Dual Tree Complex Wavelet Packet Transform based features for Malaysian Speaker and Accent Recognition

by

Rokiah Abdullah
(1431311445)

A thesis submitted in fulfilment of the requirements for the degree of Master of Science in Program (Biomedical Electronic Engineering)

Faculty of Electronic Engineering Technology
UNIVERSITI MALAYSIA PERLIS

2021

ACKNOWLEDGMENT

First and foremost, I would like to express my gratitude and thanks to my supervisor, Dr. Vikneswaran Vijean for his advice and guidance throughout my research project. His suggestions, encouragement, and dedication make me able to complete this project.

I am very grateful and big thanks to Dr. Hariharan Muthusamy for his support, discussion, ideas, and valuable comments to help me complete this work. His motivation, advice, and discussions help me with the completion of the research and thesis.

I extend my thanks to Prof Sazali Yaacob who brought in this opportunity, suggestions, encouragement, and motivation to start this work. His inputs and ideas help me to discover this research deeper.

I would also like to thank and appreciate Dr. Noriha Basir, Senior Lecturer from Centre for International Language (CIL), Universiti Malaysia Perlis (UniMAP) for her support as a consultant for the wordlist used in the study. I am exceedingly thankful to Dr. Nor Azrita Mohd Amin for being a friend, in addition to an adviser, guidance and her useful discussions.

I would like to thank my research teams, Mr. Zulkapli Abdullah and Miss Farah Nazlia Che Kassim for ideas, comments, and discussion. They are willing to help solve obstacles and problems during my studies. Special thanks also to undergraduate students of UniMAP who helped me by participating in the data collection experiment for this work.

Last but not least to my beloved husband, Mr. Jamaludin bin A. R. Rawi who always supports and understanding during my studies. His patience and encouragement have given me the strength and determination in completing this work. Thank you also to my kids, Nur Amirah Batrisyia, Muhammad Faris Fahmi, Muhammad Haziq Hilmi and Nur Eryna Husna for understanding and make me able to complete these studies.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

A/D	Analog to Digital
AID	Accent Identification
ANOVA	Analysis of variance
ApEn	Approximate entropy
ASR	Automatic Speech Recognition
CIL	Center for International Languages
CNN	Convolutional Neural Network
CV-CV	Consonant Vowel-Consonant Vowel
CVC	Consonant Vowel Consonant
CV-CVC	Consonant Vowel-Consonant Vowel Consonant
CV-CVV	Consonant Vowel-Consonant Vowel Vowel
DCT	Discrete Cosine Transform
DT-CWPT	Dual Tree- Complex Wavelet Packet Transform
DWT	Discrete Wavelet Transform
EAs	Evolutionary Algorithms
EER	Equal Error Rate
EGY	Energy
ELM	Extreme Learning Machine
FB	Filter Bank
FFNN	Feed Forward Neural Network
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GA	Genetic Algorithm
GFCC	Gamma-tone Frequency Cepstral Coefficients
HMM	Hidden Markov Model
IPA	International Phonetic Alphabet
k-NN	k- Nearest Neighbours
LPC	Linear Predictive Coding
LPCC	Linear Predictive Cepstral Coefficient
MASS	Malay Speech corpus
MATLAB	Matrix Laboratory
MBSE	Mel-Bands Spectral Energy
MFCC	Mel Frequency Cepstral Coefficient

NN	Neural Network
PC	Personal Computer
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
PR	Perfect Reconstruction
RASTA	Relative Spectra
RBF	Radial Basis Function
ReEn	Renyi Entropy
RS	Rank Selection
RWS	Roulette Wheel Selection
SampEn	Sample Entropy
SCFE	Subtractive Clustering based Feature Enhancement
SD	Standard Deviation
ShEn	Shannon Entropy
SLFNs	Single Hidden Layer Feedforward Networks
SLP	Speech Language Pathologists
SRM	Structural Risk Minimization
STFT	Short Time Fourier Transform
SUS	Stochastic Universal Selection
SVM	Support Vector Machine
TDNN	Time Delay Neural Networks
TS	Tournament Selection
TsEn	Tsallis Entropy
TSR	Total Success Rate
UniMAP	Universiti Malaysia Perlis
VC	Vapnic-Chervonenkis
V-VC	Vowel-Consonant Vowel
WPT	Wavelet Packet Transform
WT	Wavelet Transform

LIST OF SYMBOLS

e(n)	prediction error
f	frequency
h0[n]	low pass filter
h1[n]	high pass filter
j	number of decomposition
k	wavelet packet node
M(f)	Mel frequency
m	pattern length
N_f	cardinality of the selected features
r	similarity coefficients
r(m)	Autocorelation function
ŝ(n)	estimated sample
W(n)	Hamming window
X (n)	input signal
Y(n)	output signal
π	constant, 3.142
α	pre-emphasis filtering control parameter

Ciri berasaskan Dual Tree Complex Wavelet Packet Transform untuk Pengecaman Penutur dan Loghat di Malaysia

