CHAPTER 1
INTRODUCTION

This chapter introduces the roles of Automatic Speech Recognition (ASR), automatic gender classification and machine learning (ML) and highlights the main motivation and objectives of doing this thesis. As well, problem of the thesis is discussed.

1.1 Introduction

Speech recognition is the knowledge base subfield of linguistics that develops methodologies and technologies that permit the recognition and translation of spoken communication into text by computers. It's conjointly called ASR, computer speech recognition or speech to text (STT). It incorporates data and analysis within the linguistics, computer science, and electrical engineering fields. Some speech recognition systems need "training" (also referred to as "enrollment") wherever a private speaker reads text or isolated vocabulary into the system. The system analyzes the person's specific voice and uses it to fine-tune the identification of that person's speech, leading to inflated accuracy. Systems that don't use training are referred to as "speaker independent" systems. Systems that use training are referred to as "speaker dependent" [1].

Like all computer software, speech recognition employs algorithms that work using acoustic and language modeling. Acoustic modeling represents the intermediary between linguistic units of a speech and audio signals whereas language modeling matches sound produced with word sequences to help distinguish between words that are familiar. This has made speech recognition to have a wide range of application such as call routing, speech-to-text processing, voice dialing, voice search and also is applied in simple data entry practices.

Automatic gender classification from speech is widely applied in speech recognition. Many applications including speaker identification, speaker segmentation, smart human computer interaction, biometrics social robots, and audios or videos content

indexing, etc., use gender classification. Gender identification can improve the prediction of other speaker traits such as age and emotion, by jointly modeling gender with age (or emotion) or either by together modeling gender with age (or emotion) or during a pipelined manner. Speaker verification systems additionally implicitly or expressly use gender information. Generally humans can easily identify a person's age, gender and emotion by hearing to this person's voice. In some circumstances like conversations over the telephone, the genders of adults are easy to identify, but the genders of children are difficult to identify. The acoustic and linguistic characteristics of child speech are particularly different from those of adult speech. Children generally have higher elementary and format frequencies than those of adults, thanks to a shorter vocal tract, smaller vocal folds, developing articulators (e.g. tongue size and movement) [2]. Therefore, there is no significant difference in their acoustic-phonetic properties in both male and female child. This gender classification system examines the features efficient in discriminating the gender from children's speech with the help of feature extraction algorithms. The extraction of the relevant and important information from the speech signals of the human voice is an important task to produce a latter recognition performance. And then the system makes performance analysis about classification by using different ML algorithms [3].

In computer science, ML identifies to a type of data analysis that applies algorithms that learn from data. It is a type of artificial intelligence (AI) that provides systems with the ability to learn without being especially programmed. This implements computers to find data within data without human intervention. The main importance to know about ML is that data is being used to make predictions, not code. Data is dynamic so ML permits the system to learn and evolve with experience and the more data that is analyzed [4].

ML tasks are classified into several broad categories. In supervised learning, the algorithm builds a mathematical model from a set of data that contains both the inputs and the desired outputs. Semi-supervised learning algorithms develop mathematical models from incomplete training data, where a portion of the sample input doesn't have labels [5]. Classification algorithms and regression algorithms are types of supervised learning. Classification algorithms are used when the outputs are restricted to a limited

set of values. Regression algorithms are named for their continuous outputs, meaning they may have any value within a range. In unsupervised learning, the algorithm builds a mathematical model from a set of data which contains only inputs and no desired output labels. Unsupervised learning algorithms are used to find structure in the data, like grouping or clustering of data points. Active learning algorithms access the desired outputs (training labels) for a limited set of inputs based on a budget, and optimize the choice of inputs for which it will acquire training labels. When used interactively, these can be presented to a human user for labeling. Reinforcement learning algorithms are given feedback in the form of positive or negative reinforcement in a dynamic environment, and are used in autonomous vehicles or in learning to play a game against a human opponent.

In this thesis, children gender classification system is proposed to solve the problem that gender identification of children is difficult than adults; it is confusing to identify whether the speaking child is male or female by combining ASR, feature extraction and ML algorithms. The proposed system is implemented by applying Python programming language and the experimental result has been analyzed.

1.2 Aims and Objectives

The main goals of the thesis are describes as follows:

1. To study the algorithm of feature extraction and machine learning algorithms for classification
2. To understand the speech recognition operations in details
3. To implement a gender classifier that can automatically predict the gender of the speaker.

1.3 Motivation

A voice contains linguistic information of a speaker and with this human beings can verify even an unknown speaker (gender, age, origin). Therefore, voice features are considered as voiceprints [6]. This proposed work extracts features from a voice signal

and these features are then used to determine the gender of the speaker. This system of gender recognition can be used upon in some very useful applications, for instance it can provide more targeted services based on gender interoperability. Moreover in Human-computer interaction (HCI), this system could get the user interface scope and develop the experience in most Internet of Things (IoT) applications. These information can be used for gender customizations in such IoT apps and extend the security in these applications [7].First of all, the interactive information system can be developed with user gender identification as it can automatically select the proper interaction service for different genders. In human machine interaction, the system service is unlike for both male and female, for example, interface styles and color. For this reason, the differentiation of gender using such voice authentication would improve the overall efficiency and outcome of HCI. Second of all, in social networking system, access can be controlled through the gender determination of user. This can also reduce the advertising cost by narrowing down and targeting the exact market [6].

1.4 Research Problem

Computer recognition of children's speech is particularly difficult. This is due to the fact that children have large differences in both the acoustic and the linguistic aspects of speech compared to adults. Children's speech has shorter transition duration and larger spectral difference between consonant and vowel in the consonant-vowel pair than those of adults' speech. Differences in the pitch, the formant frequencies, the average phone duration, the speaking rate, the glottal flow parameters, pronunciation and grammar are the various acoustic and linguistic differences. As reported, children have different values of mean and variance of the acoustic features of speech than those of adults. For example, the area of the formant ellipses is larger for children than for adults for most vowel phonemes and children speech contains more dis-fluencies and extraneous speech. As all children undergo rapid development and with varying rates, it is difficult to model their constantly changing speech characteristics. Also as children grow, their speech production organs change and so their anatomy and physiology keep changing quite significantly.

1.5 Organization of the Thesis

In this thesis, the following chapters will be discussed.

Chapter one describes brief introduction, aims and objectives, motivation, and problem statement of the thesis.

The literature reviews on the speech recognition, feature extraction and ML systems that are currently using around the world are discussed in chapter two.

In chapter three, details of speech recognition, steps of feature extraction algorithm and classification models are demonstrated with their theoretical background.

Chapter four mentions the proposed system with its design, detailed workflow, applied methodologies, the implementation and performance evaluations.

Finally, discussion, conclusion as well as limitations and future works are presented in chapter five.