2002

Tone classification of syllable -segmented Thai speech based on multilayer perceptron

Nuttavudh Satravaha
West Virginia University

Tone Classification of Syllable-Segmented Thai Speech Based on Multilayer Perceptron

Nuttavudh Satravaha

Dissertation Submitted to the
College of Engineering and Mineral Resources
at West Virginia University
In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
In
Electrical Engineering

Powsiri Klinkhachorn, Ph.D., Chair
Ali Feliachi, Ph.D.
Mark A. Jerabek, Ph.D.
Norman J. Lass, Ph.D.
Ronald L. Klein, Ph.D.

Lane Department of Computer Science and Electrical Engineering

Morgantown, West Virginia

2002

Keywords: Thai speech, tone classification, syllable segmentation, Tonal coarticulation, stress, intonation, multilayer perceptron

ABSTRACT

Tone Classification of Syllable-Segmented Thai Speech Based on Multilayer Perceptron

Nuttavudh Satravaha

Thai is a monosyllabic and tonal language. Thai makes use of tone to convey lexical information about the meaning of a syllable. Thai has five distinctive tones and each tone is well represented by a single *F0* contour pattern. In general, a Thai syllable with a different tone has a different lexical meaning. Thus, to completely recognize a spoken Thai syllable, a speech recognition system has not only to recognize a base syllable but also to correctly identify a tone. Hence, tone classification of Thai speech is an essential part of a Thai speech recognition system.

In this study, a tone classification of syllable-segmented Thai speech which incorporates the effects of tonal coarticulation, stress and intonation was developed. Automatic syllable segmentation, which performs the segmentation on the training and test utterances into syllable units, was also developed. The acoustical features including fundamental frequency (*F0*), duration, and energy extracted from the processing syllable and neighboring syllables were used as the main discriminating features. A multilayer perceptron (MLP) trained by backpropagation method was employed to classify these features. The proposed system was evaluated on 920 test utterances spoken by five male and three female Thai speakers who also uttered the training speech. The proposed system achieved an average accuracy rate of 91.36%.

Acknowledgments

First and foremost, I would like to express my sincere appreciation to my research advisor, Dr. Powsiri Klinkhachorn, for his continuous guidance, support, encouragement and valuable suggestion throughout the years of my research. An equally important contributor to the success of this dissertation is Dr. Norman J. Lass who willingly assumes the responsibilities of a co-advisor. I am specially grateful to him for his kindness and generosity. I am immensely indebted to both of them for their expertise and assistance. Without them, this dissertation would not have materialized.

My thanks and sincere appreciation also go to the members of my Advisory and Examining Committee: Dr. Ali Feliachi, Dr. Mark A. Jerabek, and Dr. Ronald L. Klein, for their valuable comments and suggestions. I am truly grateful to Sharon Santos, whose proof reading and numerous suggestions on how to improve the writing of this dissertation was of immense help.

I am grateful to the Telephone Organization of Thailand for providing the financial support throughout my graduate years. I also thank all Thai students who participated in this work, especially Pawalai who always helped and supported since the beginning of this research.

Lastly, I would like to thank my parents for their love and encouragement throughout my life. I am truly grateful to my mother-in-law, Mae Jook, for her encouragement and support during the years. My deepest gratitude especially goes to my wife, Job, for her undying love, patience and support throughout our marriage. I appreciate the sacrifices she has made for me which have enabled me to reach this special milestone in my life. I also thank to my beloved daughter and son: Tam and Thew, for their love and understanding.

Table of Contents

Chapter 1 Introduction and Background

During the past decade, speech recognition technology has made rapid advances, supported by the progress in computer technology and advances in speech understanding and linguistics. Recently, there has been much progress in bringing speech recognition technology to commercial applications. The potential applications of speech recognition systems can range from simple tasks, such as speech-to-text, to more complicated tasks, such as language translator. Although much progress has been made in speech recognition technology, most of the existing methods are developed mainly for spoken English. Due to numerous potential applications, many researchers are currently developing speech recognition for their own languages [28], [48], [49], [51], [63], [96], [107], [112].

Thai is a monosyllabic and tonal language [1]. Thai uses tone to convey a lexical meaning of a syllable. Thus, to completely recognize a spoken Thai syllable, the speech recognition system has not only to recognize a base syllable but also to correctly identify a tone. Although tone recognition of Thai speech is an essential part of a Thai speech recognition system, only a few research studies which focus on classifying Thai tones are available.

Thai has five distinctive tones and each tone is well represented by its fundamental frequency (*F0*) pattern [59]. Several factors, including tonal coarticulation, stress, and intonation, may affect the tone patterns of Thai speech. The *F0* pattern of a syllable affected by the *F0* patterns of neighboring syllables is referred to as tonal coarticulation [27]. The *F0* contour patterns of the unstressed syllables are generally different from the stressed ones [74]. The intonation effect makes the *F0* contour of the utterances decline gradually [76]. Currently, no research on tone classification exists which incorporates tonal coarticulation, stress

and intonation effects in their studies; thus, these factors are taken into consideration in this study which makes this research a pioneer work in this area. In addition, a method to perform automatic segmentation on Thai speech into syllable units is currently not available [76], automatic syllable segmentation is developed in this research study as well.

Chapter 1 is organized as follows: A brief introduction to standard Thai is given in Section 1.1. The factors that affect the tone pattern of Thai are described in Section 1.2. Section 1.3 provides the problem statement, objectives and contributions of this research study.

1.1 Overview of Standard Thai

Standard Thai is the national language of Thailand, and the dialect spoken throughout the central region of the country, including that spoken in the city of Bangkok. Standard Thai is used in broadcasting and in conducting official business and legal matters. In general, Thai is a monosyllabic and tonal language. A syllable is assigned a tone and each spoken syllable with a different tone will have a different lexical meaning [1]. The details of Thai syllable structures are listed in Appendix A. The following contains a brief survey of standard Thai based on Luksaneeyanawin and Naksakul studies [59], [62].

1.1.1 Tones

The most important feature of the Thai language is its use of tone to convey the lexical meaning of the syllables. Thai has five contrasting lexical tones: mid, low, falling, high, and rising. The characteristics of Thai tones are well represented by their fundamental frequency (*F0*) pattern [59]. Thai tones can be divided into two groups: static and dynamic. The static group consists of three tones (mid, low, and high) whereas the dynamic group consists of two tones (falling and rising). The dynamic group is characterized by sharp *F0* contours as opposed to the relatively

smooth *F0* movement of the static group. The phonemic transcription uses the diacritics /^{0}/, /^{1}/, /^{2}/, /^{3}/ and /^{4}/ as tone markers for mid, low, falling, high and rising tones, respectively. Traditionally, the five Thai tones are in the following order as listed in Table 1.1.

Tone order	0	1	2	3	4
Tone in Thai	“สามัญ”	“เอก”	“โท”	“ตรี”	“จัตวา”
Pronunciation	/saa^{4} man^{0}/	/?ek^{1}/	/thoo^{0}/	/trii^{0}/	/cat^{1} ta^{1} waa^{0}/
Tone name	mid	low	falling	high	rising

Table 1.1 Order of traditionally Thai five tones.

As already mentioned, a Thai syllable having any one of five tones produces a different lexical meaning. The examples in Table 1.2 demonstrate the effect of tones on meaning.

Thai syllable	Pronunciation	Meaning
“นา”	/ naa^{0}/	“field”
“หน่า”	/ naa^{1}/	“custard apple”
“หน้า”	/ naa^{2}/	“face”
“น้า”	/ naa^{3}/	“mother’s younger sibling”
“หนา”	/ naa^{4}/	“thick”

Table 1.2 Examples of Thai syllables with different tones having different meanings.

The tonality of a monosyllable is mainly characterized by the shape of its *F0* contour, especially *F0* height and *F0* slope or direction. The average *F0* contours of the five Thai tones when an isolated monosyllabic word was spoken are presented

in Figure 1.1. As shown in Figure 1.1, it has been found that the beginning portion of high and falling tones are higher than low and rising tones, with mid tone being intermediate. The ending portion (of the high and rising tones) has been found to be higher than low and falling tones, with mid tone being intermediate. *F0* contours associated with mid and low tones fall steadily throughout, whereas those associated with the falling, high, and rising tones change abruptly in slope, approximately halfway through their duration. The falling tone rises slightly and then falls sharply; the high and rising tones fall slightly and then rise. Luksaneeyanawin [59] concluded that the two main features which can be used to distinguish the five tones from each other are the *F0* height and *F0* direction.

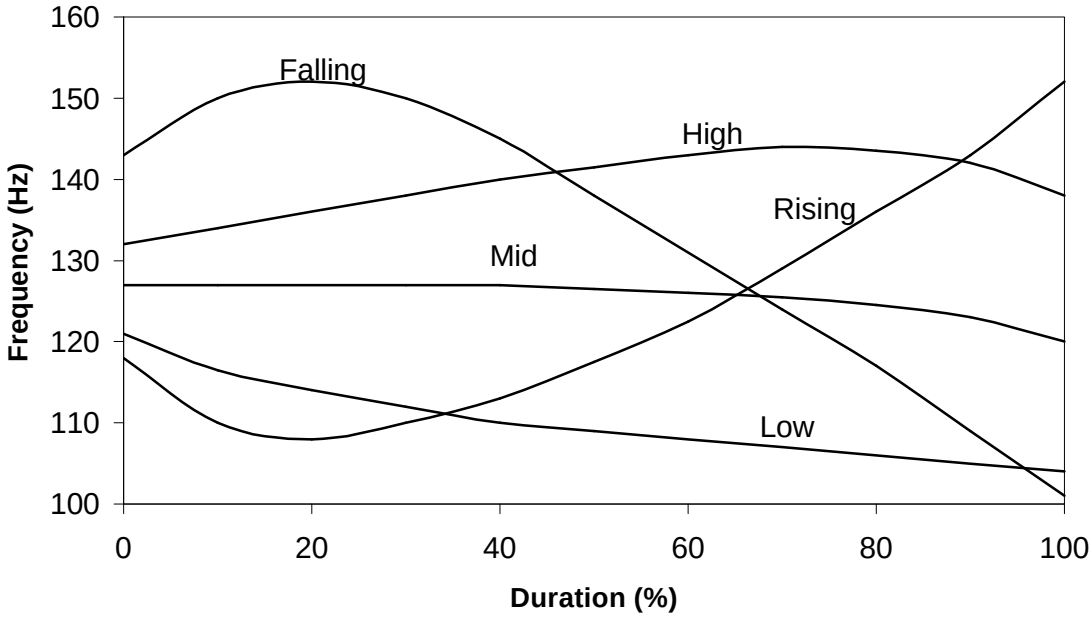

Figure 1.1 Average *F0* contours of five Thai tones when an isolated monosyllable word was spoken (adapted from Abramson, [1]).

1.1.2 Stress

The syllable in a word produced with a higher degree of respiratory effort is referred to as “stress.” The stressed syllables are usually louder, longer, and higher in pitch than unstressed syllables. The placement of stress on a word in Thai is

linguistically significant and governed by rules including the monosyllabic word rule and the polysyllabic word rule. For monosyllabic words, all content words are stressed, whereas all grammatical words are unstressed. However, monosyllabic unstressed words when spoken in isolation or emphasized can be stressed as well. For polysyllabic words, stress placements are determined by the number of syllables as well as the structure of the component syllables in the word. The primary stress falls on the final syllable of a word. The secondary stress is determined by the position of the remaining syllables and whether or not they are linker or non-linker syllables [59].

1.1.3 Vowels

Thai vowel system is divided into 18 monophthongs and six diphthongs. Thai monophthongs consist of nine short or single vowels: /i/, /e/, /x/, /v/, /q/, /a/, /u/, /o/, and /@/, where each short vowel has a corresponding long vowel represented by double letters: /ii/, /ee/, /xx/, /vv/, /qq/, /aa/, /uu/, /oo/, and /@@/. The short and long pair are quantitatively different but qualitatively similar. The eighteen Thai monophthongs, classified according to tongue heights and positions, are depicted in Table 1.3 [1].

The six Thai diphthongs consist of /ia/, /iia/, /va/, /vva/, /uua/, and /ua/. The first member of diphthongs can be one of /i, ii, v, vv, u, uu/ and always ends with /a/. Hence these eighteen monophthongs together with six diphthongs are the core of Thai syllables [59].

Tongue height	Tongue advancement					
	Front		Central		Back	
	short	long	short	long	short	long
high	/i/ “ิ”	/ii/ “ี”	/v/ “ึ”	/vv/ “ื”	/u/ “ ุ”	/uu/ “ ู”
mid	/e/ “เ_ะ”	/ee/ “เ_”	/q/ “เ_อะ”	/qq/ “เ_อ”	/o/ “โ_ะ”	/oo/ “โ_”
low	/x/ “แ_ะ”	/xx/ “แ_”	/a/ “_ะ”	/aa/ “_า”	/@/ “เ_าะ”	/@@/ “_อ”

Table 1.3 Thai monophthongs classified according to tongue heights and positions

1.1.4 Consonants

Although Thai has 44 consonant letters, there are only 21 consonantal phonemes: / p t c k ? ph th ch kh b d m n ng l r f s h w j /. It is implied that several Thai consonant letters could be represented by the same consonantal phoneme. For example, four fricative Thai consonants <ซ,ศ,ษ,ส> are represented by the consonantal phoneme /s/. Thai consonants classified according to the voicing, the manner of articulation and the place of articulation are listed in Table 1.4.

Consonants		Place of articulation					
		bilabial	labio and dental	dental and alveolar	palatal	velar	glottal
Plosive	voiceless unaspirated	/p/ <ป>		/t/ <ต,ฏ>	/c/ <จ>	/k/ <ก>	/?/ <อ>
	voiceless aspirated	/ph/ <พ,ภ,ผ>		/th/ <ท ,ธ,ฒ,ฑ, ถ,ฐ>	/ch/ <ช,ฌ,ฉ>	/kh/ <ค,ฆ,ข>	
	voice	/b/ <บ>		/d/ <ด,ฎ,ฑ>			
Nasal		/m/ <ม>		/n/ <น,ณ>		/ng/ <ง>	
Lateral				/l/ <ล,ฬ>			
Rolled				/r/ <ร,ฤ>			
Fricative			/f/ <ฟ,ฝ>	/s/ <ซ,ศ,ษ,ส>			/h/ <ห,ฮ>
Semi-vowel		/w/ <ว>			/j/ < ,ย>		

Table 1.4 Classification of Thai consonants according to voicing, the manner of articulation and the place of articulation. Note that the letters in parentheses are Thai consonant letters.

1.2 Important Factors that Affect Thai Tones

The tone patterns of Thai speech are subject to various modifications resulting from the effects of tonal coarticulation, stress, and intonation. Since these factors are taken into consideration in this study, the following sections provide a brief report of these factors. The definition of tonal coarticulation is given in the Section 1.2.1. Section 1.2.2 describes the acoustical characteristics of stress. The intonation is described in Section 1.2.3. The explanation of the acoustical variability

occurring between different speakers and within the same speaker is given in Section 1.2.4

1.2.1 Tonal Coarticulation

The *F0* contour shape of a syllable affected by the *F0* contour patterns of adjacent syllables is commonly known as “tonal coarticulation.” Two types of tonal coarticulation are anticipatory and carryover coarticulation. The *F0* contour shape of a syllable that is influenced by succeeding phonemes is called “anticipatory coarticulation.” The carryover coarticulation occurs when the preceding sound influences the succeeding sound.

Gandour *et al*., [27] investigated tonal coarticulation in Thai by analyzing tone contours in terms of the height and slope of the fundamental frequency (*F0*). Gandour employed the utterances from ten speakers, each of whom produced 25 possible sequences of two tones from the five tones of Thai embedded in a carrier sentence. Both syllables in the two-tone sequences were stressed. Gandour reported that the coarticulation effects of Thai tones were asymmetric. Thai tones were more influenced by carryover than by anticipatory coarticulation. In addition, their results showed that tonal coarticulation in Thai does not necessarily affect the whole *F0* contour of adjacent syllables. For instance, the carryover effects of *F0* height extend forward to about 75% of the duration of the succeeding syllable whereas the anticipatory effects extend backward to about 50% of the duration of the preceding syllable. Tonal coarticulation in Thai affected primarily *F0* height while *F0* slope or direction was relatively unaffected. In addition, the carryover coarticulation affected a greater number of Thai tones than anticipatory coarticulation. In terms of *F0* height, carryover coarticulation effects were relatively absent in the rising tone only, while anticipatory coarticulation affected the low and falling tones as shown in Figure 1.6 and Figure 1.7. In terms of *F0* slope, anticipatory coarticulation affected none of the five tones while carryover coarticulation affected mid and low tones only.

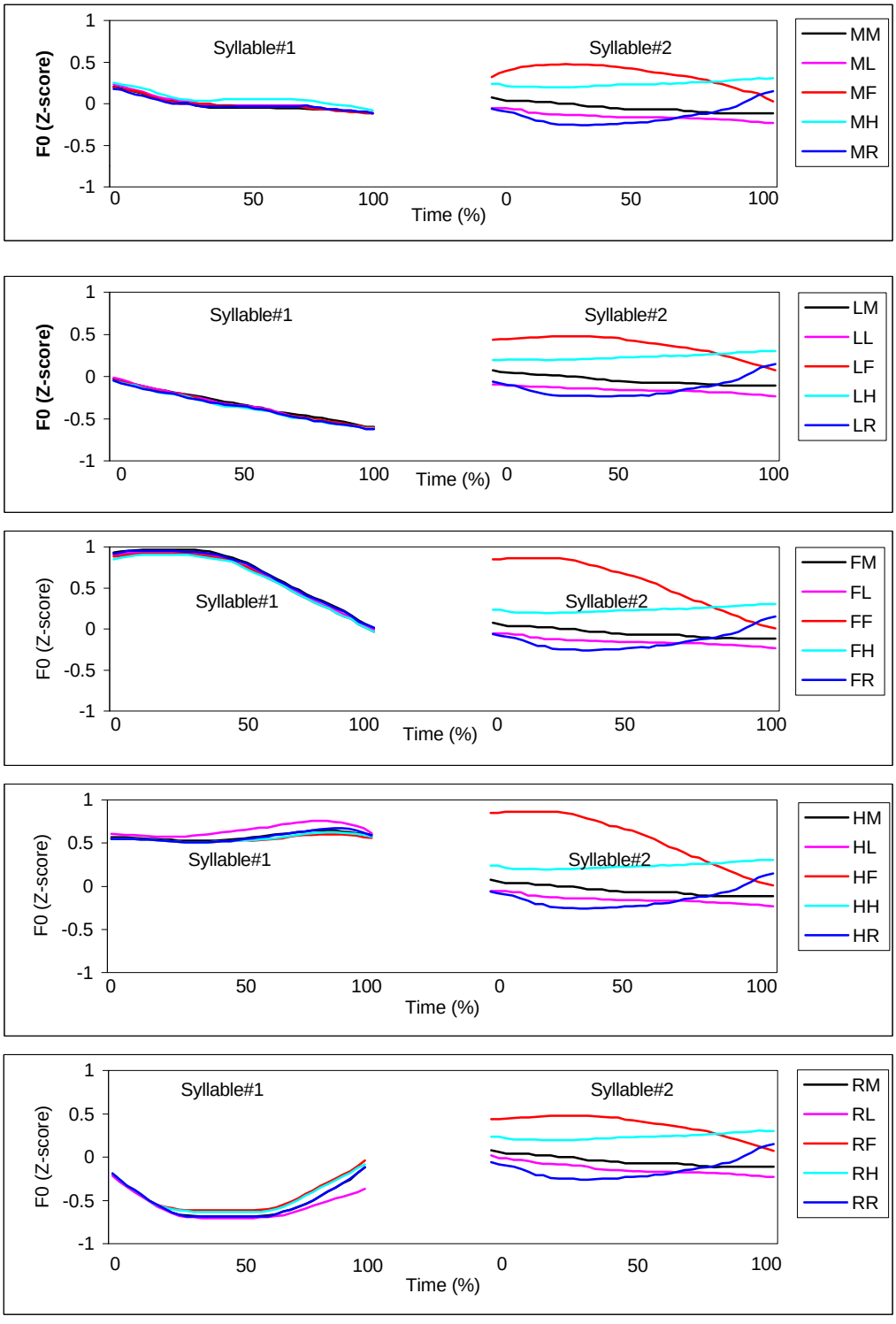

Figure 1.6 *F0* contours of two-tone test utterances when affected by anticipatory coarticulation (adapted from Gandour *et al*., [27]).

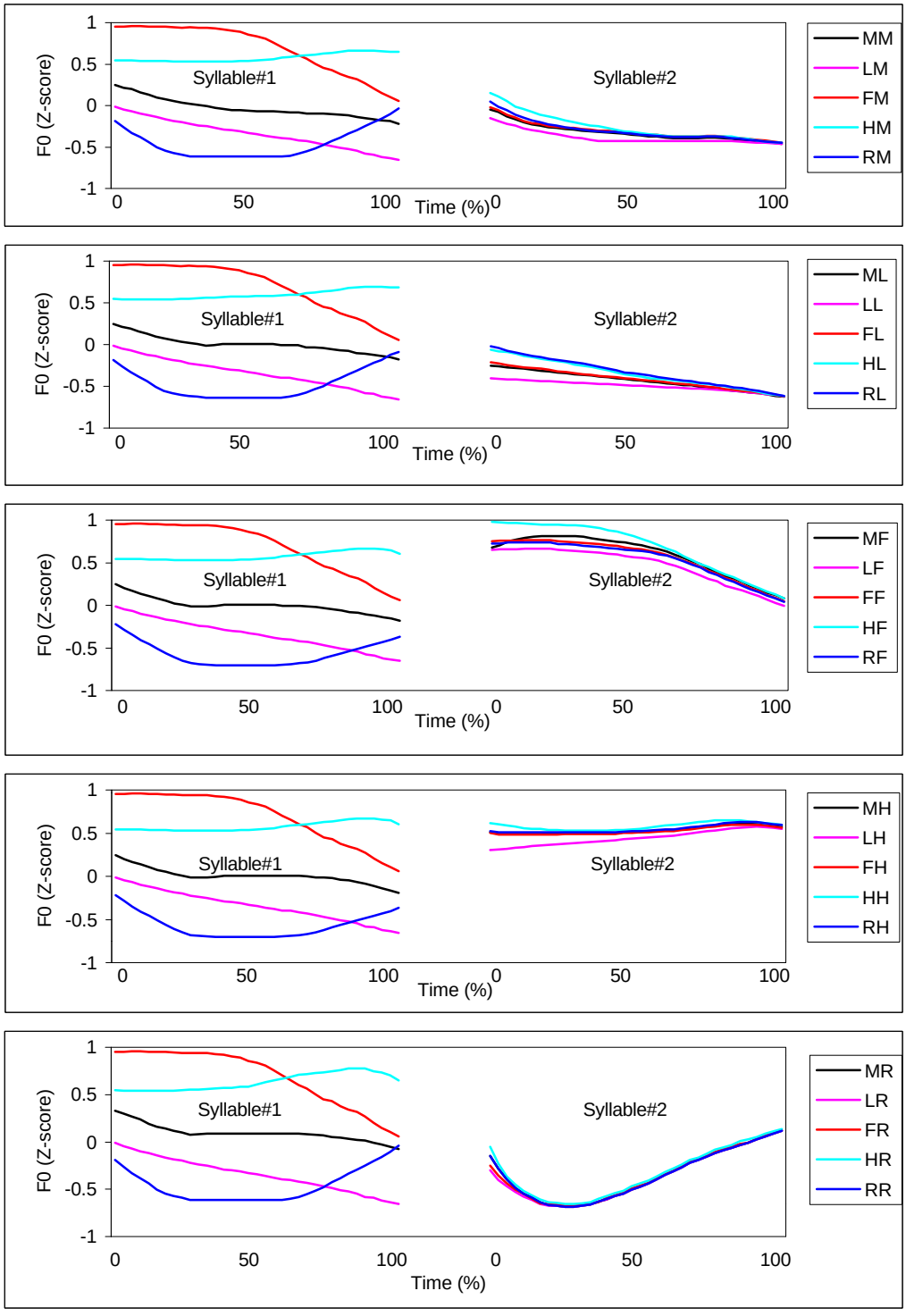

Figure 1.7 *F0* contours of two-tone test utterances when affected by carryover coarticulation (adapted from Gandour *et al.*, [27]).

It is important to note that the anticipatory and carryover coarticulation effects in the two-tone sequences were measured at the first and second syllable of the two-tone sequences. The left parts of the panels in Figure 1.6 and Figure 1.7 represent the fundamental frequency (z-score) of the first syllable of the two-tone sequence whereas the right parts of panels of Figure 1.6 and Figure 1.7 represent the second syllable of the two-tone sequence. The phonemic transcription uses the diacritics /M/, /L/, /F/, /H/ and /R/ as tone markers for the mid, low, falling, high and rising tones, respectively. For example, MM, ML, MF, MH, and MR represent the two-tone sequences of mid-mid, mid-low, mid-falling, mid-high, and mid-rising, respectively.

1.2.2 Stress Effects

Stress is a suprasegmental feature relating to the production of a syllable. The perception of stress involves the three acoustic parameters of intensity, duration, and fundamental frequency. Stressed syllables are more prominent than unstressed syllables. This prominence is due to an increased physical effort in production, with the result that stressed syllables are usually louder, longer, and higher in pitch than unstressed ones. The acoustic characteristics of the stressed and unstressed syllables have been investigated at length by Phinicharom [69] and Potisuk *et al*., [74]. The following brief reviews are based on their studies.

Phinicharom [69] investigated the *F0* of tones and the duration of voiced segments in the stressed and unstressed Thai syllables. She reported that tones in unstressed syllables differ from those in stressed tones in three ways. First, unstressed syllables maintain the *F0* height and contour of original tones but the range of the *F0* movement of rising and falling tones is narrower than stressed ones for all syllable structures. Second, all five tones are neutralized toward mid-level tones. Third, the *F0* height and contour of dynamic tones (falling and rising tones) are changed to be similar to a high tone. In addition, unstressed syllables are found to have a much shorter duration than those in stressed syllables. The unstressed

syllables primarily have an insignificantly different duration among different types of syllable structures and different speakers.

Potisuk *et al*. [74] studied the relative contribution of fundamental frequency (*F0*), duration, and intensity in signaling stress in Thai and determined the usefulness of combining acoustic parameters to classify stressed and unstressed syllables. In their study, other factors affecting *F0* contour such as speaking rate, tonal coarticulation, and intonation in terms of declination were held constant. The spoken utterances were collected from five speakers and each speaker produced 25 pairs of sentences at normal speaking rate. Each pair of sentences contained a two-syllable sequence and the first member of each pair had a speaker who spoke with a stressed-stressed pattern, whereas the second sentence had an unstressed-stressed pattern. Five prosodic dimensions of the rhyme portion of the target syllable (duration, average *F0*, *F0* standard deviation, average intensity and intensity standard deviation) were measured.

Potisuk reported that the tonal *F0* contours differed in shape between stressed and unstressed syllables. In stressed syllables, the contrastive relationship among the five tones replicated earlier findings on Thai tones [1]. In unstressed syllables, the excursion size of *F0* movement of all five tones was reduced. The falling, high and rising tones were primarily affected whereas mid and low tones were relatively unaffected. The *F0* contours of stressed and unstressed syllables for all five tones are presented in Figure 1.8

In addition, the stressed syllables were significantly longer than the unstressed syllables for all five lexical tones, while the intensity contours did not vary significantly between stressed and unstressed syllables across all five tones. Potisuk concluded that duration is the predominant cue in distinguishing stressed and unstressed syllables in Thai. In addition, the *F0* shape is different between the stressed and unstressed syllable, primarily falling, high and rising tone.

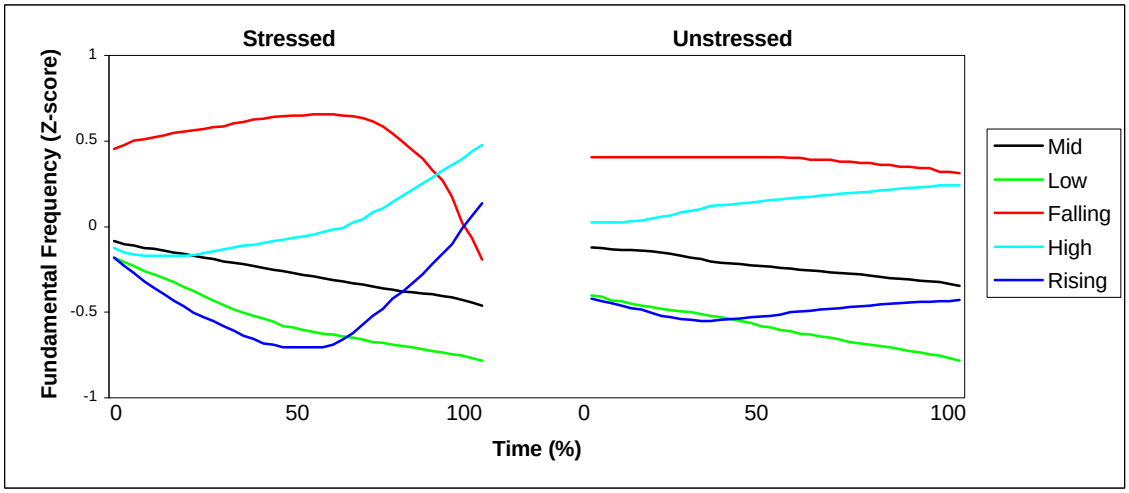

Figure 1.8 *F0* contours of stressed and unstressed syllables for all five tones (adapted from Potisuk *et al*., [74]).

1.2.3 Intonation

Intonation is a suprasegmental feature of a language that appears simultaneously with the segmental feature and superimposed on a segment of speech unit such as a phrase or sentence [20]. Its function is to express differences in the speaker's intended meaning. All languages, including tonal languages, make use of intonation. A non-tonal language such as English uses a rise-fall intonation pattern to signal a declarative statement. It is believed that the interaction between intonation and tone manifests itself in terms of intonation patterns being superimposed on the tones. The important characteristic of intonation in Thai is declination, which refers to a gradual modification over a segment of speech unit such as a phrase or sentence and makes the *F0* contour of the utterance decline gradually [76].

1.2.4 Inter-speaker and Intra-speaker Variability

Acoustic variability can occur both between different speakers and within the same speaker. This variability creates a problem for speech perception and speech

synthesis, as well as speech recognition. In addition, the acoustic variability problem also extends to suprasegmental features, segmental features, and the production of lexical tones.

Gandour *et al*. [26] investigated interspeaker and intraspeaker variability in fundamental frequency of Thai tones. A comparison of interspeaker and intraspeaker variability in the production of five Thai tones was made. The results indicated that the degree of interspeaker variability in *F0* was greater than intraspeaker variability across all five tones. Young and old speakers exhibited the same pattern of variability; thus age-related effects appear to be minimal. Since the average *F0* is determined largely by the size of the vocal folds, Thai females primarily produce a higher average *F0* than Thai males for all five tones. However, variability mainly shifts *F0* contour upward or downward instead of changing *F0* shape. The degree of variability in tone production varies depending on the lexical tone. For instance, the dynamic tones (falling and rising tones) exhibited a smaller degree of variability than the static tones (mid, low, and high tones). The fundamental frequency of five tones of Thai male and female speakers are shown in Figure 1.9 and Figure 1.10.

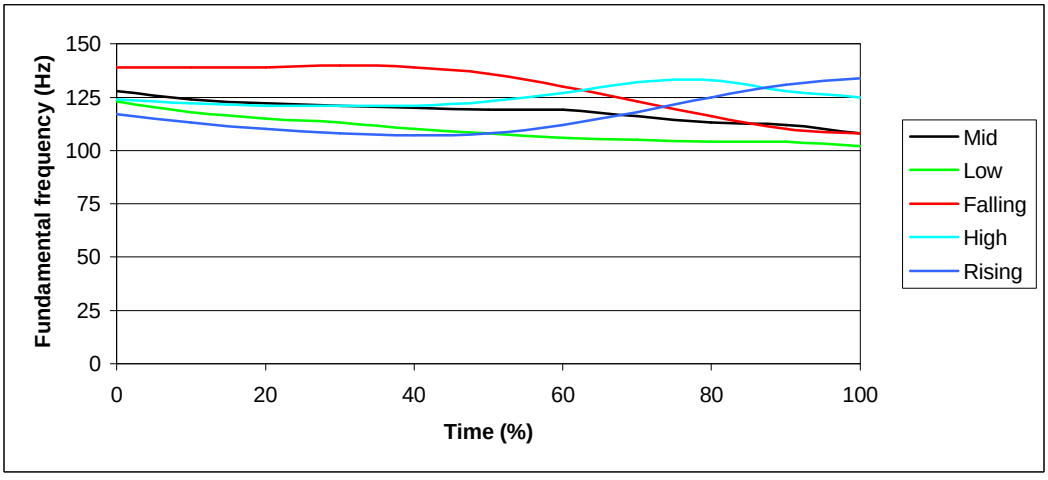

Figure 1.9 *F0* contours of five tones of Thai male speaker (after Gandour *et al*., [26]).