ABSTRAK

Pendekatan menggunakan tenaga dan entropi berasaskan Dual Tree Complex Wavelet Packet Transform (DT-CWPT) telah dicadangkan untuk pengecam penutur dan loghat di Malaysia. Kaedah transformasi frekuensi menggunakan transformasi Fourier bukanlah kaedah yang tepat untuk menganalisis isyarat kerana maklumat berkaitan masa telah hilang. Frekuensi-masa, menggunakan pendekatan wavelet adalah kaedah yang baik untuk analisis isyarat tidak statik dalam skala waktu dan frekuensi. Untuk menguji ketepatan kaedah yang dicadangkan, isyarat pertuturan diuraikan menjadi 5 tahap dari tenaga dan entropi yang berasal dari WPT dan DT-CWPT. Prestasi tersebut dibandingkan dengan ciri Mel Frequency Cepstral Coefficients (MFCC) dan Linear Predictive Coding (LPC) konvensional. Tiga pengklasifikasi yang berbeza, seperti k-Nearest Neighbours (k-NN), Support Vector Machine (SVM) dan Extreme Learning Machine (ELM) digunakan untuk menilai prestasi pengecaman penutur dan loghat. Pangkalan data baru dibangunkan menggunakan digit bahasa Inggeris (0-9) dan perkataan Melayu yang diucapkan oleh 75 pelajar sarjana Universiti Malaysia Perlis (UniMAP), yang terdiri daripada tiga loghat utama di Malaysia, iaitu bahasa Melayu, Cina dan India. Eksperimen ini dijalankan menggunakan ciri-ciri secara berasingan dan tergabung. Didapati melalui pemerhatian eksperimen bahawa hasil menggunakan tenaga dan wavelet berdasarkan entropi adalah meyakinkan dan setara. Kadar pengiktirafan terbaik yang dicapai adalah 91.05%, yang dikira dari wavelet dan tenaga berdasarkan entropi (DT-CWPT) untuk pengecaman penutur menggunakan perkataan Melayu. Untuk pengecaman loghat, kadar pengiktirafan terbaik pada 94.84% diperoleh dari ciri MFCC menggunakan perkataan Melayu. Untuk penggabungan ciri, pengecaman penutur menggunakan kata-kata Melayu mencapai 97.67%. Sementara untuk pengecaman loghat, pengiktirafan tertinggi untuk gabungan ciri yang diperoleh adalah 98.13%. Walaupun hasil pengiktirafan memberikan hasil yang baik, ia memerlukan banyak ciri dan masa pengiraan yang panjang. Pemilihan ciri, iaitu Binary Genetic Algorithm (GA) diaplikasikan untuk memilih subset terbaik dari ciri asal untuk mengurangkan bilangan ciri dan masa pengiraan yang panjang. Hasil kajian menunjukkan bahawa jumlah ciri berkurang lebih dari 70%. Masa pengiraan menggunakan klasifikasi k-NN, SVM, ELM dikurangkan sekurang-kurangnya sebanyak 69.5%, 53% dan 14%. Telah diperhatikan bahawa GA mengurangkan masa pengiraan dengan peratusan yang signifikan sambil mengekalkan kadar pengecaman pada angka yang setara dengan hanya sedikit perbezaan dalam lingkungan 3%. Hasil pengecaman loghat menunjukkan bahawa terdapat perbezaan yang signifikan antara ketiga-tiga loghat di Malaysia menggunakan bahasa Melayu. Dapat disimpulkan bahawa antara tiga loghat di Malaysia, penutur asli (loghat Melayu) lebih baik diikuti dengan loghat India dan loghat Cina. Ini kerana sebagai penutur asli, penutur bahasa Melayu dapat menyebut perkataan Melayu dengan lebih tepat, berbanding Cina dan India. Secara keseluruhan, hasil dari kata-kata Melayu menunjukkan prestasi yang lebih baik berbanding dengan angka bahasa Inggeris untuk pengecam penutur dan loghat.

Dual Tree Complex Wavelet Packet Transform based features for Malaysian Speaker and Accent Recognition

ABSTRACT

An approach using energy and entropy based on Dual Tree Complex Wavelet Packet Transform (DT-CWPT) has been proposed for Malaysian speaker and accent recognition. The frequency transformation method using Fourier transform is not a very useful tool to analyse the signal as the time-information is lost. Time-frequency, using the wavelet approach is a good tool for the analysis of nonstationary signals both in time and frequency scale. In order to test the accuracy of the proposed method, speech signals are decomposed into 5 levels from energy and entropy derived from WPT and DT-CWPT. The performance is compared with conventional Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) features. Three different classifiers, such as k-Nearest Neighbors (k-NN), Support Vector Machine (SVM) and Extreme Learning Machine (ELM) are used to evaluate the performance of speaker and accent recognition. The new database was developed using English digits (0-9) and Malay words uttered by 75 undergraduate students of Universiti Malaysia Perlis (UniMAP), consisting the three main accents in Malaysia, which are Malay, Chinese and Indian. The experiments were carried out individually and by consolidating all the features. It was found through the experimental observations that the results employing the energy and entropy-based wavelet are promising and comparable. The best recognition rate achieved was 91.05%, which was computed from the energy and entropy-based wavelet (DT-CWPT) for speaker recognition using Malay words. For accent recognition, the best recognition rate at 94.84% was obtained from MFCC features using Malay words. For combined features, the speaker recognition using Malay words achieved 97.67%. While for accent recognition, the highest recognition for combined features obtained was 98.13%. Although the recognition result yielded a good result, it suffers from large features and long computation time. The feature selection, namely Binary Genetic Algorithm (GA) was applied to select the best subset from original features to reduce the number of features and long computation time. The results demonstrated that the number of features decreased by more than 70%. The computation time using k-NN, SVM, ELM classifier was reduced by at least by 69.5%, 53% and 14%, respectively. It was observed that GA reduces the computation time with a significant percentage while maintaining the recognition rate at a comparable figure with only a slight difference of within 3%. The accent recognition results show that there are significant differences between the three accents in Malaysia using the Malay language. It can be concluded that between the three accents in Malaysia, the native speakers (Malay accent) performed better followed by the Indian accent and Chinese accent. This is because as the native speakers, Malay speakers can pronounce the Malay words more precise, compared to Chinese and Indian. Overall, the results from Malay words show better performance compared to English digits for speaker and accent recognition.

