Chapter 1

Introduction

Speech is the most natural way of human communication. The human uses the production system to generate the speech and uses the auditory system to understand the speech. Sound is a longitudinal pressure wave formed by compressions and rarefaction of air molecules, in a direction parallel to that of the application of energy. It can be described by a sine wave. The crests of sine-wave corresponds to moments of maximal compression and the troughs to moments of maximal rarefaction.

In the production system, speech is the air-pressure waves emanating from the mouth and the nostrils of a speaker (Demeechai *et al.,* 2001). The human speech production apparatus consists of: lungs, vocal cords (larynx), velum (soft palate), hard palate, tongue, teeth, and lips. The most fundamental distinction between sound types in speech is the voiced/unvoiced distinction. Voiced sounds, including vowels, have a roughly regular pattern in their time and frequencies structure. Voiced sounds typically have more energy. When the vocal folds vibrate during phoneme articulation, the phoneme is considered voiced. Vowels are voiced throughout their duration. The distinct vowel timbres are created by using tongue and lips to shape the main oral resonance cavity in different ways. The rate of cycling of the vocal folds in the larynx during phonation of voiced sounds is called the fundamental frequency (F0). It can be as low as 60 Hz for a large man and as high as 300 Hz or higher for a small woman or child. The fundamental frequency contributes more than any other single factor to the perception of pitch in speech. Since the glottal wave is periodic, consisting of fundamental frequency and a number of harmonics. The resonance of the vocal tract is excited by the glottal energy. Harmonics near the resonance is emphasized, and in speech, the resonance of the cavities that is typical of particular articulator configurations are called formant.

In the auditory perception system there are two major components: the peripheral auditory organs (ears) and the auditory nervous system (brain). The ear processes the signal by first transforming it into mechanical vibration pattern on the basilar membrane, and then representing the pattern by a series of pulses to the auditory nerve. The human ear has three sections: the outer ear, the middle ear and the inner ear. The relevant structure of the inner ear for sound perception is the cochlea. The cochlea can be roughly regarded as a filter bank, whose outputs are ordered by location, so that a frequency-to-place transformation is accomplished.

In most of the world's languages, the inventory of phonemes can be split into two basic classes: consonants and vowels. The tongue shape and positioning in the oral cavity do not form a major constriction of air-flown during vowel articulation. The major resonance of the oral and pharyngeal cavities for vowels is called F1 and F2 – the first and the second formant, respectively. They determine the characteristics timbre or quality of the vowel. Consonants, as opposed to vowels, are characterized by significant constriction or obstruction in the pharygeal and oral cavities. Some are voiced; others are not. The oral, nasal, pharyngeal, and glottal mechanisms actually make available a much wider range of effects than English happens to use. The primary dimension lacking in English that is exploited by a large subset of the world's language is pitch variation. Many of the huge language families of Asia and Africa are tonal, including all varieties of Chinese. To be considered tonal, a language should have lexical meaning contrasts cued by pitch. For example, Mandarin Chinese has four primary tones, and Thai has five primary tones.

1.1 Motivations

During the past decade, speech recognition technology has undergone significant progress. Several applications of speech recognition to human-computer interface have been developed, since speech is the most natural way of human communication and interaction. Most existing methods for speech recognition have been developed mainly for spoken English, and some of them have been adapted to be applicable to Thai language. However, unlike English, Thai language is a tonal language. In such a language, the referential meaning of an utterance is dependent on the lexical tones. Therefore a tone classifier is an essential component of a speech recognition system of a tone language.

Tone classification is not only necessary but also difficult. The tone is determined by the F0 of the speech data. Then the pitch detection algorithm is the first step for tone classification. Except a robust pitch detection algorithm, how to reduce the effects of every interacted factor is also a big issue.

- Because of the formant structure in the speech, the pitch detection often is error-prone. Also the quality of speech data affects the pitch detection. A robust and high precision pitch detection algorithm is important for tone classification.

- Speech data not only carry the language information also carry the characteristics of the speaker and the emotion of the speaker and so on. Although all of information carried by the speech is useful for human understanding. But for tone classification, the information like the speaker, emotion, age, gender gives more confusing information. Generally, the F0 of female is higher than the male's even though it expresses the same tone.
- The spoken mode of speech data significantly affects the tone classification. Isolated speech is easier for analysis and classification. But for continuous speech, the auditory feature of one word will be affected by its context, especially for tone. Then to consider the context interaction is also an important issue for tone classification.

1.2 Goal and Expected Outcome:

The general goal of this thesis is to find a suitable design of tone classification for Thai speech. Specifically, this thesis tried to accomplish the following tasks:

- To study the phonetic knowledge of tone and the current researches on tone classification;
- To design a tone classifiers that can be used to improve the performance of Thai speech recognition systems;
- To improve speech recognition accuracy by means of improving tone classification performance.

In order to achieve the thesis goal, the expected outcome of thesis should be:

- A tone classifier of Thai vowels that can be integrated into and improved the performance of Thai speech recognition systems.
- Knowledge of feature extraction techniques and classifier design which are suitable for tone classification.

1.3 Outline

The remainder of this thesis is organized into 7 chapters.

Chapter 2 introduces the design principles of two pitch detection algorithms. The evaluation experiments were done on a small Thai digit speech database collected under office environments. The discussions and conclusions are presented.

Chapter 3 discusses the possible factors that affect the performance of the tone classification and describes the relative techniques for reducing these factors. The evaluation experiments is done on the speech data taken from Thai large vocabulary corpus (Thongprasert *et al.*, 2002).

In Chapter 4, the proposed method of classifying the tone according to the different final consonant is introduced. The evaluation experiment is done on continuous Thai speech databased.

Chapter 5 introduces a proposed method using 2-stage NN method for tone classification in order to improve the performance. The experiments are presented and the results are discussed.

Finally the summary of the work done in this thesis and the contributions of this research work are concluded. The future work and topic are introduced.