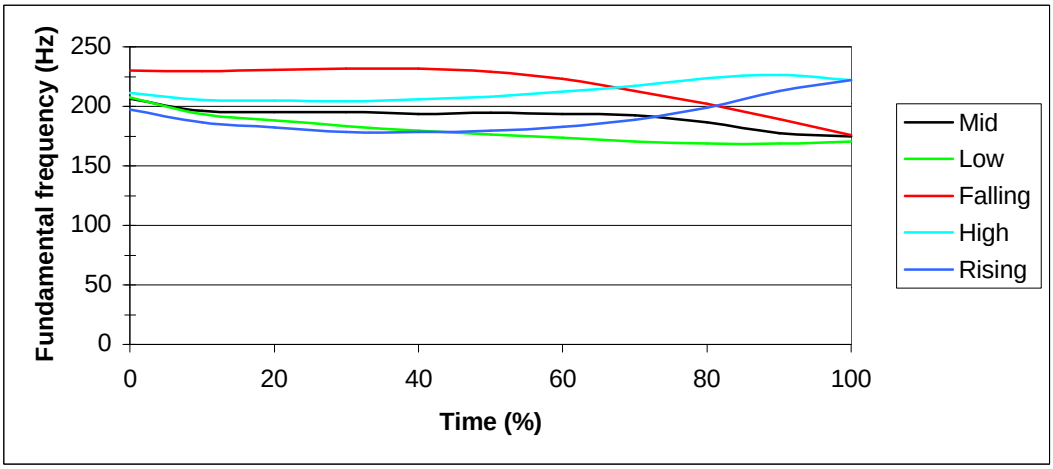

Figure 1.10 *F0* contours of five tones of Thai female speaker (after Gandour *et al.*, [26]).

1.3 Problem Statement and Objectives

In this section, the problem statement is addressed, and the objectives and assumptions of this research study are presented. The contributions of this research study are also outlined.

1.3.1 Problem Statement

This dissertation addresses the problems of tone classification in Thai speech. The problems assigned in this study are complex, particularly in three ways:

1. Due to the unavailability of automatic syllable segmentation of Thai speech, automatic syllable segmentation is developed in this study. Syllable segmentation is a difficult task because of the variation in duration and amplitudes for different sounds. In addition, the performance of endpoint detectors may degrade dramatically when spoken words or syllables are coupled together, the result of which is that it is difficult to locate a word or syllable boundaries accurately.

2. As mentioned earlier, Thai has five distinctive tones and each tone is characterized by its fundamental frequency (*F0*) pattern [59]. Several factors, including tonal coarticulation, stress, and intonation, may affect the tone patterns of Thai speech. The *F0* pattern of a syllable affected by the *F0* patterns of neighboring syllables is referred to as tonal coarticulation [27]. The *F0* contour patterns of the unstressed syllables are generally different from the stressed ones [74]. Other than tonal coarticulation and stress, intonation also plays an important role in the production of Thai speech. The important characteristic of intonation is declination, which refers to a gradual modification over a segment of speech unit such as a phrase or sentence and makes the *F0* contour of the utterance decline gradually [76]. The declination effect is superimposed on the *F0* contour which is already influenced by other factors such as tonal coarticulation and stress. Hence, the effects of tonal coarticulation, stress, and intonation make the tone classification of Thai speech a complicated task.

3. There are perceptible differences in the average *F0* and *F0* range of different speakers as well as within the same speaker. Due to the smaller size of vocal folds, Thai females primarily produce a higher average *F0* than Thai males for all five tones. Thus, it is possible for a Thai male's high tone to have a lower *F0* than a Thai female's low tone. Since there are differences in the excursion size of *F0* movements related to differences in voice range between speakers as well as the differences in height of *F0* movements between speakers, a *F0* normalization procedure is required to overcome these problems before the tone classification is performed.

1.3.2 Research Objectives

The objectives of this research study are to develop a tone classification of Thai syllables in which tonal coarticulation, intonation, stressed and unstressed syllables are taken into consideration. In addition, methods are developed for automatic syllable segmentation and stress detection as well.

In summary, the research objectives are:

- Automatic syllable segmentation of Thai speech has been developed in this study due to the unavailability of a method to perform segmentation of Thai speech into syllables [76]. The syllable segmentation makes use of the relationships between peaks and valleys in the modified energy contour to locate the starting and ending points of the spoken syllables.

- A stress detector based on the fuzzy logic technique is implemented. The stress detector utilizes the duration and energy of a syllable to determine a stress degree of a syllable.

- A tone classification method for Thai speech based on multilayer perceptron (MLP) is implemented. The tone classifier is trained by using the training vectors which contain the acoustical features extracted from the neighboring and processing syllables in order to deal with the effects of tonal coarticulation, intonation, and stressed and unstressed syllables.

- A proposed tone classification system is speaker-dependent in which the system is tested by the speaker who also trained the system. The proposed tone classification system is tested on 920 test utterances which are spoken by five male and three female Thai speakers. The test utterances contain all stress patterns and carry all five tones. The performance of the proposed method is evaluated in terms of an accuracy rate.

1.3.3 Assumptions

The assumptions made in this research study are summarized as follows:

1. The speakers are natives of Thai and ranged from 20 to 40 years of age.
2. The speakers use the standard Thai dialect.
3. The training and test utterances are spoken by five male and three female speakers.
4. It is assumed that the training and test utterances are recorded in a quiet environment.
5. A proposed tone classification system is implemented using MATLAB.

1.3.4 Research Contributions

This dissertation has made the following contributions to the literature on Thai speech recognition.

These contributions can be divided into two areas:

- The first area is based on the implementation of automatic syllable segmentation of Thai speech. Manual speech segmentation is a tedious and time-consuming task, and the results lack reproducibility because of the subjective decisions involved. Since performance of speech recognition systems relies on the accuracy of the endpoint detection, the computation for processing speech is minimum when the endpoints are accurately located. However, the performance of endpoint detectors may degrade dramatically when applied to the input speech in which spoken words or syllables are coupled together and it is difficult to locate word or syllable boundaries. Since a method to perform automatic segmentation on Thai speech into syllables is not available [76], the automatic method of syllable segmentation developed in this study should be beneficial to future research on Thai speech recognition.

- Although tone classification is an essential part of Thai speech recognition, there have been only a few research studies conducted on tone classification. In addition, these studies did not consider the effects of tonal coarticulation, stress and intonation [30], [89]. However, a recent study incorporated tonal coarticulation and intonation, but considered only the stressed syllables [76]. Both stressed and unstressed syllables are part of the grammar of the Thai language, and Thai people generally speak both stressed and unstressed syllables in everyday lives. Since no research on tone classification which incorporates tonal coarticulation, intonation, as well as stressed and unstressed syllables is available, this dissertation is considered to be a pioneer work in these areas.

1.4 Summary

This chapter gives a brief overview of standard Thai, as well as a description of the important factors that affect Thai tones. This description will help readers better understand the problems and the objectives of this research. Finally, the assumptions and research contributions are presented.

The organization of this dissertation is as follows. Chapter Two, entitled “Literature Survey”, provides a brief review of several tools and techniques related to this research study. Also, previous research studies of Thai speech recognition and other tonal languages (e.g., Mandarin, Cantonese) related to this research are reviewed. Chapter Three, entitled “Analysis of Thai Tones”, presents the method of collecting speech data and the statistical data involving the stressed and unstressed syllables, intonation and tonal coarticulation. Chapter Four, entitled “Thai Tone Classification System”, provides the details of implementation of the proposed method. Chapter Five, “Results and Observation”, presents the experimental results and a discussion of the important findings. Finally, in Chapter Six, the research is summarized and suggestions for future work are listed.

Chapter 2 Literature Survey

In this chapter, several speech recognition methods are first presented. The important aspects of a fuzzy inference system are also outlined. The literature with special emphasis on previous studies on Thai speech recognition is then reported. Although this research focuses on tone classification of Thai speech, some previous research on other tonal languages (e.g., Mandarin and Cantonese) is also reviewed.

2.1 Speech Recognition Methods

Research in speech recognition has been investigated for four decades with most research being devoted to spoken English [39], [40], [46], [47], [70], [82], [85], [86], [87]. However, there are several factors that create problems in speech recognition. These difficulties are due to the acoustical variability of speech seen in speaker characteristics and environmental conditions, for example. Presently, speech researchers have made tremendous gains in developing recognition systems, especially in several recognition techniques which have been employed to deal with the acoustical variability of speech. These methods include dynamic time warping (DTW), hidden Markov model (HMM), and neural network (NN). However, HMM and NN approaches are currently the most popular methods in speech recognition applications. The brief reviews of these methods are presented here.

2.1.1 Dynamic Time Warping (DTW)

Early research on speech recognition focused mainly on the area of isolated word or discrete utterance recognition. The most popular techniques were the ones based on the template matching method. This technique relied on the idea that the unknown input pattern was compared to the reference patterns via some distortion measures. The best matches between the reference patterns and the input patterns

yielded the smallest dissimilarity or distance. Finally, the decision was made upon the pattern that yielded the minimum distortion. A block diagram of a template matching method is depicted in Figure 2.1.

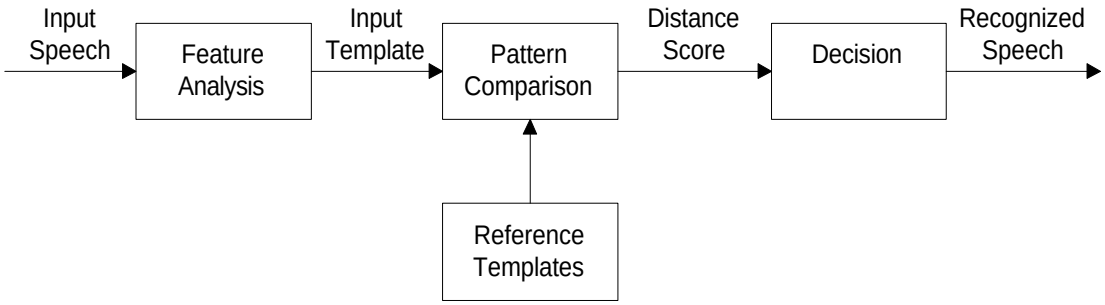

Figure 2.1 Block diagram of template matching method

In this approach, speech signal was first analyzed and a time series of feature vectors that characterize the spectral components of speech were extracted. The feature vectors were typically the short-time spectra of a speech signal obtained from an analysis via a bank of bandpass filters or through a linear predictive coding (LPC) analysis [88]. The feature vectors of input speech were compared to a set of reference templates using distance measurements such as log spectral deviation, Itakura-Saito distortion, Itakura distortion, model distortion and likelihood ratio distortion [29]. The distance scores of an input template and each reference template were determined. However, due to the fact that speakers did not always have the same speaking rate for the input and reference speech, the result was that the input and reference templates did not have the same length, which made it difficult for the algorithm to perform the pattern comparison. A dynamic time warping (DTW) algorithm proposed by Sakoe and Chiba [21] was employed to solve this problem. The DTW algorithm performed the time alignment and time normalization between the input templates and reference templates and finally, the recognition decision was based on the smallest distance scores obtained from pattern comparison.

Although the DTW technique was very successful in isolated speech recognition, research on speech recognition shifted toward the hidden Markov model (HMM) approach due to some limitations of the DTW [46], [84]. The computational cost of the DTW approach is high with some difficulties in being extended to other tasks such as connected speech and continuous speech. In addition, a more robust parametric model to represent the speech is desired instead of the nonparametric template used by DTW approach. The following section contains a brief review of the HMM technique.

2.1.2 Hidden Markov Model (HMM)

The basic theory of the hidden Markov model (HMM) was published by Buam and Petrie in the mid 1960s, but the methodology of HMM was only well known in a few laboratories (IBM and Dragon system) [88]. The reason that HMM was not known among speech researchers at that time was because the basic theory of the hidden Markov model was published in mathematical journals not generally read by speech researchers [86]. However, after several tutorial papers were published in the mid 1980s, which provided sufficient information for speech researchers to understand the methods and theory of the hidden Markov model, the HMM technique has since been widely applied in speech recognition [39], [40], [70], [82], [83], [84], [85].

The basic idea of HMM relies on the assumption that the speech signal can be well characterized as a parametric random process, and the parameters of the stochastic process can be determined in a precise, well-defined manner. HMM does not require explicit time alignment whereas DTW does. Instead, a probabilistic transition and observation structure were defined for each reference word. The structure of a Markov model included a state transition probability matrix, an initial probability vector, and an observation probability matrix. For each given reference model, the observation probability of the test utterance was computed during the recognition phase. Finally, the model with the maximum observation probability was

selected to be the recognition result. The HMM modeled speech as a stochastic process that transitions through a network of interconnected states. Each of these states represented an important aspect of the speech sound. The HMM structure with 4-state is shown in Figure 2.2.

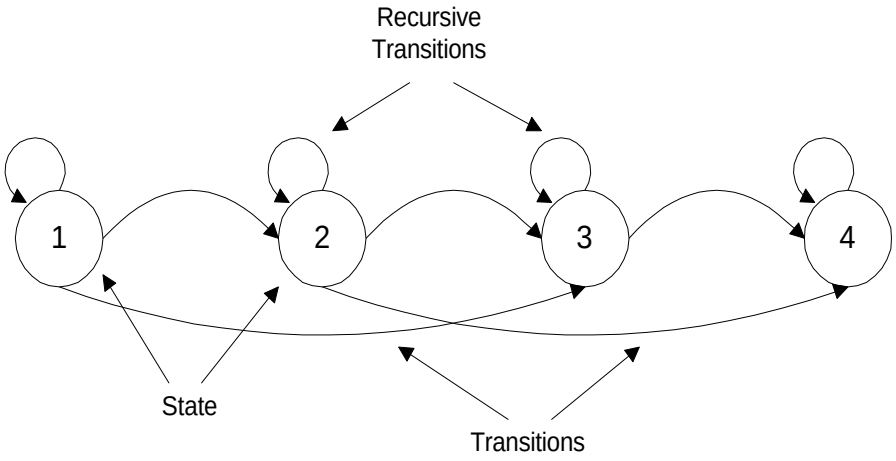

Figure 2.2 A 4-state hidden Markov model structure

HMM gained much popularity among speech researchers since the late 1980s and appeared in much of the literature [18], [34], [35], [47], [52], [57], [58], [63], [90]. The applications of HMM can range from isolated, connected and continuous speech to speaker-dependent and speaker-independent tasks. Many HMM-based speech recognition applications have been conducted and several techniques have been proposed to improve the performance of HMM-based systems [82], [83], [84], [85], [87].

2.1.3 Neural Network (NN)

The neural network approach was quite new in speech recognition. However, the idea of neural networks was not new; it was first introduced by McCulloch, Pitts, and Hebb in the 1940s. The basic idea of neural networks was motivated by a desire to both understand the brain and to emulate some of its strengths. The

neural network models consisted of many nonlinear computational elements or nodes operating in a parallel manner. These nodes were connected via weights that were typically adapted during use to improve performance [56].

Neural networks had the greatest potential in speech recognition problems due to their abilities to perform the high computation rates by simultaneously employing massive parallel nets. In addition, most neural networks had the ability to adapt connection weights in time to improve the performance based on current results. The ability of speech recognition systems to adapt to new speakers, new words and a new environment was of concern to many speech researchers. The ability of neural network to adapt and continue learning made it a suitable technique for speech recognition.

Neural networks were introduced into speech recognition problems in the late 1980s and have remained popular up to the present time. Several types of neural networks, including mutilayer perceptron [51], self-organizing neural network [56], time delay neural network [106], learning vector quantization [19], recurrent neural network [92], and hierarchical neural network [14], have been developed. Since my research study employed multilayer perceptron (MLP) trained by backpropagation method as a tone classifier, the following section provides a brief discussion on the important aspects of multilayer perceptron.

Multilayer perceptron is a feedforward net with one or more layers of nodes between the input and output layer. The additional layers contain hidden units or nodes that are not directly connected to either the input or output nodes. However, the output units and the hidden units may have biases which act like weights on the connections from the unit whose output is always 1. A multilayer perceptron with one hidden layer is depicted in Figure 2.3. The details of a multilayer perceptron trained by standard backpropagation are given as follows.

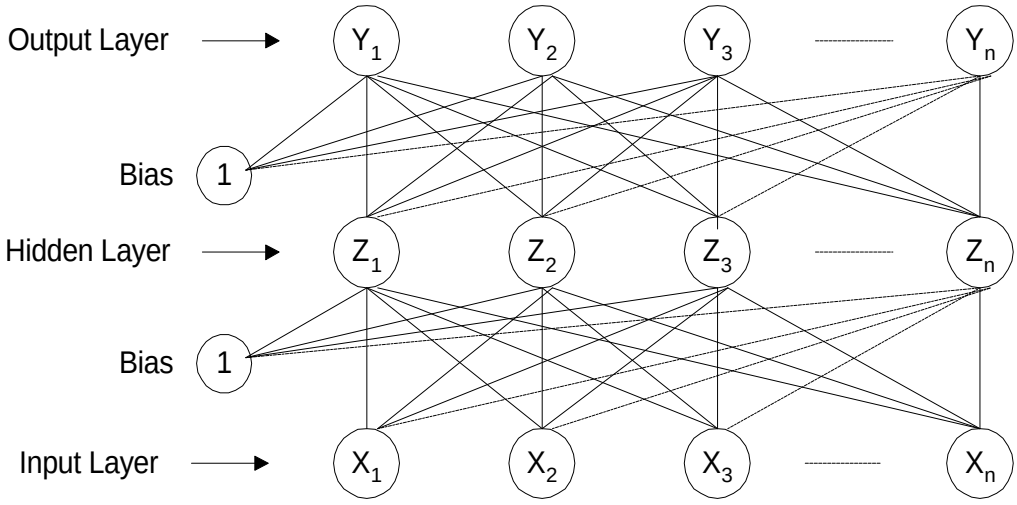

Figure 2.3 Multilayer perceptron with one hidden layer

The training of a multilayer perceptron by standard backpropagation involves three stages: the feedforward of the input training pattern, the backpropagation of associated error, and the adjustment of the weights. However, when the training is completed, the application of the net involves only the computation of the feedforward phase. Even if the training is slow, a trained net can produce its output very rapidly [19].

The standard backpropagation may take too long for the algorithm to converge in some situations. The performances of the standard backpropagation algorithm may be improved by modifying the weight update procedure and using the variable learning rate. The algorithm may converge faster if a momentum term is added during the weight adjustment process. In standard backpropagation, the learning rate is generally held constant throughout the training process where the proper setting of the learning rate has impacted the performance of the algorithm. If the learning rate is set too high, the algorithm may oscillate and become unstable whereas setting too small a learning rate may take the algorithm too long to converge. The adaptive learning rate can improve the performance of

backpropagation by changing the learning rate during the training process, and the learning step is kept as large as possible while keeping learning stable [19].

2.2 Fuzzy Inference System (FIS)

Since the proposed stress detector employs a fuzzy inference system (FIS) to identify a stress degree of the syllable, a brief discussion on the important aspects of fuzzy inference systems is presented in this section.

Fuzzy inference is the process of mapping from a given input to an output based on a set of fuzzy if-then rules. The processes of the fuzzy inference system involve input fuzzification based on input membership functions, implication process for each if-then rule, combining the output of each rule by aggregation process, and defuzzification process to resolve an output value. The fuzzy inference process is given of Figure 2.4.

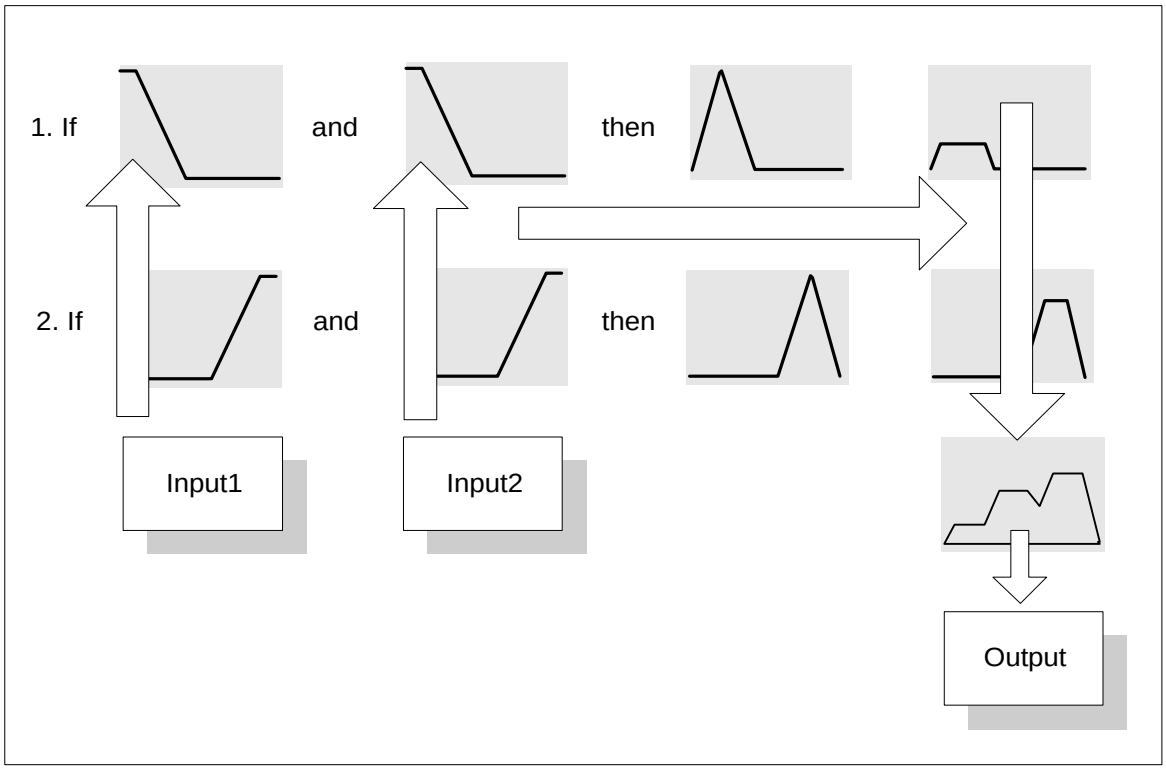

Figure 2.4 Fuzzy inference process

A fuzzy set is completely characterized by its membership function. A membership function is a curve that defines how each point in the input space is mapped to a degree of membership between 0 and 1. Several types of membership functions, including triangular, trapezoidal, and Gaussian, have been widely used. [37]. Due to their simplicity and computational efficiency, both triangular and trapezoidal membership functions have been widely used in many applications. However, the curves of both membership functions are not smooth because they are formed by straight lines. Due to the smoothness of its curve and ability to contain non-zero at all points, the Gaussian membership function is becoming increasingly popular for specifying fuzzy sets.

The fuzzy if-then rule also plays an important role in the fuzzy inference system by being used to formulate the conditional statements that comprise fuzzy logic. A simple fuzzy if-then rule has the form "If *x* is *A* then *y* is *B,*" where *A* and *B* are linguistic values defined by fuzzy sets on the ranges X and Y, respectively. The first part of the fuzzy if-then statement ("*x* is *A*") is called the antecedent whereas the second part ("*y* is *B*") is called the consequent part. The antecedent of a rule defines a fuzzy region in the input space, whereas the consequent part specifies the output in the fuzzy region. Some applications, however, may have more complicated if-then statements wherein their rules may contain a statement that has more than one antecedent part in the if-then statements. In the process of interpreting these if-then rules, all parts of the antecedent are calculated simultaneously and resolved to a single number using the logical operators described in the if-then statement.

The defuzzification process is required in order to produce a crisp value that best represents the output value. Several defuzzification methods have been widely used, including centroid of area, bisector of area, mean of maximum, smallest of maximum, and largest of maximum [37]. The most widely adopted defuzzification method is the centroid of area which provides the center of area under the curve.

2.3 Research on Thai Speech Recognition

The acoustic phonetic study of Thai started in the early 1960s, with the study of Thai vowels and tones by Abramson [1]. However, there have been additional studies on the acoustic aspect of Thai segmentals and suprasegmentals from the 1970s to the present time. The studies of Thai acoustic phonetics included Thai rhythm, Thai intonation, Thai speech pause, Thai non-stop consonants, Thai stop consonants, acoustic characteristics of unstressed syllables in Thai, and acoustic characteristics signaling syllable boundary in Thai-connected speech [59], [69], [100].

Speech analyses were considered the first step in speech technology in Thailand and these studies contributed a strong theoretical background for the development of speech computing and speech technology. However, the first application of a speech computing system was not begun in Thailand until the 1980s. All earlier studies focused mainly on isolated speech, such as digit recognition by Pekanan, [66], and syllable recognition by Prathoomthan, [77] and template matching was the main recognition technique. In the 1990s, numeral and syllable recognition became the topics of interest. Hidden Markov models (HMMs) and neural networks (NNs) were the popular techniques. The following sections briefly report several previous studies.

The 1980s was the decade in which speech recognition was started in Thailand and the template matching method was the main recognition technique. In 1982, the first research on Thai speech recognition system was conducted by Pekanan [66]. His research concentrated on the construction of a digital encoder from a voice signal. The system was evaluated on Thai digit word (0-9), and a recognition rate of 60% was obtained.

In 1986, Prathumthan [77] developed a syllable-based Thai speech recognition system. There are many interesting tokens in his study where the

system failed to recognize the speech. All of them were the unstressed syllables with vowels. He also suggested that the acoustic description of the variation of unstressed syllables still needs to be done for the development of speech recognition systems.

In the 1990s, several research studies on Thai speech recognition were conducted, mainly concentrating on numeral and syllable recognition. Dynamic time warping (DTW), hidden Markov model (HMM), and neural network (NN) were employed as the recognition methods.

Thamphothong [103] developed an isolated speech recognition system in 1990. The proposed system used two parameters: reflection coefficients and pitch values to create the reference patterns. A dynamic programming technique was employed to find the distance between the input signals and the reference templates. The K-nearest neighbor (KNN) technique was applied for the decision task. A recognition rate of 85.8% was obtained for words with initial bilabial stops and nasal, and a recognition rate of 94.3% was obtained in the recognition of randomized words in the second set.

In 1995, Pensiri [67] constructed a Thai numeral voice recognition using the dynamic time warping (DTW) technique. The feature vectors were extracted from isolated words by using a discrete Hartley Transform technique, and then the distance between the test pattern and the reference pattern was calculated. The results indicated that good speech recognition depends on speech parameter selection and suggested that DTW was not suitable for a speech recognition system having so many recognized patterns. Phatrapornnant [68] developed an isolated Thai vowel recognition by using the dynamic time warping (DTW) technique. The 24 Thai vowel reference templates were created from ten speakers. The recognition rate of 84.44% was achieved from an unclassified reference group while the recognition rates of a classified reference group were 90.83%. This system can also

classify five tones of three vowel words (/a:/, /I:/, /u:/) with an accuracy rate of 81.00%. Areepongsa [7] proposed a Thai numeral recognition based on Hidden Markov Model (HMM) and vector quantization. Her study showed the relationship between accuracy and number of training sets. She concluded that the increment of the number of training sets increased the recognition rate of HMM.

In 1996, Pornsukjantra [72] developed a Thai numeral speech recognition using the LPC and a backpropagation neural network. The feature vectors were extracted by linear predictive coding (LPC) and sets of LPC coefficients were used as input vectors to the neural network. The result indicated 89.4% accuracy for one syllable speech and 84.7% accuracy for two and three syllable speech.

In 1997, Ahkuputra *et al*. [4] developed an algorithm for Thai polysyllabic word recognition. This method was based on the discrete hidden Markov model in conjunction with vector quantization. The testing set was comprised of single, double, triple syllables with the last set consisting of ten Thai numeral words, zero to nine. The experimental results showed that the increase in the number of codebooks and the number of model states have a major effect on the recognition rate. The average recognition rate of 89.91% was obtained from the testing set .

In 1998, Wutiwatchai *et al*., [109] developed a Thai polysyllabic word recognition based on neural networks and a fuzzy technique. The fuzzy features converted from actual features were used as the inputs of multilayer perceptron (MLP). The binary desired-outputs were used during training. This system was evaluated on 70 Thai words, and achieved an average recognition rate of 94.4%. Wutiwatchai *et al*., [110] also proposed a fuzzy-neural network for Thai numeral speech recognition. This method employed the fuzzy membership input with conventional binary desired output. The proposed method achieved a recognition rate of 90.8%, compared with a recognition rate of 86.0% obtained from conventional neural networks.

Thubthong and Kijsirikul [105] implemented a syllable-based Thai digit speech recognition system in 1999. The system was based on neural network and duration modeling. The feature vectors consisted of MEL cepstral coefficients and perceptual linear prediction (PLP). The experimental results showed that a nine-frame span of speech is appropriate for containing sizable parts of syllables. The duration modeling with a duration ratio as well as sentence-matching algorithm could improve the recognition performance.

Up to the present time, there are only a few studies on tone classification of Thai speech. Ramalingam [89] developed a tone extraction in isolated Thai speech based on subharmonic summation technique and vector quantization. Hasan [30] implemented tone recognition in isolated Thai speech using hidden Markov model (HMM). Recently, Potisuk [76] proposed a tone classification in syllable-segmented speech based on the analysis by synthesis method. The following section contains a brief report of these studies.

Ramalingam [89] constructed a tone extraction algorithm for isolated Thai syllables in 1995. The effects of tonal coarticulation, stress, and intonation were not taken into consideration in his study. The speech data was collected from four speakers who uttered both training and testing sets. The subharmonic summation technique was employed as a pitch determination algorithm to estimate the pitch contour of the syllable. The pitch contour was given as input to the vector quantizer where a codebook was constructed to contain reference vectors corresponding to each tone. A distortion measure was computed between the test vector and each of the reference vectors. The least amount of distortion corresponding to the reference vector was identified as the tone. His experimental results showed that the system correctly identified 100% of tone recognition for a noise-free environment, and recognition rates of 98, 97, 97, 98, 93 and 93% were obtained for signal to noise ratios of 40 dB, 30 dB, 20 dB, 10 dB, 5 dB, and 2 dB, respectively.

In 1997, Hasan [30] implemented tone recognition of isolated Thai speech based on a semi-continuous hidden Markov model (HMM). The effects of tonal coarticulation, stress, and intonation were not taken into consideration in his study. His research used two male and two female speakers to utter six isolated syllables for training the reference model. The testing results showed recognition accuracy of 97.3%, 99%, and 97.8%, respectively for three, four, and five-state hidden Markov model.

In 1999, the latest research on Thai tone classification was conducted by Potisuk *et al*., [76]. The analysis by synthesis method, which was an extension of Fujisaki's model, was proposed to classify Thai tone sequences in syllable-segmented speech. The autocorrelation with a three-level center clipping method was employed as a pitch detector. In their studies, the effects of tonal coarticulation and intonation were taken into account but used only the stressed syllables in the training and testing utterances. Both training and testing sets of utterances were spoken by the same five speakers.

The algorithm consisted of two modules, analysis and synthesis modules. First, the analysis module generates hypothesized tone sequences and then the synthesis module generates predicted *F0* contour to match against the input contour. A classification test was performed on a set of 55 test utterances consisting of 11 sentences with varying tone sequences. Potisuk reported that the algorithm incorrectly identified six of the 55 test utterances and the proposed tone classification achieved the recognition rate of 89.1%.

Potisuk suggested several further work on tone classification of Thai speech, e.g., a method to perform segmentation of Thai utterances into syllable units needed to be developed, and the training set needed to be expanded to cover the effects of stressed and unstressed syllables.

2.4 Research on Other Tonal Languages

The existing research on tonal languages focused mostly on Mandarin and Cantonese. Although tone recognition is a mandatory part of tonal speech recognition, it was found that most research concentrated on syllable recognition, with only a few studies focusing on tone recognition separately. The hidden Markov model (HMM) and neural network (NN) have been mainly applied to Mandarin and Cantonese speech recognition [48], [49], [50], [51], [57], [107], [111]. The following section presents several research studies with special emphasis on tone recognition of Mandarin and Cantonese speech.

The first tone recognition in isolated Mandarin speech was developed by Yang [111] in 1988. The proposed system was based on a combination of hidden Markov model (HMM) and vector quantization. A logarithmic pitch interval and its first derivative were used as parameters. The speech database was provided by seven male and seven female speakers. This system can achieve an accuracy of 96.53%.

In 1993, Lee *et al*. [49] implemented a real-time Mandarin dictation machine which recognizes Mandarin speech with a very large vocabulary. The dictation machine was speaker-dependent and sequences of isolated syllables were used as input speech. The machine consisted of two subsystems. The hidden Markov model technique was used in the first subsystem to recognize the 408 very confusing syllables (disregarding the tones) and five different tones were classified by special feature vectors. Then the second subsystem identified the exact characters from the syllables and corrected the errors in syllable recognition by finding all possible word hypotheses and forming a word lattice through a lexical access process. The output sentence was obtained by determining the best path in the lattice with the maximum likelihood. This machine took only about 0.45 second to dictate a syllable and achieved a recognition rate on the order of 90%.

In 1995, tone recognition of isolated Mandarin monosyllables based on multilayer perceptron (MLP) was proposed by Chang, Sun and Chen [11]. The back propagation learning rule was employed to train the MLP. The feature vectors consisted of energy, pitch mean, and pitch slope. The recognition rate of 93.8% was obtained in this study. The first tone recognition of isolated Cantonese syllables was implemented by Lee *et al*. [50]. Cantonese is a commonly used dialect in Southern China. The system was based on multilayer feedforward neural network. Four normalized parameters were used as the input to the neural network: duration, energy drop rate, initial pitch and final pitch. They reported that the proposed system was able to achieve the recognition rate of 89.4%.

