FEATURE EXTRACTION AND CLASSIFICATION OF MALAY SPEECH VOWELS

SHAHRUL AZMI BIN MOHD YUSOF

© This item is protected by original copyright

UNIVERSITI MALAYSIA PERLIS

2010

Feature Extraction and Classification of Malay Speech Vowels

by

Shahrul Azmi bin Mohd Yusof

(0640610049)

A Thesis submitted

In fulfillment of the requirements for the degree of

Doctor of Philosophy (Mechatronic Engineering)

School of Mechatronic Engineering

UNIVERSITI MALAYSIA PERLIS

2010

© This item is protected by original copyright

ACKNOWLEDGEMENT

I would like to thank my main supervisor Prof. Dr. Sazali Yaacob for his guidance and comments during my work on the thesis. He has always been so patient with me, continuously giving me encouragement and always confident that I will complete this journey successfully. I have always respected you as my supervisor, my boss, my mentor and my friend. Next, I would like to thank my next supervisor, Assoc. Prof. Dr. Paulraj Murugesa Pandiyan for all the knowledge he has shared with me. He has always treated me more like a colleague than his student. For that, I sincerely thank you from the bottom of my heart.

I wish to thank my employer, Universiti Utara Malaysia, for funding me to further my studies. I also wish to express gratitude to my colleagues from Universiti Utara Malaysia, namely Assoc. Prof. Fadzilah Siraj, Assoc. Prof. Dr. Lim Kong Teong, Dr. Nur Idayu Mahat, Dr. Azizan Saaban for sharing their valuable knowledge in their expert fields of Artificial Intelligence, Statistics, Mathematics and Pattern Recognition. I will always cherish the wonderful discussions we had and that I am really grateful you had find the time to discuss with me on my work. I would like to also thank distinguished individuals like Prof. Dr. Nagarajan Ramachandran and Prof. Dr. Farid Ghani from UniMAP for giving me useful tips on how to improve my work.

Work is only one part of life. The other part is personal. It is impossible to work with a light heart if the personal life is disturbed. Therefore, I would like to thank all of my family members and friends for being so supportive during this long journey. I would also like to thank my parents and my family in giving me the support and confidence in

© This item is protected by original copyright

taking this journey. I have also been blessed with a number of friends with whom I have enjoyed their company in making my PhD trip less stressful and more bearable. I thank them for making my life as wonderful as it is.

In the end, I would like to thank my beautiful wife, Normiyah Jusoh, for walking beside me in this journey, and for making every step in the path of life worth taking. Her spirit, devotion and sacrifices were the fuel which kept me going. I would also like to thank my six beautiful children, Muhammad Faris, Muhammad Hakimi, Nur Sarah Aisyah, Aziz Imran, Sufiya Najihah and Shaza Fatini for filling my life with happiness and joy. I could have never achieved this without all your love, support and encouragement.

© This item is protected by original copyright

TABLE OF CONTENTS

© This item is protected by original copyright

© This item is protected by original copyright

© This item is protected by original copyright

LIST OF TABLES

© This item is protected by original copyright

© This item is protected by original copyright

LIST OF FIGURES

© This item is protected by original copyright

© This item is protected by original copyright

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
BRKI	Bark Intensity
df	Degree of Freedom
F1BW	First Formant Bandwidth
FFB	Fixed Frequency Band
FFD	Formant Frequency Difference
FFT	Fast Fourier Transform
KNN	K-Nearest Neighbours
LDA	Linear Discriminant Analysis
LM	Levenberg-Marquardt (LM) Network
LPC	Linear Predictive Coding
MFCCs	Single Framed Mel-frequency cepstrum coefficients
MFCCf	Multi-Framed Mel-frequency cepstrum coefficients
MLR	Multinomial Logistic Regression
SNR	Signal-to-Noise Ratio
SPSS	A computer program used for statistical analysis
SpD	Spectral Delta
ZCR	Zero-Crossing Rate

