CHAPTER 1 INTRODUCTION

Automatic speech recognition (ASR) is utilized in many applications, for example, voice activation in GPS systems and portable equipment for disabled people, and the user interface in mobile phones.

In non-tonal languages such as English and French, tone or pitch variation called intonation is used to express emphasis, contrast, and emotion. Tonal languages such as Thai, Chinese, and Vietnamese have different linguistic features as tones compared with other non tonal languages. In tonal languages, each tone represents a different grammatical meaning. A tone recognition system is therefore an important part of automatic tonal speech recognition (ATSR).

This thesis aims to reduce the calculation complexity of tone recognition in real-time ATSR so that the tone recognition can be applicable for portable electronic equipment.

In this chapter, motivation, contribution, and thesis outline are described. Motivation shows the important of the tone recognition in ATSR and advantages of low calculation complexity tone recognition which are provided by real-time ATSR. Contribution presents a problem analysis of tone recognition for ATSR and thesis statement. Finally, the outline of the thesis is introduced.

1.1 Motivation

In automatic tonal speech recognition (ATSR), tone can cause serious problems in speech recognition accuracy, especially in the case of a group of words having the same phonemes but different tones.

Table 1 shows the tone influence of words having same phonemes but different tones in Mandarin Chinese and Thai. Each word of both sentences has the same consonant and vowel phonemes but different tones. In Mandarin Chinese [1], the sentence "*māma mà mǎ de ma*" means "Is mum complaining about her horse?". The sentence "*shī shì shí shī shǐ*" means Lion-eating poet in the stone den. In Thai, the sentence "*mǎi mài mâi mái*" means "Does new silk burn?".

Table 1.1 Tone influence by using the same consonant and vowel phonemes but different tones

Language	Script	International Phonetic Alphabet	English Translation
Mandarin Chinese	妈妈骂马的吗? / 媽媽罵馬的嗎?	māma mà mǎ de ma	Is mum complaining about her horse?
Mandarin Chinese	施氏食狮史/ 施氏食獅史	shī shì shí shī shǐ	Lion-eating poet in the stone den.
Thai	ไหมใหม่ไหม้มั้ย	mǎi mài mâi mái	Does new silk burn?

Due to different tones, each syllable provides different grammatical meaning. This results in a low recognition rate of automatic tonal speech recognition (ATSR). Therefore, tone recognition becomes important in ATSR.

Different types of tone can be classified by using the trajectory of fundamental frequency (F_0). Figure 1 indicates standard F_0 contours of four Mandarin Chinese tones [2]. Tone 1 is a high-level tone (¯). Tone 2 is a high-rising tone (´). Tone 3 and Tone 4 are low-dipping (ˇ) and high-falling (` ) tones. Figure 2 shows standard $F_0$ contours of five Thai tones [3]. Tone 1 is a mid tone ( ¯ ). Tone 2 is a low tone ( `). Tone 3 is a falling tone (ˆ). Tone 4 and Tone 5 are high (´) and rising tones (ˇ).

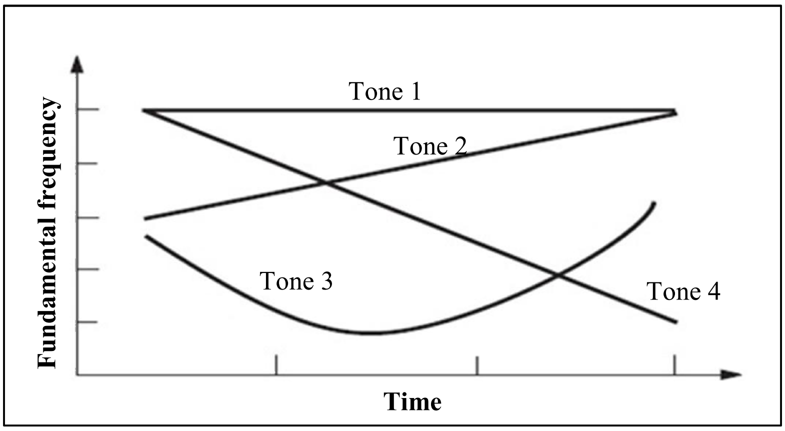

Figure 1.1 Standard F_0 contours of four Mandarin Chinese tones [2]

