CHAPTER 1

INTRODUCTION

1.1 Introduction

Speech is a vital medium for human interaction. It is natural ability for humans to detect and discern speakers and rich information contained in speech signals. However, to attain similar task of recognizing speakers and speech, machines must emulate human’s auditory and perceptive capability, which is a complex tasks that encompasses multiple disciplines including digital signal processing, pattern recognition and linguistics. This project explores a gender and accent identifier that is able to distinguish between different ethnic groups speaking Malaysian English.

A shared problem between speaker recognition and automatic speech recognition (ASR) is the analysis and modeling of speaker variability [1]. Speaker information one can derives from speech signals in general fall into two groups: linguistic (accent, dialect, language and text contents) and speaker-specific or paralinguistic (gender, age, emotion and identity). This information allows speaker profiling, which can be applied in many fields encompassing voice biometrics for physical security, forensic and Interactive voice response (IVR) systems.
The complexity lies in the huge variableness of speech characteristics, which can vary immensely across speakers set due to idiosyncrasies [2]. Gender [3-5] and accent [6-9] are among the most substantial speaker variables [10]. The fusion of these two variables also yielded in increased complexity of speech variations as concluded by Martland, et al. [11]. Dissimilar speech characteristics between male and female are caused by physiological and acoustical dissimilarities. Male and female have different thickness of vocal fold and length of vocal tract [4, 12], which lead to formant frequencies and fundamental frequency (pitch) differences. On average, female exhibit higher formant frequencies and pitch than male, which has the lowest frequency among male, female and children [5].

Accent variation is evident and measureable in acoustic phonetics [10, 13, 14]. Variation occurred due to many interacting factors such as second language acquisition, geographical background, ethnicity and socio-economic class [8, 10, 12, 15]. As speaker’s backgrounds influenced the formation of accent in speech, accent becomes a major indication that characterizes a speaker belonging to a group or at large, a community. Such is the case of Malaysian English, which is a standard non-native variation of English, introduced during the British colonial in Malaysia [16, 17]. Development of Malaysian English as a localized English standard is largely influenced by other local languages namely Malay, Chinese and Indian [13, 14, 18-20]. Among the three, Malay, the national language of Malaysia, is noted to be the most prominent cause of distinctive variations of Malaysian English [13], meanwhile Chinese is the most significant group of Malaysian English speaker due to its large number of speaker [21].

Performance of speaker-independent and text-independent recognition system deteriorates when speaker possesses features that the system is not trained with [6, 7, 22, 23]. This is especially prominent among non-native speakers as current commercial recognition systems are homogenous and poorly accommodate speaker variability on weighty consideration [9, 22]. Effectiveness of speaker and
speech recognition can be enhanced through deploying classifiers that identify speaker’s gender and accent information preceding to the recognition process [7, 10]. Preempt knowledge of the speaker enables more reliable selection of acoustic and lexical model [8, 12]. However, speaker information extraction and classification remains a challenging task due to complexity of the domain [1-3, 10]. Speech corpus used for training, features extraction algorithm and statistical modeling algorithm are all crucial factors to baseline performance of the recognition system.

1.2 Background of Study

Compared to other variation of English, Malaysian English has not received much attention in studies of speech processing. Speech modeling of languages in Malaysia is only beginning to gain attention in past ten years [24].

Formant frequencies (specifically the first three frequencies: F0, F1, and F2), energy (expanded energy to produces a stressed syllable), and fundamental frequency (also known as pitch) are three prominent speech characteristics investigated and adapted in current proposed and deployed speaker and speech recognition systems [6, 25]. Formant frequencies and pitch are two evident features that differentiate male and female voicing on interacting physiological and phonetic factors [4, 5, 12]. While energy is most notable in differentiating speaking styles inter-gender and inter-language [25, 26]. Energy is the stress of syllables by applying more muscular energy, which is an important indication of differentiated speaking styles of a language from parent norms [25, 27]. The accent identification system proposed by Yunista et al. recorded 94.2% as its highest identification rate using a combination of formant analysis and log energy vector. Researchers have investigated the combination of features to achieve as high as possible the robustness of the

In the case of Malaysian English, a majority of speakers learn the language in formal educations after already acquiring a first language, usually associated to their ethnicity [18]. Differentiation in pronunciation is eminently dictated by prosodic and phonological deviation during language learning. On a phonemic level, phonological coding becomes inconsistent from native standards due to unfamiliarity with phonological change, hence phonemes substitution occurred [10]. Hence classifying vowels are substantive to Malaysian English accent identification. Vowel classification is utilized by Paulraj, et al. [20] in adapting Malaysian English pronunciation to ASR. Additionally, Mohd Yusof and Yaacob [24] also proposed a vowel classification system based on first three formant frequencies (F0, F1 and F2), which are spectrum peaks resulting from resonant frequencies. On the other hand, Yunista et al. [9] also use formant frequencies to identify accent.