CHAPTER 1 : INTRODUCTION

1.1 Overview

Automatic speech recognition (ASR) is the use of hardware-based techniques and computer software to identify and process the human voice. It is used to identify the words spoken by a person or to verify the identity of the person speaking into the system. The field of ASR includes speech, speaker and accent recognition. Speaker recognition refers to the recognition of a person while accent recognition refers to the recognition of an accent based on the information such as dialect, speaking style and other factors from the speech features. To date, the studies about the speaker, speech and accent recognition are actively conducted. The study contributes to significant applications and has been successfully applied in the biometric field.

This research aims to develop Malaysian speaker and accent recognition employing wavelet-based energy and entropies derived from Dual Tree Complex Wavelet Packet Transform (DT-CWPT). The results were compared with baseline techniques, such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and wavelet-based energy and entropies derived from Wavelet Packet Transform (WPT). Although many works have been reported in the literature concerning speech recognition, the use of wavelet-based energy and entropies to investigate the characteristics of Malay and English utterances are still scarce. Therefore, this research undertakes the task of evaluating the performance of DT-CWPT based features for speaker and accent recognition.

Most speaker and accent recognition in Malaysia was conducted using an English database and there are limited studies that use Malay words. In this work, it concentrates to improve the performance of speaker and accent recognition employing Malay words. The research will not only benefit Malaysians who speak the language but also other people in the South-east Asian region, such as Indonesia, South of Thailand and Southern Philippines who also speak this language (Hanifa, Isa, & Mohamad, 2017).

Since Malaysian is a muti-racial country, the accent differs from each race such as in Malay, Chinese and Indian. The accent of 3 major races in Malaysian produces different pronunciations due to the different styles of speaking and mother tongue. This work contains a new speech database uttered by Malaysian speakers from three major races used to evaluate speaker and accent recognition using DT-CWPT, MFCC, LPC and WPT features. The performance of the combined features using different classifiers was studied using a new database containing English digits (0-9) and Malay words from the major races in Malaysia.

1.2 Problem statement

To date, there are many ASR related studies that use Malay language, however, most of them are focused on speech recognition. Only a few focusing on speaker and accent recognition. Most of the studies in Malay language ASR using the Malay language, focused on Malay digits, certain words and articles. In the study, the research used Malay digits, which were digit zero (‘kosong’) until nine (‘sembilan’) and several other studies using certain Malay words were also researched in the past (Zourmand & Nong, 2018; Jamal, Shanta, Rosdi, & Ibrahim, 2018; Ghani & Porle, 2019).

Since Malaysia are multi-racial country, it faces many challenges in speaker/accent recognition. Due to various accents from different parts of Malaysia, the speaker and accent/recognition algorithm and database used should be able to address the broad nature of this diversity of ethnic/origin.

English language is one of the language that were used by various race in Malaysia. However, English accent either British or American is only suitable for their native speakers. On the other hand, non-native speakers from a different country are different from one another because they were affected by their mother tongue and style of pronunciation. The standard or uniform English accent may not be suitable for Malaysian, which does not use English as the first language. This research aims to develop Malaysian speaker and accent recognition using Malay words that consist of the combination of a monosyllable, bi-syllable and also Malay diphthongs. Subjects were taken from the major races in Malaysia, which are Malay, Chinese and Indian. English digits (0-9) were also applied in this study to compare the performance between English digits and Malay words.

The voice/speech signal is a highly non-stationary signal, the Fourier transform is not a very useful tool to analyse the signal as the time-information is lost. Time-frequency/scale analysis using the wavelet transform approach is a good tool in time and frequency scale for the analysis of non-stationary signals. In recent years, many studies have been conducted on ASR using different types of wavelet (Ali, Hariharan, Yaacob, & Adom, 2015; Nazid, Muthusamy, Vijean & Yaacob, 2015; Hariharan et al., 2018; Lim et al., 2018).

In order to overcome the problems with time information lost and since wavelet-based approach have reportedly been useful for analyzing the speech samples, it was suggested that wavelet-based energy and entropy derived from WPT and DT-CWPT to be employed to enhance the performance of Malaysian speaker and accent recognition. The time-frequency information can be obtained in wavelet approaches (Kamra, Singh, & Dhaliwal, 2020; Niwatkar & Kanse, 2020; Vani & Anusuya, 2020). The advantage of having time information retained is that it provides greater specificity when events of interest occur in time series (Stoneback et al., 2013). Previous methods employing frequency transformation, such as MFCC and LPC, transform speech signal from time-based to frequency-based domain. This opened new opportunities to study and investigate the performance of wavelet-based energy and entropies using Malay words and English digits.