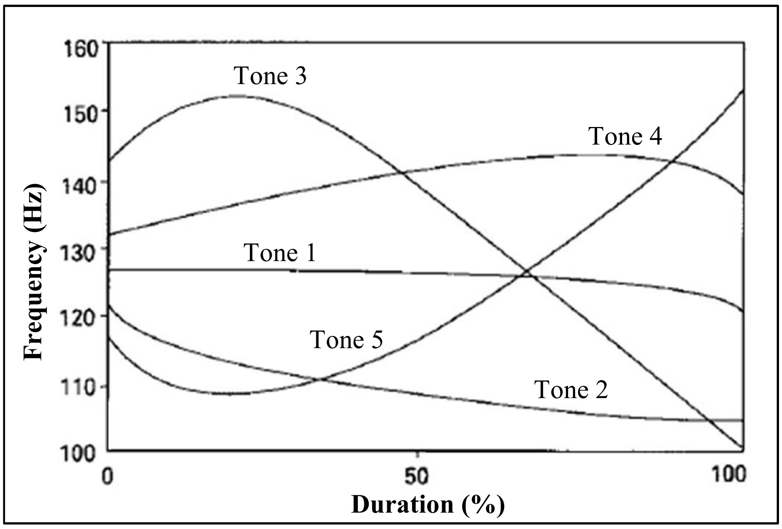

Figure 1.2 Standard F_0 contours of five Thai tones [3]

There are two different kinds of speech recognition systems: network-based and non-network-based. In the network-based system, the speech recognition system is located on a network server and almost every process is executed by software. This system is suitable for continuous speech recognition and natural language speech recognition. In addition, it is only supported in some languages. This may result in misunderstanding when it is applied by non-native speaker.

The non-network-based speech recognition system is mainly applied to command-based speech recognition and word/phrase speech recognition. When we apply ASR and ATSR to a non network environment, ASR and ATSR systems have to be implemented on mobile equipment. In this case, the systems should be real-time processing and with low calculation complexity.

Tone recognition is generally developed in both isolated speech and continuous speech. Continuous speech is more applicable for natural language which is typically used for communication. In continuous speech, the concentrated F_0 is generally varied by the negative influence from the neighboring F_0, and recognition accuracy is reduced due to incorrect F_0 features. To develop a highly efficient tone recognition, researchers should enhance the F_0 contour so that it is the same as the ideal F_0 contours of each tonal language and/or develop high recognition accuracy tone recognition.

There are many techniques for extracting F_0 features in time domain and frequency domain. While frequency domain provides high performance, it provides more complex calculation than in time domain. We therefore focus on tone feature extraction in time domain.

However, some techniques require many multiplications which are one of the most time and area-consuming operations in digital signal processing. In addition, they find F_0 feature extraction, one of the highest calculation processes, from the entire input syllable. These techniques also are operated in series. The series tone recognition operates the number of arithmetic operations in sequence, which results in an elevated computation cost, especially in the case of large vocabularies.

In order to develop tone recognition for portable equipment, this thesis designs tone recognition for automatic tonal speech recognition with real-time and low calculation complexity. To reduce tone recognition complexity, we reduce the number of input frames and determine the F_0 feature extraction process without multiplications. In addition, the tone recognition is designed by parallel and pipeline architecture to achieve real-time processing.

1.2 Contribution

To develop a real-time and low calculation complexity tone recognition, this thesis proposes a study on low calculation complexity for automatic tonal speech recognition in real-time applications.

The proposed tone recognition estimates F_0 from only estimated vowel signals, called vowel magnitude difference function, vowel-MDF (V_{MDF}). This results in a reduction by approximately half in the number of input frames, compared with techniques used by the entire input syllable and also provides no adjacent syllable negative influence. Vowel segmentation provides high accuracy when we find the vowel threshold from the average energy of the entire syllable. However, the average energy threshold calculation starts after the end of the utterance. This results in a processing time delay response. This thesis finds the vowel threshold by adapting the average energy value. We assume that the maximum energy value occurs in the middle frame and the threshold is half of the energy value of this frame. Since vowel signals can be estimated by time-synchronous calculation, the tone model can be processed in real-time processing.

In addition, the V_{MDF} method can be processed without multiplication. Due to the reduction in frame number and absence of multiplication, the tone recognition provides low complexity.

This thesis develops the tone recognition in pipeline and parallel architecture in order to achieve a high throughput and accelerate the tone recognition. The architecture is operated in 32 parallelism of three process elements. Each process element (PE) is executed in pipelined processing and this results in a reduction of the number of clock cycles. Due to the reduction of the clock rate, the proposed architecture achieves real-time and low calculation complexity making it more applicable for portable equipment.

1.3 Thesis Outline

The thesis outline is as follows: Chapter 2 provides the background and related work of this thesis. We discuss and compare the related works with our proposed method. Chapter 3 presents the theory and solution of low complexity tone recognition. We provide an overview of tone recognition which shows three tone recognition processes. Chapter 4 presents a low calculation complexity of tone recognition. In this chapter, we describe the low complexity tone recognition and its parallel architecture which makes the tone recognition achieve real-time and low calculation complexity. Chapter 5 shows the results and evaluation of this thesis with regard to recognition accuracy, number of frame reduction, and processing time, compared with the conventional method. Chapter 6 discusses the experimental results from chapter 5. Finally, chapter 7 presents conclusions and future work. This chapter states the thesis statement, summarizes contributions, and prospects for future work.