WEB-BASED MEL FREQUENCY CEPSTRAL COEFFICIENTS AND HIDDEN MARKOV MODEL FOR MANDARIN SPEECH RECOGNITION SYSTEM

TSEU GEANNA AFERA

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2022

UMS
UNIVERSITI MALAYSIA SABAH

WEB-BASED MEL FREQUENCY CEPSTRAL COEFFICIENTS AND HIDDEN MARKOV MODEL FOR MANDARIN SPEECH RECOGNITION SYSTEM

TSEU GEANNA AFERA

THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF BACHELOR OF COMPUTER SCIENCE WITH HONOURS (NETWORK ENGINEERING)

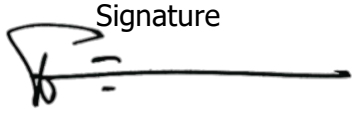
FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITI MALAYSIA SABAH
2022

UMS
UNIVERSITI MALAYSIA SABAH

NAME	**:** TSEU GEANNA AFERA
MATRIC NUMBER	**:** BI18110240
TITLE	**:** WEB-BASED MEL FREQUENCY CEPSTRAL COEFFICIENTS AND HIDDEN MARKOV MODEL FOR MANDARIN SPEECH RECOGNITION SYSTEM
DEGREE	**:** BACHELOR OF COMPUTER SCIENCE WITH HONOURS (NETWORK ENGINEERING)
VIVA'S DATE	**:** 24 JANUARY 2022

CERTIFIED BY;

1. **SUPERVISOR** Signature
   Dr. Samry @ Mohd. Shamrie Bin Sainin



DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, equations, summaries and references, which have been duly acknowledged.

24 JANUARY 2022

TSEU GEANNA AFERA
BI18110240

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr Samry for accepting to supervise me throughout this entire final year project journey, providing guidance and immense knowledge with unconditional patience and giving me the chance to do this specific project on speech recognition, a topic that I have been interested in for some time.

Besides that, I would like to thank the examiners, Professor Dr Jason Teo Tze Wi and Madam Salmah Binti Fattah for their comments and feedback to help improve this project further and point out the aspects that I can focus on for future improvements.

My sincere thanks also go to the faculty's final year project coordinators, Dr Tan Soo Fun, Dr Chin Pei Yee and Dr Lai Po Hung for the constant support and encouragement throughout the entire journey.

Not to forget, Madam Chee Shiaw Phin and Madam Ernahwatikah Nasir from the Centre for the Promotion of Knowledge and Language Learning (PPIB), Universiti Malaysia Sabah and also Mr Lim Hwan Yih from SJK(C) Chung Hwa, W.P. Labuan for responding to my request to ask for Mandarin speech samples for this project. My appreciation also extends to my friends and acquaintances who have agreed to provide me with their speech samples and agreed to participate in the testing phase of this project.

I would also like to thank my family, for being the source of motivation for me. I cannot repay the immense patience and support that they have given during my entire journey from the day I started my studies here.

Finally, I thank my friends, especially those in my batch for always spreading positive energy and helping me in multiple times of crisis.

Tseu Geanna Afera

24 January 2022

ABSTRACT

Learning a new language may be difficult for adults, especially the communication aspect of the process because pronunciation accuracy can be a challenge to master. Mandarin Chinese is a language that depends on its tone, where each character has a tone associated with it out of the available five tones. One of the most effective ways to learn the correct pronunciation in Mandarin is to practice reading and speaking with a teacher who will be able to listen and give feedback. However, this is not a practical method as the teachers cannot be available to attend to every student in the classroom, especially if it is a large class. Therefore, a Web-based Mandarin speech recognition system is proposed to help with tackling this problem. The objectives of this project are to investigate Mel Frequency Cepstral Coefficient (MFCC) and Hidden Markov Model for the proposed web-based Mandarin speech recognition system, to develop and evaluate the prototype for the proposed web-based Mandarin speech recognition system. The proposed system is targeted at students in Universiti Malaysia Sabah who are currently taking beginner levels Mandarin language. The prototype consists of several phrases to detect the speakers' pronunciation. The Mandarin speech recognition system applies the Hidden Markov Model (HMM) as the machine learning model and is implemented in PHP as well as Python in a web-based application. The project allows the user to record their speech and see the accuracy through the prediction output. Based on the training and testing, the MFCC and HMM produce speech model that yields an accuracy probability ranging from 0.80 to 0.94 and for the testing, the phrase prediction may yield different phrases than the spoken phrase but overall the accuracy probability ranges from 0.53 to 1.0, therefore it can be used as basic speech recognition. Future works include gamification of the system to make it more interesting and to train the speech model with word phonemes to enhance the performance of the system.