In 1997, Wang *et al*. [107] reported the recognition of Mandarin speech with a very large vocabulary using limited training data. The total number of phonologically allowed different syllables is 1345 but when the differences among the syllables caused by tones are disregarded, the total number of syllables is reduced to only 416 base syllables. This scheme implied that it is helpful to recognize tones and base syllables separately. The hidden Markov model (HMM) was employed as a recognition method in this study. The HMM obtained from the training data was used as the initial models and further trained by the modified segmental K-means algorithm, in which the HMM after each iteration were linearly interpolated with the initial HMM. The proposed scheme has been successfully implemented on a Sparc 20 workstation, namely, Golden Mandarin (III) 3.0 workstation version. In practical use, the recognition error can be corrected either by the user choosing the desired tonal syllables or characters from the candidate lists shown in a window using the mouse, or directly by voice. The overall time required to recognize an utterance on the prototype machine was about 1.3 times of duration of speech utterance.

In 1998, Chen and Lio [13] proposed a modular recurrent neural network (MRNN)-based speech recognition method that recognized the entire vocabulary of 1280 highly confusing Mandarin syllables. This approach first split the complicated

tasks into several subtasks involving subsyllable and tone recognition, and then used two weighting RNNs to generate several dynamic weighting functions to integrate the subsolutions into a complete solution. A priori linguistic knowledge of structures of Mandarin syllables was used in the architecture design of the MRNN. The experimental results showed that the MRNN outperformed the HMM method.

In 1999, the Cantonese syllable recognition system based on neural networks was conducted by Lee and Ching [51]. This scheme was divided into two parts; tone recognition and base-syllable recognition. Tone recognition was considered as a static pattern recognition problem. It made use of a set of suprasegmental features to describe the pitch and energy profiles of the syllable. A multilayer perceptron (MLP) was employed to classify the normalized feature parameters in a tone recognition task and each output neuron represents a particular tone. For a base-syllable recognition task, each Cantonese base syllable was represented by a dedicatedly trained RNN and a multipass selection by elimination algorithm was developed to determine the most likely base syllable model. The recognized output was based on N-best outputs of the two sub-recognizer. The proposed scheme was evaluated on speaker-dependent with 40-syllable vocabulary and expanded progressively to 200-syllable vocabulary. In a 200-syllable task, a top-1 recognition rate of 81.8% was obtained, whereas the top-3 recognition rate was 95.2%.

Chapter 3 Analysis of Thai Tones

In this chapter, the method of collecting speech data and the analysis of Thai tones involving the stressed and unstressed syllables, intonation and tonal coarticulation are presented. It begins with a discussion of the method of collecting speech data, and then is followed by the analysis of Thai tones.

3.1 Speech Data

The input speech of both training and test sets was uttered by five male and three female speakers. All speakers are Thai natives and have a standard Thai dialect. Speakers range between 20 and 40 years of age. Several details of each speaker are listed in the Table 3.1.

Speaker	Gender	Age
SPK1	Male	24
SPK2	Male	27
SPK3	Male	27
SPK4	Male	26
SPK5	Male	36
SPK6	Female	24
SPK7	Female	35
SPK8	Female	30

Table 3.1 Details of speakers who participated in the preparation of speech data

The input speech of each speaker was recorded in a quiet environment with all speakers uttering two training sets and one test set. To complete a training set, each speaker uttered 100 sentences, each of which consisted of four monosyllabic words. The sentence structure of each training sentence was “subject + verb + object + post-verb auxiliary.” The first and last word of each sentence varied across

all five tones while the middle two words were varied to give all 25 two-tone combinations. The stress pattern of the two middle words was variant to cover all possible stress patterns: unstressed-unstressed, unstressed-stressed, stressed-unstressed, and stressed-stressed, as listed in Table 3.2. For a test set, each speaker uttered 115 sentences having the same structure as the training sets. The two middle words of the test sentences contained all possible stress patterns as described in the training sets. Although speakers were asked to utter the two middle words of the sentences with the stress patterns described above, in actuality, some speakers spoke all four stressed monosyllabic words for stressed-stressed pattern, and some speakers uttered all four unstressed monosyllabic words for unstressed-unstressed pattern. The details of the training and test sentences are listed in Appendix B.

The input speech was recorded and stored in a personal computer using a sound blaster card and a headset microphone. The recorded input speech had a ".wav" format. The operation details used for conversion of the data to digital format in the sound card were as follows:

Number of bits per sample :	16
Sampling rate :	11.025 kHz

According to the sampling theorem, the sampling rate is required to be greater than twice the highest signal frequency to permit exact signal recovery [64]. The typical range of human frequency is between 0 to 3.3 KHz, for which the sampling rate of 8 KHz or higher should be appropriate; however, most speech processing and speech recognition applications use a sampling rate of 10 KHz or higher [11], [12], [13], [14], [76], [78], [89].

Stress Patterns			
Unstressed-Unstressed	Unstressed-Stressed	Stressed-Unstressed	Stressed-Stressed
1. $w_1U^0U^0w_4$	1. $w_1U^0S^0w_4$	1. $w_1S^0U^0w_4$	1. $w_1S^0S^0w_4$
2. $w_1U^0U^1w_4$	2. $w_1U^0S^1w_4$	2. $w_1S^0U^1w_4$	2. $w_1S^0S^1w_4$
3. $w_1U^0U^2w_4$	3. $w_1U^0S^2w_4$	3. $w_1S^0U^2w_4$	3. $w_1S^0S^2w_4$
4. $w_1U^0U^3w_4$	4. $w_1U^0S^3w_4$	4. $w_1S^0U^3w_4$	4. $w_1S^0S^3w_4$
5. $w_1U^0U^4w_4$	5. $w_1U^0S^4w_4$	5. $w_1S^0U^4w_4$	5. $w_1S^0S^4w_4$
6. $w_1U^1U^0w_4$	6. $w_1U^1S^0w_4$	6. $w_1S^1U^0w_4$	6. $w_1S^1S^0w_4$
7. $w_1U^1U^1w_4$	7. $w_1U^1S^1w_4$	7. $w_1S^1U^1w_4$	7. $w_1S^1S^1w_4$
8. $w_1U^1U^2w_4$	8. $w_1U^1S^2w_4$	8. $w_1S^1U^2w_4$	8. $w_1S^1S^2w_4$
9. $w_1U^1U^3w_4$	9. $w_1U^1S^3w_4$	9. $w_1S^1U^3w_4$	9. $w_1S^1S^3w_4$
10. $w_1U^1U^4w_4$	10. $w_1U^1S^4w_4$	10. $w_1S^1U^4w_4$	10. $w_1S^1S^4w_4$
11. $w_1U^2U^0w_4$	11. $w_1U^2S^0w_4$	11. $w_1S^2U^0w_4$	11. $w_1S^2S^0w_4$
12. $w_1U^2U^1w_4$	12. $w_1U^2S^1w_4$	12. $w_1S^2U^1w_4$	12. $w_1S^2S^1w_4$
13. $w_1U^2U^2w_4$	13. $w_1U^2S^2w_4$	13. $w_1S^2U^2w_4$	13. $w_1S^2S^2w_4$
14. $w_1U^2U^3w_4$	14. $w_1U^2S^3w_4$	14. $w_1S^2U^3w_4$	14. $w_1S^2S^3w_4$
15. $w_1U^2U^4w_4$	15. $w_1U^2S^4w_4$	15. $w_1S^2U^4w_4$	15. $w_1S^2S^4w_4$
16. $w_1U^3U^0w_4$	16. $w_1U^3S^0w_4$	16. $w_1S^3U^0w_4$	16. $w_1S^3S^0w_4$
17. $w_1U^3U^1w_4$	17. $w_1U^3S^1w_4$	17. $w_1S^3U^1w_4$	17. $w_1S^3S^1w_4$
18. $w_1U^3U^2w_4$	18. $w_1U^3S^2w_4$	18. $w_1S^3U^2w_4$	18. $w_1S^3S^2w_4$
19. $w_1U^3U^3w_4$	19. $w_1U^3S^3w_4$	19. $w_1S^3U^3w_4$	19. $w_1S^3S^3w_4$
20. $w_1U^3U^4w_4$	20. $w_1U^3S^4w_4$	20. $w_1S^3U^4w_4$	20. $w_1S^3S^4w_4$
21. $w_1U^4U^0w_4$	21. $w_1U^4S^0w_4$	21. $w_1S^4U^0w_4$	21. $w_1S^4S^0w_4$
22. $w_1U^4U^1w_4$	22. $w_1U^4S^1w_4$	22. $w_1S^4U^1w_4$	22. $w_1S^4S^1w_4$
23. $w_1U^4U^2w_4$	23. $w_1U^4S^2w_4$	23. $w_1S^4U^2w_4$	23. $w_1S^4S^2w_4$
24. $w_1U^4U^3w_4$	24. $w_1U^4S^3w_4$	24. $w_1S^4U^3w_4$	24. $w_1S^4S^3w_4$
25. $w_1U^4U^4w_4$	25. $w_1U^4S^4w_4$	25. $w_1S^4U^4w_4$	25. $w_1S^4S^4w_4$

Table 3.2 shows the structure of training and test sentences with their stress patterns of the two middle words. The superscripts /^{0}/, /^{1}/, /^{2}/, /^{3}/ and /^{4}/ are the tone markers of the mid, low, falling, high and rising tones, respectively. The /w_1/ is the beginning syllable whereas /w_4/ is the ending syllable of the sentence. The letter /S/ and /U/ are the stressed and unstressed syllable, respectively.

3.1.1 Statistical Data

In this section, the acoustical features, including duration, energy, and fundamental frequency (*F0*) of stressed and unstressed syllables are first presented. The effect of intonation in Thai tones is then discussed, followed by a discussion of the effects of tonal coarticulation. The statistical data analysis presented in this section is computed from the training sets uttered by all speakers.

3.1.1.1 Stressed and Unstressed Syllables

The mean duration of stressed and unstressed syllables for each Thai tone (mid, low, falling, high, and rising) is expressed in Table 3.3. It is noted that the duration of stressed and unstressed syllables was measured using the two middle words of each training sentence.

Speaker	Mean Duration (msec)									
	Stressed Syllable					Unstressed Syllable				
	Mid	Low	Falling	High	Rising	Mid	Low	Falling	High	Rising
SPK1	625.5	563	565	528	588	428	381.5	444.5	295	430
SPK2	570.5	552	571.5	562	636.5	468.5	410.5	465	335	439.5
SPK3	617.5	576	558	534.5	598	397.5	369	412	245.5	379
SPK4	608	526.5	564	568.5	594.5	364.5	321	399.5	296	395.5
SPK5	618	572	563	518	553.5	372.5	342	381	278	360
SPK6	582	552	546	544	532	439	375	444	278	372
SPK7	578	586	532	541	569.5	381.5	358.5	387	290.4	390.5
SPK8	622	627	604	633	679	398	340	422	374	421
Average	602.7	569.3	562.9	553.6	593.9	406.2	362.2	419.4	293.9	398.4

Table 3.3 Mean duration of the stressed and unstressed syllables for each speaker

To evaluate the effectiveness of duration for signaling the distinction between stressed and unstressed syllables, analyses of variance (ANOVA) were conducted on the average duration of each tone in both stressed and unstressed syllables. The results of ANOVA analysis are expressed in Table 3.4 and Figure 3.1.

ANOVA analysis of duration					
Source	SS	df	MS	*F*	Prob>F
Groups	105935.6	1	105935.6	69.72	0.000032
Error	12155.8	8	1519.5		
Total	118091.4	9			

Table 3.4 ANOVA analysis of duration in stressed and unstressed syllables.

It should be noted that the first, second, third, fourth, fifth, and sixth columns of the ANOVA table show the source of the variability, the sum of squares (SS) due to each source, the degree of freedom (df) associated with each source, the mean squares (MS) for each source, the *F* statistic, and the p-value which was derived from the cumulative distribution function of *F*, respectively.

The very small p-value of 0.000032 expressed in the sixth column of Table 3.4 indicates that the differences between the duration of unstressed and stressed syllables are highly significant. The stressed syllables of all five tones are significantly longer than are the unstressed syllables. As seen in Figure 3.1, the box plot confirms this fact graphically and shows the major role of duration in signaling the stress effects. From these results, it is concluded that duration is a significant feature for separating the stressed syllable from the unstressed ones.

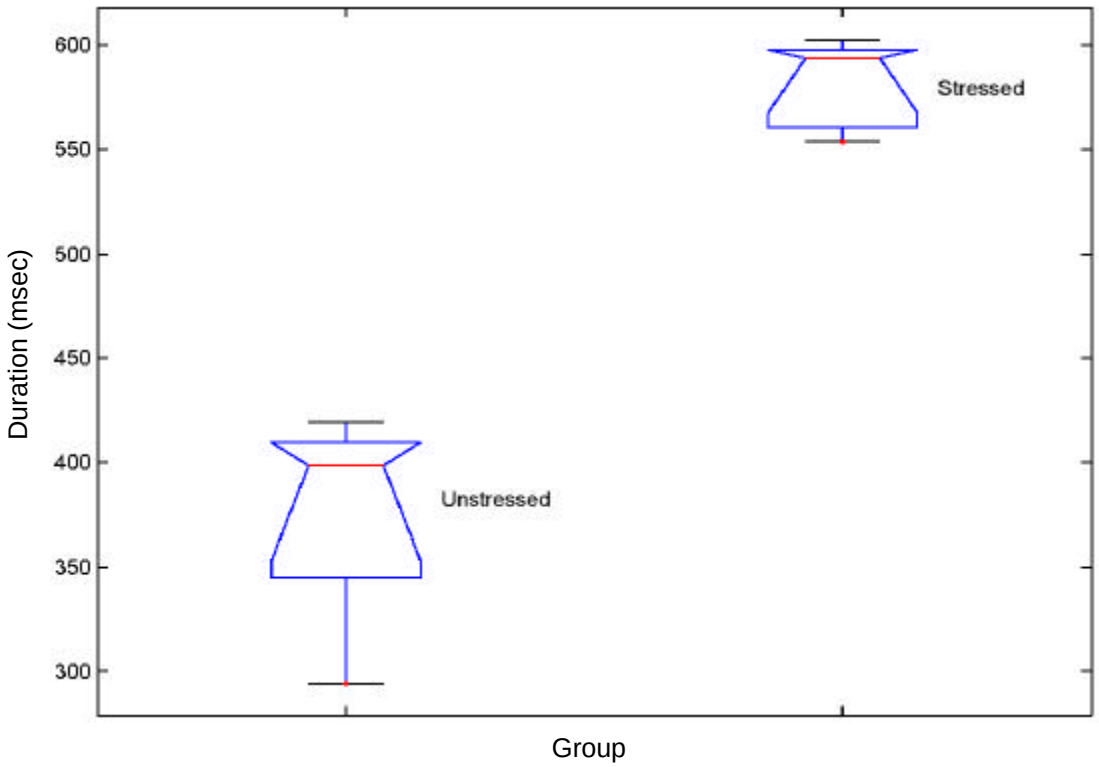

Figure 3.1 A comparison of mean duration in stressed and unstressed syllables

The normalized energy of the stressed and unstressed syllables was also computed in this study. To compute the normalized energy, the speech signal was first blocked into frames, after which an energy calculation of each frame was performed by computing the sum of the squares of the sample data in each frame. The energy of each frame was then divided by the maximum energy of the energy contour. The details of the normalized energy computation appear in Section 4.1.1 of Chapter Four. The mean normalized energy in stressed and unstressed syllables for each speaker are expressed in Table 3.5.

Speaker	Mean Normalized Energy									
	Stressed Syllable					Unstressed Syllable				
	Mid	Low	Falling	High	Rising	Mid	Low	Falling	High	Rising
SPK1	12.83	9.58	11.84	14.90	14.78	6.93	7.89	8.58	7.14	8.28
SPK2	13.49	13.49	14.33	15.61	16.77	6.25	9.66	7.48	6.19	8.05
SPK3	13.12	9.69	11.96	15.33	13.44	6.47	8.07	8.39	6.03	6.44
SPK4	12.38	11.48	14.47	17.98	12.49	6.21	6.16	7.53	6.86	5.76
SPK5	14.37	13.21	14.18	15.72	12.76	5.75	7.57	6.79	5.96	5.01
SPK6	14.99	11.10	14.74	16.01	13.48	6.94	6.53	8.35	5.38	7.64
SPK7	11.48	12.21	13.99	15.18	12.98	5.67	5.72	7.21	6.52	5.22
SPK8	15.27	11.99	16.65	18.92	16.69	6.88	6.97	8.26	9.45	7.15
Average	13.49	11.59	14.06	16.21	14.17	6.39	7.32	7.82	6.69	6.69

Table 3.5 Mean normalized energy of stressed and unstressed syllables

The average normalized energy of all five tones in stressed and unstressed syllables was used in the ANOVA analysis for evaluating the effectiveness of normalized energy to signal the distinction between stressed and unstressed syllables. The results of ANOVA analysis are expressed in Table 3.6 and Figure 3.2.

ANOVA analysis of normalized energy					
Source	SS	df	MS	F	Prob>F
Groups	119.785	1	119.785	78.06	0.000021
Error	12.276	8	1.535		
Total	132.061	9			

Table 3.6 ANOVA analysis of normalized energy in stressed and unstressed syllables.

It is consistent with the duration, the ANOVA analysis yields the very small p-value of 0.000021 which indicates that differences between the energy of the stressed and unstressed syllables are highly significant. The normalized energy in stressed syllables is significantly higher than unstressed syllables for all five tones as displayed in Figure 3.2. It is concluded that energy is also a significant feature in being able to discriminate between the stressed syllables and unstressed ones.

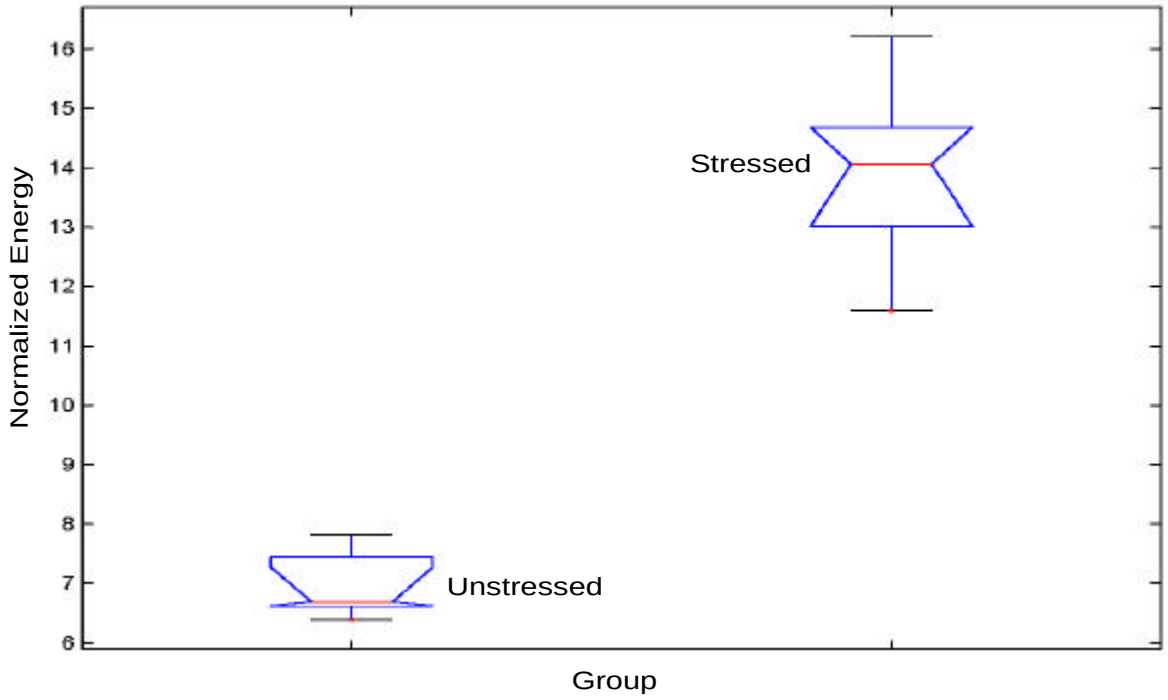

Figure 3.2 A comparison of normalized energy in stressed and unstressed syllables

To investigate whether or not fundamental frequency (*F0)* plays an important role in signaling the stress of Thai syllables, the *F0* of stressed and unstressed syllables was also computed. The *F0* extraction was performed by using the autocorrelation with three-level center clipping method (see details in Section 4.1.1). The extracted *F0* was then normalized by using a z-score transformation in order to eliminate the undesirable time and speaker variation (see details in Section 4.1.3.1). The z-score transformation expressed an observed data value as a multiple of a measure of dispersion away from a mean value as defined in (3.1).

$$Z = \frac{X - \overline{X}}{s} \quad (3.1)$$

where X is raw *F0* and s is a standard deviation about the mean ($\overline{X}$). The mean and standard deviation of *F0* were pre-computed from the training speech uttered by each speaker. The means normalized *F0* in stressed and unstressed syllables for each speaker are shown in Table 3.7.

Speaker	Mean Normalized *F0* (z-score)									
	Stressed Syllable					Unstressed Syllable				
	Mid	Low	Falling	High	Rising	Mid	Low	Falling	High	Rising
SPK1	-0.12	-0.75	1.19	0.36	-0.52	-0.21	-0.89	1.15	0.28	-0.74
SPK2	-0.17	-0.95	1.21	0.44	-0.21	-0.39	-0.93	0.93	0.34	-0.43
SPK3	-0.06	-0.78	1.74	0.71	-0.29	-0.41	-1.05	1.36	0.48	-0.56
SPK4	-0.24	-0.77	1.43	0.61	-0.69	-0.22	-0.87	1.37	0.16	-0.73
SPK5	-0.23	-1.16	1.16	0.29	-0.65	-0.68	-1.36	0.74	0.28	-0.88
SPK6	-0.22	-0.88	1.67	0.41	-0.86	-0.04	-0.67	1.52	0.11	-0.89
SPK7	0.12	-0.63	1.85	0.35	-0.46	-0.08	-0.98	1.63	0.27	-0.71
SPK8	-0.22	-0.98	1.44	0.79	-0.85	-0.22	-0.56	1.43	0.18	-0.82
Average	-0.14	-0.86	1.46	0.49	-0.57	-0.28	-0.91	1.27	0.26	-0.72

Table 3.7 Means normalized *F0* of the stressed and unstressed syllables for each tone

In Figure 3.3, the average normalized *F0* values are displayed for each tone in stressed and unstressed syllables. It is found that the low tone has the lowest mean normalized *F0* in stressed and unstressed syllables whereas the falling tone has the highest mean normalized *F0* in stressed and unstressed syllables. Within each tone, the normalized *F0* of the unstressed syllable is lower than the stressed syllables. However, the normalized *F0* is not considered as a significant feature for distinguishing between stressed syllables and unstressed ones because the normalized *F0* of stressed syllables is higher than unstressed syllable only within the

same tone but when compared with other tones, the normalized *F0* of unstressed syllables might be higher than the stressed ones. For example, the normalized *F0* of the falling tone in unstressed syllables is higher than the normalized *F0* of the high tone in the stressed syllables.

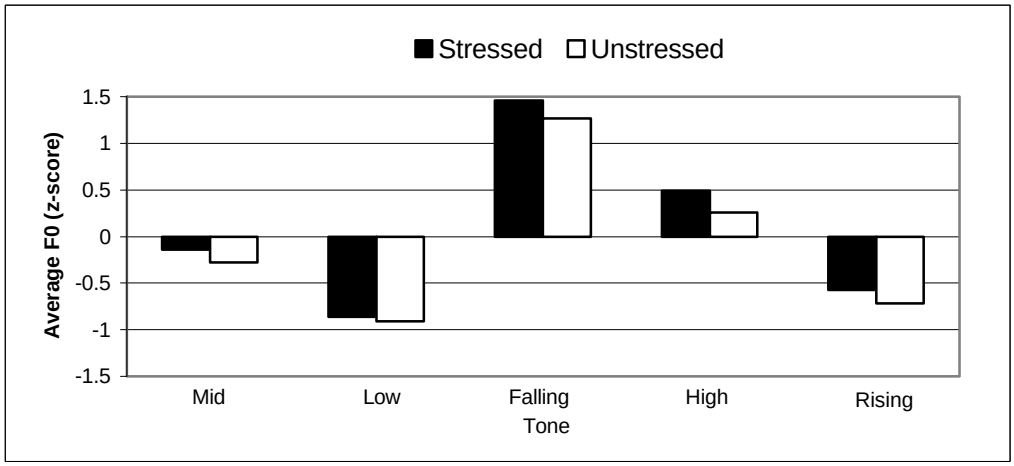

Figure 3.3 Average normalized *F0* of stressed and unstressed syllables for each tone.

To investigate the conclusion stated previously that normalized *F0* is not a significant feature to distinguish between the stressed and unstressed syllables, the ANOVA analysis was conducted on the average normalized *F0* of each tone in stressed and unstressed syllables. The results are expressed in Table 3.8 and Figure 3.4.

ANOVA analysis of mean normalized *F0*					
Source	SS	df	MS	F	Prob>F
Columns	0.05776	1	0.05776	0.07	0.7966
Error	6.50344	8	0.81293		
Total	6.5612	9			

Table 3.8 ANOVA analysis of normalized *F0* in stressed and unstressed syllables.

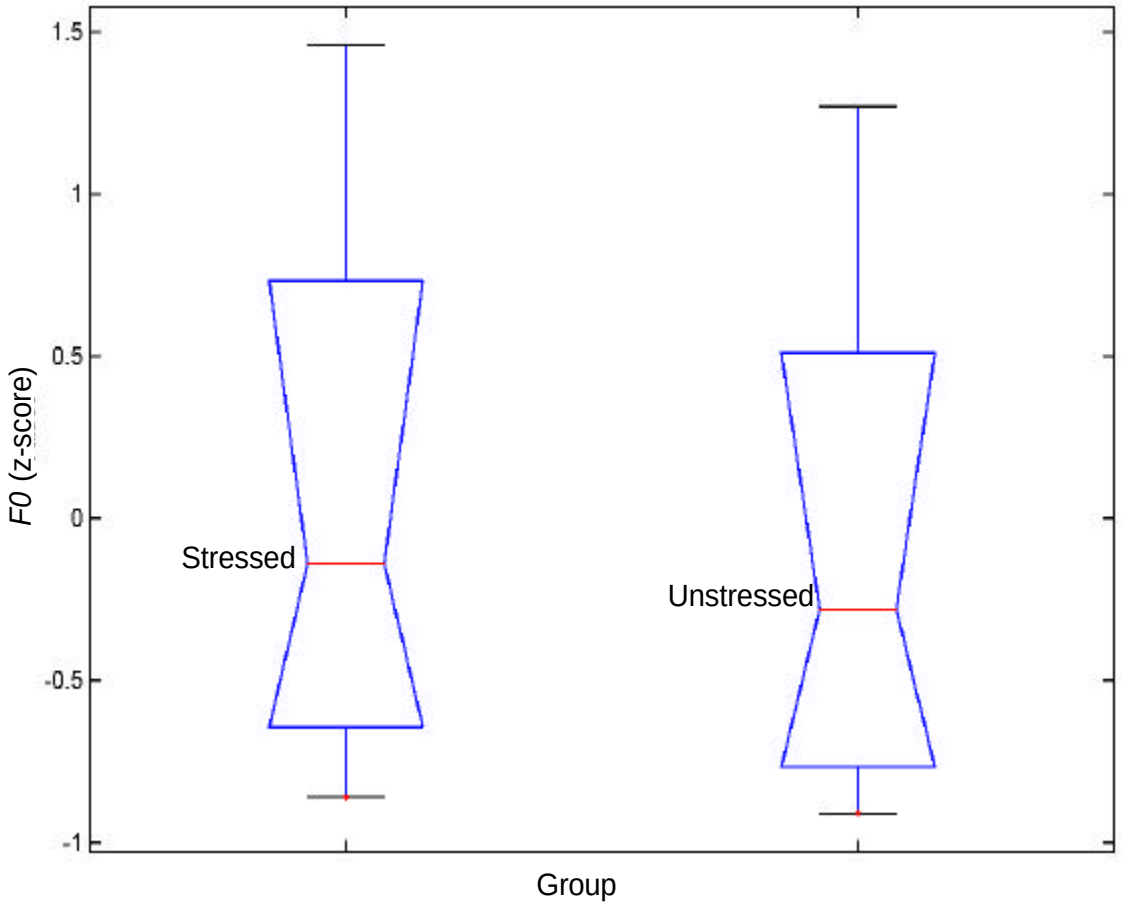

Figure 3.4 A comparison of normalized *F0* in stressed and unstressed syllables

In contrast to the duration and normalized energy, the ANOVA analysis yielded a high p-value of 0.79 in Table 3.8 which indicates that the differences between means normalized *F0* in stressed and unstressed syllables are not significant. It is clearly seen in Figure 3.4 that there are not significant differences between the normalized *F0* of unstressed and stressed syllables. Therefore, it is concluded that the normalized *F0* is not an effective feature to distinguish between the stressed and unstressed syllables.

To investigate the impact of the changes in *F0* contours among Thai tones, the average *F0* contours for the five Thai tones in stressed and unstressed syllables are displayed in Figure 3.5. As seen in the figure, it is evident that the significant

differences in *F0* contours shape between stressed and unstressed syllables appear in the falling, high, and rising tones only. The ending portion of a falling tone of an unstressed syllable was higher than the ending portion of a falling tone of a stressed syllable. The ending portion of the high tone of unstressed syllable was lower than the ending portion of the high tone of stressed syllables. The ending portion of the rising tone of unstressed syllables was lower than the ending portion of the rising tone of stressed syllables. For mid and low tones, their *F0* contours do not have significant differences between stressed and unstressed syllables.

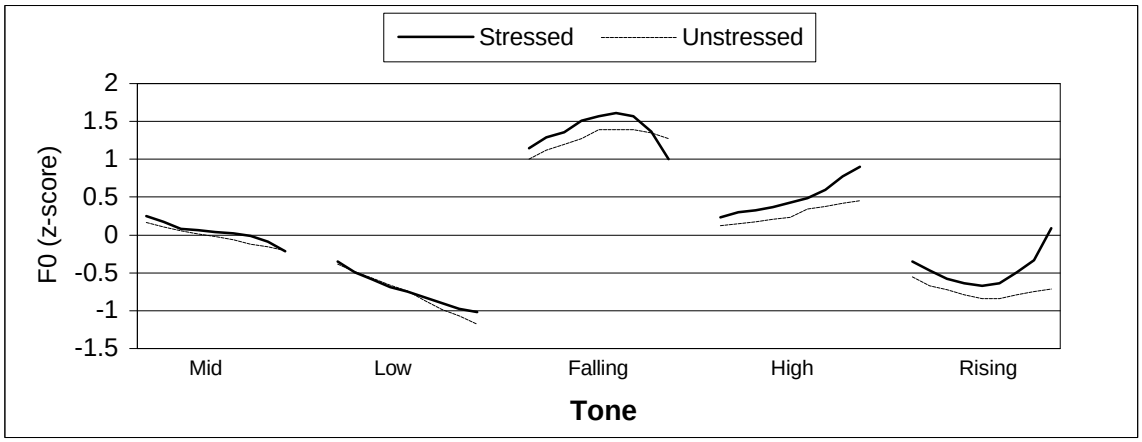

Figure 3.5 *F0* contours of stressed and unstressed syllable.

3.1.1.2 Intonation

The important characteristic of intonation in Thai is the gradual declination of the *F0* contour. To evaluate the effects of intonation of all five Thai tones, the average *F0* contours of the syllable at each location in the sentence are displayed in Figure 3.6.