© This item is protected by original copyright

ABSTRAK

EKSTRAKSI CIRI DAN KLASIFIKASI VOWAL SUARA MELAYU

Dalam bahasa manusia, fonem ialah unit struktural terkecil yang membezakan makna. Biasanya, bahasa seperti bahasa Inggeris umumnya menggabungkan fonem untuk membentuk sesuatu perkataan. Dalam banyak bahasa, unit konsonan-vowal (CV) mempunyai frekuensi kejadian yang tertinggi di antara pelbagai bentuk unit sub-perkataan. Oleh kerana itu, pengecaman unit CV dengan ketepatan yang baik adalah sangat penting untuk pembangunan sistem pengenalan suara. Ada juga banyak aplikasi yang berdasarkan kepada fonem vowal. Diantaranya ialah sistem terapi bicara yang dapat meningkatkan sebutan perkataan terutama kepada anak-anak. Ada juga sistem yang mengajar pesakit cacat pendengaran untuk bercakap dengan sebutan betul dengan mengucapkan kata-kata pada tahap kefahaman yang tinggi. Semua sistem ini memerlukan kebolehan pengenalan vowal yang sememangnya menjadi fokus di dalam tesis ini. Tesis ini menyumbangkan empat kaedah ekstraksi ciri yang diperbaikan untuk pengecaman vowal berdasarkan intensiti jalur frekuensi saringan. Cadangan-cadangan yang baru itu adalah Jalur Lebar Forman Pertama (F1BW), Jalur Frekuensi Forman Tetap (FFB), Intensiti Bark (BrKI), Delta Spektral (SpD) dan Perbezaan Frekuensi Forman (FFD). Kemampuan keempat-empat kaedah ini akan dibandingkan dengan tiga kaedah ekstraksi ciri konvensional iaitu Koefisien Cepstal Frekuensi Mel satu frem (MFCCs), Koefisien Cepstal Frekuensi Mel banyak frem (MFCCf) dan 3-Forman Pertama. Klasifikasi-klasifikasi yang dianalisa adalah Regresi Logistik Multinomial (MLR), Rangkaian Levenberg-Marquardt (LM), Jiran Terhampir-k (KNN) dan Analisa Diskriminan Linear (LDA). Ada empat sumbangan utama dari tesis ini. Pertama adalah korpus vowal baru yang terdiri daripada lebih dari 1300 huruf vowal dirakam dari 100 individu Malaysia. Kedua adalah lima kaedah ekstraksi ciri yang telah menunjukkan prestasi lebih baik berbanding MFCC jika dianalisa dengan frame tunggal. Ketiga adalah prestasi dan analisa kerobusan menggunakan klasifikasi yang berbeza dan tahap kebisingan Gaussian yang berbeza. Sumbangan keempat adalah kriteria untuk melakukan analisis vowal terpencil.

© This item is protected by original copyright

ABSTRACT

FEATURE EXTRACTION AND CLASSIFICATION OF MALAY SPEECH VOWELS

In human language, a phoneme is the smallest structural unit that distinguishes meaning. Normally, language like English commonly combines phonemes to form a word. In many languages, the Consonant-Vowel (CV) units have the highest frequency of occurrence among different forms of subword units. Therefore, recognition of CV units with a good accuracy is crucial for development of a speech recognition system. There are also many applications that use vowels phonemes. Among them are speech therapy systems that improve utterances of word pronunciation especially to children. There are also systems that teach hearing impaired person to speak properly by pronouncing words with a good degree of intelligibility. All of these systems require high degree of vowel recognition capability in which this study focuses on. This thesis contributes five modified feature extraction methods for vowel recognition based on intensities of the Frequency Filter Bands. They are First Formant Bandwidth (F1BW), Fixed Formant Frequency Band (FFB), Spectral Delta (SpD), Bark Intensity (BrKI) and Formant Frequency Difference (FFD). The performance of these five proposed methods are compared with performance of three conventional feature extraction methods of single frame Mel-frequency cepstrum coefficients (MFCCs), multiple frame Mel-frequency cepstrum coefficients (MFCCf) and the first three formant features. The classifiers analysed in this study were Multinomial Logistic Regression (MLR), Levenberg-Marquardt (LM) network, k-Nearest Neighbors (KNN) and Linear Discriminant Analysis (LDA). There are four main contributions of this thesis. First is the new vowel corpus consisting of more than 1300 recorded vowels from 100 Malaysian speakers. Second are the five improved feature extraction methods which perform better than MFCC on single frame analysis. The third is the performance and robustness analysis using different classifiers and different Gaussian noise level. The fourth contribution is the frame analysis criteria for isolated vowel analysis.