Keywords: Speech Recognition, Machine Learning, Hidden Markov Model, Mel Frequency Cepstral Coefficients, Mandarin Chinese

ABSTRAK

PEKALI CEPSTRAL FREKUENSI MEL DAN MODEL MARKOV TERSEMBUNYI UNTUK SISTEM PENGECAMAN PERTUTURAN BAHASA MANDARIN BERASASKAN WEB

Mempelajari bahasa baharu mungkin sukar untuk orang dewasa, terutamanya aspek komunikasi kerana ketepatan sebutan boleh menjadi suatu cabaran untuk dikuasai. Bahasa Mandarin ialah bahasa yang bergantung pada nadanya, di mana setiap aksara mempunyai nada yang dikaitkan dengannya daripada lima nada yang tersedia. Salah satu cara yang paling berkesan untuk mempelajari sebutan yang betul dalam bahasa Mandarin adalah dengan berlatih membaca dan bercakap dengan seorang guru yang boleh mendengar dan memberi maklum balas. Walau bagaimanapun, ini bukan kaedah yang praktikal kerana seseorang guru tidak boleh ada untuk setiap pelajar di dalam bilik darjah, terutamanya jika ia adalah kelas yang besar. Oleh itu, sistem pengecaman pertuturan berasaskan web untuk Bahasa Mandarin dicadangkan untuk membantu menangani masalah ini. Objektif projek ini adalah untuk menyiasat Pekali Cepstral Frekuensi Mel (MFCC) dan Model Markov Tersembunyi untuk sistem pengecaman pertuturan Mandarin berasaskan web yang dicadangkan, untuk membangunkan dan menguji prototaip untuk sistem pengecaman pertuturan Mandarin berasaskan web yang dicadangkan. Sistem yang dicadangkan itu disasarkan kepada pelajar di Universiti Malaysia Sabah yang kini mengambil bahasa Mandarin peringkat awal. Prototaip terdiri daripada beberapa frasa untuk mengesan sebutan penutur. Sistem pengecaman pertuturan Mandarin menggunakan Model Markov Tersembunyi (HMM) sebagai model pembelajaran mesin dan dilaksanakan dalam PHP serta Python dalam aplikasi berasaskan web. Projek ini membolehkan pengguna merakam pertuturan mereka dan melihat ketepatan melalui output ramalan. Berdasarkan latihan dan ujian, MFCC dan HMM menghasilkan model pertuturan yang menghasilkan kebarangkalian ketepatan antara 0.80 hingga 0.94 dan untuk ujian, ramalan frasa boleh memberikan frasa yang salah tetapi secara keseluruhan kebarangkalian ketepatan berjulat dari 0.53 hingga 1.0, dan boleh digunakan sebagai sistem pengecaman pertuturan asas. Penambahbaikan pada masa hadapan termasuk gamifikasi sistem untuk menjadikan sistem lebih menarik dan melatih model pertuturan dengan fonem perkataan untuk meningkatkan prestasi sistem.

Kata Kunci: *Pengecaman Pertuturan, Pembelajaran Mesin, Model Markov Tersembunyi, Pekali Cepstral Frekuensi Mel, Bahasa Mandarin*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Background

This project is done on the basis that learning a foreign language is difficult, especially for adults. Nowadays, many students will want to take up a foreign language as a plus point for their resumes or just to fulfil the university's requirements. One of the most sought after language courses in Universiti Malaysia Sabah (UMS) among undergraduates is Mandarin Chinese as it is also widely used in Malaysia. Learning a new language in adulthood is good in the sense that they will have more motivation but mastering the language is not easy as compared to younger children.