1.3 Research questions

In this study, the following research questions are postulated:

1. Would there be difference in terms of accent between the three different races of populations in Malaysia?
2. Can the difference in accent be detected by using selected Malay words and English digits?
3. Would the DT-CWPT based features efficient in detecting the accent and speaker differences compared with the conventional methods?

4. How is the performance of combined features based on Dual Tree Complex Wavelet Packet Transform and baseline MFCC, LPC and WPT using different classifiers?

1.4 Research objective

The purpose of this research is to study the speaker and accent recognition algorithm for Malaysian from its major races; Malay, Chinese and Indian. The objectives of this research are described as follows:

- To develop a database for speaker and accent recognition using English digits and Malay words for Malaysian.

- To apply DT-CWPT based features, and make a comparison with conventional methods for speaker and accent recognition using English digits and Malay words.

- To evaluate the performance of combined features based on DT-CWPT with baseline MFCC, LPC and WPT using different classifiers namely k-NN, SVM and ELM.

1.5 Scope of project

Specifically, this study was conducted in two phases. Phase one is an initial study and the database speech corpus in used were collected from 39 speakers (males and females), wherein, all groups have a balanced number of speakers. Phase two is a continuation experiment from phase one with an additional speaker which totaled up to 75 speakers. The database was collected from three main ethnics of Malaysia, which are Malay, Chinese and Indians, who originated from the peninsular of the country.

The wordlist employed was English digits (0-9) and Malay words, which used vowel of /a/, /i/, /o/, /u/ and /ɛ/. These vowels are categorized as "front" vowels, "back" vowels and a combination of vowel /a/ and /u/ is one of the diphthongs from Standard Malay (SM). Vowel /e/ which is the only "mid vowel" in SM, was avoided in this new database. Focus on recognition was given to the front and back vowel since they are the majority sources of sound in SM.

The front vowels are the sounds /i/, /e/, /ɛ/, /a/ and back vowels consists of /u/, /o/, /ɔ/. Diphthongs involved a movement or glide from one vowel to another. Front vowels and back vowels are divided into four parts such as close, close-mid, open-mid and open category, where it will have a variety of significant differences in terms of pronunciation words and of course in accent. Moreover, the speaking style of accent in Peninsular of Malaysia is different in every state.

This research covers the study of the speaker and accent recognition algorithm using energy and entropy features derived from DT-CWPT compared to the conventional MFCC, LPC and energy and entropy features from WPT. The performance of the algorithm was evaluated using three different classifiers, namely k-Nearest Neighbours (k-NN), Support Vector Machine (SVM) and Extreme Learning Machine (ELM). GA feature optimization method was applied in this research to select optimal features and reducing computation time. The results before using GA optimization were compared with after using GA.

1.6 Contribution of Thesis

The new database using Malay words and English digits was developed for Malaysian speaker and accent recognition. The performance of feature extraction method such as traditional MFCC, LPC, wavelet-based energy and entropies (WPT, DT-CWPT) was evaluated using this database. These features are analyzed using three different classifiers, k-NN, SVM and ELM. Three observations from the study are listed below:

1. The recognition rate for speaker and accent recognition using Malay words gave slighty better results compared to English digits.
2. Wavelet based energy and entropy derived from DT-CWPT gave the best accuracy compared to other features.
3. For consolidation of all features, the accuracy has been improved in all classifier results.

1.7 Structure of dissertation

This dissertation explores the topic of the speaker and accent recognition for Malaysians. The research works were presented in five chapters and each chapter is described as follows:

Chapter 2 addresses the literature review of the speaker and accent recognition developed by previous researchers were presented. Previous works using various feature extraction methods and different classifiers were surveyed and discussed.

Chapter 3 discusses the methodology of the speaker and accent recognition. The experimental setup, pre-processing, feature extraction methods and type of classification that were used in this research was described.

Chapter 4 presents the results of the proposed method. The results of different feature extraction methods and classifiers were discussed.

Chapter 5 summarizes the contribution of this research and the future recommendation for upcoming research works.

CHAPTER 2 : LITERATURE REVIEW

2.1 Introduction

This chapter describes the importance of studying the speaker and accent recognition, the challenges to the systems and review of the previous studies on speaker and accent recognition. A comprehensive literature review of different journals was deliberated to obtain useful information to provide critical evaluation for this research. Next, a review of speech material, feature extraction method, classifier algorithm and the significance of the current study were discussed. Finally, the end of this chapter summarizes the analysis of the whole research.

2.2 The importance of speaker and accent recognition

A wide range of applications, such as transaction authentication, access control, voice-based information retrieval, recognition of perpetrator in forensic analysis and personalization of user devices have been used in ASR, including the speaker, speech and accent recognition (Sopon, Suksamer, Polpinij, & Chamchong, 2017; Algabri, Mathkour, Bencherif, Alsulaiman & Mekhtiche, 2017; Dimaunahan, Ballado, Cruz, & Cruz, 2017; Nainan, Ramesh, Gohil, & Chaudhary, 2018; Sunny, 2018; Machado, Vieira Filho, & de Oliveira, 2019; Pal, Arpnikanondt, Funilkul, & Varadarajan, 2019; Sim et al., 2019).