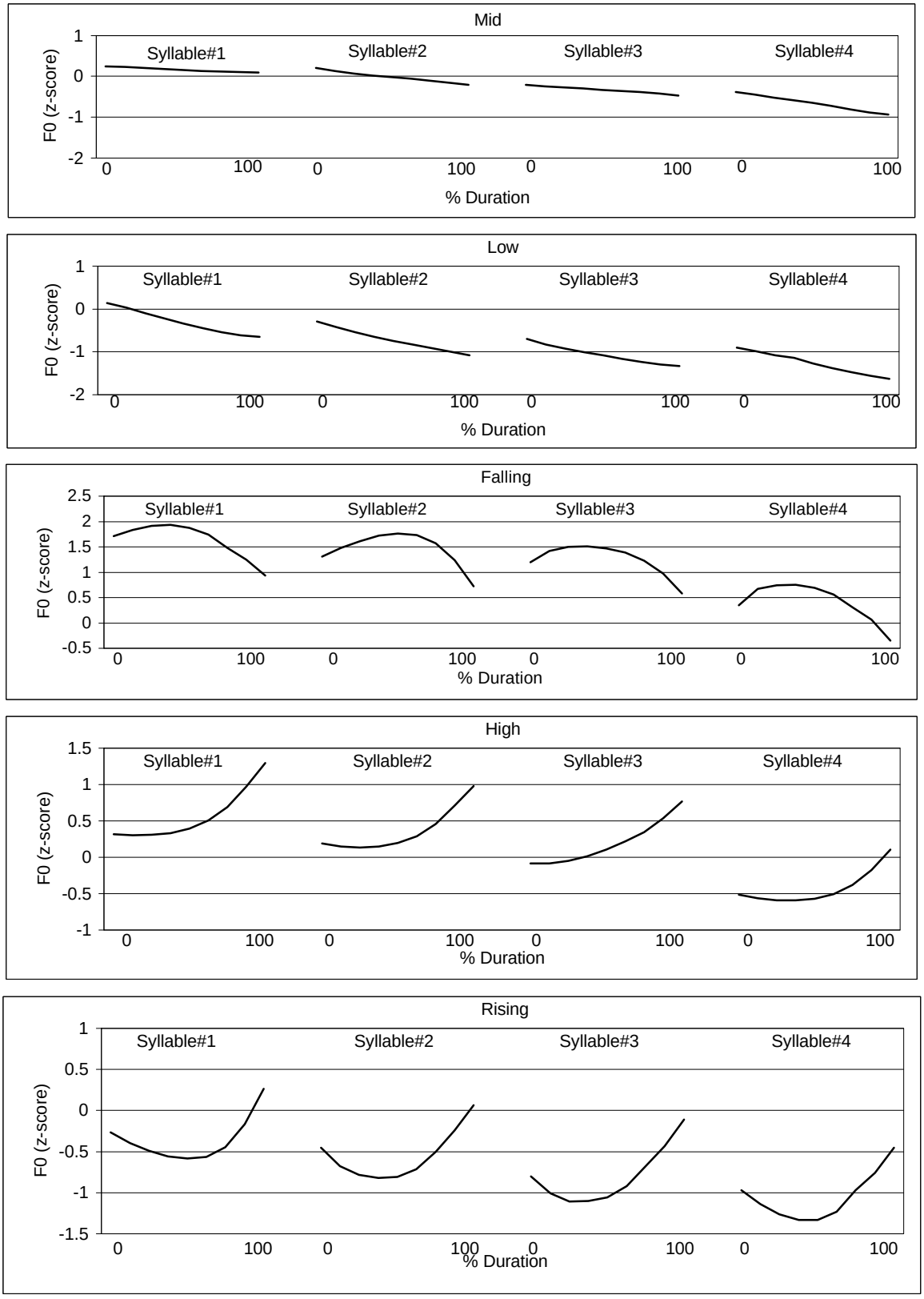

Figure 3.6 Average *F0* contours of each tone at different location in the sentence.

It is noted that the syllable#1, syllable#2, syllable#3, and syllable#4 are the first, second, third, and fourth syllable of the sentence, respectively.

As seen in Figure 3.6, the declination effects made the *F0* contours of the succeeding syllables lower than the preceding syllables for all five tones. *F0* height of the ending syllable (syllable#4) is much lower than *F0* height of the beginning syllable (syllable#1). Means *F0* of each tone at each location in the sentence for each speaker as expressed in Table 3.9.

Speaker	Tone	Syllable #1	Syllable #2	Syllable #3	Syllable #4
		Mean *F0* (z-score)	Mean *F0* (z-score)	Mean *F0* (z-score)	Mean *F0* (z-score)
SPK1	Mid	0.27	-0.05	-0.27	-0.76
	Low	-0.36	-0.67	-0.97	-1.23
	Falling	1.47	1.33	1.01	0.26
	High	0.61	0.39	0.23	-0.47
	Rising	-0.34	-0.51	-0.76	-1.08
SPK2	Mid	-0.16	-0.08	-0.49	-0.69
	Low	-0.41	-0.82	-1.06	-1.14
	Falling	0.98	1.16	0.97	0.21
	High	0.33	0.31	0.17	-0.23
	Rising	-0.51	-0.13	-0.52	-0.53
SPK3	Mid	0.16	-0.14	-0.33	-0.72
	Low	-0.24	-0.73	-0.99	-1.21
	Falling	1.46	1.76	1.34	0.83
	High	0.74	0.52	0.37	-0.18
	Rising	-0.31	-0.39	-0.46	-0.86
SPK4	Mid	0.31	-0.02	-0.44	-0.72
	Low	-0.09	-0.58	-0.87	-0.95
	Falling	1.87	1.45	1.34	0.16
	High	0.96	0.24	0.53	-0.64
	Rising	-0.22	-0.64	-0.77	-1.24

SPK5	Mid	-0.29	-0.34	-0.57	-0.9
	Low	-0.53	-1.27	-1.25	-1.67
	Falling	1.23	0.95	0.95	0.15
	High	0.29	0.45	0.12	-0.32
	Rising	-0.81	-0.57	-0.97	-1.11
SPK6	Mid	0.21	0.04	-0.31	-0.63
	Low	-0.41	-0.57	-0.98	-0.85
	Falling	1.96	1.71	1.48	0.29
	High	0.48	0.33	0.11	-0.55
	Rising	-0.63	-0.69	-1.06	-1.11
SPK7	Mid	0.41	0.23	-0.03	-0.39
	Low	-0.22	-0.57	-0.83	-1.04
	Falling	1.97	1.81	1.68	0.63
	High	0.57	0.28	0.14	-0.46
	Rising	-0.46	-0.28	-0.88	-0.62
SPK8	Mid	0.41	0.24	-0.23	-0.45
	Low	-0.18	-0.51	-0.77	-0.97
	Falling	1.81	1.53	1.33	0.74
	High	0.55	0.36	0.41	-0.48
	Rising	-0.37	-0.79	-0.87	-1.21

Table 3.9 Mean *F0* of each tone for each syllable location

The comparison of mean *F0* of each tone in each syllable location is given in Figure 3.7. As seen in the figure, within each tone, it is found that the means *F0* of the first syllable are higher than the second syllable's means *F0* while the second syllable's means *F0* are higher than the third syllable's, and so on. From these results, it is suggested that mean *F0* within each tone can be used to deal with the intonation effect.

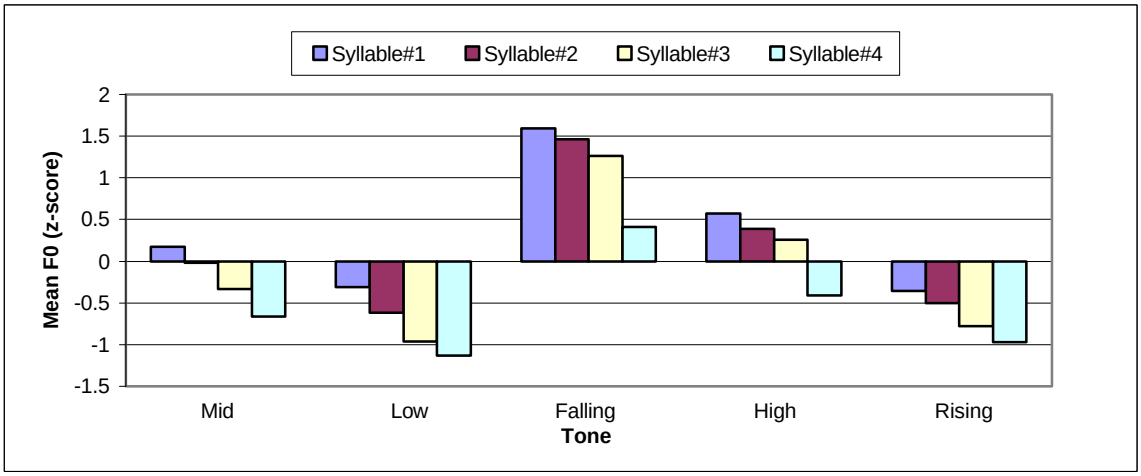

Figure 3.7 Mean *F0* of each tone for the first, second, third, and fourth syllable of the sentence.

3.1.1.3 Tonal coarticulation

There are two types of tonal coarticulation; anticipatory and carryover coarticulation. The *F0* contour of a syllable affected by the succeeding syllable is referred to as "anticipatory coarticulation" while carryover coarticulation occurs when the *F0* contour of preceding syllable influences the succeeding one. To evaluate the anticipatory and carryover coarticulation effects in stressed and unstressed syllables, the normalized *F0* contours of the two middle words of the training sentences are displayed in Figure 3.8, 3.9, 3.10, and 3.11. It is noted that the left parts of the panels in Figure 3.8, 3.9, 3.10 and 3.11 represent the normalized *F0* (z-score) of the first syllable of the two middle words whereas the right parts of panels represent the normalized *F0* of the second syllable of the two middle words of the sentences. The phonemic transcription uses the diacritics /M/, /L/, /F/, /H/ and /R/ as tone markers for the mid, low, falling, high and rising tones, respectively. For example, MM, ML, MF, MH, and MR represent the two-tone sequences of mid-mid, mid-low, mid-falling, mid-high, and mid-rising, respectively.

Figure 3.8 shows the anticipatory coarticulation effects in stressed syllables. The mid tone was significantly greater in height when followed by the rising or high tone than when followed by the mid or low tone. The falling tone was significantly greater in height when followed by the rising tone than when followed by the mid or high tone. The high tone was significantly greater in height when followed by the rising tone than when followed by the low tone. The rising tone was significantly greater in height when followed by the mid or rising tone than when followed by the falling or high tone. The rising tone following the target tone raised the height of the target tone when compared to the low and mid tones.

Figure 3.9 shows the carryover coarticulation effects in stressed syllables. The mid tone was significantly greater in height when preceded by the high or rising tone than when preceded by the low or mid tone. The low tone was significantly greater in height when preceded by the high or rising tone than when preceded by the low, mid, or falling tone. The falling tone was significantly greater in height when preceded by the high tone than when preceded by the low, falling, or rising tone. The high tone was significantly greater in height when preceded by the rising tone than when preceded by the mid or low tone. The rising tone preceding the target tone raised the height of the target tone when compared to the low and mid tone.

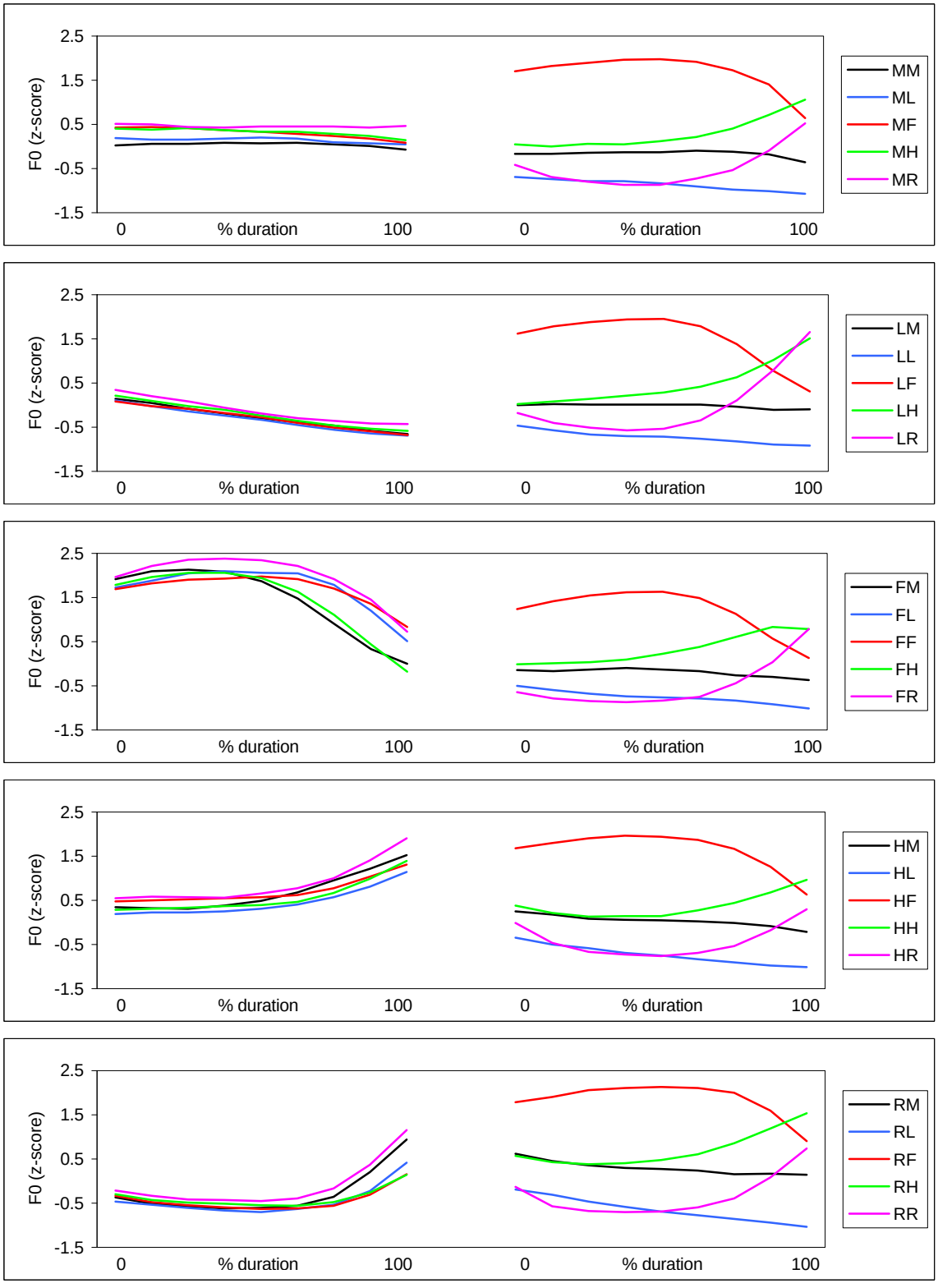

Figure 3.8 Anticipatory coarticulation in stressed syllables

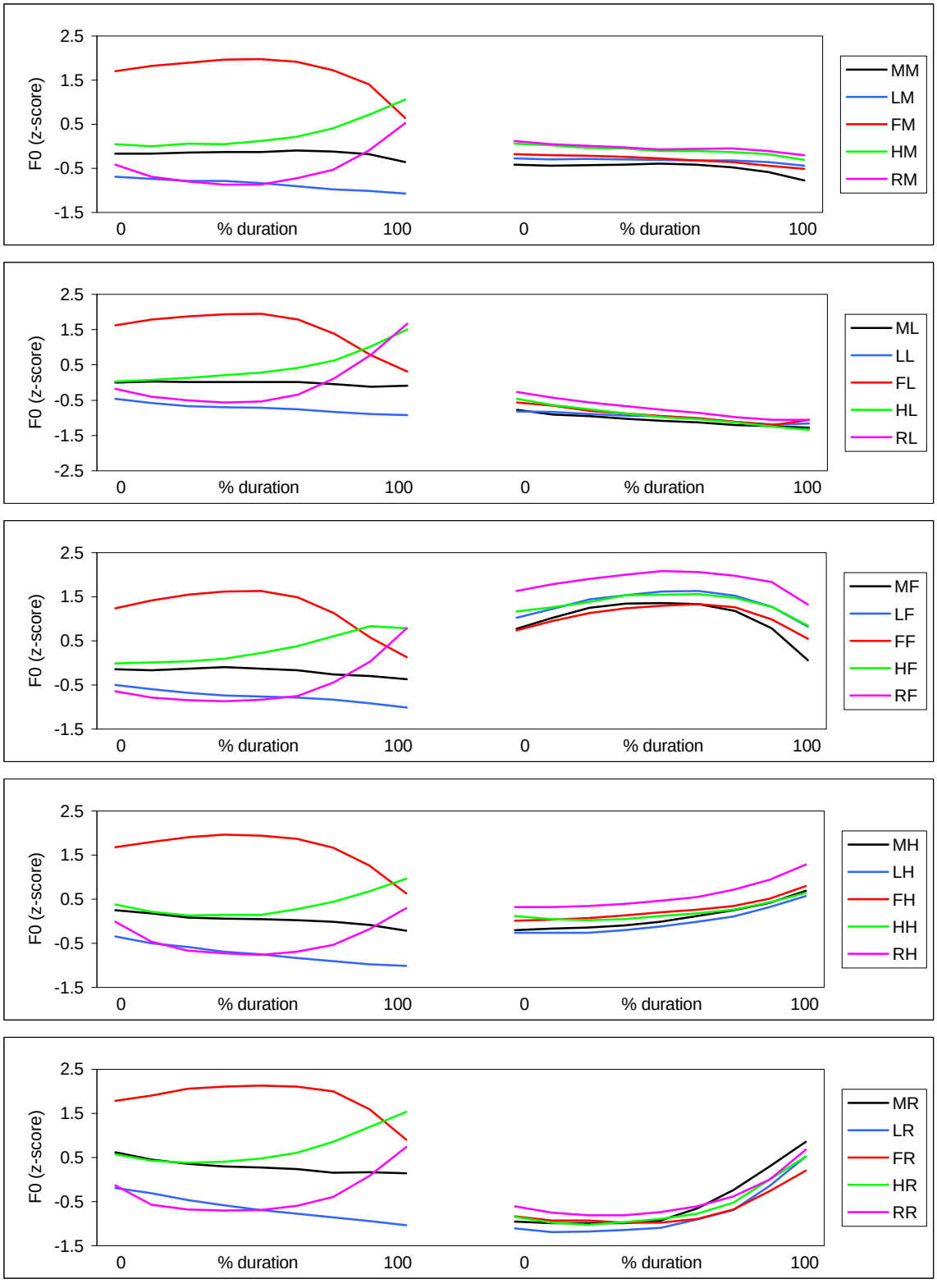

Figure 3.9 Carryover coarticulation in stressed syllables

Figure 3.10 shows the anticipatory coarticulation effects in unstressed syllables. The mid and low tones were significantly greater in height when followed by the rising or high tone than when followed by the mid or low tone. The falling tone was significantly greater in height when followed by the rising tone. The falling was significantly lower in height when preceded by the falling tone. The high tone was significantly greater in height when followed by the rising tone than when followed by the falling or high tone. The rising tone following the target tone raised the height of the target tone when compared to the low and mid tones.

Figure 3.11 shows the carryover coarticulation effects in unstressed syllables. The mid tone was significantly greater in height when preceded by the rising tone than when preceded by the low or mid tone. The carryover effects are absent in the low tone. The falling tone was significantly greater in height when preceded by the rising tone. The falling was significantly lower in height when preceded by the falling tone. The high tone was significantly greater in height when preceded by the rising tone than when preceded by the high or falling tone. The rising tone preceding the mid, falling and high tone raised the height of these tones.

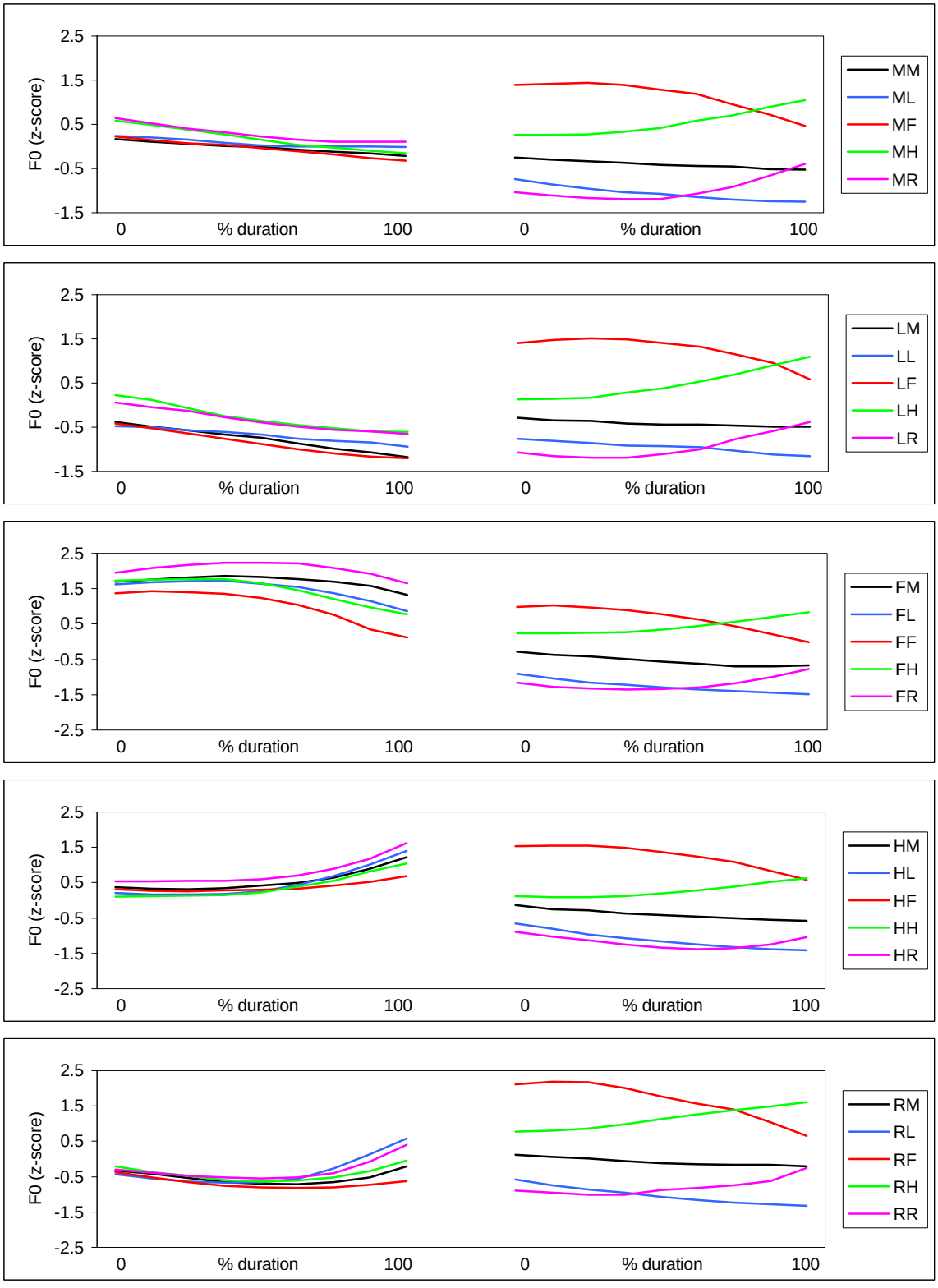

Figure 3.10 Anticipatory coarticulation in unstressed syllables

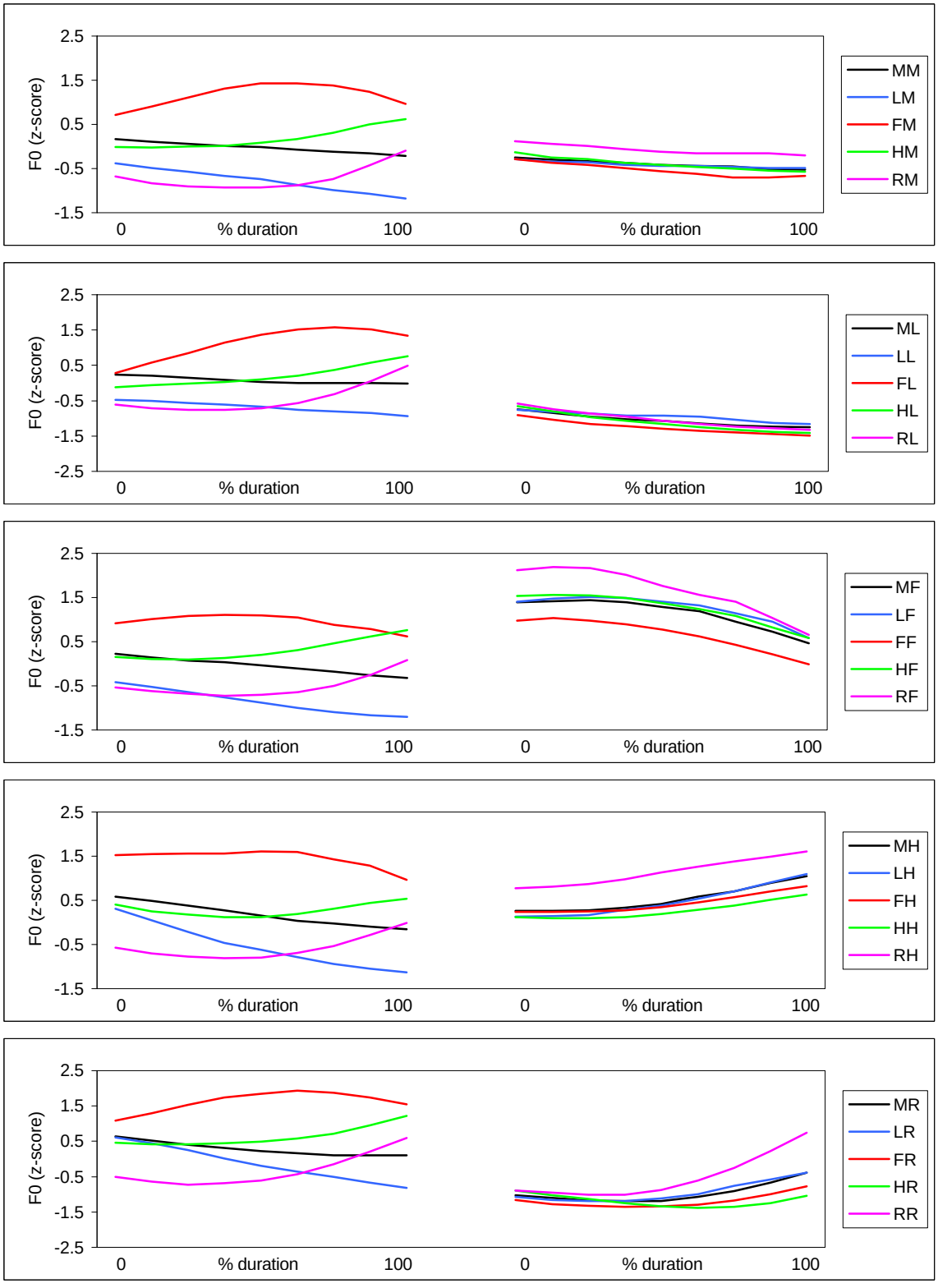

Figure 3.11 Carryover coarticulation in unstressed syllables

3.2 Summary

In this chapter, the methods for collecting and analyzing speech data of Thai tone involving stressed and unstressed syllables, intonation, and tonal coarticulation are given. Duration and normalized energy are the effective features for distinguishing between the stressed and unstressed syllables whereas the mean normalized *F0* did not signal the stress function of the syllables. Intonation affects the tone patterns by making the tone pattern decline gradually. Within each tone, the mean *F0* of the preceding syllable is higher than the succeeding syllable and the mean *F0* is lowest at the ending syllable of the sentence. This datum suggests that the mean *F0* can be used to deal with the intonation effect. The tone pattern of a syllable is also affected by tone patterns of the neighboring syllables due to the anticipatory and carryover coarticulation. It is evident that *F0* contours of both stressed and unstressed syllables are subject to modification by the preceding and succeeding syllables.

Chapter 4 Thai Tone Classification System

In this chapter, the details regarding the implementation of a proposed tone classification system are presented. The chapter begins with a discussion of the preprocessing stage. A syllable segmentation method is then outlined. The feature extraction module, including data normalization and stress detector, is also described. Finally, the operation of a proposed tone classifier is reviewed.

4.1 System Implementation

A block diagram of the proposed system is given in Figure 4.1. The summary of system implementation is described as follows. The system consists of four modules: preprocessing, syllable segmentation, feature extraction, and tone classifier. In the preprocessing module, a speech signal was first low-pass filtered and then blocked into frames. The pitch detection and energy computation were also performed in the preprocessing stage. The output of the preprocessing module was identified as the modified energy contour, which was next presented to a syllable segmentation module. In the syllable segmentation module, the beginning and ending points of syllables were determined based on the relationships between the peaks and valleys in the modified energy contour. The duration, energy, and *F0* were extracted from the segmented syllables by the feature extraction module. The *F0* contour was then normalized by using the z-score transformation. The duration as well as energy of a segmented syllable were presented to a stress detector where a degree of stress of the syllable was determined.

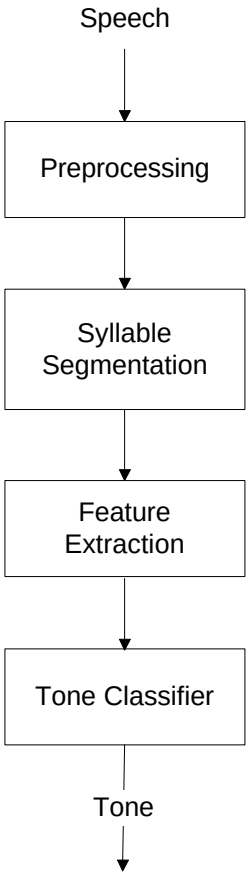

Figure 4.1 Tone Classification Model

Finally, the normalized *F0*, *F0* variation, mean *F0*, a degree of stress, and syllable ordering number of the processing and neighboring syllables were presented to a tone classifier which was a multilayer perceptron (MLP) trained by a backpropagation method. The tone classifier was trained by using speech data from two training sets for each speaker. The performances of the tone classifier was evaluated on one test set which contained 115 test sentences for each speaker. All algorithms in this study were implemented using MATLAB. The details of system implementation are presented in the next sections.

4.1.1 Preprocessing

A block diagram of the preprocessing stage is given in Figure 4.2. In this stage, the input speech was first low-pass filtered by the low-pass digital filter with a passband of 0 to 900 Hz. In general, the fundamental frequency (*F0*) range was between 50 Hz (20-msec) to 400 Hz (2.5-msec), which was suggested by Dubnowski [16] and Rabiner [79]; thus, this low-pass filtered signal should be appropriate for the extraction of *F0*. The low-pass filtered signal was then blocked into frames of 40-msec, with adjacent frames being overlapped by 20-msec. For example, the second frame began 20-msec after the first frame, and overlapped it by half of the frame length. This process continued until all the speech was accounted for within one or more frames. The reason for using 40-msec frame length in this study was because this frame length ensured that there were at least two complete cycles within a frame and was considered as a suitable frame length for the extraction of fundamental frequency of a speech signal [41], [76].

After the low-pass filtered speech signal was blocked into frames, the energy of each frame was computed. The energy *E(n)* was computed as

$$\underset{n=1}{\overset{N}{E}}(n) = \sum_{i=1}^{L} x^2(i) \tag{4.1}$$

where *x* is the sample in each frame, *N* is a number of frames and *L* is a number of samples in a frame.

The algorithm searched for the maximum energy; $E_{\max}$, of the energy contour. The energy contour was then normalized by its maximum energy. The normalized energy $\overline{E}(n)$ was computed by dividing the energy of each frame by $E_{\max}$ as defined in (4.2)

$$\underset{n=1}{\overset{N}{\overline{E}}}(n) = \frac{\overset{N}{E}(n)_{n=1}}{E_{\max}} \tag{4.2}$$

where *N* is a number of frames.

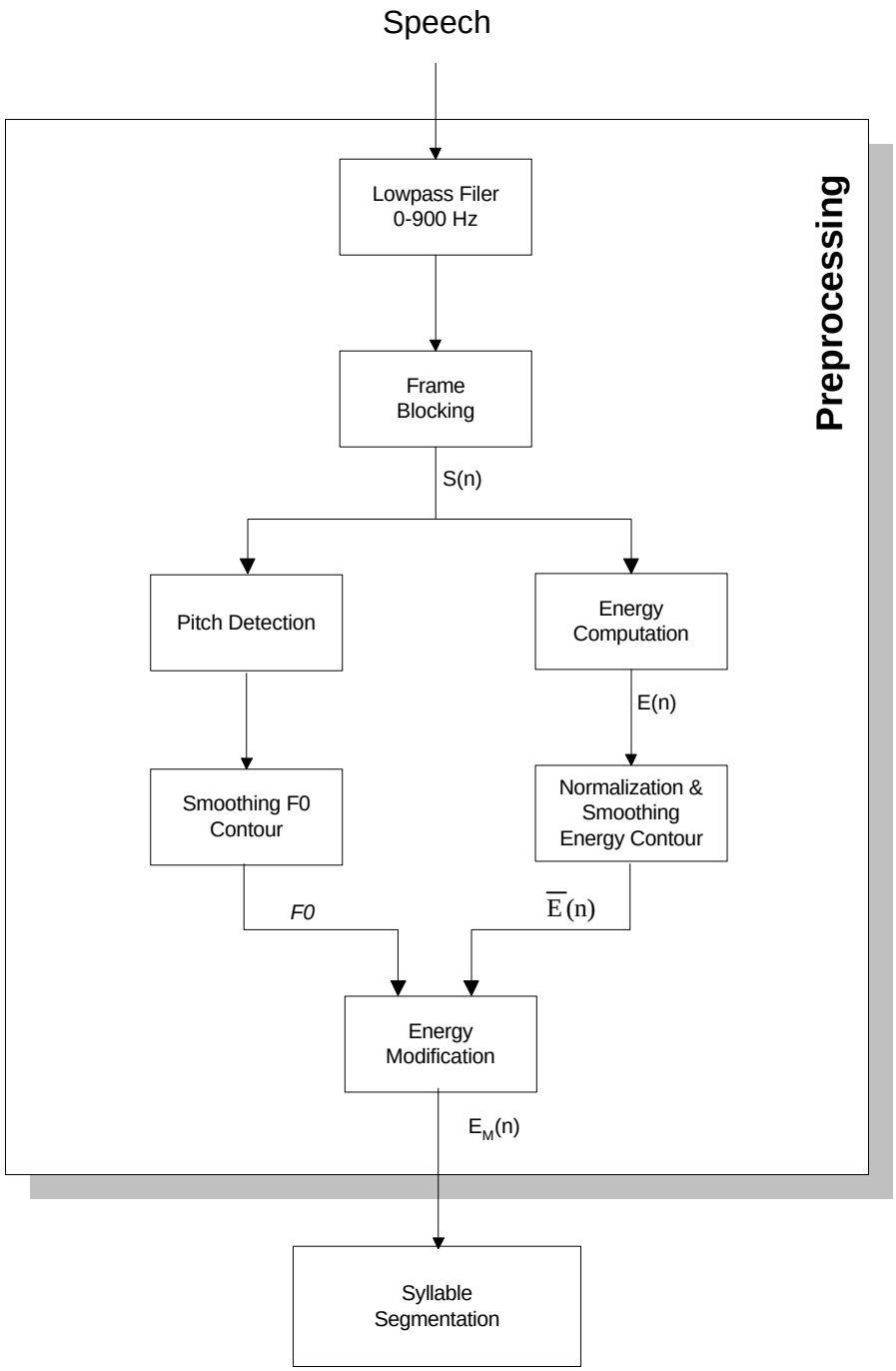

Figure 4.2 Preprocessing Stage

As a result of energy normalization, all input speech with different loudness (amplitude) could be represented on a scale from 0 to 1. Figure 4.3. shows a speech waveform and the normalized energy contour of speech utterance.