Learning a language consists of learning to write and speak. The characters in Mandarin can be mastered by constantly practising to write while referring to the textbooks. However, the most difficult parts of learning a new foreign language is how to make sure that the pronunciation of a certain phrase or word is spoken correctly. the most effective way to learn the correct pronunciation in Mandarin is to practice reading and speaking with a teacher who will be able to listen and give feedback. Unlike learning to write, the teacher will not be able to help to correct each student with their pronunciations for every phrase they taught and the students may still need to have a reliable guide to practice themselves at their own time as the books cannot confirm the accuracy of the speakers' pronunciation as well.

Mandarin Chinese is a language that is heavily stressed on its tone, which means that each character in the language is associated with one of the five tones (J. Chen *et al*, 2007). The tone of a character can be different based on various contexts, thus it becomes a troubling issue for total beginners to pronounce each character in a sentence accurately (J. Chen *et al*, 2007).

It is also important to note that sometimes if a character is spoken incorrectly, it may be confusing to those who speak Mandarin or rather gives a different meaning to the sentence altogether. Hence, there is a need to create a system that can help beginners with their pronunciation.

1.2 Problem Statement

When taking up a foreign language as a beginner, like Mandarin, for instance, there is always a challenge to get used to writing and speaking the language. The main purpose of most people taking up a foreign language is always to learn how to communicate with others in that language. Most of the time, people only realize the importance of learning a foreign language when they are in secondary school or university because then they will have an idea of what benefits they can get in the society after graduation, for example, if they were to open a new business, learning how to communicate in Mandarin can help attract the Chinese citizens to support your business. However, one of the struggles of a beginner to learn to speak in a new language with those who are used to speaking the language is that the pronunciation of certain phrases may sound inaccurate.

Usually, beginners who are aiming to learn a new language will buy books and sign up for language classes so that they can have a space to interact and practice in the language. However, even though some books and dictionaries have *pinyin* to guide the beginners on how to pronounce the words, without the teacher to listen and give feedback to them or a guide for them to compare their pronunciation, there is a possibility that their pronunciation may not be accurate.

To have a better grasp of the pronunciation of basic Mandarin phrases, a system is being developed to help detect whether the speaker is speaking the phrases in the right way. This is because even when there is a teacher to teach them how to speak, the beginner will still need to practice on their own and their teacher will not be always available to correct their pronunciation, especially in a class with more than 20 students. Hence, there is a need to have a web-based system that can be used for practice at their own time.

1.3 Project Objectives

1. To investigate Mel Frequency Predictive Cepstral Coefficient and Hidden Markov Model for Mandarin speech recognition system.
2. To develop the web-based Mandarin speech recognition system with MFCC and HMM.
3. To evaluate the prototype of the Mandarin speech recognition system and the functionality of the web-based system.

1.4 Project Scope

This project is built for educational purposes to aid UMS students who are taking beginner level Mandarin Language in the Centre for the Promotion of Knowledge and Language Learning (PPIB), Universiti Malaysia Sabah or want to learn basic Mandarin phrases. This project will feature several phrases that will be used to help the speakers practice their pronunciation. The system is a web-based system and can be used by all UMS students. The modules of the system are suggested as below:

1. Register/Login

   In this module, all users of the system will register to create their accounts and log into their created accounts.

2. List of basic Mandarin keywords

   In this module, all users will be able to see the list of basic Mandarin keywords and see the *pinyin* or pronunciation of each phrase. After that, they will be able to record their pronunciation and obtain the predicted speech with the accuracy probability to the predicted phrase.

1.5 Conclusion

This web-based Mandarin Speech Recognition System is made to aid and guide Mandarin learners who are at the beginner level to get used to the pronunciation of ten selected phrases for this system.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Speech is a method of spoken communication among human beings. R. Lotfidereshgi and P. Gournay in 2017 also mentioned that speech is a basic way of communication. As technology advances, the communication between humans and machines has included speech as its interface so that more convenience can be brought to the users. As explained by P. K. Sahu and D. S. Ganesh in 2015, the process of recognizing speech and understanding them needs a lot of knowledge on the various speech features.