The voice/speech signal was used for discriminating a person while accent recognition conveys the information of a person's social and linguistic background. One of the major applications in using voice signal is voice biometric, which is employed in the security system, where the voice was used to authenticate the identity of the person. It is noteworthy that speaker recognition enables easy biometric remote testing and verifying. This makes speaker recognition to be useful and suitable to be used in real-world applications. Among such applications are financial, forensic, access control and security as well as teleconferencing. Accent studies also have similar importance in ASR. It can improve ASR functionality by recognizing the speaker's accents, hence, the speaker's geological origin. It can then be used to activate a certain set of properties related to the region. Accent recognition can also be used in a language training system where it can be used to train proper utterances of words.

Although English accents studies are very well-known and popular, the standard or uniform English accent may not be suitable for Malaysians due to the heavy influence of Malay and lack of use of English. Moreover, Malaysians English accent among Malaysians is not similar to the standard English language. Therefore, this work focused on studies of Malaysian accents using Malay words and English digits, which was employed as a comparison. In addition, the effectiveness of the Malay accents system can further be explored and enhanced using the proposed method applied in this work.

2.3 Challenges to an automatic speaker and accent recognition systems

The performance of ASR degrades due to many factors, such as differences in pronunciation and intonation of speech and accent, environment-related factors (channel mismatch, background noise and multiple-speakers) as well as suitable feature extraction and classifiers (Seresht, Ahadi, & Seyedin, 2017; Rahman, Dipta, & Hasan, 2018; Bansal & Imam, 2018; Mnassri, Bennasr, & Adnane, 2019; Srinivasan, Illa, & Ghosh, 2019; Ibrahim et al., 2019; Kadyan, Mantri, & Aggarwal, 2020). Therefore, many research studies used different approaches to address this problem to improve the efficiency level in ASR.

2.4 Automatic speaker recognition

In recent years, many research works have been done in ASR using speech signals. Different feature extraction and classification algorithms were proposed for speaker recognition fields.

Speaker verification using MFCC and SVM was implemented by Rusli, Ahmad and Ilyas (2016). A Malay spoken digit database was employed using five types of MFCC order (5, 10, 15, 20, 25) and was classified using SVM. The best Total Success Rate (TSR) was 95.67% and 4.44% of EER achieved from the 20th and 25th order of MFCC. The result showed that the development of speaker verification for text-dependent achieved a high accuracy.

Fook, Hariharan,Yaacob and Adom (2016) investigated the performance of MFCC with a new hybrid method using Subtractive Clustering based Feature Enhancement and Probabilistic Neural Network (SCFE-PNN) technique for isolated Malay speech recognition. 16 feature vectors of MFCC were extracted to discriminate eleven Malay words ("arang", "bola", "empat", "isnin", "emak", "bulan", "merah", "bumi", "bohong", "university", "enak") and validated by using speaker-dependent and speaker-independent. The experimental results show the effectiveness of the proposed SCFE-PNN technique with a promising average result of 99.61% for speaker dependent and 96.21% for speaker independent.

Hanifa, Isa and Mohamad (2017) researched voice recognition program using Malay speech from four main ethnis in Malaysia such as Malay, Chinese, Indian and Other using Visual C#. 13 subjects from different ethnics backgrounds were chosen and tested with the developed program for smart wheelchair where their voices were used to ease their movement from one location to another using twelve Malay words such as muzik" (*music*), "makan" (*eat*), "dahaga" (*thirsty*), "mandi" *(bathe)*, "letih*" (tired)*, "radio" *(radio)*, "tv" *(tv)*, "tidur" *(sleep)*, "baring" *(lie down)*, "ubat" *(medicine)*, "wuduk" *(ablution)* and "gosok gigi" *(brush teeth)*. It was found that Malay as the native speakers gives better recognition rate than other non-native speakers. The results indicate that volume, voice and accent and word emphasis are the important part influencing the quality of recognition.

Rosdi, Mustafa and Salim (2017) investigated the potential of ASR application in measuring speech intelligibility of speech- impaired for Malaysian speakers. A speaker-independent system was developed using 30 speech impared children with different diagnosis and severity levels aged between 8 and 12 years old. The speakers uttered 51 short, simple and meaningful sentences. A panel of Speech Language Pathologists (SLP) involved in accessing the speech intelligibility subjectively to provide the benchmark scores. In this research, the MFCC feature extraction method was used and classification was made using HMM. From the results, it was found that the intelligibility scores are not as reliable as compared to the scores of SLPs. This indicates that the importance of incorporating speech knowledge such as speech features to detect the abnormal variations in impared speeches in order to give reliable scores in classifying speech intelligibility.

Zourmand & Nong (2018) proposed a fuzzy evaluation approach in an intelligent speech assessment system using Malay vowels (/a/, /e/, /ɛ/, /i/, /o/ and /u/) from Malay children. MFCC, HMM was applied as a feature extraction method and classification algorithm of speech recognition in the system. In the vowel quality spoken by the speakers involved in the decision-making process for the proposed system, distance of the formant frequencies, number of correct recognition by HMM and log-probability of HMM play the important rule. The proposed fuzzy-based evaluation system shows the performance and efficiency of the training session for each client in terms of Malay vowel pronunciation. The result of the proposed fuzzy method increases the accuracy of the evaluation system by more than 89%.