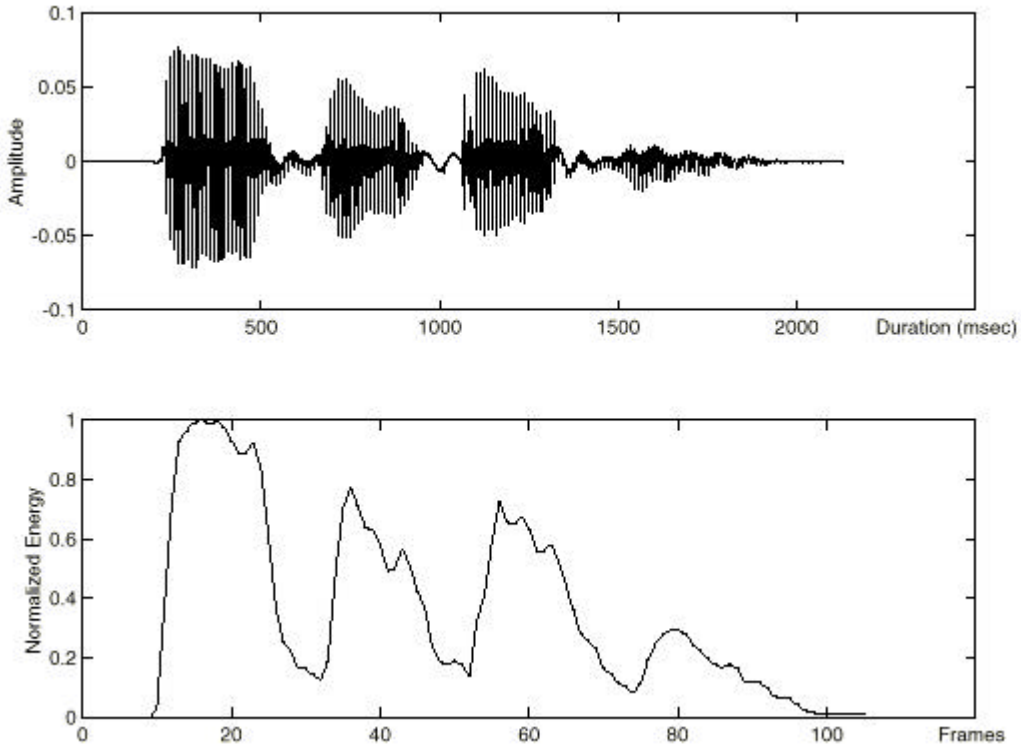

Figure 4.3 A speech waveform and its normalized energy contour

The normalized energy contour was smoothed by a median smoother, which eliminates the small dips that appeared around the peaks energy. The reason for smoothing the energy contour was because the peaks and valleys of the energy contour will be used to locate the endpoints of spoken syllables; thus, a clear indication of peaks and valleys was desired. In this step, a 7-point median smoother was applied to the normalized energy contour by finding the median of seven consecutive values. Figure 4.4 demonstrates the speech waveform and normalized energy contour before and after the smoothing process, respectively. This median smoother was employed because it provided the best results in which most of the

excessive dips around the peaks were eliminated and a clear indication of peaks and valleys was retained (bottom panel).

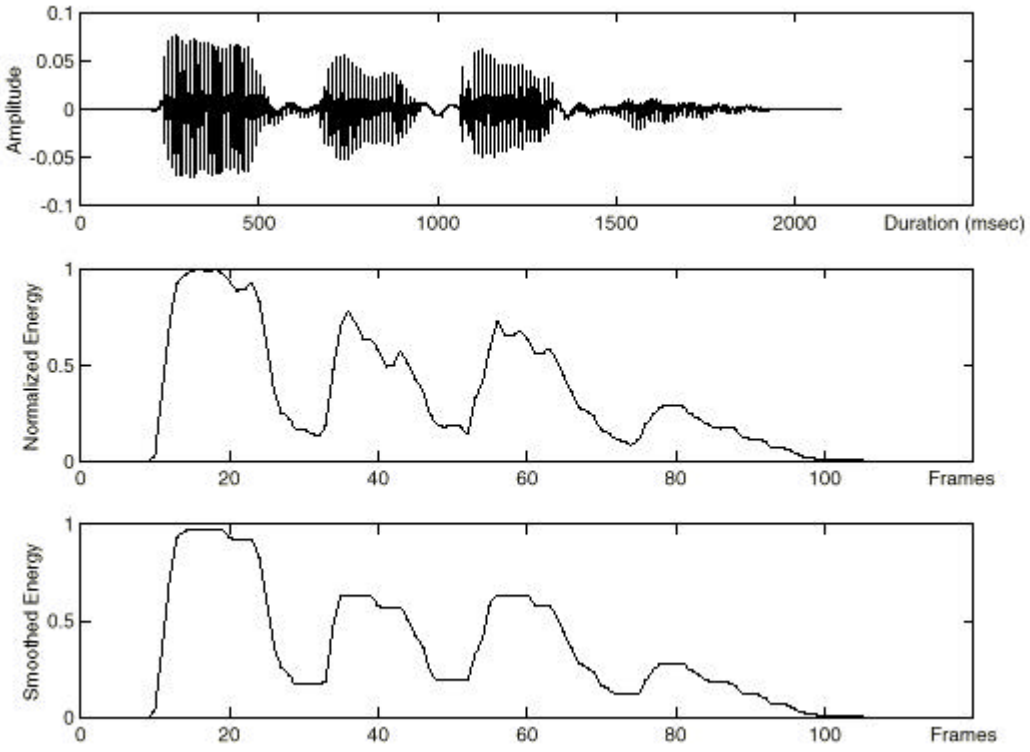

Figure 4.4 Energy contour before and after smoothing

The next step of the preprocessing stage was to extract fundamental frequency (*F0*) from a speech signal *S*(n). The autocorrelation method using a three-level center clipping (AUTOC) was employed as a pitch or *F0* detector. This *F0* extraction method was proposed by Dubnowski *et al*. [16] and is one of the most robust and widely used methods in speech processing applications [11], [41], [76]. A block diagram of pitch detector is given in Figure 4.5. The details of this method are described as follows.

The first stage of pitch detection for each 40-msec frame was the computation of the clipping level (C_L). The way in which the clipping level was chosen can be described as follows. Each frame was divided evenly into three consecutive

sections. The algorithm found the maximum absolute peak levels of the first section (PK_1) and third section (PK_2) of the frame. The clipping level within each frame was then set at 64% of the minimum of these two maximum levels [16]. The selected clipping level had an impact on the performance of the pitch detection. For example, if the clipping level was set too high, much of the waveform would fall below the clipping level and be lost. On the other hand, if the clipping level was set too low, more peaks passed through the clipper and the autocorrelation computation became more complex. The appropriate clipping level suggested by Rabiner [80] was 60-80% of the minimum value between PK_1 and PK_2.

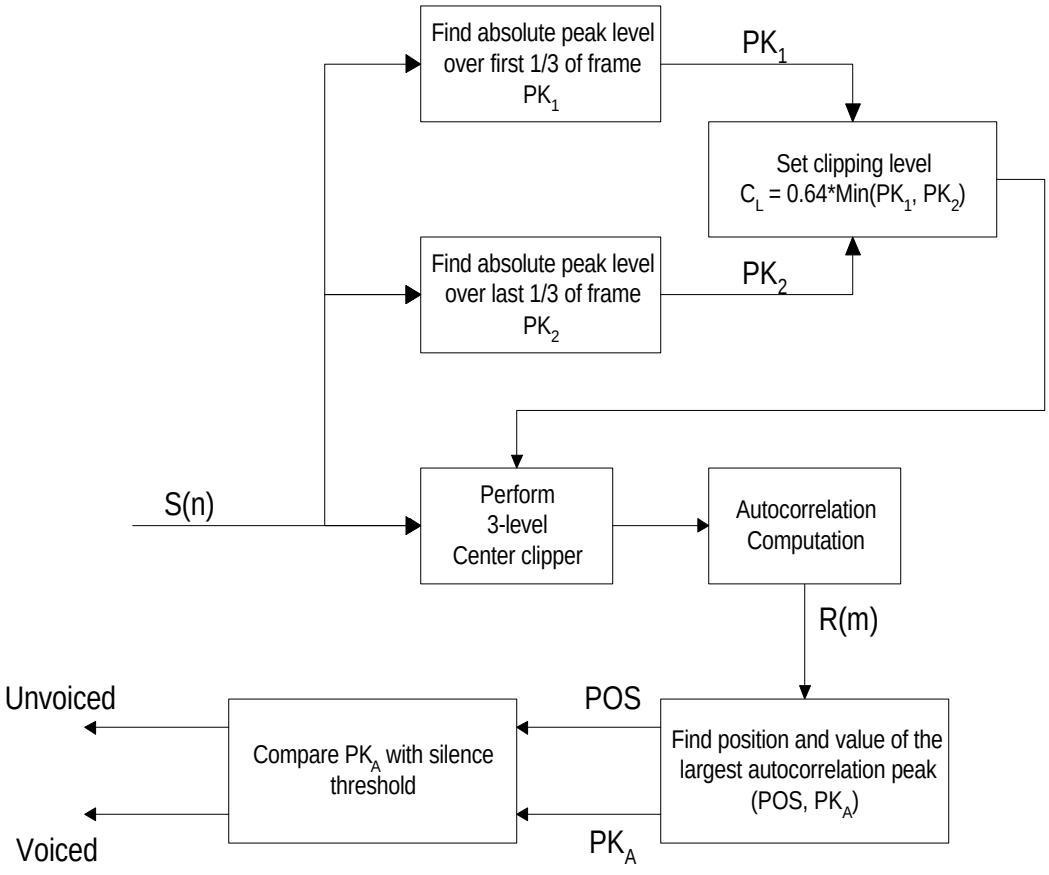

Figure 4.5 Pitch detector

After the clipping level was determined, the part of the sample which exceeded the positive clipping level ($+C_L$) was set to +1; the part of the sample which fell below the negative clipping level ($-C_L$) was set to -1; and the part of the sample which fell between $+C_L$ and $-C_L$ was assumed to be 0. Figure 4.6 shows a plot of the input-output characteristic for the three-level center clipper.

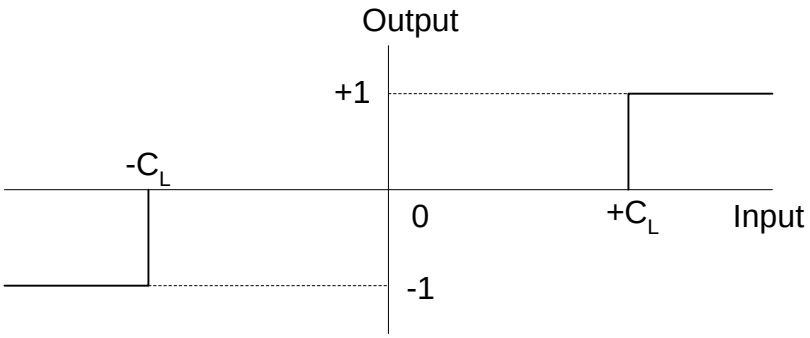

Figure 4.6 three-level center clipping function

A three-level center clipper processed the speech signal *S*(n), and the autocorrelation function was computed. The computation of the autocorrelation function *R*(*m*) for a 3-level center clipped signal was defined as

$$R(m) = \sum_{n=0}^{L-m-1} y(n)y(n+m) \qquad m = 0,1,2,...,L-1 \tag{4.3}$$

where *y* were the clipped samples in each frame, *L* was the number of samples in a frame (for a 40-msec frame with a sampling rate of 11,025 Hz, *L*=441 samples). Demonstration of the autocorrelation computation of the clipped samples within a frame is displayed in Figure 4.7.

Following the autocorrelation computation, the autocorrelation function was then searched for the position (POS) and the maximum peak (PK_A) over the anticipated range of pitch periods. Typical values of the pitch range suggested by Rabiner [79] and Dubnowski [16] were 400 Hz down to 50 Hz, which were

corresponding to the autocorrelation function $R(m)$ in the interval m=27 to m=220 in (4.3).

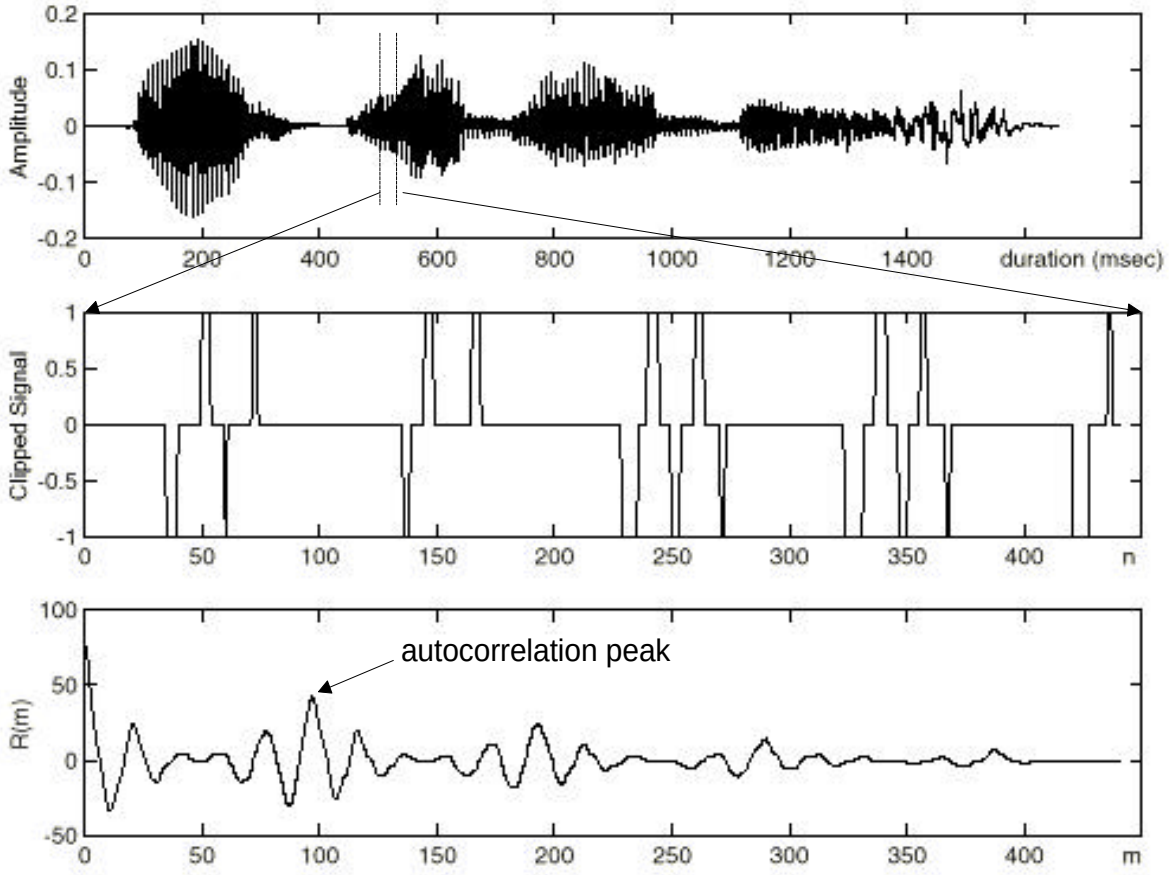

Figure 4.7 An example of the autocorrelation computation of the clipped samples within a frame.

The energy of the clipped signal was then computed and the silence threshold was set at a certain percent of the energy of the clipped signal (e.g., 30% of energy [16]). The autocorrelation peak (PK_A) was compared to the silence threshold. If the autocorrelation peak was higher than the silence threshold, this frame was classified as voiced and the position of the autocorrelation peak (POS) was defined as a pitch period. If the autocorrelation peak fell below the silence threshold, the frame was classified as unvoiced. For example, if a frame was classified as voiced and its position of autocorrelation peak was located (e.g.,

POS=98 in Figure 4.7), the pitch period was converted to fundamental frequency (*F0*) by dividing the sampling rate by POS (e.g., for sampling rate of 11,025 Hz, $F0 = \frac{11025}{98} Hz$).

After the fundamental frequency (*F0*) was extracted from the input speech, the next step was to smooth the *F0* contour. Because of the noise-like components of a signal and the instantaneous variations in the *F0* contour, a smoothing method was required to remove those errors in the *F0* contour. In this step, a 3-point median smoother was applied to the *F0* contour by finding the median of three consecutive values. Examples of speech waveforms and their smoothed *F0* contours are given in Figures 4.8a and 4.8b.

The final step of the preprocessing stage was to modify a smoothed energy contour with its *F0* contour. As a result of this modification process, the modified energy contour had the same starting and ending points of energy pulses as did the *F0* pulses. The benefits of using the *F0* contour to modify the energy contour can be described as follows. First, the input utterances might contain not only the spoken syllables but also other components such as lip click and breath noise which often occurred during speech production. The energy pulses of lip click and breath noise might be mistakenly selected as a spoken word or syllables by the endpoint detector. However, these undesirable components were considered as unvoiced and were eliminated during the pitch detection process; thus, the modified energy contour contains only the syllable pulses. Secondly, a Thai syllable always carries a tone with its *F0* pattern corresponding to a spoken syllable; thus, using a modified energy contour will help the syllable segmentation algorithm locate the spoken syllables more accurately. Furthermore, it was guaranteed that the segmented syllables contained *F0* which was the desirable feature for a tone classifier. Lastly, by using this modification process alone, the utterance was automatically segmented into syllables in some cases.

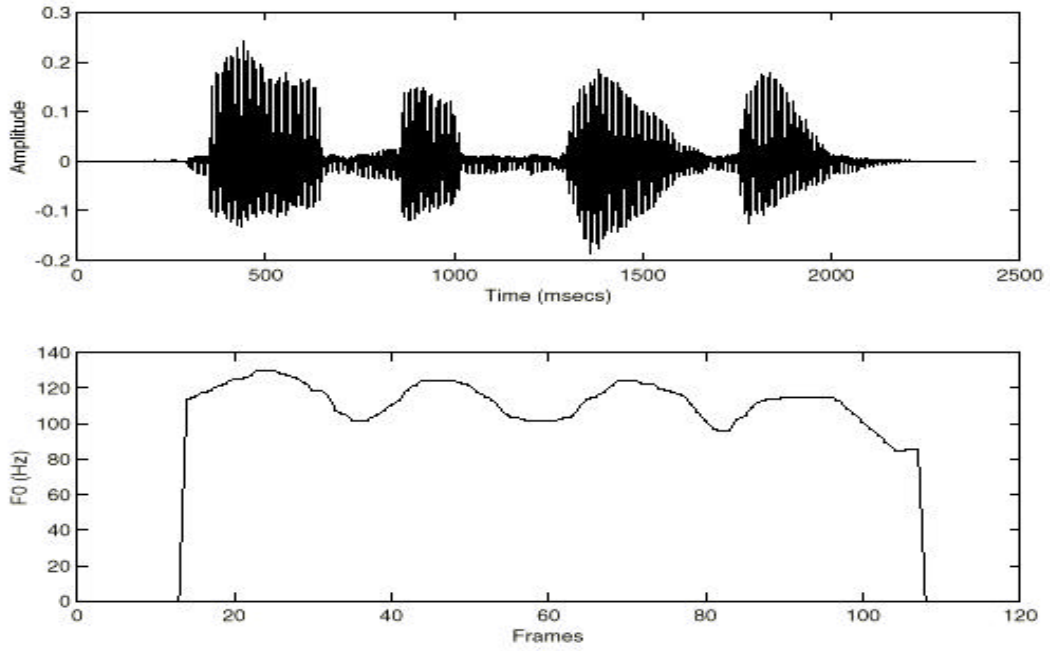

Figure 4.8a *F0* contour of a “falling-falling-falling-falling” tone sequence spoken by male speaker

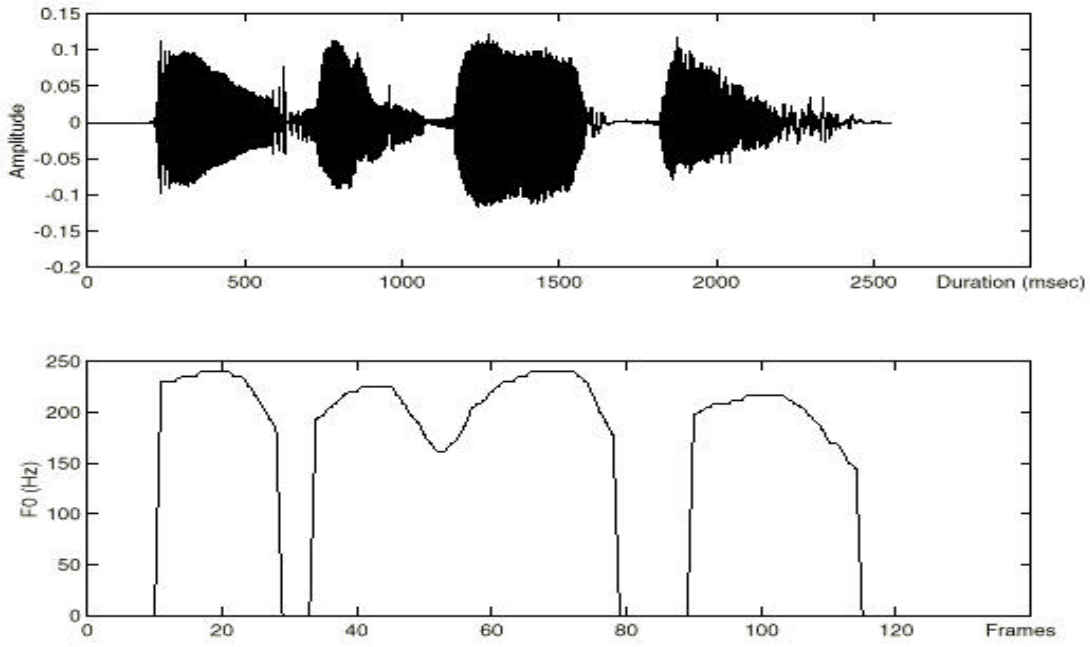

Figure 4.8b *F0* contour of a “falling-falling-falling-falling” tone sequence spoken by female speaker

The energy modification process can be done by scanning both *F0* and smoothed energy contour from the beginning to the ending of the utterance. When the *F0* value at the current frame was absent, the energy at that frame was set to 0. An example of the energy modification process is depicted in Figure 4.9. As seen in Figure 4.9, the energy pulses of the third and fourth syllable of the utterance were connected without a valley between them as shown in the second panel, the energy pulses of the third and fourth syllable were then segmented by the modification process as shown in the bottom panel.

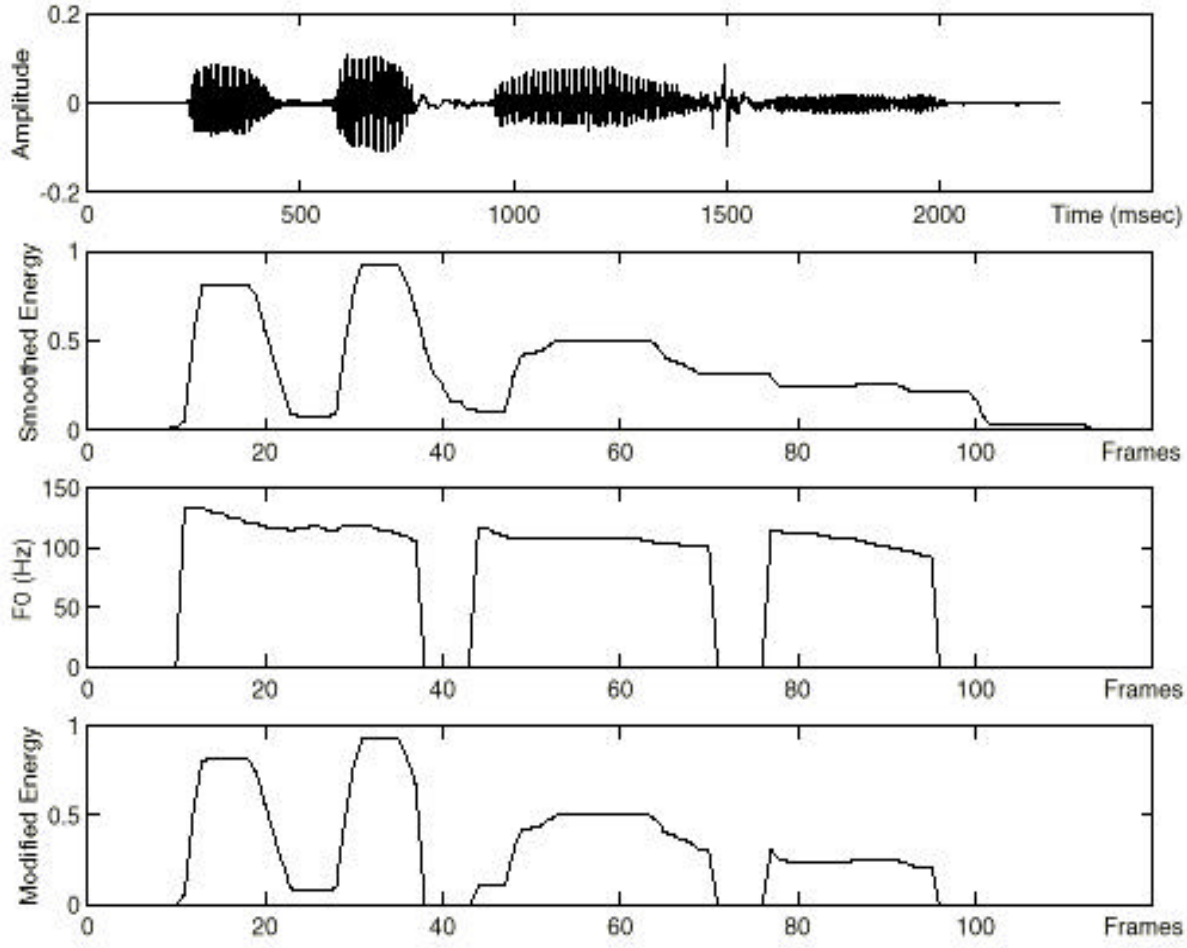

Figure 4.9 Smoothed energy contour before and after modification. The panels from top to bottom show a speech waveform, smoothed energy contour before modification, *F0* contour, and modified energy contour (E_M), respectively.

4.1.2 Syllable Segmentation

Since an automatic method to perform syllable segmentation of Thai speech is currently not available, automatic syllable segmentation was developed in this study. The key idea of this method was to scan the modified energy contour (E_M) and search for the valleys in the modified energy contour because these valleys, which appeared between the peaks, were most likely the endpoints of the syllables. If a valley existed, the algorithm then compared the ratio of the neighboring peaks energy and valley with the thresholds in order to decide whether or not it was a potential valley and then the starting and ending points of the syllable were determined.

Before the details of the syllable segmentation algorithm are given, the following describes the characteristics of peaks and valleys in the modified energy contour. The modified energy contours were on a scale from 0 to 1, which resulted from the energy normalization process. The energy of the valleys was classified into three levels: low (less than 0.25), medium (between 0.25 and 0.5), and high (more than 0.5) as displayed in Figure 4.10. In order to study the relationships between peaks and valleys, the ratios of the preceding and succeeding peaks energy to the valleys energy were computed for all modified energy contours. Table 4.1 provides the relationships between peaks and valleys in the modified energy contours obtained from the training speech uttered by all speakers. The minimum ratios of the preceding and succeeding peaks to the valleys energy level are given in the second and third column of Table 4.1, respectively. For example, in a low energy valley, the preceding peak energy was at least equal or higher than 1.63 times the valley energy whereas the succeeding peak energy was at least equal or higher than 1.31 times the valley energy. Based on these relationships, the syllable segmentation algorithm employed these ratios to determine the thresholds in order to locate the starting and ending points of the syllables.

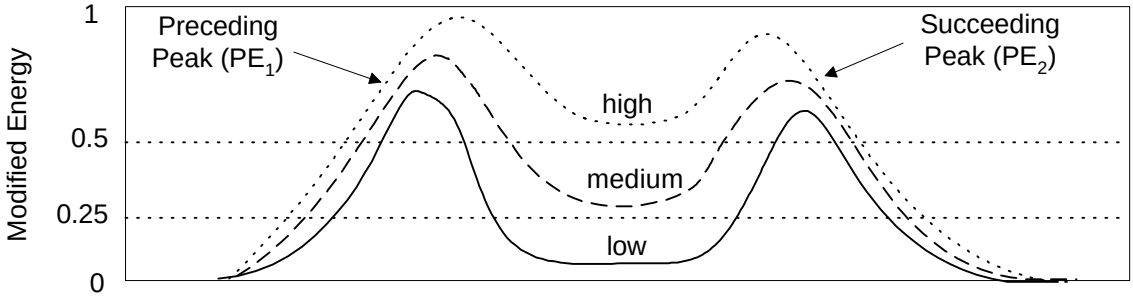

Figure 4.10 Three types of valleys energy level.

Valley Energy Level (V)	Preceding Peak/Valley Ratio (PE_1/V)	Succeeding Peak/Valley Ratio (PE_2/V)
Low (Less than 0.25)	1.63	1.31
Medium (Between 0.25 and 0.5)	1.42	1.23
High (More than 0.5)	1.25	1.22

Table 4.1 The relationship between peaks and valley energy level

A block diagram of the proposed syllable segmentation is given in Figure 4.11. The inputs of the syllable segmentation module are the modified energy contour $E_M(k)$, k=1,2,…,N, where N is a number of frames. The outputs are the starting and ending pairs of the segmented syllables, which are next processed by the feature extraction module. The details of the proposed method are described as follows.

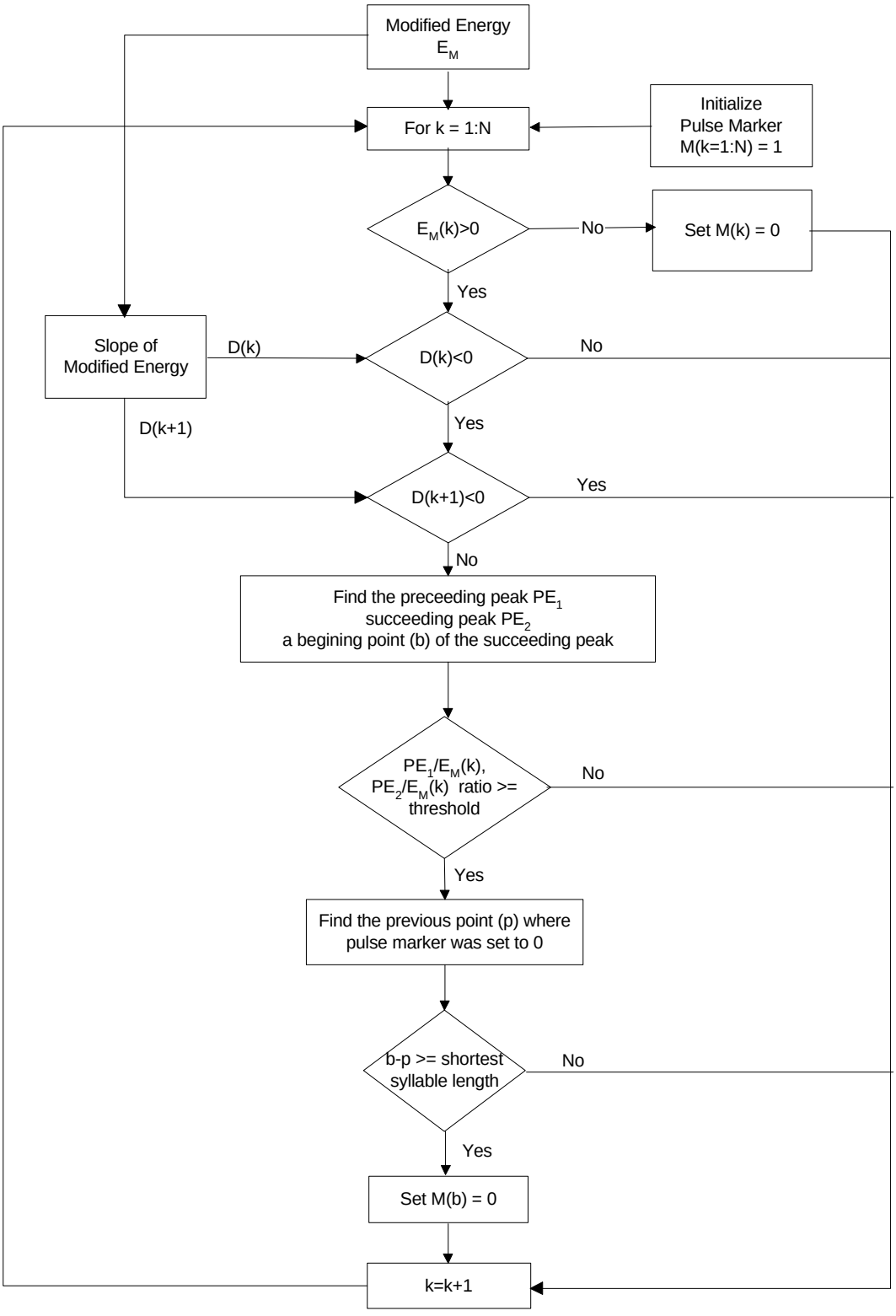

Figure 4.11 Flow chart of the syllable segmentation algorithm

The pulse marker M(k=1,2,…,N) of all frames was first set to +1, where N was a number of modified energy frames. The E_M contour was scanned and when E_M(k) was absent, the pulse marker M(k) was set to 0. When E_M(k) was present (e.g., E_M(k)>0), the slope of modified energy contour at the current frame D(k) was considered before the value of M(k) was determined. If D(k) was an upward slope (e.g., point A in Figure 4.12), or the slope was level, or D(k) was a downward slope (e.g., point B) and a slope of the next frame D(k+1) was also a downward slope, the algorithm continued on to the next frame. If D(k) was a downward slope (e.g., point C) and a slope of the next frame D(k+1) was level or upward, then the algorithm searched for a preceding peak (PE_1), a succeeding peak (PE_2), and a beginning frame of the succeeding peak (e.g., point D) and further tests were made on the energy level of the current frame E_M(k) and the energy levels of both peaks (PE_1 and PE_2). An example of the preceding peak (PE_1), succeeding peak (PE_2), and a valley of the normalized energy contour is depicted in Figure 4.12.