Mel Frequency Cepstrum Coefficients (MFCC) and Linear Predictive Coefficients (LPC) are two algorithms dominantly used in speech analysis. LPC is used in all three Malaysian English accent classification systems reported [9, 20, 24]. It is also applied in Teixeira, et al. [7]’s accent identification system encompassing six European English accents. A common point among these systems is small database (less than 100 speakers). Meanwhile, MFCC are applied on investigation with relatively bigger databases (up to or more than 1000 speakers) [1, 28]. Training and decision making methods adopted are different among proposed Malaysian English accent classification systems, with Paulraj, et al. [20] and Mohd Yusof and Yaacob [24] using neural network while Yunsita, et al. [9] adapted a classifier based on K-Nearest Neighbor (KNN). Both Li, et al. [26] and Metze, et al. [28] investigated on multiple algorithms and proposed fusions of algorithms. Li, et al. [23] proposed a MFCC and Gaussian Mixture Model (GMM) on preliminary levels
and decision making based on Support Vector Machine (SVM) are deployed later. Meanwhile, Metze, *et al.* [26] proposed a parallel phone recognizer using MFCC, LPC, Linear Discriminant Analysis (LDA), and Bayesian networks. Additionally, Hidden Markov Model (HMM) is also a popular training model, alternative to GMM and neural network, employed by Liu and Fung [25] and Kumpf and King [29].

1.3 Problem Statement

Malaysian English, a formal second language to the Malaysian community is only beginning to gain attention in local research community. To date, researches of Malaysian English focused on accent identification without considering the gender factor. However, a gender-dependent accent identification system is unreported. Therefore, it is meaningful to extend previously proposed accent identification system to adapt gender and accent, the two substantive speaker variables, and establish baseline recognizer for gender-dependent Malaysian English accents identification.

1.4 Objectives of Study

The goal of this project is to establish and evaluate gender and accent identification technique for Malaysian English using Gaussian Mixture Model (GMM) and Mel Frequency Cepstrum Coefficients (MFCC).
This goal is supported by the following objectives:

(i) To identify corpus text suitable to elicit gender and accent features in speech.
(ii) To establish a gender and accent baseline recognizer.
(iii) To evaluate the performance of the established baseline recognizer.

1.5 Research Questions

This project addresses the following research questions:

(i) What are the gender-sensitive and accent-sensitive speech features?
(ii) Does MFCC and GMM-based algorithm increase the performance of gender and accent identification system?
(iii) What are the performance of the MFCC and GMM-based gender and accent identifier?

1.6 Scope of Study

Scopes of this project are listed as follow:

(i) Ethnicity and gender are the two concerned variability in this project.
(ii) Ethnic group involved are Chinese and Indian. Speech data is collected from both male and female. Background, age and geographical location are not considered.
(iii) This project studies speaker/speech recognition on the acoustic and phonetics aspects. Phonology attributions are not discussed.

(iv) Recording texts for both training and testing data are excerpted from “Comma Gets a Cure” passage, which is specially designed for pronunciation study [30, 31].

(v) Level of literacy is not considered, however speakers must be able to read firmly and clearly to maintain relevancy and intelligibility of the speech recording. To foster smooth recording, speakers are given time to read the script before recording.

(vi) Speech recording environment and speakers are random.

(vii) Speech sample is recorded with Sony ICD-AX412F digital recorder; storage of speech sample is in mp3 format. Noise-cut function on device is enabled.

(viii) Pre-processing of audio file includes format, sampling rate and channel conversion. Manual pre-processing is performed using Sony Sound Forge 9.0.

1.7 Contribution of Study

Researches on speaker and speech recognition are majorly focused on the British and American English variations, while relevant studies involving Malaysian English are limited. Findings on Malaysian English accent and speaker gender identification can be extended to application in voice biometric and forensic investigation, where success rate of identification or verification is affected by the diffusion of voice characteristics of speakers set.
1.8 Organization of Report

Chapter 1 of this report introduces the project aim, background of study and problem statement. A detail report of literature review is covered in Chapter 2, followed by discussion on research methodology in Chapter 3. Framework of gender and accent identification is explained in Chapter 4, meanwhile results and findings are discussed in Chapter 5. Conclusion of this study is presented in Chapter 6.