© This item is protected by original copyright

1. INTRODUCTION

1.1 Overview

Automatic speech recognition (ASR) has made great strides with the development of digital signal processing hardware and software especially using English as the language of choice. Despite of all these advances, machines cannot match the performance of their human counterparts in terms of accuracy and speed, especially in case of speaker independent speech recognition. Victor Zue in 2004 stated that within 5 to 10 years, systems that can handle more complicated human-to-computer interactions, like processing a request for movie tickets at a particular theatre via speech recognition should be in use (Hoffman, 2009). Today, significant portion of speech recognition research focuses on speaker independent speech recognition problem. The reasons are its wide range of applications, and limitations of available techniques of speech recognition.

Speech processing analyses and processes speech signals for information retrieval, giving commands and speaker recognition. Due to the heavy dependency on digital signals, speech processing can be placed in the area of digital signal processing and natural language processing. Speech processing covers a broad area that relates to the following important research directions like speaker recognition (Campbell Jr, 1997), Speech enhancement (Ephraim, 1992), Speech coding (Furui, 2001), Voice analysis (Jilek, Marienhagen, & Hacki, 2004), Speech synthesis (Furui, 2001) and Speech recognition (Rabiner & Juang, 1993).

Automatic Speech Recognition (ASR) is the automatic conversion of speech sound waves to text and this automatic interaction between man and machine. This has

© This item is protected by original copyright

been an increasingly interesting problem for many years due to its vast potential like cell phone voice-dialing, bank transaction with a computer, or software dictating letters to word. The language of choice is usually English but researchers from different countries are focusing on doing speech recognition using their native language.

Speech recognition system performs two fundamental operations: signal modeling and pattern matching (Picone, Inc, & Dallas, 1993). Signal modeling represents process of converting speech signal into a set of parameters. Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal. The signal modeling involves four basic operations: spectral shaping, feature extraction, parametric transformation, and statistical modeling (Picone *et al*., 1993).

Spectral shaping is the process of converting the speech signal from sound pressure wave to a digital signal; and emphasizing important frequency components in the signal. Feature extraction is the process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal. Parameter transformation is the process of converting these features into signal parameters through process of differentiation and concatenation. Statistical modelling involves conversion of parameters in signal observation vectors.

© This item is protected by original copyright

1.2 Problem Statement and Its Significance

In human language, a phoneme is the smallest structural unit that distinguishes meaning. Normally, language like English commonly combines phonemes to form a word. In many languages, the Consonant-Vowel (CV) units have the highest frequency

of occurrence among different forms of subword units. Therefore, recognition of CV units with a good accuracy is crucial for development of a speech recognition system. Recognition of these subword units is a large class set pattern classification problem because of the large number (typically, a few thousands) of units (Sekhar, Takeda, & Itakura, 2002). In this case, if ASR recognizes the vowel with a good accuracy, system can reduce region of search and improve accuracy and time.

English uses a combination of phonemes to form words which may not exactly follow the characters of the words. Because of this, a large database of vocabulary is needed in order to represent each individual word. Standard Malay (SM) on the other hand can be uttered properly based on the combination of CV phonemes. One advantage the Bahasa Malaysia has over English is the numbers of vowel phoneme that need to be considered. The proper Bahasa Malaysia has only 6 vowels phonemes which are /a/, /e/, /i/, /o/, /u/ and /ə/ (Maris, 1966) whereas typical American English has 20 vowel phonemes (Power, 2009).

There are many speech recognition applications that can help physically handicapped or paralyzed individuals in their daily life. Activities such as opening doors or pressing a button may prove very difficult for these individuals. Voice command system may help to lessen their burden. There are limitless applications for voice command systems such as opening doors, ordering and purchasing items or even turning on electrical appliances. Current new technologies are enabling robots to assist human via voice command. Most of these applications use English speech recognition capability. In Malaysia, there are still many individuals who are unable to converse well in English. Their pronunciation in English may not be suitable for these speech

© This item is protected by original copyright

recognition systems which normally works well with American or British English spoken individuals. Systems that work well with Malaysians spoken English are still limited and there are even lesser systems that utilize Malay language.