Jamal, Shanta, Rosdi and Ibrahim (2018) investigated the first three formant frequencies of /a/, /u/ and /i/ vowel for normal Malay speakers of young adults and elderly including male and female. The mean age for young adults are 23 years while for elderly speakers range from 61 to 79 years old were evaluated. The speakers uttered the Malay digit words (*empat, tujuh* and *sembilan*) replicated three times. The results demonstrate that the Malay female have higher formant frequencies compare to male speakers and aged people have lower formant values due to their slower speaking rate. It also observed that that formant pattern ambiguity of vowel sounds for /i/ and /u/. The reason why this happens is because that the formant values of /i/ and /u/ vowels keep continuously changing due to the influence of the neighbouring phonetic sounds.

Meanwhile (Singh & Tan, 2018) investigated a new baseline for Malay-English code-switched corpus using Time Delay Neural Networks (TDNN). Two types of datasets employed in this study were for training, Pure Malay speech corpus (MASS) and for testing purposes Malay-English code-switched corpus. For feature extraction method, MFCC was used and GMM-HMM was employed for comparison with TDNN. The acoustic model based GMM-HMM system was using monophone training and triphone training. Among all triphone models, tri5b gave the best and outperformed the monophone model with a relative percentage reduction of 29.83% in terms of WER. Meanwhile, the best WER achieved by the TDNN is 28.07%.

Ghani and Porle (2019) proposed the performance of the Malay language using Feed Forward Neural Network (FFNN) was employed. Ten Malay words were obtained from Malay online news and the MFCC feature extraction method was used to extract the feature. Ten Malay words that were used such as “dan”, “akan”, “yang”, “itu”, “ini”, “pada”, “tidak”, “untuk”, “bagi”, “dalam”. The sizes of the hidden layer from 10 to 50 was analyzed to identify which hidden layer that can give the best performance. The hidden layer with size 40 has shown the highest accuracy compared to others. Overall, 94% of the samples for ten classes are correctly classified and 6% were wrongly classified.

Hanifa, Isa and Mohamad (2020) presented different cepstral features for speaker identification recognition using MFCC and Gammatone Frequency Cepstral Coefficients (GFCC). As for the classification part, SVM was employed using four ethnics; Malay, Chinese, Indian and Bumiputera. The speech corpus was developed from 15 speakers (8 male and 7 female) aged from 20 to 40 years old by reading different texts in the Malay language acquired from the local news website. Four models were performed in the study where model 1 employed MFCC features, model 2 employing MFCC with pitch, model 3 using GFCC features while model 4 using GFCC with pitch. The results have shown that model 4 through combination GFCC with pitch as a feature vector gave the highest accuracy rate of 86.1%.

Table 2.1 depicts the summary of several significant research works conducted on speaker recognition. As described in the aforementioned literature, it can be concluded that the major drawbacks found include long computation time, small data size and mostly working with MFCC methods, where it has a problem with the time-domain information while performing the frequency transformation. It was also found that in previous studies that were using Malay words, the studies focused on the Malay digit, certain word for certain purpose and articles from online news. There is a need to have other Malay words to support the development of research in this field to explore new Malay words that consist of different frequencies that exist in a front vowel and back vowel in the selected vowel employed to evaluate the performance for speaker and accent recognition.

Table 2.1: Summary of research works on speaker recognition

Reference	Database	Methodology	Significant Observations
(Rusli, Ahmad, & Ilyas , 2016)	Malay digit (0-9)	Feature extraction: MFCC Classifier: SVM	20th and 25th order of Mel-Frequency Cepstral Coefficients give the best total success rate(TSR) and Equal Error Rate (EER).TSR: 95.67%
(Fook, Hariharan,Yaacob, & Adom 2016)	Self collected	Feature extraction: MFCC Classifier: PNN	The proposed method shows promising average results of 99.61% (Speaker Dependent) and 96.21% (Speaker Independent)
(Hanifa, Isa, & Mohamad, 2017)	Self collected	-	Malay as the native speakers perform better recognition compared to other non-native speakers.
Rosdi, Mustafa, & Salim, 2017)	Self collected	Feature extraction: MFCC Classifier: HMM	Word recognition accuracy increased with the increment of intelligibility scores.
(Zourmand & Nong, 2018)	Self collected	Feature extraction: MFCC Classifier: HMM	The accuracy of the proposed fuzzy expert evaluation system increase up to 89%.
(Jamal, Shanta, Rosdi, & Ibrahim 2018)	Self collected	Feature extraction: - Classifier: -	Malay female groups have higher formant frequencies compared to male groups.
(Singh & Tan, 2018)	Pure Malay speech corpus (MASS) and Malay-English code-switched corpus	Feature extraction: MFCC Classifier: TDNN, GMM-HMM	The best WER achieved by the TDNN is 28.07%.
(Ghani & Porle, 2019)	Ten Malay words from online news (dan, akan, yang, itu, ini, pada, tidak, untuk, bagi, dalam)	Feature extraction: MFCC Classifier: FFNN	The hidden layer with size 40 has shown the highest accuracy compared to others. 94% of the samples for ten classes are correctly classified and 6% were wrongly classified.
(Hanifa et al., 2020)	Local news website	Feature extraction: MFCC & GFCC Classifier: SVM	The accuracy rate using combination of GFCC and pitch as the feature vectors (Model 4) produced the highest accuracy rate of 86.1%.