Speech recognition systems can be divided into speaker-dependent and speaker-independent systems where the first one involves training the system with a certain speaker but the system will become tailored to only a trained set of voices and the latter can be used by many different untrained voices but there will be some inaccuracies (P. K. Sahu and D. S. Ganesh, 2015). There are three basic phases in speech recognition, which are signal pre-processing, feature extraction and speech classification (K. Naithani *et al.*, 2018).

2.2 Review of Similar Speech Recognition Systems

Learning how to pronounce Mandarin characters accurately is a challenge for those who had just begun their studies in this language or those who are older. Based on the research by J.-C. Chen *et al.*, 2007, proposed some techniques to build a Computer-Assisted Pronunciation Training (CAPT) system that not only judge the quality of the pronunciation but also gives comments for improvements based on

several criteria, which are phoneme, tone, intensity and rhythm. Their system was able to detect the wrong pronunciation and provides scores to the student who is using the system.

Mel Frequency Cepstral Coefficient (MFCC) is used as it is often deemed as one of the feature extraction techniques that can produce stable and better results in most speech recognition systems. Pedronan, R. *et al.,* 2017, in their research to detect accurate English pronunciation, has gone through a speech recognition process and used several feature extraction techniques that includes Mel Frequency Cepstral Coefficient (MFCC) that produces a high recognition rate of correctly pronounced words. However, to achieve a more enhanced rate, they also considered Linear Predictive Coding (LPC) in combination with Mel Frequency Cepstral Coefficient (MFCC) in their research (Pedonan, R. *et al*., 2017). K. Chakraborty *et al*. in their study has also applied Mel Frequency Cepstral Coefficient (MFCC) as an algorithm for voice recognition and mentioned that it was a popular technique due to the ease of implementation (K. Chakraborty *et al*., 2014).

The ideas of having a system that can detect language pronunciation accuracy are not entirely new and are still in process of research as technology advances. A. Lee *et al*. (2013) has demonstrated a mispronunciation recognition system with the use of Dynamic Time Warping on Deep Belief Network posteriograms as the input and it can be compared with a Mel Frequency Cepstral Coefficient-based system to increase the system performance. Another research by S. Jenne and N. T. Vu in 2019 explained a channel that can point out errors made in non-native pronunciation by gathering acoustic features and phonological categories from visual cues. The input for their study is collected through spectrograms and Mel Frequency Cepstral Coefficient (MFCC) that can produce excellent outcomes but more studies are still needed to find out how to expand the approach towards the use in education (S. Jenne and N. T. Vu, 2019).

Besides the Mel Frequency Cepstral Coefficient which is applied by many researchers, it is noted that some of these researches uses another feature extraction technique or used it with another technique, for instance, Z. Ge *et al*. in 2016, has introduced a word pronunciation recognition system that used the Mel Frequency Cepstral Coefficient-Hidden Markov Model (MFCC-HMM) framework to detect

mispronunciation and the speech data is warped into Mel scale as it has features that can match the sensitivity of the human ear and it is the best scale to test the accuracy for recognition. In addition to that, they applied Adaptive Frequency Cepstral Coefficients (AFCC) to further enhance the adaptivity using a word adaptive frequency scale (Z. Ge *et al.*, 2016).

Next, Y. Wang and X. Zhang (2010) have also introduced a Mandarin Continuous Digits Speech Recognition System that utilizes SphinxBase that mainly uses Mel Frequency Cepstral Coefficients (MFCC) to carry out the front-end feature extraction in a series of stages and train the acoustic model using a semi-continuous Hidden Markov Model (HMM).

Similarly, K. Naithani *et al.* (2018) also apply Mel Frequency Cepstral Coefficient and Hidden Markov Model techniques to measure English speech recognition accuracy. In their paper, they can use their program to detect the differences between multiple users speaking commands and commonly used English words in various accents under unsupervised learning (K. Naithani *et al.*, 2018). Then, G. J. Jang *et al.* in 2012 used the Hidden Markov Model toolkit to train their acoustic model and test the speech recognition for their interactive Chinese speech recognition for embedded devices. D. Liu *et al.* in 2021 also uses Mel Frequency Cepstral Coefficients to develop and train their multilingual acoustic model but also apply the Automatic Speech Recognition (ASR) language identification for their multilingual speech recognition system.