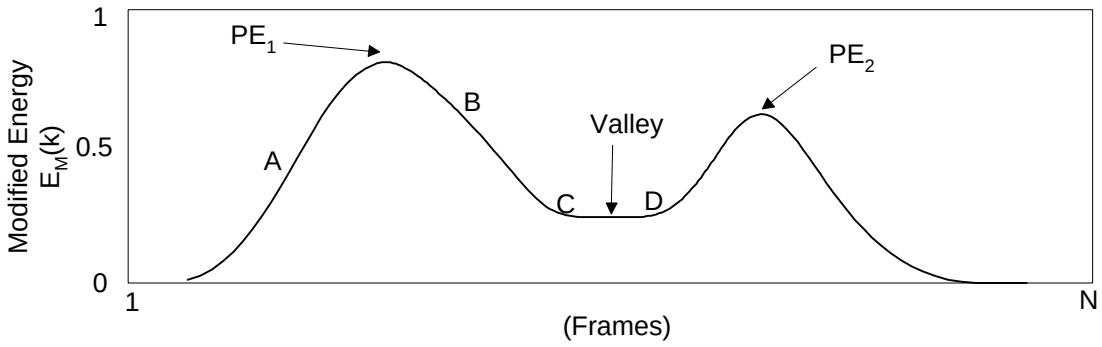

Figure 4.12 Example of the preceding peak (PE_1), succeeding peak (PE_2), and a valley of the modified energy contour.

If E_M(k) was less than 0.25, and PE_1/E_M(k) ratio was equal or higher than 1.6 or PE_2/E_M(k) ratio was equal or higher than 1.3, then the algorithm searched the pulse marker backward for the previous marked point (p) where its value was equal to 0 and measured the length between the previous marked point and the beginning point (b) of the succeeding peak. If this length (b-p) was equal or longer than the shortest syllable length (as observed from the speech data spoken by all speakers at

140 milliseconds as expressed in Table 4.2), then the pulse marker was set to 0 at the beginning point of the succeeding peak; otherwise, the algorithm continued on to the next frame.

In the same manner, if E_M(k) was between 0.25 and 0.5, and PE_1/E_M(k) ratio was equal or higher than 1.4 or PE_2/E_M(k) ratio was equal or higher than 1.2, or if E_M(k) was higher than 0.5 and PE_1/E_M(k) and PE_2/E_M(k) ratios were equal or higher than 1.2, and the length of the expected syllable was equal or longer than the shortest syllable length, then the pulse marker was set to 0 at the beginning point of the succeeding peak; otherwise, the algorithm continued on to the next frame.

Speaker	Syllable Length		
	Shortest (msec)	Longest (msec)	Average (msec)
SPK1	220	820	473.60
SPK2	220	840	506.83
SPK3	140	880	460.5
SPK4	140	860	463.85
SPK5	140	900	452.25
SPK6	180	700	459.60
SPK7	160	820	488.07
SPK8	220	860	530.50

Table 4.2 Syllable length obtained from the training sets

The reason that the energy level of the preceding peak (PE_1) and succeeding peak (PE_2) were compared with the energy level of current point E_M(k) described previously, was to decide whether or not the considered point was a real valley. It should be noted that the several thresholds used in this algorithm resulted from the study of relationships between energy of the peaks and valleys shown in Table 4.1. An example of syllable segmentation is displayed in Figure 4.13. From the top to

bottom panel of the figure show the speech waveform, modified energy contour, and pulse marker, respectively.

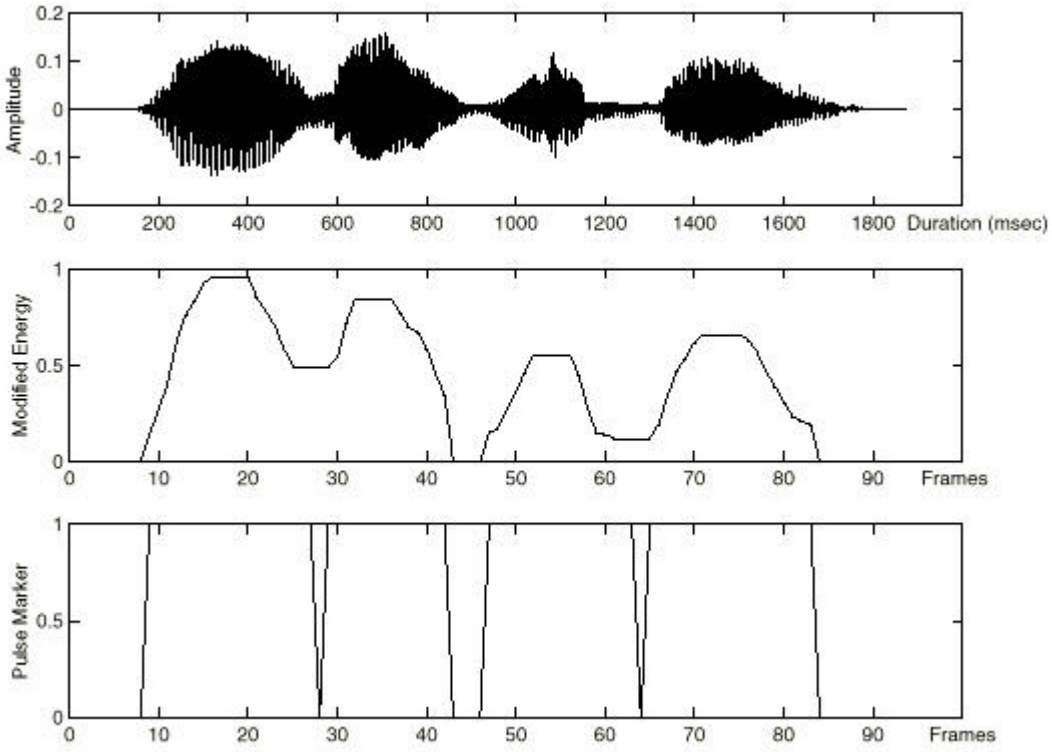

Figure 4.13 An example of syllable segmentation process

The syllable segmentation algorithm continued until a whole modified energy contour was processed, and the pulse marker was then ready for final processing. In the final process of syllable segmentation, the pulse marker was scanned and algorithm searched for the starting and ending points of each syllable by determining the point where the pulse marker changed its level. For example, the pulse marker changed its level from 0 to +1 and set it as a starting point, then found the closest point where the pulse marker changed its level from +1 to 0 and set it as an ending point. This process continued until a whole pulse marker contour was completely scanned. Finally, the starting and ending pairs of the syllables were produced and this information was passed to the feature extraction module in order to extract the acoustical features of the segmented syllables.

4.1.3 Feature Extraction

In the feature extraction module, the acoustical features, including *F0*, energy, and duration, were extracted according to each starting and ending pair obtained from the syllable segmentation module. The feature extraction module consisted of two submodules: data normalization and stress detector. The data normalization was employed to eliminate time and speaker variations whereas duration and energy were passed to the stress detector to determine the degree of stress of each segmented syllable. The details of data normalization and stress detector are described in the following section.

4.1.3.1 Data Normalization

Normalization of the feature parameters was a necessary task because it eliminated the undesirable time and speaker variation of the feature parameters. For instance, there were perceptible differences in the fundamental frequencies of different speakers due to the differences of the vocal chord sizes. In general, a male had longer and more massive chords than had a female, with the result that male speakers had lower *F0* levels. This phenomenon created a problem for tone classification; thus normalization of *F0* was necessary before tone classification could be performed. The normalization technique used in this study was z-score transform. This transform expressed an observed data value as a multiple of a measure of dispersion away from a mean value.

$$Z = \frac{X - \overline{X}}{s} \tag{4.4}$$

where X is raw data and s is a standard deviation about the mean ($\overline{X}$). This technique has been successfully applied in speech recognition by several speech researchers [26], [27], [76].

The parameters of z-score transform were mean and standard deviation of *F0* which were pre-computed from the training speech uttered by each speaker. Table

4.3 provides the details of mean and standard deviation of *F0* of each speaker. As seen in Table 4.3, all female speakers (SPK6, SPK7, and SPK8) had a higher average *F0* (in the range of 180 to 200 Hz) than the male speakers (SPK1, SPK2, SPK3, SPK4, and SPK5). The male speakers had an average *F0* in the range of 100-120 Hz except SPK4, whose average *F0* was lower than 100 Hz.

Speaker	*F0* (Hz)	
	Mean	STD
SPK1	111.79	14.01
SPK2	107.53	12.56
SPK3	126.22	17.02
SPK4	97.52	14.32
SPK5	118.11	18.43
SPK6	187.71	25.08
SPK7	197.68	27.59
SPK8	184.61	23.58

Table 4.3 Mean and standard deviation of *F0* for each speaker.

Due to the fact that the spoken syllables were not always equal in length, despite being spoken by the same speaker, a method was needed to represent the spoken syllables so that all spoken syllables would be the same length. In this study, the normalized *F0* value at 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of syllable duration were computed to represent a whole normalized *F0* contour. Therefore, a normalized *F0* contour was then represented by *F0*(0%), *F0*(10%), *F0*(20%), *F0*(30%), *F0*(40%), *F0*(50%), *F0*(60%), *F0*(70%), *F0*(80%), *F0*(90%), *F0*(100%). By doing this, all *F0* contours can be represented by the same length no matter what the length of the *F0* contour was. However, a number of errors occurred at the beginning and ending of the *F0* contour. Some of these errors were the result of the *F0* extraction process. Other errors were possibly introduced by the syllable segmentation process. Only the normalized *F0* value at 10% up to 90% of syllable duration; e.g., *F0*(10%), *F0*(20%), ……, *F0*(80%), *F0*(90%), were

employed. Figure 4.14 shows a tone sequence of four monosyllabic words as well as the normalized *F0* of each syllable. In addition, mean normalized *F0* of each segmented syllable was calculated as well. The final process of this stage was the *F0* variations computation, which were calculated by finding the differences between each *F0* value; e.g., {*F0*(30%)-*F0*(10%), *F0*(50%)-*F0*(30%), *F0*(70%)-*F0*(50%), *F0*(90%)-*F0*(70%)}. The *F0* variations represented the trend and movement within *F0* contour and were also used as the features to identify tones.

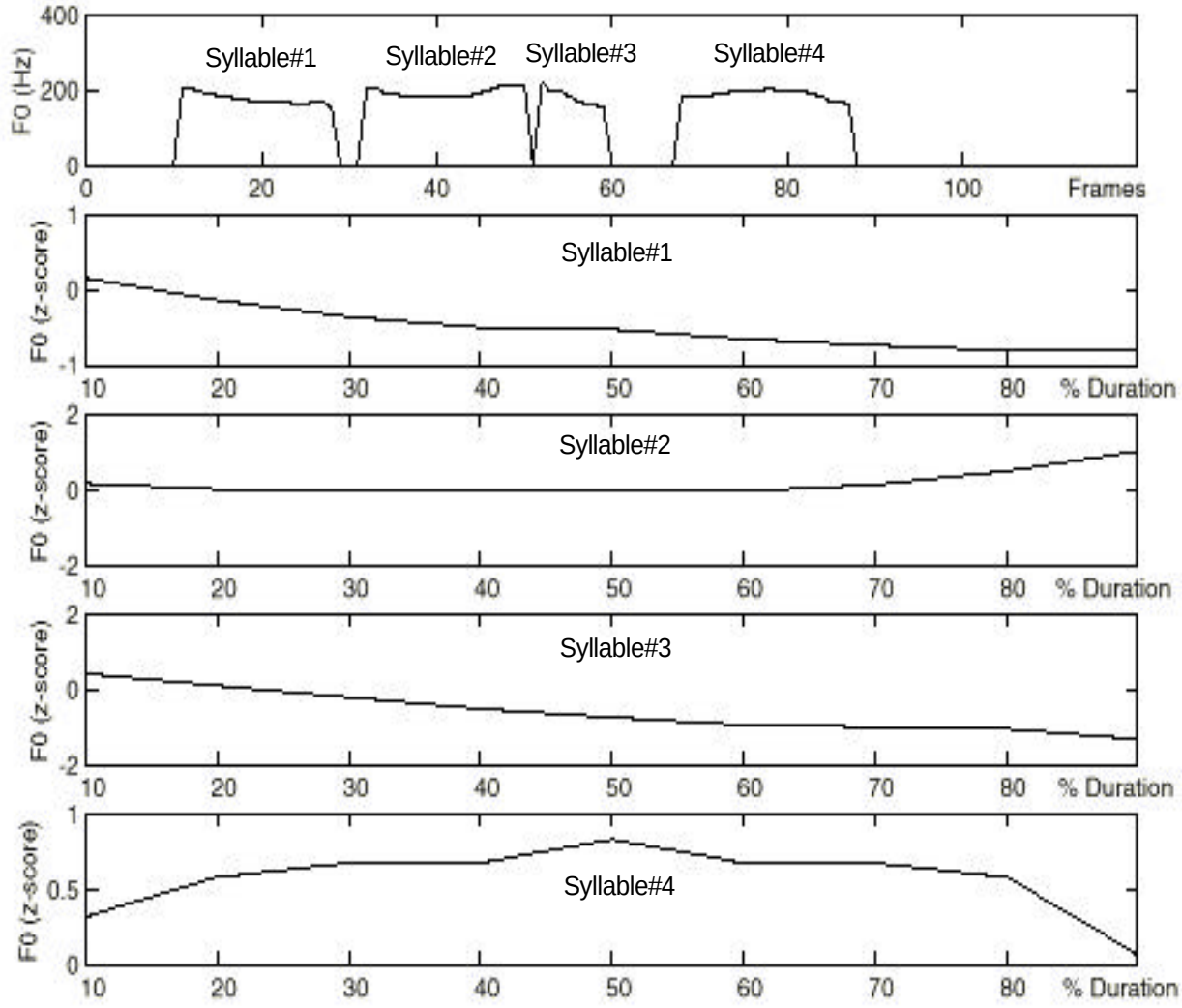

Figure 4.14 shows an example of the normalized *F0* contour for each syllable. The top panel is the *F0* contour of a “low-high-low-falling” tone sequence. Panels two

through five show the normalized *F0* contour (z-score) of the first, second, third, and fourth syllable, respectively.

4.1.3.2 Stress Detector

In this stage, duration and energy of syllables, which were the effective parameters in distinguishing between stressed and unstressed syllables based on the analysis of Thai tones in Chapter Three, were presented to a stress detector to determine the degree of stress of a syllable. The stress detector was implemented by using a fuzzy inference system (*FIS*), which was based on the concepts of fuzzy set theory, fuzzy if-then rules, and fuzzy reasoning. The *FIS* performed a nonlinear mapping from its input space to output space. This mapping was accomplished by a number of fuzzy if-then rules, each of which described the local behavior of the mapping. *FIS* has been successfully applied in a wide variety of applications, such as automatic control, data classification, and pattern recognition [37].

The basic structure of the stress detector is shown in Figure 4.15. From the figure, it is evident that the information flows from left to right and from two inputs (duration and energy) to a single output (degree of stress). The membership function of each input was constructed based on the mean duration and energy of stressed and unstressed syllables which were pre-computed from the training sets of each speaker. Mean duration and energy of stressed and unstressed syllables for each speaker are presented in Table 4.4. The output of the stress detector, which was in the interval ranging from 0 to 1, was computed based on nine rules as shown in Figure 4.15.

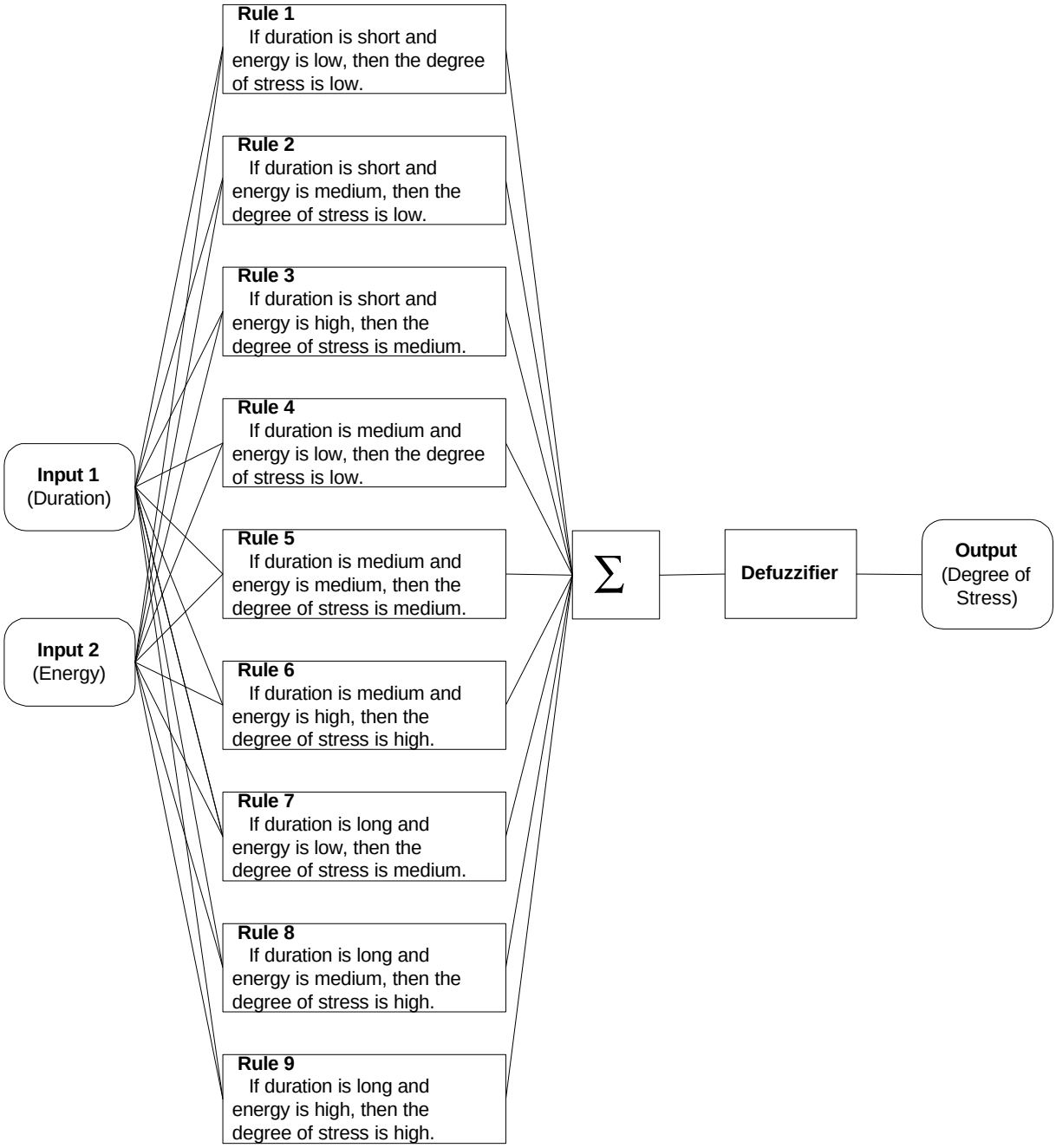

Figure 4.15 Basic structure of a stress detector

Speaker		Unstressed Syllable	Stressed Syllable
		Mean	Mean
SPK1	duration (msec)	394.0	572.0
	energy	7.72	13.03
SPK2	duration (msec)	410.2	582.4
	energy	7.02	15.33
SPK3	duration (msec)	399.2	574.6
	energy	7.71	12.15
SPK4	duration (msec)	347.2	578.0
	energy	6.42	13.61
SPK5	duration (msec)	356.2	559.6
	energy	6.28	13.71
SPK6	duration (msec)	381.6	551.2
	energy	6.97	14.06
SPK7	duration (msec)	360.6	552.2
	energy	6.24	13.69
SPK8	duration (msec)	391.0	633.0
	energy	7.74	15.91

Table 4.4 Mean duration and energy of stressed and unstressed syllables for each speaker

Figure 4.16a, b, and c show the membership functions of two inputs and one output, respectively. From the figures, the duration was classified as short, medium, and long whereas energy was classified as low, medium, and high. The output (degree of stress) was classified as low, medium, and high.

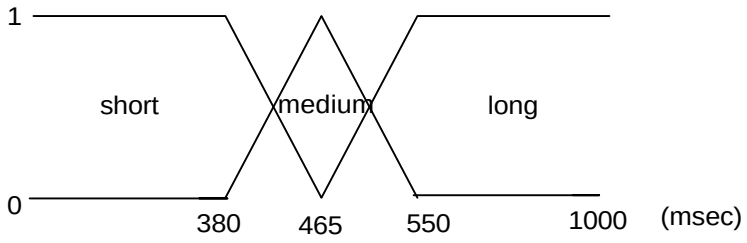

Figure 4.16a Example of duration membership function

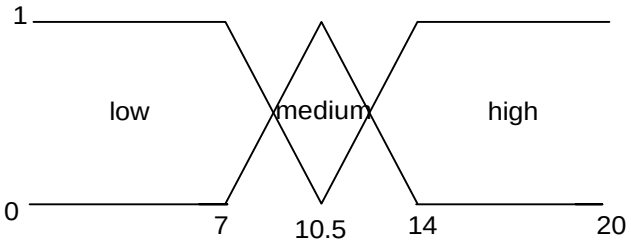

Figure 4.16b Example of energy membership function

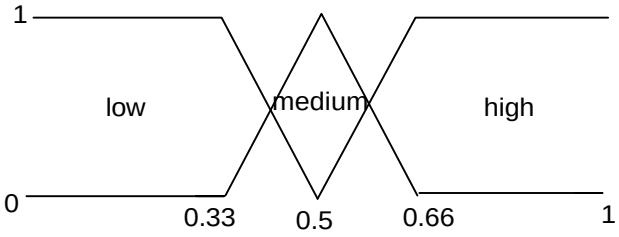

Figure 4.16c Output membership function

The procedure of a proposed stress detector is summarized as follows. The first step of stress detection was the input fuzzification. This step was implemented by taking the inputs (duration and energy) and determining the degree which they belonged to each of the appropriate fuzzy sets via membership functions. The input was generally a crisp numerical value in the limited range of the input variable. For example, the interval between 0 to 1000 milliseconds was a range of duration whereas the interval ranges from 0 to 20 was a range of energy input. The outputs of input fuzzification were a degree of membership of each input having a value between 0 and 1.

The fuzzy operator was then applied to the fuzzified inputs which were obtained from the first step. The antecedent part of the if-then rule was evaluated by applying the fuzzy operator to obtain a single value that represented the antecedent for each rule. For example, in the fifth if-then rule ("If the duration is medium and energy is high then the output is stressed."), the antecedent of this rule was evaluated by applying fuzzy "AND" operator which selected the minimum of the two fuzzified inputs.

Next, the implication process was implemented for each rule. The input of the implication process was a single number given by the antecedent and the result of implication was a fuzzy set for each rule. The output functions produced by the implication process for each rule were then combined into a single fuzzy set by the aggregation process, resulting in a fuzzy set for each output variable. Finally, the defuzzification process determined a single crisp value that best represents the output of the system according to the inputs and associated rules. In this study, the centroid of area was employed as a defuzzification method to produce the output of stress detector, which was in the interval ranges of 0 to 1.

4.1.4 Tone Classifier

Before the details of the tone classifier are given, the acoustical features employed to train the tone classifier are first discussed. For each speaker, a tone classifier was trained by using two training sets, and then tested with one test set. Each training set contained 100 sentences which covered all five Thai tones and stressed patterns whereas a test set contained 115 test sentences.

Based on the analysis of Thai tones in Chapter Three, *F0* patterns are subjected to several modifications due to the effects of tonal coarticulation, stress, and intonation. The *F0* pattern of the syllable in process was influenced by the preceding and succeeding syllables due to tonal coarticulation; hence, in order to deal with this effect, the training of tone classifier was carried out with a sequence of

F0 patterns in which the normalized *F0* contour of the processing, preceding and succeeding syllable were employed in the training vectors. In addition, the tone patterns were also affected by the intonation in the form of declination. However, the study of intonation effects suggested that the mean *F0* of each syllable could be used to deal with the intonation effects, thus the mean *F0* of each syllable and syllable ordered number in the sentence were also included in the training vector. The *F0* patterns of stressed and unstressed syllables were different, especially the falling, high, and rising tones. The stressed and unstressed syllables can be distinguished by their duration and energy. The stress detector made use of these features to identify the degree of stress of the syllable; thus the degree of stress of the syllable was also included in the training vector to help the tone classifier to distinguish between stressed and unstressed syllable. Based on these training features, the tone classifier was properly trained to deal with the effects of tonal coarticulation, stress, and intonation.

In this study, a multilayer perceptron (MLP) trained by the backpropagation method was employed as a tone classifier. The structure of the tone classifier consisted of one input layer, one hidden layer and one output layer as shown in Figure 4.17. The output layer had five nodes in which each output node represented a particular Thai tone. The hidden layer contained 65 nodes whereas the input layer consisted of 48 nodes in which each input node represented an element of the extracted features as follows.

1. Normalized *F0*, *F0* variations, mean *F0*, and a degree of stress of the syllable in process (9+4+1+1=15 nodes)
2. Normalized *F0*, *F0* variations, mean *F0*, and a degree of stress of the preceding syllable (9+4+1+1=15 nodes)
3. Normalized *F0*, *F0* variations, mean *F0*, and a degree of stress of the succeeding syllable (9+4+1+1=15 nodes)
4. Syllable ordered number of the syllable in process in the sentential utterance (3 nodes)

It should be noted that all 15 elements (nodes) of the preceding syllable were set to -1 when the syllable in process was the first syllable of the sentence. In the same manner, all nodes of the succeeding syllable were set to -1 when the syllable in process was the last syllable of the sentence.

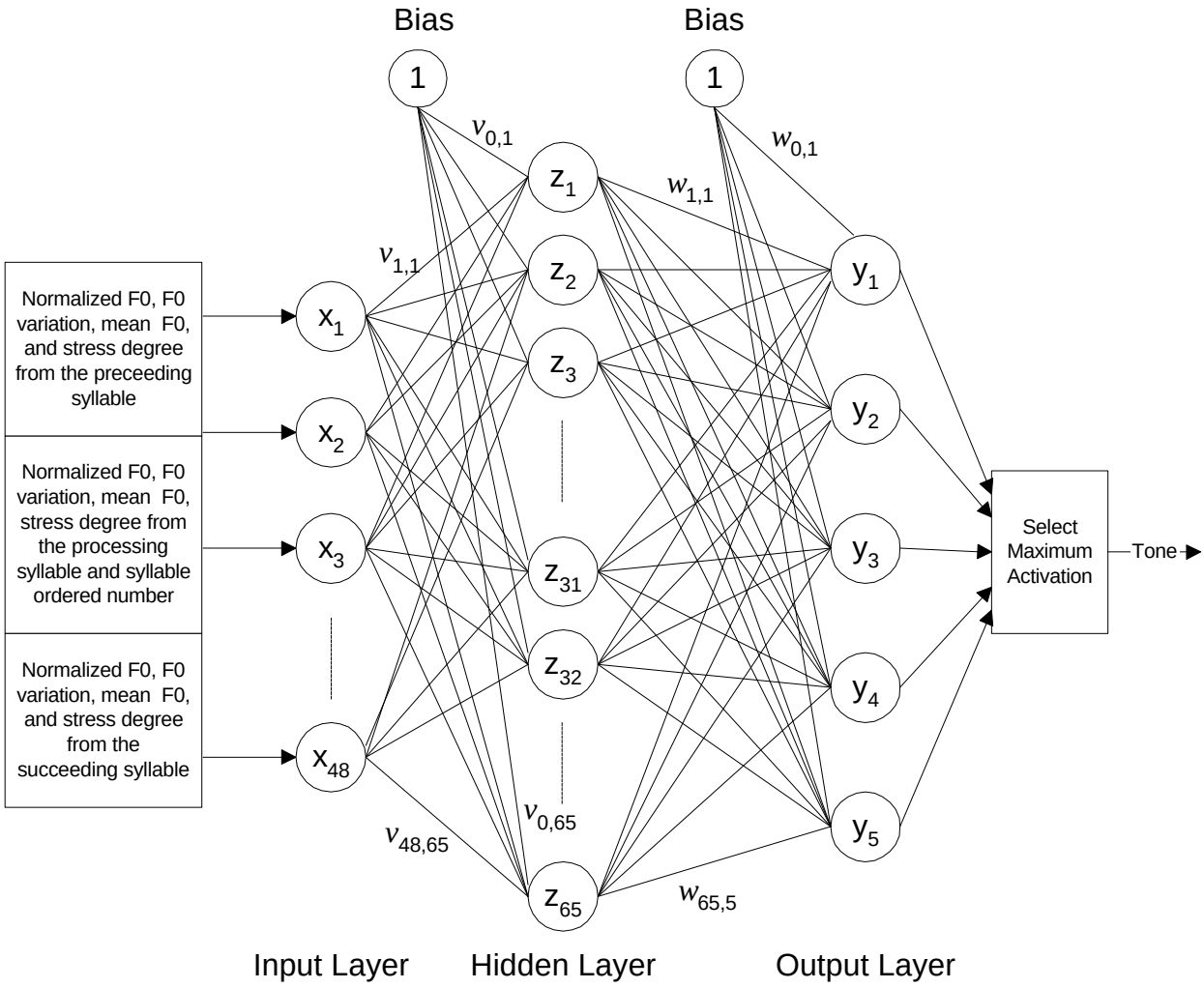

Figure 4.17 Tone Classifier

To train a MLP tone classifier by using a backpropagation method, the features derived from the training utterances as described previously were presented to the input nodes (X_1,…., X_{48}) repetitively. The connection weights (*v*, *w*)

between each layer were then adjusted so as to reduce the errors between the actual output patterns and the desired output patterns. The training process continued until the average error of the training data reached an assigned accuracy, then the tone classifier was regarded as being properly trained and the training process was terminated.

When the MLP tone classifier was completely trained, then its performance was evaluated on one test set which contained 115 sentences for each speaker. Each test utterance was automatically segmented into syllable units before the acoustical features as described previously were extracted from the segmented syllables. These acoustical features were then presented to the input nodes of the tone classifier and the outputs of the tone classifier were generated using the connection weights obtained from the training process. Finally, the recognized tone was selected based on the maximum activation value produced by the output nodes. For example, if the *k*th output node yielded the maximum activation value, then the tone pattern of that syllable was identified as the *k*th tone

4.2 Summary

In this chapter, the details of Thai tone classification system are presented. The proposed system consisted of four modules: preprocessing, syllable segmentation, feature extraction, and tone classifier. The speech signal was first processed by the preprocessing module where the *F0* and modified energy contours were extracted. The syllable segmentation module located the endpoints of the spoken syllables by utilizing the relationships between the peaks and valleys in the modified energy contour. The normalized *F0* and a degree of stress were then extracted from the segmented syllables by the feature extraction module. A multilayer perceptron (MLP) trained by the backpropagation method was employed as a tone classifier. The structure of the tone classifier consisted of 48 input nodes, 65 hidden nodes and five output nodes. In order to deal with tonal coarticulation,

stress, and intonation effects, the normalized *F0*, *F0* variation, mean *F0*, syllable ordering number, and a degree of stress of the preceding, processing, and succeeding syllables were employed to train the tone classifier. For each speaker, a tone classifier was trained by using two training sets, and then tested by one test set. Each training set contained 100 sentences which covered all five Thai tones and stressed patterns whereas a test set contained 115 test sentences. The recognized tone was selected from the output nodes which produced the maximum activation value.

