteristics or the time waveform and classifying in different speech spoken. For instance, speech is varying with time and speed; vary in pitch, the environment noise, the unlimited number of speech. Consequently, the optimum technique for the speech recognition system had not been decided through the efforts of researchers.

1.3 Objectives

This project will focus on designing the speech recognition model by applying the techniques of feature extraction and feature matching.
The objective of the project is:

1) To develop a speech recognition system for forensic application.
2) To design a speech recognition model using Mel Frequency Cepstrum Coefficient (MFCC) extraction technique.

3) To develop the speech recognition model using Vector Quantization model (VQ) technique for the stage of recognition.

1.4 Scope of the project

The scope of the project is to develop a speech recognition model for forensic by using MFCC extraction technique and also with Vector Quantization model. The numbers of people of different gender are invited to record the speech for the training session and testing session. These files will be recorded in Microsoft WAV format. The focus of the project is to develop the algorithm using MATLAB. There have to analyze the feature of the speech wave.

1.5 Expected Result

An accurate and reliable speech recognition system has been developed for forensic application by using MFCC extraction technique and also with Vector Quantization model.

CHAPTER 2

LITERATURE REVIEW

2.1 Automatic Speech Recognition System (ASR)

The speech processing is the processing methods of the input signal and the research of speech signals. The signals are processed in a digital representation therefore speech processing can be observed as the interaction of natural language processing and digital signal processing. Natural language processing is a branch of artificial intelligence and linguistics [5]. It examines the issues of human understanding of natural language which is automatically generated. Natural language understanding systems convert samples of human language into more formal representations that are easier for computer programs to manipulate, and conversion of natural language generation system from computer database information into normal-sounding human language.

2.1.1 Speech Recognition

Speech recognition is the process that computers recognize the spoken speech. The meaning is that talking to the computer then correctly recognizes that the speech

uttered. There are several ways to perform the recognition; however, the basic principle is to extract certain important feature from the spoken speech and after that taking those features as the key to recognize the speech when it is spoken again.

2.1.2 Speaker Dependence

There are speaker dependent and speaker independent in Automatic Speech Recognition (ASR). The system of speaker dependent is trained with one speaker and then the recognition is completed for that speaker whereas the system of the speaker independent is trained with one set of speakers.

2.2 Speech Analyzer

There are also called as feature extraction or front-end analysis. This speech analysis is important step and also the first step in the automatic speech recognition system. The objective of this process is to extract the acoustic features from the speech waveform. The output of the front end of the analysis is compressed and high efficiency set as seen from the input speech signal with the acoustic parameters and for subsequent used by the acoustic modeling. There are three major types of front-end processing techniques which are Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC). These two techniques are most commonly used in state-of-the-art ASR systems [3].

2.2.1 Linear predictive coding

Linear Predictive Coding (LPC) is one of the techniques of compression that models the process of speech production. Linear Predictive Coding models the process as linear sum of earlier samples using a digital filter in putting an excitement signal. The explanation is that the LPC able to predict the future value of the input

signal based on past signal. LPC "…models speech as an autoregressive process, and sends the parameters of the process as opposed to sending the speech itself" [4]. This method was first proposed as a technique for encoding human speech by United States Department of Defense in federal standard, published in 1984.

Linear Predictive Coding (LPC) has two important processes which is analysis or encoding and synthesis or decoding. The analysis part of LPC involves breaking the signal down into segments or blocks. In addition, it is examining the speech signal.

- Is voiced or voiceless segment?
- What is the pitch of the segment?
- What parameter is needed for building a filter that modeling the vocal tract for the current segment?

The analysis of LPC is conducted by the sender who is answering the questions and then transmits the answers onto the receiver. By using the receiver receives the answer for the establishment of an LPC synthesis filter set to the correct input source will be able to accurately reproduce the original speech signal. Therefore, LPC synthesis is able to try to imitate human speech production.

2.2.2 Mel Frequency Cepstrum Coefficients

MFCCs have been the dominant features used for speech recognition for some time (e.g. (Young, Woodland & Byrne 1993)). Their success is due to their representatives in a compact speech amplitude spectrum capacity. Each step in the process of creating MFCC features are driven by emotional or computational cost.
These are derived from one type of audio clips which is represented in cepstral. Mel Frequency Cepstrum and cepstrum can be distinguished in MFC which is the band is positioned on a logarithmic (Mel scale), it is closer to the human auditory system in response to the interval is not obtained directly from the FFT linear band and DCT. This may allow better handling of data, for example, in the audio compression.

However, MFCCs are missing ear model, therefore, does not represent an accurate perception of loudness. MFCCs are usually derived as follows:

I. The Fourier transform of (a windowed excerpt of) a signal is taken.
II. The log amplitudes of the spectrum is mapped which obtained above onto the Mel scale by using triangular overlapping windows.
III. As if it were a signal, to take the Discrete Cosine Transform of the list of Mel log-amplitudes.
IV. The MFCC is the amplitudes of the resulting spectrum.

.

2.3 Speech Classifier

The pattern recognition is the common topic in engineering and scientific of automatic speech recognition. The objective of pattern recognition is to identify objects of interest into one of a number of categories or classes. By using the technique which described in the previous section, the sequences of acoustic vectors which are extracted from the spoken speech are the object of interest calling pattern. The individual speakers are referred with the classes. The classification is also referred to as feature since it is applied on extracted features. The state-of-the-art in feature matching techniques used in speech recognition includes Vector Quantization (VQ), Hidden Markov Modelling (HMM), and Dynamic Time Warping (DTW).

2.3.1 Dynamic Time Warping

Dynamic Time Warping (DTW) is to find the best alignment between two time series, if a time series are "warped" non-linear stretching or shrinking technology along its time axis. The warping that is between two time series can be used to determine the resemblance between two time series or to find relevant regions between two time series. Dynamic time warping is usually used in speech recognition to determine when two waveforms are representing the same spoken speech. The duration of every spoken speech and the interval between sounds are allowed to change in a speech waveform, however, the whole speech waveforms

have to be similar. Furthermore, in speech recognition, dynamic time warping has been found beneficial in many other branch of learning [6], and also including gesture recognition data, medicine, robotics, mining and manufacturing. Dynamic time warping is usually used in data mining as the distance measure between the time series.

2.3.2 Hidden Markov Model

The fundamental principle of HMM is to evaluate speech or word into probabilistic models where in the various phonemes which link to the speech or word which represent the state of the Hidden Markov Model when the transition probabilities can be the probabilities of the next phoneme being spoken (ideally 1.0). The models for the speech are the part of the vocabulary that is produced in the training phase. For instance, when the user speaks a speech in the recognition phase then it is split up into phonemes as done before, after that, it‟s HMM is produced. The most probable phoneme to be followed after speaking some particular phoneme is found from the models that had been produced by comparing the phoneme with the newly formed model. Finally, the most probable speech or word is stored in this chain from a phoneme to other phoneme. Consequently, the recognition is performed in the finite vocabulary system. This probabilistic system is more efficient than only cepstral analysis when these are some amount of agility in terms of how the speeches are spoken by the users.

2.3.3 Vector Quantization

Vector Quantization is the process of mapping vectors from a large vector space to the finite number of regions in the space. . Every region is called a cluster and represented by its center called the centroid. The collection of all codeword is called a codebook [1]. Figure 2.1 shows a conceptual diagram to illustrate this recognition process. There are only two speakers and two dimensions of the acoustic space shown in this figure. The circles are representing to the acoustic vectors from