2.4.1 Speech materials for speaker recognition

The speech materials for speaker recognition was summarized below in Table 2.2. The table presents a summary of some previous studies in ASR regarding subjects/speaker's details, type of speech, type of accents, the purpose of study and database/resources. In the last recent years, Malay words such as digits, vowels, speech and sentences have been proposed on the studies. Most of the studies used database from online sources and are self-collected.

From the past studies as tabulated in Table 2.2, it can be concluded that a different method and algorithms have been used to investigate, improve and enhance the performance of speaker recognition. A variety of speech, such as text-independent, isolated words, digits and database from a multi-source was used. It was observed that mostly these studies used Malay and followed with the English language. However, their studies focused on digits, certain words and speech sentence. Furthermore, the research focused more speech recognition field in ASR rather than speaker recognition field.

Table 2.2: Summary of speech material for speaker recognition

Reference	No of subjects	Types of speech	Types of accents	Purpose of the study	Database
(Rusli, Ahmad, & Ilyas , 2016)	27 speakers	Digits (0-9)	Malay	Test performance of SVM by selecting the best order of MFCC Coefficients	Malay corpus data from Faculty of Language and Linguistic, University Malaya
(Fook, Hariharan,Yaacob, & Adom 2016)	24 speakers	Words	Malay	Investigate the performance of MFCC in distinguishing the selected Malay words and the effectiveness of the SCFE-PNN technique	Self-collected
(Hanifa, Isa, & Mohamad, 2017)	13 speakers	Words	Malay	Studying the effect of accents by multi-racial speakers on the result of the speech recognition system for the Malay language.	Self-collected
(Rosdi, Mustafa, & Salim, 2017)	30 speakers	Speech	Malay	Investigate the potential of ASR application in measuring the speech intelligibility of speech impaired speakers.	Self-collected
(Zourmand & Nong, 2018)	360 speakers	Vowel	Malay	Develop a modified feature extraction method to improve the speaker-independent vowel recognition with the fuzzy-based SLP-independent performance evaluator.	Self-collected
(Jamal, Shanta, Rosdi, & Ibrahim 2018)	16 speakers	Digits (empat, tujuh and sembilan)	Malay	Investigated the first three formant frequencies of /a/, /u/ and /i/ vowels for normal Malay young adult and aged people speakers.	Self-collected
(Singh & Tan, 2018)	200 speakers	Sentences	English and Malay	Develop a new baseline for Malay-English code-switched speech corpus which is constructed using a factored form of time delay neural networks (TDNN-F)	Pure Malay speech corpus (MASS) and Malay-English code-switched
(Ghani & Porle, 2019)	-	10 words	Malay	Investigate the performance of the conversion speech to text using Malay words.	Online Malay news
(Hanifa et al., 2020)	15 speakers	Text	Malay	Investigate the performance of different Cepstral Features for speaker identification recognition	Local news website

2.5 Automatic accent recognition

This section deals with some of the previously published works on accent recognition. Yusnita, Paulraj, Sazali Yaacob, R and Fadzilah (2015) presented statistical descriptors of Mel-Bands Spectral Energy (MBSE) features with feature reduction Principal Component Analysis (PCA) for robust accent recognition in Malaysian English. Two types of utterances i.e isolated words and continuous speech in the form of sentences were used to identify the performance of automatic accent recognition. The robustness of MBSE and PCA-MBSE was compared with the standard MFCC and LPC features using k-NN classifier in clean and noisy conditions. The experimental results show that the best accuracy rate for males was 92.7% and for females was 93.0% using the proposed PCA-MBSE features.

Mannepalli, Sastry and Suman (2016) proposed an accent recognition system for Telugu speech signals using MFCC-GMM. The features were extracted from MFCC and GMM, which were used in the classification of speech. The recognition efficiency of the accent recognition system was performed using MFCC-GMM and Prosodic-Nearest Neighbour Classifier (NNC). The recognition accuracy of Coastal Andhra accent using the MFCC-GMM algorithm was 88%, Rayalaseema and Telangana accents were 92% and the overall percentage recognition accuracy was 91%. All main accents showed an improvement using the MFCC-GMM algorithm. The percentage of recognition accuracy for Coastal Andhra was slightly lower than the other two accents. This may be due to a mismatch language and the influence of other languages from neighbouring states.

Tverdokhleb, Dobrovolskyi, Keberle and Myronova (2017) developed accent recognition methods subsystems using Wildcat Corpus of Native and Foreign-Accented English for e-learning systems. The corpus contains 1342 audio recordings consisting of accent English: proper English pronunciation; Korean, Japanese, Chinese, Italian, Indian and Irish. The feature extraction, namely MFCC and LPCC were used and the performance of these subsystems was evaluated using Neural Network (NN) and GMM. Massive open online courses system, such as Moodle and activity module accesses an external service written in Python, which can be plugged in the accent recognition subsystem. The service gets the student utterance based on the proposed text and the result returns to the accent recognition using neural networks. The average recognition rate for the MFCC-NN method was 91.43%, the LPCC-NN method was 78.73% and for MFCC-GMM method was 87.53%. From this study, it was proven that the accent recognition subsystem for e-learning system based on native and foreign-accented English was promising using the MFCC feature. Nevertheless, using LPCC techniques, the recognition was poorer compared to MFCC.