Chapter 5 Results and Observations

In this chapter, the experimental results of syllable segmentation, stress detection, and tone classification are reported. The chapter begins with a discussion of the results of syllable segmentation and stress detection, followed by the results of the tone classification.

5.1 Syllable Segmentation

As mentioned earlier, automatic syllable segmentation was developed in this study due to the unavailability of syllable segmentation of Thai speech. The syllable segmentation method used the modified energy contour and the relationships between the energy of the peaks and valleys to locate the starting and ending points of the syllables. Syllable segmentation was performed on two training sets and one test set; there were a total of 315 sentences per speaker. The results of this automatic syllable segmentation are shown in Table 5.1.

Speaker	Accuracy
SPK1	98.41%
SPK2	97.77%
SPK3	95.87%
SPK4	96.82%
SPK5	98.09%
SPK6	96.51%
SPK7	97.46%
SPK8	96.19%

Table 5.1 The experimental results of automatic syllable segmentation

As seen in Table 5.1, the highest accuracy rate of 98.41% was obtained from SPK1 while the lowest accuracy rate of 95.87% was provided by SPK3. The high accuracy rates achieved by the proposed syllable segmentation resulted mainly from the strategy that employed the modified energy contour and the relationship between peaks and valleys of energy to locate the starting and ending points of the syllables. In many cases, it was evident that speech utterances were automatically segmented into syllables by using an energy modification process alone. Figure 5.1 shows an example of the advantage of using energy modification process. As seen in Figure 5.1, the normalized energy contour contained four energy pulses which were connected but after the energy modification process the energy pulses were segmented into each syllable as shown in the bottom panel.

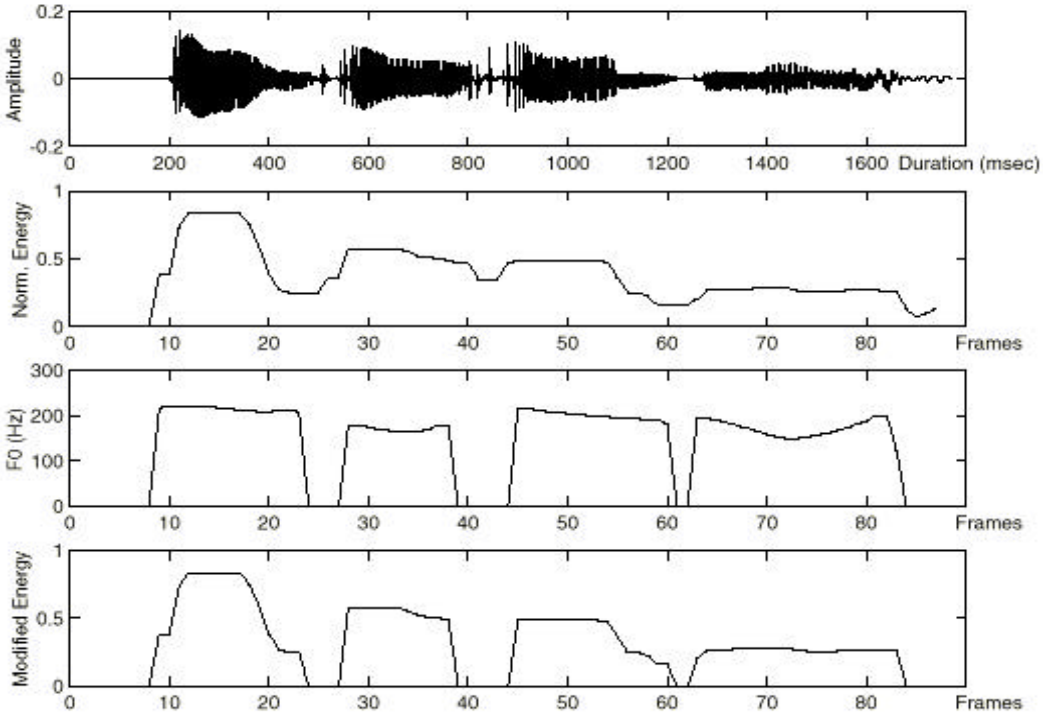

Figure 5.1 A speech utterance was automatically segmented into syllables resulting from the energy modification process.

Other than the least complex case where the syllables were automatically segmented by the energy modification process as described previously, the more complex and the most complex cases that dealt with the syllable segmentation algorithm are displayed in Figures 5.2 and 5.3. From the top to bottom panel of the figures show the speech waveform, modified energy contour, and pulse marker, respectively. Figure 5.3 shows the more complex case where speech utterance contained both the connected and not connected modified energy pulses. Figure 5.4 shows the most complex case where the modified energy pulses were coupled and connected without gaps between them. For these cases, the syllable segmentation algorithm had to find the valleys that appeared between the peaks and marked them as the starting and ending points of the syllables.

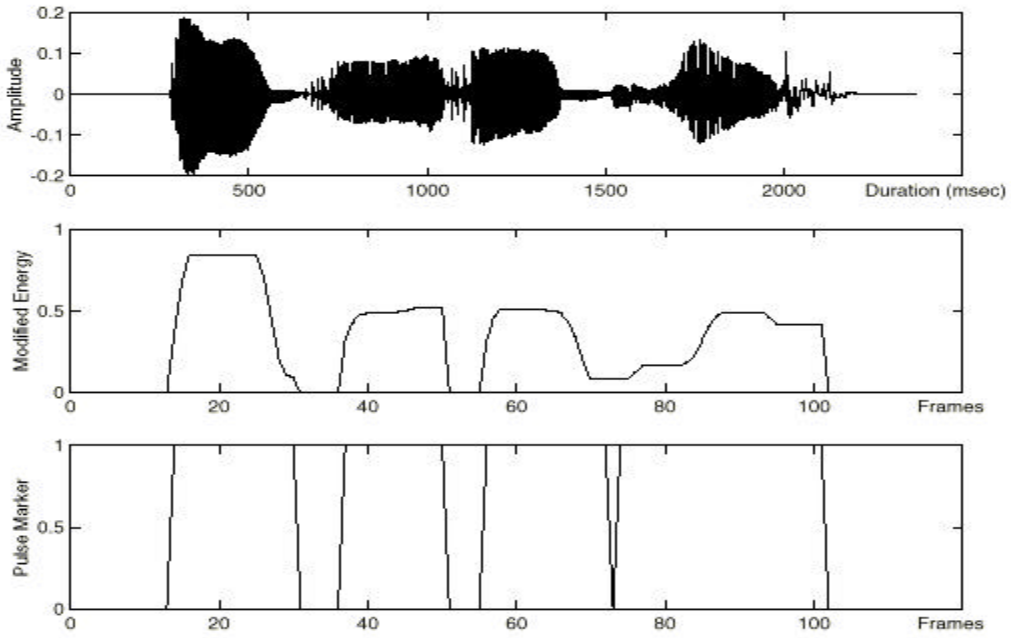

Figure 5.2 An example of a more complex case where the modified energy contour contained both the connected and not connected energy pulses.

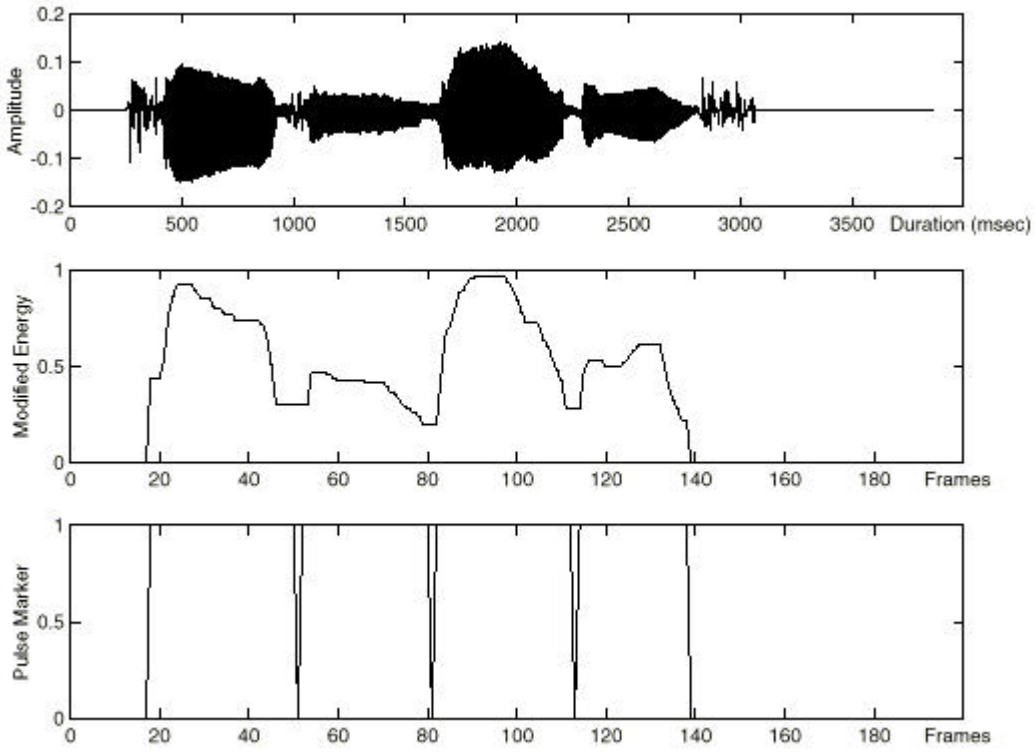

Figure 5.3 An example of the most complex case where the modified energy contour contained all connected energy pulses.

Although syllable segmentation worked very well for most cases, there were some cases in which the syllable segmentation algorithm incorrectly located the endpoints of the syllables. An example of this case can be seen in Figure 5.4 where the algorithm oversegmented the second syllable of the sentence. In this scenario, the speaker stressed the second syllable too much and pronounced it as if there were two syllables instead of one syllable. The result of this exaggeration was that there were two peaks energy for the second syllable (as shown in the second panel of Figure 5.4) which made the syllable segmentation algorithm oversegment the second syllable (as shown in the bottom panel).

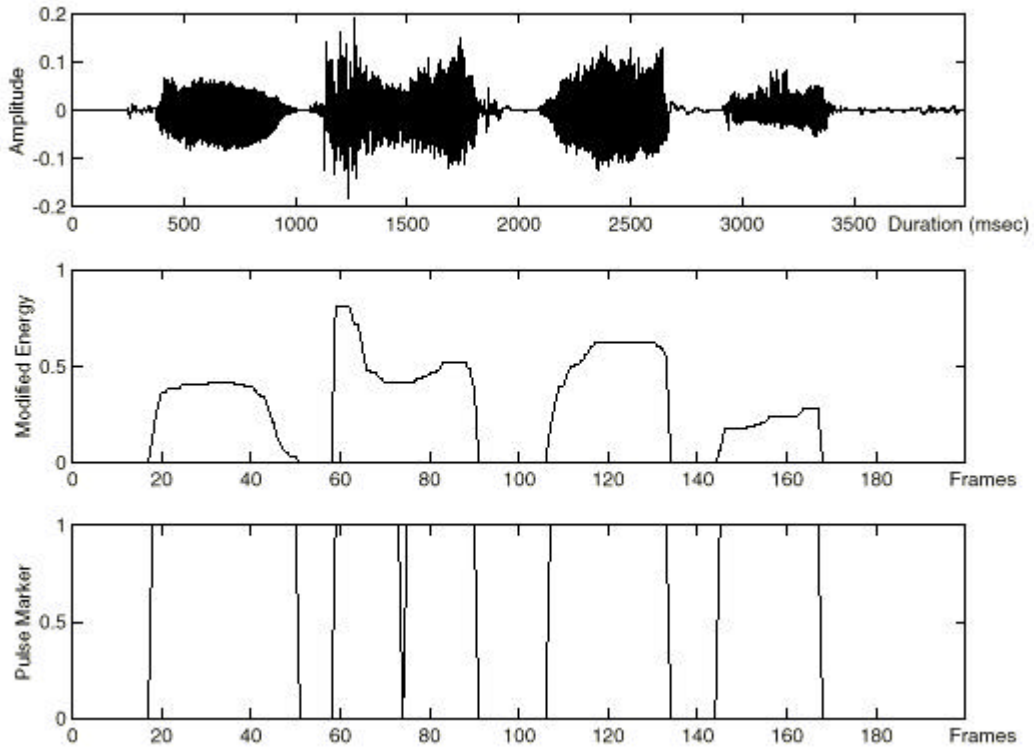

Figure 5.4 The syllable segmentation oversegmented the second syllable of the sentence.

5.2 Stress Detection

As stated in Chapter Two, each speaker was asked to speak the two middle words of a sentence with stressed-stressed, stressed-unstressed, unstressed-stressed, and unstressed-unstressed patterns for each training and test set. The energy and duration were employed as the parameters by the stress detector to determine the degree of stress of the syllables. The performance of the stress detector was evaluated on two training sets and one test set for each speaker. The results of the stress detector are shown in Table 5.2. It should be noted that the results shown in Table 5.2 were obtained from the comparison between the degree of stress of a syllable as determined by the stress detector and the stress pattern of the syllable as uttered by each speaker.

Speaker	Accuracy Rate	
	Unstressed Syllables	Stressed Syllables
SPK1	98.33%	98.67%
SPK2	97.33%	97.67%
SPK3	96.33%	97.0%
SPK4	99.33%	98.67%
SPK5	98.67%	99.0%
SPK6	98.67%	97.33%
SPK7	98.67%	98.67%
SPK8	98.33%	98.0%
Average	98.21%	98.13%

Table 5.2 The experimental results of stress detector

As seen in Table 5.2, the stress detector performed very well for both stressed and unstressed syllables. The stress detector correctly identified the unstressed syllables with the average accuracy rate of 98.21% and correctly identified the stressed syllables with the average accuracy rate of 98.13%. For unstressed syllables, the stress detector achieved the highest accuracy rate of 99.33% for SPK4 while the lowest accuracy rate of 96.33% was obtained from SPK3. For stressed syllables, the stress detector achieved the highest accuracy rate of 99.0% for SPK5 while the lowest accuracy rate of 97.0% was obtained from SPK3. From the high accuracy rates achieved by the stress detector, it was proved that the duration and energy of syllables employed by the stress detector were the effective parameters for discriminating between stressed and unstressed syllables. The source of errors mainly came from the way each speaker uttered the syllables. For example, speakers sometimes did not speak the stressed pattern as they were asked to speak; as in speakers intended to speak stressed syllables but spoke unstressed syllables instead, and vice versa. Another source of error might have come from the result of incorrect syllable segmentation; for example, one stressed syllable was segmented into two unstressed syllables.

5.3 Tone Classification

The performances of a proposed tone classification system are presented in this section. For each speaker, the performance of the tone classifier was evaluated on 115 test sentences. Examples of test sentences which demonstrate how tones have an effect on the meaning of Thai sentences are listed in Table 5.3.

Test Sentences	Pronunciation	Meaning
1. นานเลนวาวมัง	/naan^{2} len^{2} waaw^{2} mang^{2}/	“Nan is playing kite”
2. หมอหนีหมีไหว	/m@@^{4} nii^{4} mii^{4} waj^{4}/	“Doctor can run away from a bear“
3. ตอฆาแกะมาก	/t@@^{1} khaa^{2} kx^{1} maak^{2}/	“Taw killed a lot of sheep”
4. ตอค้าแกะมาก	/t@@^{1} khaa^{3} kx^{1} maak^{2}/	“Taw sold a lot of sheep”
5. โอจับขาอยู่	/?oo^{4} cab^{1} khaa^{4} ?uu^{1}/	“O is holding her leg”
6. โอจับข้าอยู่	/?oo^{4} cab^{1} khaa^{2} ?uu^{1}/	“O is holding me”
7. แมมีน้าดี	/mxx^{2} mii^{0} naa^{3} dii^{0}/	“Mom has a nice aunt”
8. แมมีหน้าดี	/mxx^{2} mii^{0} naa^{2} dii^{0}/	“Mom has a pretty face”
9. โอขึ้นเขาได้	/?oo^{4} khvn^{2} kh@w^{4} daj^{2}/	“O can hike the mountain”
10. โอขึ้นเข่าได้	/?oo^{4} khvn^{2} kh@w^{1} daj^{2}/	“O can knee”
11. อาเข้ามาได้	/?aa^{0} kh@w^{2} maa^{0} daj^{2}/	“Uncle can come in”
12. นานขี่ม้าเกง	/naan^{2} khii^{1} maa^{3} keng^{1}/	“Nan is a good horse rider”
13. นานขี่หมาเกง	/naan^{2} kii^{1} maa^{4} keng^{1}/	“Nan is a good dog rider”
14. หมางับขาแล้ว	/maa^{4} ngab^{3} khaa^{4} lxxw^{3}/	“A dog bit my leg”
15. หมางับข้าแล้ว	/maa^{4} ngab^{3} khaa^{2} lxxw^{3}/	“A dog bit me”

Table 5.3 Examples of test sentences It is noted that the phonemic transcription uses the diacritics /^{0}/, /^{1}/, /^{2}/, /^{3}/ and /^{4}/ as tone markers for mid, low, falling, high and rising tones, respectively.

Table 5.4 shows the experimental results of a proposed tone classification system for each speaker. As seen in Table 5.4, SPK4 had the highest accuracy rate of 93.04% whereas the lowest accuracy rate of 90.0% was obtained from SPK3. The average accuracy rate of 91.36% was achieved by the proposed tone classification system. There were no significant differences in accuracy rates between male (SPK1, SPK2, SPK3, SPK4, SPK5) and female (SPK6, SPK7, SPK8) speakers.

Speaker	Accuracy Rate
SPK1	91.08%
SPK2	90.43%
SPK3	90.0%
SPK4	93.04%
SPK5	91.73%
SPK6	91.52%
SPK7	92.17%
SPK8	90.87%
Average	91.36%

Table 5.4 The experimental results of tone classification system

However, it would be useful to know when the algorithm incorrectly classified tones and identified them as what specific tones. Table 5.5a-h provides the details on how each tone was identified by the tone classifier for each speaker. The first column of each table represents the actual input tones to the system; the second to sixth columns represent the recognized tones, and the total column shows the total number of test syllables for each tone. For example, in Table 5.5a the system correctly recognized the mid tone 73 times from 86 test syllables, identified the mid tone as the low tone 3 times, and identified the mid tone as the falling tone 10 times for SPK1.

Actual Tone	SPK1					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	73	3	10	-	-	86	84.88%
Low	14	79	-	-	-	93	84.94%
Falling	2	-	98	-	-	100	98.0%
High	3	-	1	83	-	87	95.40%
Rising	-	3	-	5	86	94	91.49%
						460	91.08%

(a)

Actual Tone	SPK2					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	73	8	3	2	-	86	84.88%
Low	11	80	-	-	2	93	86.02%
Falling	1	-	97	2	-	100	97.0%
High	3	-	-	79	5	87	90.80%
Rising	1	2	-	4	87	94	92.55%
						460	90.43%

(b)

Actual Tone	SPK3					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	76	7	-	3	-	86	88.37%
Low	11	80	-	-	2	93	86.02%
Falling	1	-	97	2	-	100	97.0%
High	3	-	-	79	5	87	90.80%
Rising	1	6	-	4	83	94	88.29%
						460	90.0%

(c)

Actual Tone	SPK4					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	76	9	-	1	-	86	88.37%
Low	7	84	-	-	2	93	90.32%
Falling	-	-	100	-	-	100	100.0%
High	1	-	-	81	5	87	93.10%
Rising	1	4	-	2	87	94	92.55%
						460	93.04%

(d)

Actual Tone	SPK5					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	78	6	2	-	-	86	90.69%
Low	13	80	-	-	-	93	86.02%
Falling	-	-	99	1	-	100	99.0%
High	4	-	-	82	1	87	94.25%
Rising	1	4	-	6	83	94	88.29%
						460	91.73%

(e)

Actual Tone	SPK6					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	74	7	3	1	1	86	86.05%
Low	5	84	2	-	2	93	90.32%
Falling	1	1	98	-	-	100	98.0%
High	4	-	-	79	4	87	90.81%
Rising	-	4	-	4	86	94	91.49%
						460	91.52%

(f)

Actual Tone	SPK7					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	75	11	-	-	-	86	87.21%
Low	5	87	-	-	1	93	93.55%
Falling	2	-	98	-	-	100	98.0%
High	3	1	-	81	2	87	93.10%
Rising	-	4	-	7	83	94	88.29%
						460	92.17%

(g)

Actual Tone	SPK8					Number of Syllables	Accuracy Rate
	Recognized Tone						
	Mid	Low	Falling	High	Rising		
Mid	74	7	2	3	-	86	86.05%
Low	7	84	-	1	1	93	90.32%
Falling	-	-	99	1	-	100	99.0%
High	7	-	-	75	5	87	86.21%
Rising	-	7	-	1	86	94	91.49%
						460	90.87%

(h)

Table 5.5a-h shows the details of the recognized tones for SPK1, SPK2, SPK3, SPK4, SPK5, SPK6, SPK7, and SPK8, respectively.

As shown in Table 5.5a-h, six speakers obtained the lowest accuracy rate obtained from the mid tone while two speakers had the lowest accuracy rate obtained from the low tone. All speakers achieved the highest accuracy rate from the falling tone. The lowest accuracy rates obtained from the mid tone were 84.88%, 84.88%, 88.37%, 86.05%, 87.21%, and 86.05%, for SPK1, SPK2, SPK4, SPK6, SPK7, and SPK8, respectively. The lowest accuracy rates obtained from the low tone were 86.02% and 86.02%, for SPK3 and SPK5, respectively. The falling tone had the greatest accuracy rate: 98%, 97%, 97%, 100%, 99%, 98%, 99%, and 98% for SPK1, SPK2, SPK3, SPK4, SPK5, SPK6, SPK7, and SPK8, respectively.

For SPK2, SPK4, SPK6, SPK7, and SPK8, the lowest accuracy rates were given by the mid tone where the algorithm most incorrectly identified it as the low tone whereas only for SPK1 did the algorithm most often incorrectly identify it as the falling tone. For SPK3 and SPK5, the lowest accuracy rate was obtained from the low tone where the algorithm mostly incorrectly identified it as the mid tone. This scenario implied that there was a relationship between the mid and low tones when the algorithm incorrectly classified the mid tone and likely recognized the mid tone as the low tone, and vice versa. This situation was due to the fact that the *F0* patterns of the mid and low tones were quite similar in nature and some speakers might have uttered the mid tone very close to the low tone, and vice versa. In addition, the characteristics of both tones were even closer, especially when they were influenced by tonal coarticulation and intonation effects. Regarding the highest accuracy rate, all speakers achieved the highest accuracy rates from the falling tone. The reason that the highest accuracy rates were obtained from the falling tone for all speakers might have resulted from a clearly distinctive characteristic of a falling tone which made it easier for the tone classifier to discriminate between the falling tone and other tones. A comparison of the average accuracy rates of each tone are displayed in Figure 5.5.

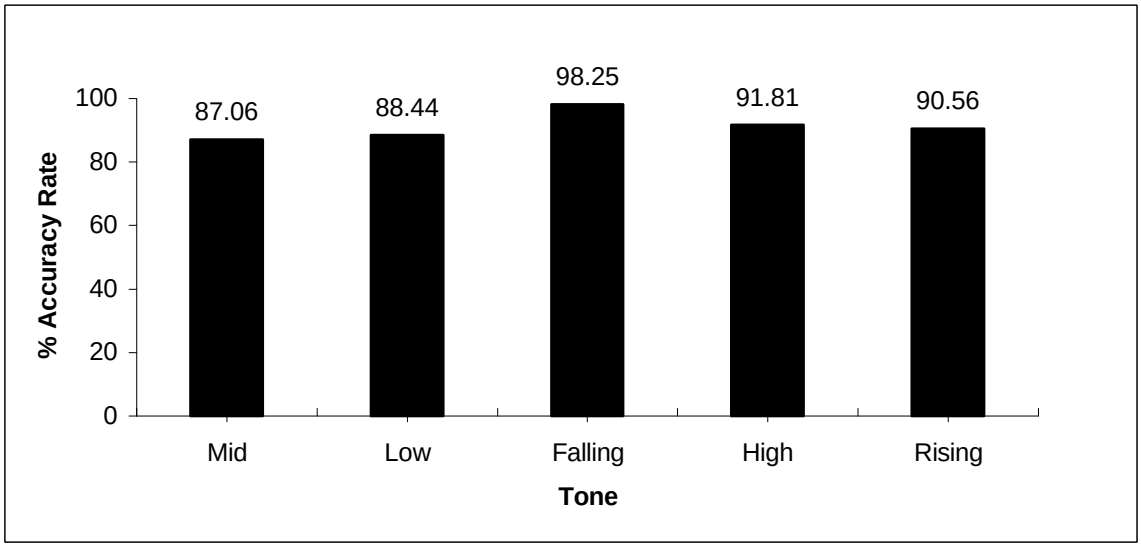

Figure 5.5 Comparison of average accuracy rates for each tone

5.4 Summary

In this chapter, the experimental results of syllable segmentation, stress detection, and tone classification are presented. The syllable segmentation achieved the highest accuracy rate of 98.41% from SPK1 while the lowest accuracy rate of 95.87% was provided by SPK3. The stress detector correctly identified the stressed syllables with an average accuracy rate of 98.13% whereas the average accuracy rate of 98.21% was achieved for the unstressed syllables. These results showed that the duration and energy of syllables employed by the stress detector were effective parameters for distinguishing between stressed and unstressed syllables. The tone classifier achieved an average accuracy rate of 91.36%. For individual speakers, SPK4 had the highest accuracy rate of 93.04% whereas the lowest accuracy rate of 90.0% was obtained from SPK3. The mid tone was the most incorrectly classified tone for six speakers while only two speakers had the lowest accuracy rates for the low tone. All speakers achieved the highest accuracy rates for the falling tone.

Chapter 6 Summary, Conclusions, and Future Work

6.1 Summary and Conclusions

In this study, a tone classification system of syllable-segmented Thai speech based on a multilayer perceptron was developed. This dissertation has made two important contributions to the research on Thai speech recognition. First, automatic syllable segmentation was developed because of its unavailability in Thai speech [76]. Manual syllable segmentation is a tedious and time-consuming task; thus the developed automatic syllable segmentation supported not only this research but should be beneficial to future research on Thai speech recognition as well. Secondly, a proposed tone classification system took into consideration the effects of tonal coarticulation, intonation, as well as stressed and unstressed syllables. Since no previous research on tone classification incorporating these factors, this dissertation contributes pioneering work in this area.

The structure and operation of the proposed tone classification system can be summarized as follows. The tone classification system consisted of four modules: preprocessing, syllable segmentation, feature extraction, and tone classifier. The preprocessing module computed the normalized energy and then the fundamental frequency (*F0*) was extracted from the speech signal. The *F0* extraction was implemented based on the autocorrelation with a three-level center clipping method. The normalized energy contour was adjusted by its *F0* contour in order to produce the modified energy contour which was employed by the syllable segmentation module to locate the starting and ending points of the syllables. The key idea of this method was based on the relationships between the energy of the peaks and valleys in the modified energy contour. The valleys, which appear between the peaks energy, were likely to be the syllable boundaries.

After speech input was segmented into syllables, the *F0*, duration, and energy were extracted from the segmented syllable by the feature extraction module. The extracted *F0* was normalized by using the z-score transformation in order to eliminate the undesirable time and speaker variation. The duration and energy of the segmented syllable were then presented to a stress detector. The stress detector, included in the feature extraction module, was implemented based on a fuzzy inference system (FIS). The membership functions and fuzzy if-then rules were constructed based on the mean of the duration and energy of stressed and unstressed syllables, which were precomputed from the training sets of each speaker. The output of a stress detector was a degree of stress of the syllable which ranged in the interval between 0 and 1.

A tone classifier was developed based on the multilayer perceptron (MLP) which was trained by the backpropagation method. The structure of the tone classifier consisted of 48 input nodes, 65 hidden nodes, and five output nodes each of which represents a particular Thai tone. The tone classifier provided a recognized tone according to the largest activation generated by the output nodes. A tone classifier was trained by using the training sets, which were spoken by five male (SPK1, SPK2, SPK3, SPK4, and SPK5) and three female (SPK6, SPK7, and SPK8) Thai speakers. Each training set consisted of 100 speech sentences in which each sentence contained four monosyllabic words. The two middle words of the training sentences were varied to cover all two-tone combinations and all stressed patterns: unstressed-unstressed, unstressed-stressed, stressed-unstressed, and stressed-stressed. In order to deal with the effects of tonal coarticulation, intonation, stressed and unstressed syllables, a tone classifier was trained by using the training vectors which included the normalized *F0, F0* variation, mean *F0,* and a degree of stress of the preceding, processing, and succeeding syllable, as well as syllable ordered number of the processing syllable. After a tone classifier was completely trained, the performances of the tone classifier were evaluated on one test set (115 sentences

per speaker) which were spoken by the same group of speakers who uttered the training sets.

Although the number of speakers used in this study were small due to a limited population of Thai students, several research studies on speech recognition of Mandarin, Cantonese, and Thai employed a number of speakers ranging from 2 to 12 speakers [13], [14], [30], [50], [51], [53], [76], [89], [111]. Thus the number of speakers used in this study should be considered reasonable and adequate.

The experimental results are summarized as follows. The proposed syllable segmentation achieved the highest accuracy rate of 98.41% from SPK1 while the lowest accuracy rate of 95.87% was obtained from SPK3. The stress detector performed very well for both stressed and unstressed syllables. The stress detector correctly identified the unstressed syllables with an average accuracy rate of 98.21% and correctly identified the stressed syllables with an average accuracy rate of 98.13%. From these results, it was found that the duration and energy of syllables employed by the stress detector were the effective parameters for distinguishing between stressed and unstressed syllables.

The proposed tone classifier achieved an average accuracy rate of 91.36%. Considering the accuracy rate obtained from each speaker, SPK4 had the highest accuracy rate of 93.04% whereas the lowest accuracy rate of 90.0% was obtained from SPK3. The mid tone was the tone most incorrectly classified for six speakers (SPK2, SPK4, SPK6, SPK7, and SPK8) where the algorithm incorrectly identified it as the low tone. The low tone was the tone most incorrectly classified for two speakers (SPK3 and SPK5) where the algorithm incorrectly identified it as the mid tone. These errors were due to the fact that the *F0* patterns of the mid and low tones were quite similar in nature and some speakers actually uttered the mid tone very close to the low tone, and vice versa. In addition, the characteristics of both tones were even closer, especially when they were influenced by tonal coarticulation

and intonation effects. All speakers achieved the highest accuracy rates from the falling tone. The reason that the highest accuracy rates were obtained from the falling tone for all speakers might be the result of the clearly distinctive characteristic of a falling tone, which made it easier for the system to discriminate between the falling tone and other tones.

Currently, there have been only a few research studies conducted on the tone classification of Thai speech and the most current research on this topic was conducted by Potisuk *et al*. [76]. In their studies, tonal coarticulation and intonation effects were considered but only stressed syllables were used. Potisuk's system was evaluated on 55 test sentences uttered by five Thai speakers who also spoke the training set. The test utterances were manually segmented into syllable units before the tone classification was performed. Potisuk's system achieved a recognition rate of 89.1%.

In comparison with Potisuk's research, this study used more speakers and test sentences for system evaluation and achieved a higher recognition rate of 91.36%. In addition, the effects of tonal coarticulation and intonation as well as stressed and unstressed syllables were taken into consideration, whereas only stressed syllables were used in Potisuk's research. Furthermore, automatic syllable segmentation was developed in this study while the test utterances were manually segmented into syllables in Potisuk's system.

6.2 Future Work

1. In this study, a proposed syllable segmentation was tested by using speech which was recorded in a quiet environment. The algorithm was reasonably robust and worked well under this condition. However, this ideal condition does not always exist, so it is useful to extend this study to other conditions such as in noisy environments (e.g., an office-like environment). In general, the accurate detection of the boundaries of a speech utterance is crucial for reliable and

robust automatic speech recognition. Noisy conditions could create problems for the syllable segmentation task and degrade the tone classifier's performance dramatically. For this reason, special algorithms or methods are needed to overcome these difficulties so that the syllable segmentation is more robust to changes in the background noise environment.

2. The proposed tone classification system was trained and tested with only one type of sentence structure which was "subject + verb + object + post-verb auxiliary" and contained four monosyllabic words. A future study on this topic could be conducted by experimenting with other types of sentence structure such as a longer sentence, which contains more than four monosyllabic words. In addition, it could also be useful to extend the speaker-independent task where the tone classifier is trained by a group of speakers and then tested with a different group of speakers who do not train the tone classifier.

3. A totally automatic Thai speech recognition system is an ultimate goal for Thai speech researcher. The tone classification system developed in this research study could be incorporated in the future with a base-syllable recognition system to complete such an automatic Thai speech recognition system. Numerous potential applications await this ultimate goal, including Thai speech-to-text and Thai-English translation, would greatly benefit the Thai people.

References

[1] Abramson, A.S. (1962). The vowels and tones of standard Thai: Acoustical measurements and experiments (Publication No. 20). Bloomington, IN: Indiana University Research Center in Anthropology, Folklore, and Linguistics, International Journal of America Linguistics, 28(2), Part III.