Besides that, C. Huang *et al.* in 2008 has written about using speaker adaptive training and maximum likelihood linear regression (MLLR) to improve the automatic Mandarin pronunciation evaluation. The system uses Hidden Markov Model with multi-space distribution as this technique can detect the accuracy of the tone pronounced and gives a good outcome in their experiment.

It is noticeable in these studies and researches that more experiments and studies have been done to help language learners to improve their pronunciation. One improvement that these studies should do is to have their system be integrated into a web-based application for ease of students' use.

2.3 Benefits of a Web-Based System

The project is integrated into a web-based system because there are fewer software and hardware requirements needed for the students to access the system. As compared to mobile applications, there are concerns since different mobile operating systems will have different requirements and creating a prototype for both Android and iOS users will be a challenge. A web-based system can be accessed by both mobile and computer users regardless of which operating system they prefer, as they can access it through a search engine.

Besides that, Y. Duan *et al*. in 2005 discussed the benefits and challenges of a web-based expert system and one of the benefits they mentioned was that it can store a large amount of data when using the Internet database. With this type of system, the user interface can be made to be more user-friendly so that it is very straightforward for users to use (Y. Duan *et al*., 2005). A. Milad *et al.* in 2021 has built and written about an educational expert system for engineering students to gain site experience since they are not able to learn more about that upfront due to the pandemic. The project is similar to this project since lectures are forced to be conducted online, students taking beginner-level Mandarin courses will need to make more effort to learn on their own.

2.4 Summary

Based on the review of all the past research and studies done on speech recognition systems using different types of feature extraction methods and concepts, it is found that most of these studies have a similar gap, which the researchers did not specify how the system is implemented in the form of a product even though most of them are expected to be made for educational purposes in possibly embedded devices or software, for example, K. Naithani *et al*. (2018) has done good research on the speech recognition system with the implementation of MFCC and HMM but even though there are tests done and graphs are generated, it is not revealed how the speech recognition system that was built is going to be utilized. However, some research made it clear where their research is implemented like J.-C. Chen *et al*.

(2007) for instance, has shown that the speech recognition system is used in software created for Japanese students.

All in all, the researchers have given an idea of how speech recognition systems can be built and implemented. It is found that each study has made notes to improve from the previous speech recognition systems that were studied before.

Table 2.1: Summary of Literature Review

Author/ Date	Title	Method/ Technique	Context	Finding	Gaps
J.-C. Chen *et al.* (2007)	Automatic Pronunciation Assessment for Mandarin Chinese: Approaches and System Overview	Hidden Markov Model, Mel Frequency Cepstral Coefficients, Gaussian Mixture Model	Identify the speakers' pronunciation and give them a score. The system is primarily for Japanese students.	The overall recognition rate is about 75%.	It may be difficult to create software that is compatible with all devices.
K. Chakraborty et al. (2014)	Voice Recognition Using MFCC Algorithm	MFCC	Comparing male and female voices in different situations using MFCC	The same speaker saying the same word in different situations may have many variations.	It is not revealed how users can use the system.

Pedronan, R. et al. (2017)	Automatic Recognition of Correctly Pronounced English Words using Machine Learning	Mel Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT), Linear Predictive Coding (LPC)	Comparing the features: MFCC, FFT and MFCC + LPC.	MFCC + LPC has the best performance in recognizing less syllabled words.	It is not explained how the system can be used.
A. Lee et al. (2013)	Mispronunciation Detection Via Dynamic Timewarping On Deep Belief Network-Based Posteriorgrams	MFCC, GMM, Deep Belief Networks, Support Vector Machine	Detect mispronunciations	Support vector machines are effective in learning from the misalignment in the scores.	It is not revealed how users can use the system.
S. Jenne and N. T. Vu (2019)	Multimodal Articulation-Based Pronunciation Error Detection with Spectrogram and Acoustic Features	MFCC-CNN, MFCC-RNN	Pronunciation error detection	Spectrogram and MFCC- in isolation are capable of detecting a large proportion of errors	It is not revealed how users can use the system.