In Malay language, children are taught to spell the words using a combination of consonants (C) and vowels (V) sounds. A computer system that can read CV combinations from a person who can properly pronounce any SM words have the capability to function without a proper database. There are also many applications that use vowels phonemes. All of these systems require high degree of SM vowel recognition capability.

Although there are studies concerning Malay phoneme recognition, it is still at its infancy (Rosdi & Ainon, 2008) and multiple frame analysis is mostly in use by Malaysian researcher. Accuracy and processing time is a concern when developing speech therapy systems. More efforts are needed to be taken in order to develop Malay speech recognition system and this study is an effort to improve Malay vowel recognition.

In Malaysia, researches in vowel recognition is still lacking especially in the usage of Malay vowels, independent speaker systems, recognition robustness and algorithm speed and accuracy. There is a need to develope a better algorithm of Malay vowel recognition in terms of accuracy and robustness. This thesis will address the issues of researching on Malay word and vowel databases, independent speaker systems, and robustness analysis and algorithm improvement.

© This item is protected by original copyright

1.3 Research Objectives

This study is an effort to increase Malay vowel recognition capability by using a new speech database that consist of words uttered by Malaysian speakers from the three major races, Malay, Chinese and Indians. Robust feature extraction methods need to be developed. The main objective of this study is to increase independent speaker Malay vowel recognition capability in terms of accuracy and robustness. In order to achieve this, three sub-objectives are listed below.

i). To develop an improved feature extracting algorithms for Malay vowels using independent speaker database.

ii). To study linear and non linear classifiers in classifying vowels.

iii). To study robustness of feature extraction methods using different classifiers under different Gaussian noise level.

1.4 Motivation for the Present Work

English word pronunciation depends on a sequence of phonemes. Audio signals are broken up into acoustic components and translated into phonemes. These phoneme sequences are then compared with words from an English database that can make up of thousands of words. For Malay words, the approach is different. It is comprised of Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) combinations. It is possible that a Malay word can be spelled out by a computer similar to a human being. We believe that a computer can be taught to spell like a child and having a computer system able to translate and CV or CVC combinations into proper and understandable words.

© This item is protected by original copyright

Among the motivations for this work is to have voice command and speech recognition application which uses Malay words instead of the commonly used English words. This may allow Malaysian citizen who are not well versed in English language to be able to use this system to assist them in their daily life like purchasing items using an automated speech recognition telemarketing system. This system may also help children to improve their pronunciation of Malay words thus promoting our National Language of Bahasa Malaysia. A speech therapy system may improve utterances of word pronunciation especially to children. Hearing impaired person can learn to speak properly by pronouncing words with a good degree of intelligibility through the use of a visual therapy system and evaluate the pronunciation capability of the speaker and display the results. Even language learners from foreign countries may find it easier to learn Bahasa Malaysia with the assistance of an application that teaches Bahasa Malaysia by correcting the pronunciation of the speaker.

Another motivation for this work is to have a computer system to recognise Malay Language based on CV or CVC words. This capability may allow the speech recognition system to recognize any Malay word spoken through the sequence of consonant and vowels, independent of speaker's race, gender and age. All of this type of system requires a high degree of Standard Malay vowel recognition capability. A good application of this capability is in healthcare where foreign doctors are able to interact with a local Malaysian without the use of a translator. In tourism industry, this system may allow a foreigner to converse with a local through the use of a mobile electronic translator in order to find directions to place of interest. For example, a German tourist may be able to interact with a villager from Malaysia concerning some food delicacies from that community with an application that can recognise different

© This item is protected by original copyright

dialects. In other words, the language or even dialect barrier between different people from different culture and ethnic group can be brought down with the use of a good, reliable Malaysian speech recognition system. This will bring in more tourists and enrich the locals especially from villages and promote the tourism industry in Malaysia.