Jain et al. (2018) reported accent embeddings and multi-task learning to improve speech recognition for accented speech. Seven English accents, namely the United States, England, Australia, Canadian, Scottish, Irish and Welsh were employed as seen accents, meanwhile, New Zealand and South Asian (India, Pakistan, Sri Lanka) were categorized as unseen accents. The baseline system was adopted using feed-forward Time-Delay Neural Networks (TDNNs) with sub-sampling at intermediate layers. Two types of accent embedding, which are frame level and utterance level were carried out in the experiments. The proposed approach achieved more than 15% of relative Word Error Rates (WER) reduction with seen accents and 10% relative WER reduction for unseen accent.

The best result reported was 82.6% with TDNN seven layers. The obtained results provided evidence that unseen accent gave a low reduction of WER than seen accent due to different accents uttered by the speakers.

Patel and Barkana (2018) reported an analysis of English cornels vowel based on the first and second formant frequencies, LPCCs and MFCCs produced by Mandarin, Hindi and American accented speakers. 10 orders of LPC and 12 MFCC plus energy coefficient was employed in this work while the classification algorithm employed was the k-NN classifier. The series of experiments were conducted with different experiments using L1 (native accent) and L2 (foreign accent) speakers. Among the three speaker groups, the highest accuracy of approximately 80% was achieved from native Hindi speakers. The overall accuracy of 69.55 % was obtained using energy feature plus MFCC and the performance of MFCC is better than LPC. Nevertheless, the combination of LPCs and MFCCs did not improve the accuracy of all speaker groups. It was observed from the results that major misclassification occurred from L2 speakers due to the differences in pronunciations and foreign accents, which are not similar to the speakers.

Odulio et al. (2019) investigated an accent recognition using Philippines language, namely Bikol and Tagalog accents using Convolutional Neural Network (CNN). The different features experimented were F0, Energy, Mean Energy, Minimum and Maximum Pitch, Minimum and Maximum Energy. The study focused on two accents, namely Tagalog and Bikol. 79.28% was achieved for Tagalog accent and 78.33% for Bikol accent. The constraint of this study is the unbalance of participant's distribution with regards to gender, age-group and language.

Ahamad, Anand and Bhargava (2020) presented AccentDB database that contains samples of 4 Indian- English accent and a compilation of samples from 4 native-English, and a metropolitan Indian-English accent. 13 MFCC features and different classifier namely MLP, CNN and attention-CNN were used. The generalization of the classifier models was tested in a variety of setups of seen and unseen data. All the models performed well and CNN has a slightly better than other model in accuracy as expected.

Based on the literature, it had been inferred that the major shortcomings of accent recognition are facing a confusable accent, inter-mixed speaker of the boundary, dialect of English/native accent and small database that affect the performance of recognition. Furthermore, it can be seen that the performance of feature extraction method, especially using MFCC, LPCC and LPC, still need improvement.

Table 2.3 summarizes several of the significant research works that were conducted on accent recognition from previous literature. It addresses the database, methodology and the significant observations of the research.

Table 2.3: Summary of research on automatic accent recognition

Reference	Database	Methodology	Significant Observations
(Yusnita, Paulraj, Sazali, R & Fadzilah, 2015)	Self- collected	Feature extraction: PCA, MFCC, LPC Classifier: k-NN	Two new methods for extracting and compressing accent features of Malaysian English accented speech using MBSE and PCA-MBSE. Accuracy: 92.7% (male) and 93.0% (female)
(Mannepalli, Sastry, & Suman, 2016)	Self collected: Telugu language(39 speakers)	Feature extraction: MFCC Classifier: GMM	All main accents namely Andhra, Telangana and Rayalaseema showed an improvement using MFCC-GMM compared to the Prosodic- NNC method Accuracy: 91%
(Tverdokhleb, Dobrovolskyi, Keberle, & Myronova, 2017)	Wildcat Corpus of Native and Foreign-Accented English.	Feature extraction: MFCC, LPCC Classifier: NN, GMM	MFCC used neural network can improve the reliability of accent recognition by 14% in compared with other existing methods. MFCC-NN: 91.43% LPCC-NN: 78.73% MFCC-GMM: 87.53%
(Jain et al., 2018)	Common Voice corpus from Mozilla	Feature extraction: MFCC Classifier: feed- forward TDNN	Accent embeddings learned from a standalone network give further performance improvements. Accuracy: 82.6%
(Patel & Barkana, 2018)	English cornels vowels	Feature extraction: MFCC, LPC and F1-F2 formant frequencies Classifier: k-NN	Energy feature and MFCCs set provided better performance than the LPCs Energy feature and MFCC accuracy 69.55%
(Odulio et al., 2019)	Self- collected: Bikol and Tagalog accent (76 male speakers and 81 female speakers)	Feature extraction: F0, Mean Energy,Duration, Minimum Pitch and Maximum Pitch. Classifier: 1D-CNN	The results have shown the performance of the developed model and it reflects that the amount of correctly recognized input is more than the misrecognized input Tagalog accent: 79.28% and Bikol accent: 78.33%
(Ahamad, Anand & Bhargava, 2020)	Self-collected & Amazon Polly database (15)	Feature extraction: MFCC Classifier: MLP, CNN and attention-CNN	The model performed very well on the non-native to native accent neutralization. Accuracy: More than 85% when converting from native to non-native accents.