1.5 Contributions of the Thesis

Among the contributions of this thesis are the five improved feature extraction methods for vowel recognition based on intensities of the frequency filter bands and single frame analysis. They are First Formant Bandwidth (F1BW), Fixed Formant Frequency Band (FFB), Spectral Delta (SpD), Bark Intensity (BrKI) and Formant Frequency Difference (FFD). These features are analysed using four different classifiers of Linear Discriminant Analysis (LDA), k-Nearest Neighbours (KNN), Multinomial Logistic Regression (MLR) and Levenberg-Marquardt (LM). Robustness analysis of these features and classifiers are also contributions of this thesis in providing robustness capability of different extraction methods compared to conventional methods of Mel-frequency cepstrum coefficients (MFCC) and the first three formant frequencies. This robustness analysis also includes the capability of the classifiers under different SNR condition.

© This item is protected by original copyright

To summarize the contributions of this thesis, four main contributions are listed below. They are:

i). New vowel corpus consists of more than 1300 recorded vowels from 100 Malaysian speakers.

ii). Five improved feature extraction methods which perform better than MFCC on single frame analysis.

iii). Performance and robustness analysis of Malay Vowel Classification under different Gaussian noise level.

iv). Criteria for isolated Malay Vowel analysis.

1.6 Organization of the Thesis

This work has been divided into seven chapters:

Chapter 1 introduces the framework in which the thesis has been developed. First, overview of speech processing is presented, followed by an explanation of the Problem Statement and Significance, Research Objectives, Motivation for the present work and Contributions of the thesis.

Chapter 2 presents the literature review part of the thesis on the background studies and researches done by previous researchers in the field of speech recognition. First, an explanation of speech production and acoustics is given followed by Linear Prediction of Speech. Malay Phonemes is introduced together with a brief explanation about Standard Malay (SM). Next, an explanation of Spectral Envelope is presented followed by explanation of Frequency Scales of Bark and Mel. An explanation of current feature extraction methods based on Malay phonemes from studies done mostly by academicians from Malaysian universities is presented. A brief explanation of classification techniques is also presented.

Chapter 3 contains the methodology used in the research including the experimental setup, the feature extraction methods and experimental work. It presents the database

© This item is protected by original copyright

used in the experiments, along with an endpoint detector designed for cleaning the samples.

Chapter 4 explains the 4 new feature extraction methods of First Formant Bandwidth, Spectral Delta, Formant Frequency Difference, Bark Intensity and Fixed Frequency Band. Their performance will be compared with the first 3 formant features and Mel-frequency cepstrum coefficient features. Feature validation is also presented using ANOVA method to show the mean significance of each vowel against the features.

Chapter 5 discusses the classification results obtained in the experiments. Then, a detailed comparison between the performance of the proposed approach and that of conventional approach is offered. These results will be discussed and concluded at the end of this chapter.

Chapter 6 presents the study of robustness of the feature extraction methods and three classifiers. In terms of robustness analysis, the performance of different methods and different classifiers will be presented.

Chapter 7 makes a summary and discussion of the results from the study. An application utilising of the vowel recognizer will also be discussed and presented. Finally, the most important conclusions are extracted, and future work directions are suggested.

© This item is protected by original copyright

2 LITERATURE REVIEW

2.1 Introduction

Automatic Speech Recognition (ASR) belongs to the class of digital speech processing technologies that also includes speech synthesis (text-to-speech, language generation) and voice biometrics (speaker identification, speaker verification). In general, their aim is to allow a machine to replicate human ability to hear, identify, and utter natural human spoken language. For the past 35 years, robust ASR systems have made great progress over the years. Several factors have contributed to this rapid progress, such as the development of advanced signal processing techniques but also the continuously increasing computing power.

© This item is protected by original copyright

2.2 Automatic Speech Recognition (ASR)

The earliest attempts to devise ASR systems were made in 1950s and 1960s, when various researchers tried to exploit fundamental ideas of acoustic phonetics. Since signal processing and computer technologies were yet very primitive, most of the speech recognition systems investigated used spectral resonances during the vowel region of each utterance which were extracted from output signals of an analogue filter bank and logic circuits.

The goal of an Automatic Speech Recognition (ASR) system is to transcribe speech to text. As illustrated in Figure 2.1, the speaker's mind decides the source word sequence W that is delivered through his/her text generator. The source is passed through a noisy communication channel that consists of the speaker's vocal apparatus