[2] Abramson, A.S. (1978). Static and dynamic acoustic cues in distinctive tones,Language and Speech, 21, 319-325.

[3] Abramson, A.S. & Ren N. (1990). Distinctive vowel length: Duration V.S. Spectrum in Thai, Journal of Phonetics, 18, 79-92.

[4] Ahkuputra, V., Jitapunkul, S., Pornsukchandra, W. & Luksaneeyanawin, S. (1997). A speaker independent Thai polysyllabic word recognition system using hidden Markov model, 1997 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, xxiii, 593-599 vol.2.

[5] Arkuputra, V., Jitapunkul, S., Maneenoi, E., Kasuriya, S. & Amornkul, P. (1998). Comparison of different techniques on Thai speech recognition, 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings., xxvi, 177-180.

[6] Albesano, D., Gemello, R. & Mana, F. (2000). Hybrid HMM-NN modeling of stationary-transitional units for continuous speech recognition, Information Sciences, 123(1-2), 3-11.

[7] Areepongsa, S. (1995). Speaker independent Thai numeral speech recognition by hidden Markov model and vector quantization, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[8] Ayudhya, P.N.N., Khawparisuth, D. & Chamnongthai, K. (1999). Speaker-independent isolated Thai consonant recognition by using wavelet and simulated auditory system, 1999 IEEE International Symposium on Intelligent Signal Processing and Communication Systems. Signal Processing and Communications Beyond 2000, Thailand, xxvii, 797-800.

[9] Bahl, L. R., Jelinek, F. & Mercer, R. L. (1983). A maximum likelihood approach to continuous speech recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-5(2), 179-190.

[10] Cerf, P. L., Ma, W. & Compernolle, D. V. (1994). Multilayer perceptrons as labelers for hidden Markov models, IEEE Transactions on Speech and Audio Processing, 2(1) Part II, 185-193.

[11] Chang, P. C. & Chen, S. H., Sun, S. W. (1990). Mandarin tone recognition by multilayer perceptron, Proceeding IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP), 517-520.

[12] Chen, S. H. & Wang, Y. R. (1995). Tone recognition of continuous Mandarin speech based on neural networks, IEEE Transactions on Speech and Audio Processing, 3(2), 146-150.

[13] Chen, S. H. & Liao, Y. F. (1998). Modular recurrent neural networks for Mandarin syllable recognition, IEEE Transactions on Neural Networks, 9(6), 1430-1441.

[14] Chen, W. Y., Liao Y. F. & Chen, S. H. (1995). Speech recognition with hierarchical recurrent neural networks, Pattern Recognition, 28(6), 795-805.

[15] Davis, S. B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), 357-366.

[16] Dubnowski, J. J., Schafer R. W. & Rabiner, L. R. (1976). Real-time digital hardware pitch detector, IEEE Transactions on Acoustics, Speech and Signal Processing, 24(1), 2-8.

[17] Dugast, C., Devillers, L. & Aubert, X. (1994). Combining TDNN and HMM in a hybrid system for improved continuous-speech recognition, IEEE Transactions on Speech and Audio Processing, 2(1) Part II, 217-223.

[18] Evans, J. S. & Krishnamurthy, V. (1999). Hidden Markov model state estimation with randomly delayed observations, IEEE Transactions on Signal Processing, 47(8), 2157-2166.

[19] Fausett, L. (1994). Fundamentals of Neural Networks Architectures, Algorithms, and Applications, Prentice Hall, New Jersey.

[20] Fucci, D. & Lass, N. (1999). Fundamental of Speech Science, Allyn and Bacon, Massachusetts.

[21] Furui S. (1989). Unsupervised speaker adaptation based on hierachical spectral clustering IEEE Transactions on Acoustics, Speech and Signal Processing, 37(12), 1923-1930.

[22] Furui S. (1989). Digital Speech Processsing, Synthesis, and Recognition, Dekker, New York.

[23] Gandour, J. (1974). Consonant types and tones in Siamese, Journal of Phonetics, 2, 337-350.

[24] Gandour, J. & Harshmann, R. (1978). Cross-language study of tone perception,Linguistic Variations-Models and Methods, Academic Press, 139-147.

[25] Gandour, J. (1983). Tone perception in far eastern languages, Journal of Phonetics, 11, 149-175.

[26] Gandour, J., Potisuk, S., & Dechangkit, S. (1991). Inter-and intraspeaker variability in fundamental frequency of Thai tone, Speech Communication, 10, 355-372.

[27] Gandour, J., Potisuk, S., & Dechangkit, S. (1994). Tone coarticulation in Thai,Journal of Phonetics, 22, 477-492.

[28] Gavat, I., Zirra, M. & Cula, O. (1998). Hybrid speech recognition system with discriminative training applied for Romanian language, MELECON'98. 9th Mediterranean Electrotechnical Conference. Proceedings, xviii, 11-15 vol.1.

[29] Gray, R. M., Buzo, A., Gray jr., A. H. & Matsuyama, Y. (1980). Distortion measures for speech processing, IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), 367-376.

[30] Hasan, M. K. (1997). Tone recognition of speech using hidden Markov models, Master's Thesis, Asian Institue of Technology, Thailand.

[31] Hermes, D. & van Gestel, J. (1991). The frequency scale of speech intonation, Journal Acoustical Society of America, 90, 97-102.

[32] Huang, X. & Lee, K. F. (1993). On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition, IEEE Transactions on Speech and Audio Processing, 1(2), 150-157.

[33] Huang, E. F. & Wang, H. C. (1994). An efficient algorithm for syllable hypothesization in continuous Mandarin speech recognition, IEEE Transactions on Speech and Audio Processing, 2(3), 446-452.

[34] Huo, Q., Chan, C. & Lee, C. H. (1995). Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition, IEEE Transactions on Speech and Audio Processing, 3(5), 334-345.

[35] Huo, Q. & Lee, C. H. (1997). On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate, IEEE Transactions on Speech and Audio Processing, 5(2), 161-172.

[36] Hwang, S. H. & Chen, S. H. (1994). Neural-network-based F0 text-to-speech synthesiser for Mandarin, IEE Proceedings Vision, Image and Signal Processing, 141(6), 384-390.

[37] Jang, J. S. R, Sun, C. T. & Mizutani, E. (1997). Neuro-Fuzzy and Soft Computing, Prentice Hall, New Jercy.

[38] Jitapunkul, S., Luksaneeyanawin, S., Arkuputra, V., Maneenoi, E., Kasuriya, S. & Amornkul, P. (1998). Recent advances of Thai speech recognition in Thailand, 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings., xxvi, 173-176.

[39] Juang, B. H. (1984). Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains, AT&T Technical Journal, 64(6), 1235-1249.

[40] Juang, B. H. (1984). On the hidden Markov model and dynamic time warping for speech recognition : A unified view, AT&T Technical Journal, 63(7), 1213-1243.

[41] Jun, L., Zhu X. & Luo, Y. (1998). An approach to smooth fundamental frequencies in tone recognition, International Conference on Communication Technology Beijing, China, October 22-24.

[42] Katagiri, S. & Lee, C. H. (1993). A hybrid algorithm for speech recognition based on HMM segmentation and learning vector quantization, IEEE Transactions on Speech and Audio Processing, 1(4), 421-430.

[43] Kongkachandra, R., Tamee, K. & Kimpan, C. (1999). Improving Thai isolated word recognition by using Karhunen_Loeve transformation and learning vector quantization, 1999 IEEE International Symposium on Intelligent Signal Processing and Communication Systems. Signal Processing and Communications Beyond 2000, Thailand, xxvii, 777-780.

[44] Kongsupanich, S. (1997). The transformation of Thai morphemes to phonetic symbols for Thai speech synthesis syste, Master's Thesis, King Mongkut's Institute of Technology Ladkrabang, Thailand. (In Thai)

[45] Kundu, A. & Bayya, A. (1998). Speech recognition using hybrid hidden Markov model and NN classifier, International Journal of Speech Technology, 2(3), 227-240.

[46] Lee, K. F. & Hon H. W. (1989). Speaker-independent phone recognition using hidden Markov models, IEEE Transactions on Acoustics, Speech and Signal Processing, 37(11), 1641-1648.

[47] Lee, K. F., Hon H. W. & Reddy, R. (1990). An overview of the SPHINX speech recognition system, IEEE Transactions on Acoustics, Speech and Signal Processing, 38(1), 35-45.

[48] Lee, L. S. (1997). Voice dictation of Mandarin Chinese, IEEE Signal Processing Magazine, 14(4), 63-101.

[49] Lee, L. S., Tseng, C. Y., Gu, H. Y., Liu, F. H., Chang, C. H., Lin, Y. H., Lee, Y., Tu, S. L., Hsieh, S. H., & Chen, C. H. (1993). Golden Mandarin (I) A real-time Mandarin speech dictation machine for Chinese language with very large vocabulary, IEEE Transactions on Speech and Audio Processing, 1(2), 158-177.

[50] Lee, T., Ching, L. W., Cheng, Y. H. & Mak, B. (1995). Tone recognition of isolated Cantonese syllables, IEEE Transactions on Speech and Audio Processing, 3(3), 204-209.

[51] Lee, T. & Ching, P. C. (1999). Cantonese syllable recognition using neural networks, IEEE Transactions on Speech and Audio Processing, 7(4), 466-472.

[52] Lee, Y. & Lee, L. S. (1993). Continuous hidden Markov models integrating transitional and instantaneous features for Mandarin syllable recognition, Computer Speech and Language, 7, 247-263.

[53] Lee, Y., Lee, L. S., & Tseng, C. Y. (1997). Isolated Mandarin syllable recognition with limited training data specially considering the effect of tones, IEEE Transactions on Speech and Audio Processing, 5(1), 75-80.

[54] Levinson, S. E., Rabiner, L. R., Rosenberg, A. E. & Wilpon, J. G. (1979). Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 27(2), 134-141.

[55] Lin, C. T., Nein H. W. & Lin, W. F. (1999). Speaker adaptation of fuzzy-peceptron-based speech recognition, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 7(1), 1-30.

[56] Lippmann, R. P. (1987). An introduction to computing with neural nets, IEEE ASSP Magazine, 4, 4-22.

[57] Liu, F. H., Lee, Y. & Lee, L. S. (1993). A direct-concatenation approach to train hidden Markov models to recognize the highly confusing Mandarin syllables with very limited training data, IEEE Transactions on Speech and Audio Processing, 1(1), 113-119.

[58] Lleida, E. & Rose, R. C. (2000). Utterance verification in continuous speech recognition: decoding and training procedures, IEEE Transactions on Speech and Audio Processing, 8(2), 126-139.

[59] Luksaneeyanawin, S. (1993). Speech computing and speech technology in

Thailand, Proceedings of the 1993 Symposium on Natural Language Processing in Thailand, Chulalongkorn University, 276-321.

[60] Lyu, R. Y., Hong, J. C., Shen, J. L., Lee, M. Y. & Lee, L. S. (1998). Isolated Mandarin Base-syllable recognition based upon the segmental probability model, IEEE Transactions on Speech and Audio Processing, 6(3), 293-299.

[61] Markel, J. D. (1972). The SIFT algorithm for fundamental frequency estimation, IEEE Transactions on Audio and Electroacoustics, 20(5), 367-377.

[62] Naksakul, K. (1998). Thai Phonology, Chulalongkorn university Press, Thailand.

[63] Nieuwoudt, C. & Botha, E. C. (1999). Connected digit recognition in Afrikaans using hidden Markov models, South African Computer Journal, 23, 85-91.

[64] O'Shaughnessy, D. (1987). Speech Communication Human and Machine, Addison-Wesley, USA.

[65] Pan, K. C., Soong, F. K. & Rabiner, L. R. (1985). A vector-quantization-based preprocessor for speaker-independent isolated word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 33(3), 546-560.

[66] Pekanan, S. (1982). Development of a digital encoder from voice signal, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[67] Pensiri, R. (1995). Speaker-independent Thai numerical voice recognition by using dynamic time warping, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[68] Phatrapornnant, T. (1995). Speaker-independent isolated Thai spoken vowel recognition using spectrum distance measurement and dynamic time warping, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[69] Phinicharom, R. (1991). Acoustic characteristics of unstressed syllables in Thai, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[70] Picone, J. (1990). Continuous speech recognition using hidden Markov models, IEEE ASSP Magazine, 7, 26-41.

[71] Picone, J. (1993). Signal modeling techniques in speech recognition, Proceedings of the IEEE, 81(9), 1215-1247.

[72] Pornsukjantra, W. (1996). Speaker-independent Thai numeral speech recognition using LPC and the back propagation neural network, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[73] Potisuk, S. (1995). Prosodic disambiguation in automatic speech understanding of Thai, Docteral Dissertation, Purdue University, Indiana.

[74] Potisuk, S., Gandour, J. & Harper, M. (1996). Acoustic correlates of stress in Thai, Phonetica, 53, 200-220.

[75] Potisuk, S., Gandour, J. & Harper, M. (1997). Contextual variations in trisyllabic sequences of Thai tones, Phonetica, 54, 22-42.

[76] Potisuk, S., Harper, M. & Gandour, J. (1999). Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method, IEEE Transactions on Speech and Audio Processing, 7(1), 95-102.

[77] Prathumthan, T. (1986). Thai speech recognition using syllable units, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[78] Rabiner, L. R., Cheng, M. J., Rosenberg, A. E. & Mcgonegal C. A. (1976). A comparative performance study of several pitch detection algorithms, IEEE Transactions on Acoustics, Speech and Signal Processing, 24(5), 399-418.

[79] Rabiner, L. R. (1977). On the use of autocorrelation analysis for pitch detection, IEEE Transactions on Acoustics, Speech and Signal Processing, 25(1), 24-33.

[80] Rabiner, L. R. & Schafer, R. W. (1978). Digital Processing of Speech Signal, Prentice Hall, New Jersey.

[81] Rabiner, L. R. & Schmidt, C. E. (1980). Application of dynamic time warping to connected digit recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 28(4), 377-388.

[82] Rabiner, L. R., Juang, B. H., Levinson, S. E. & Sondhi, M. M. (1985). Recognition of isolated digits using hidden Markov models with continous mixture densities, AT&T Technical Journal, 64(6), 1211-1234.

[83] Rabiner, L. R., Juang, B. H., Levinson, S. E. & Sondhi, M. M. (1985). Some properties of continuous hidden Markov model representations, AT&T Technical Journal, 64(6), 1251-1269.

[84] Rabiner, L. R. & Juang, B. H. (1986). An introduction to hidden Markov models, IEEE ASSP Magazine, 3(1), 4-16.

[85] Rabiner, L. R., Wilpon, J. G. & Juang, B. H. (1986). A segmental k-means training procedure for connected word recognition, AT&T Technical Journal, 65(3), 21-31.

[86] Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 77(2), 257-285.

[87] Rabiner, L. R., Wilpon, J. G. & Soong, F. K. (1989). High performance connected digit recognition using hidden Markov models, IEEE Transactions on Acoustics, Speech and Signal Processing, 37(8), 1214-1225.

[88] Rabiner, L. R. & Juang, B. H. (1993). Fundamental of Speech Recognition, Prentice Hall, New Jersey.

[89] Ramalingam, H. (1995). Extraction of tones of speech : An application to the Thai language, Master's Thesis, Asian Institue of Technology, Thailand.

[90] Renals, S. & Hochberg, M. M. (1999). Start synchronous search for large vocabulary continuous speech recognition, IEEE Transactions on Speech and Audio Processing, 7(5), 542-553.

[91] Rigoll, G. (1994). Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition systems, IEEE Transactions on Speech and Audio Processing, 2(1) Part II, 175-184.

[92] Robinson, A. J. (1994). An application of recurrent nets to phone probability estimation, IEEE Transactions on Neural Networks, 5(2), 175-184.

[93] Rose, M. J., Shaffer, H. L., Cohen, A., Freudberg, R. & Manley, H. J. (1974). Average magnitude difference function pitch extractor, IEEE Transactions on Acoustics, Speech and Signal Processing, 22(5), 353-362.

[94] Rose, P. (1987). Considerations in the normalisation of the fundamental frequency of linguistic tone, Speech Communication, 6, 343-351.

[95] Sambur, M. R. & Rabiner, L. R. (1976). A statistical decision approach to the recognition of connected digits, IEEE Transactions on Acoustics, Speech and Signal Processing, 24(6), 550-558.

[96] Shen, J. L. (1998). Continuous Mandarin speech recognition for Chinese language with large vocabulary based on segmental probability model, IEE Proceedings on Vision, Image and Signal Processing, 145(5), 309-315.

[97] Shen, J. L. (1998). Segmental probability distribution model approach for isolated Mandarin syllable recognition, IEE Proceedings on Vision, Image and Signal Processing, 145(6), 384-390.

[98] Shen, X. S. (1990). Tonal coarticulation in Mandarin, Journal of Phonetics, 18, 281-295.

[99] Sondhi, M. M. (1968). New methods of pitch extraction, IEEE Transactions on Audio and Electroacoustics, 16(2), 262-266.

[100] Sriraksa, U. (1995). Acoustic charateristics signalling syllable boundary in Thai connected speech, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[101] Sripramong, T. (1994). Thai speech analysis in harmonic-frequency domain, Master's Thesis, King Mongkut's Institute of Technology Ladkrabang, Thailand. (In Thai)

[102] Stearns, S. D. & David, R. A. (1996). Signal Processing Algorithms in MATLAB, Prentice Hall, New Jersey.

[103] Thamphothong, P. (1990). Multispeaker speech recognition system, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[104] Thubthong, N. (1995). A Thai speech recognition system based on phonemic distinctive features, Master's Thesis, Chulalongkorn university, Thailand. (In Thai)

[105] Thubthong, N. & Kijsirikul, B. (1999). A syllable-based connected Thai digit speech recognition using neural network and duration modeling, 1999 IEEE International Symposium on Intelligent Signal Processing and Communication Systems. Signal Processing and Communications Beyond 2000, Thailand, xxvii, 785-788.

[106] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K. & Lang, K. J. (1989). Phoneme recognition using time-delay neural networks, IEEE Transactions on Acoustics, Speech and Signal Processing, 37(3), 328-339.

[107] Wang, H. S., Ho, T. H., Yang, R. C., Shen, J. L., Bai, B. R., Hong, J. C., Chen, W. P., Yu, T. L. & Lee, L. S. (1997). Complete recognition of continuous Mandarin speech for Chinese language with very large vocabulary using limited training data, IEEE Transactions on Speech and Audio Processing, 5(2), 195-200.

[108] Wu, C. H. (1997). Subsyllable-based discriminative segmental Bayesian network for Mandarin speech keyword spotting, IEE Proceedings on Vision, Image and Signal Processing, 144(2), 65-71.

[109] Wutiwiwatchai, C., Jitapunkul, S., Luksaneeyanawin, S., & Arkuputra, V. (1998). Thai polysyllabic word recognition using fuzzy-neural network , 1998 IEEE Second Workshop on Multimedia Signal Processing., xvii, 137-142.

[110] Wutiwiwatchai, C., Jitapunkul, S., Arkuputra, V., Maneenoi, E., Amornkul, P. & Luksaneeyanawin, S. (1998). A new strategy of fuzzy-neural network for Thai numeral speech recognition, 1998 IEEE Asia-Pacific Conference on Circuits and Systems. Microelectronics and Integrating Systems. Proceedings., xxvi, 161-164.

[111] Yang, W. J., Lee, J. C., Chang, Y. C. & Wang, H. C. (1988). Hidden Markov model for Mandarin tone recognition, IEEE Transactions on Acoustics, Speech and Signal Processing, 36(7), 988-992.

[112] Yoon, S. M., Jung, K. C., Park, M. H. & Kim, H. J. (1997). Korean speech vector quantization using a continuous hidden Markov model, Proceeding of IEEE TENCON'97. IEEE Region 10 Annual Conference. Speech and Image Technologies for Computing and Telecommunications, 2 vol.,xxiv, 249-252 vol.1.

[113] Zahorian, S. A. & Nossair, Z. B. (1999). A partitioned neural network approach for vowel classification using smoothed time/frequency features, IEEE Transactions on Speech and Audio Processing, 7(4), 414-425.

[114] Zavaliagkos, G., Zhao, Y., Schwartz, R. & Makhoul, J. (1994). A hybrid segmental neural net/hidden Markov model system for continuous speech recognition, IEEE Transactions on Speech and Audio Processing, 2(1) Part II, 151-159.

Appendix A

A.1 Thai Syllable Structure

There are five types of Thai syllable structures. A Thai syllable may occur by itself; however, it will be considered as a word if it has a meaning. The syllable structure is classified as follows [62].

1. $C(C)VV^{0-4}$
2. $C(C)VN^{0-4}$
3. $C(C)VS^{1,3}$
4. $C(C)VVN^{0-4}$
5. $C(C)VVS^{1,2}$

Note that: the phonemic transcription uses /C/, /CC/, /V/, /VV/, /N/, and /S/ as initial consonant, consonant cluster, short vowel, long vowel, ending nasal consonant, and ending stop consonant, respectively. The superscript $/^{0}/$, $/^{1}/$, $/^{2}/$, $/^{3}/$, $/^{4}/$ represent the tone marker of mid, low, falling, high, and rising, respectively.

1. $C(C)VV^{0-4}$

This syllable structure contains an initial consonant (C) or consonant cluster (CC), and a long vowel (VV) or diphthong. This syllable type can have any one of five tones. Examples of this syllable structure are shown below:

CVV^{0}	$/naa^{0}/$	“นา”
	$/phxx^{0}/$	“แพ”
$CCVV^{0}$	$/khruu^{0}/$	“ครู”
	$/plaa^{0}/$	“ปลา”
CVV^{1}	$/svv^{1}/$	“สื่อ”

	/jaa^{1}/	“อย่า”
CCVV^{1}	/tree^{1}/	“เตร่”
	/praa^{1}/	“ปร่า”
CVV^{2}	/naa^{2}/	“น้า”
	/ruua^{2}/	“รั่ว”
CCVV^{2}	/khruu^{2}/	“ครู่”
	/phlaa^{2}/	“พล่า”
CVV^{3}	/khaa^{3}/	“ค้า”
	/svv^{3}/	“ซื้อ”
CCVV^{3}	/khwaa^{3}/	“คว้า”
	/phraa^{3}/	“พร้า”
CVV^{4}	/sii^{4}/	“สี”
	/huu^{4}/	“หู”
CCVV^{4}	/khruua^{4}/	“ขรัว”
	/khlaa^{4}/	“ขลา”

2. C(C)VN$^{0-4}$

This syllable structure contains an initial consonant (C) or consonant cluster (CC), a short vowel (V), and an ending nasal or semivowel consonant (N). This syllable type can have any one of five tones. Examples of this syllable structure are shown below:

CVN^{0}	/pen^{0}/	“เป็น”
	/kham^{0}/	“คำ”

$CCVN^0$	/khlang^0/	“คลัง”
	/khwan^0/	“ควัน”
CVN^1	/tam^1/	“ต่ำ”
	/lang^1/	“หลั่ง”
$CCVN^1$	/klin^1/	“กลิ่น”
	/pleng^1/	“เปล่ง”
CVN^2	/nvng^2/	“นึ่ง”
	/ten^2/	“เต้น”
$CCVN^2$	/khrang^2/	“ครั่ง”
	/phrung^2/	“พรุ่ง”
CVN^3	/chin^3/	“ชิ้น”
	/wun^3/	“วุ้น”
$CCVN^3$	/phliw^3/	“พลิ้ว”
	/khrang^3/	“ครั้ง”
CVN^4	/hiw^4/	“หิว”
	/saj^4/	“ใส”
$CCVN^4$	/khlaw^4/	“ขวั ”
	/khlang^4/	“ขลัง”

3. $C(C)VS^{1,3}$

This syllable structure contains an initial consonant (C) or consonant cluster (CC), a short vowel (V), and an ending stop consonant (S). The tone of this syllable type is either low or high tone. Examples of this syllable structure are shown below:

CVS^{1}	/phit^{1}/	“ผิด”
	/lak^{1}/	“หลัก”
CCVS^{1}	/krup^{1}/	“กรุบ”
	/phli?^{1}/	“ผลิ”
CVS^{3}	/khvk^{3}/	“คึก”
	/rat^{3}/	“รัด”
CCVS^{3}	/phlik^{3}/	“พลิก”
	/khwak^{3}/	“ควัก”

4. C(C)VVN$^{0-4}$

This syllable structure contains an initial consonant (C) or consonant cluster (CC), a long vowel (VV) or diphthong, and an ending nasal or semivowel consonant (N). This syllable type can have any one of five tones. Examples of this syllable structure are shown below:

CVVN^{0}	/phaan^{0}/	“พาน”
	/laaj^{0}/	“ลาย”
CCVVN^{0}	/plaaj^{0}/	“ปลาย”
	/phlaang^{0}/	“พลาง”
CVVN^{1}	/paan^{1}/	“ป่าน”
	/dvvm^{1}/	“ด่วน”
CCVVN^{1}	/plian^{1}/	“เปลี่ยน”
	/pliaw^{1}/	“เปลี่ยว”
CVVN^{2}	/ruang^{2}/	“ร่วง”

	/laam^{2}/	“ล่าม”
CCVVN^{2}	/khlvvan^{2}/	“เคลื่อน”
	/khrvvang^{2}/	“เครื่อง”
CVVN^{3}	/laang^{3}/	“ล้าง”
	/th@@ng^{3}/	“ท้อง”
CCVVN^{3}	/khwaan^{3}/	“ค้วาน”
	/phr@@m^{3}/	“พร้อม”
CVVN^{4}	/haaj^{4}/	“หาย”
	/khiaw^{4}/	“เขียว”
CCVVN^{4}	/khwaan^{4}/	“ขวาน”
	/khwxxng^{4}/	“แขวง”

5. C(C)VVS^{1,2}

This syllable structure contains an initial consonant (C) or consonant cluster (CC), a long vowel or diphthong (VV), and an ending stop consonant (S). Tone of this syllable type is either low or falling tone. Examples of this syllable structure are shown below:

CVVS^{1}	/laak^{1}/	“หลาก”
	/buak^{1}/	“บวก”
CCVVS^{1}	/praap^{1}/	“ปราบ”
	/pleet^{1}/	“เปรต”
CVVS^{2}	/riap^{2}/	“เรียบ”
	/luak^{2}/	“ลวก”

CCVVS^{2}	/khlaat^{2}/	“คลาด”
	/khraap^{2}/	“คราบ”

The consonant in the initial position of Thai syllables can be a double initial consonant as well. There are about 2,600 Thai words with double initial consonants. The double initial consonants can be classified into four different types: true cluster, leading consonant characters, parallel consonant characters, and pseudo cluster [59].

1. True cluster

   The true cluster is a double initial consonant in which the first consonant letter is represented by one of the following consonants; /p, ph, k, kh, t, th/ and the second letter can only be /r/ or /l/ or /w/. However, there are only 12 admissible true clusters in Thai; /pr, phr, pl, phl, tr, thr, kr, khr, kl, khl, kw, khw/. The examples of true clusters are listed as follows:

/praan^{0}/	“ปราณ”
/phraan^{0}/	“พราน”
/plii^{0}/	“ปลี”
/phlii^{0}/	“พลี”
/traj^{0}/	“ไตร”
/kraj^{0}/	“ไกร”
/khraj^{2}/	“ใคร่”
/klaaj^{0}/	“คลาย”
/kwaan^{2}/	“กว้าน”
/khwaaj^{0}/	“ควาย”
/traa^{0}/	“ตรา”

2. Parallel consonant characters

   These double consonant characters represent two consonantal sounds; the first letter represents the first consonantal sound with an intrusive /a/ vowel forming the first syllable, and the second consonantal sound functions as the initial consonant of the second syllable.

3. Leading consonant characters

   The double consonant characters are used to represent two consonantal sounds the same way as in the parallel consonant characters, except that the tonal assignment rules for the second syllable are determined by the consonant class of the first consonant letter. For example, in the word <tlok> the first letter <t> has a low tone according to the initial middle class consonant, and <l> in the second syllable <lok> has a high tone according to the initial low class consonant and is pronounced /lok^{3}/ when it occurs by itself. But because of the leading consonant character rule, the second syllable tone is changed from high tone to low tone, the same tone as the first consonant letter, so <tlok> is pronounced as /ta^{1} lok^{1}> instead of /ta^{1} lok^{3}/.

4. Pseudo cluster

   These words are initialed with double consonant letters but they represent only single consonantal sounds. For instance, <thr> in the word <thraap> is a pseudo cluster because <thr> represent a single sound /s/, and the word is pronounced as /saap^{0}/.

Appendix B

B.1 Training Sentences

Stress Patterns			
Unstressed-Unstressed	Unstressed-Stressed	Stressed-Unstressed	Stressed-Stressed
1.ตองกวนอามดี	1.ตองกวนอามดี	1.ตองมองอามดี	1.ตองมองอามดี
2.โองกวนไกดี	2.โองกวนหนาดี	2.โองมองไกดี	2.โองมองหนาดี
3.ออกวนอวนดี	3.ออกวนมากดี	3.ออมองอวนดี	3.ออมองมากดี
4.น้องกวนตุ๊กดี	4.น้องกวนอ๊าตดี	4.น้องมองตุ๊กดี	4.น้องมองอ๊าตดี
5.หนองกวนต่องดี	5.หนองกวนออดี	5.หนองมองต่องดี	5.หนองมองออดี
6.ตองบอกอามอยู	6.ตองบอกอามอยู	6.ตองแอบอามอยู	6.ตองแอบอามอยู
7.โองบอกไกอยู	7.โองบอกหนาอยู	7.โองแอบไกอยู	7.โองแอบหนาอยู
8.ออบอกอวนอยู	8.ออบอกมากอยู	8.ออแอบอวนอยู	8.ออแอบมากอยู
9.น้องบอกตุ๊กอยู	9.น้องบอกอ๊าตอยู	9.น้องแอบตุ๊กอยู	9.น้องแอบอ๊าตอยู
10.หนองบอกต่องอยู	10.หนองบอกออออยู	10.หนองแอบต่องอยู	10.หนองแอบอออยู
11.ตองไลอามไต้	11.ตองไลอามไต้	11.ตองอ้างอามไต้	11.ตองอ้างอามไต้
12.โองไลไกไต้	12.โองไลหนาไต้	12.โองอ้างไกไต้	12.โองอ้างหนาไต้
13.ออไลอ้วนไต้	13.ออไลมากไต้	13.ออ้างอ้วนไต้	13.ออ้างมากไต้
14.น้องไลตุ๊กไต้	14.น้องไลอ๊าตไต้	14.น้องอ้างตุ๊กไต้	14.น้องอ้างอ๊าตไต้
15.หนองไลต่องไต้	15. หนองไลออไต้	15.หนองอ้างต่องไต้	15. หนองอ้างออไต้

16.ต้องนัดอ้ามแล้ว	16.ต้องนัดอ้ามแล้ว	16.ต้องล้ออ้ามแล้ว	16.ต้องล้ออ้ามแล้ว
17.โอ้งนัดไก้แล้ว	17.โอ้งนัดหน้าแล้ว	17.โอ้งล้อไก้แล้ว	17.โอ้งล้อหน้าแล้ว
18.อ้อนัดอ้วนแล้ว	18.อ้อนัดม้ากแล้ว	18.อ้อล้ออ้วนแล้ว	18.อ้อล้อม้ากแล้ว
19.น้องนัดตุ๊กแล้ว	19.,น้องนัดอ๊าตแล้ว	19.น้องล้อตุ๊กแล้ว	19.น้องล้ออ๊าตแล้ว
20.หน้องนัดต้องแล้ว	20.หน้องนัดออแล้ว	20.หน้องล้อต้องแล้ว	20.หน้องล้อออแล้ว
21.ต้องโออ้ามเดี๊ยว	21.ต้องโออ้ามเดี๊ยว	21.ต้องหาอ้ามเดี๊ยว	21.ต้องหาอ้ามเดี๊ยว
22.โอ้งโอ้ไก้เดี๊ยว	22.โอ้งโอหน้าเดี๊ยว	22.โอ้งหาไก้เดี๊ยว	22.โอ้งหาหน้าเดี๊ยว
23.อ้อโออ้วนเดี๊ยว	23.อ้อโอม้ากเดี๊ยว	23.อ้อหาอ้วนเดี๊ยว	23.อ้อหาม้ากเดี๊ยว
24.น้องโอ้ตุ๊กเดี๊ยว	24.น้องโออ๊าตเดี๊ยว	24.น้องหาตุ๊กเดี๊ยว	24.น้องหาอ๊าตเดี๊ยว
25.หน้องโอ้ต้องเดี๊ยว	25.หน้องโอออเดี๊ยว	25.หน้องหาต้องเดี๊ยว	25.หน้องหาออเดี๊ยว

B.2 Test Sentences

1.ต้องกวนอ้ามดี๊	30.หน้องกวนออดี๊	59.น้องแอบตุ๊กอยู้	88.ออ้อ้างม้ากได้
2.โอ้งกวนไก้ดี๊	31.ต้องบอกอ้ามอยู้	60.หน้องแอบต้องอยู้	89.น้องอ้างอ๊าตได้
3.อ้อกวนอ้วนดี๊	32.โอ้งบอกหน้าอยู้	61.ต้องอ้างอ้ามได้	90. หน้องอ้างออได้
4.น้องกวนตุ๊กดี๊	33.ออบอกม้ากอยู้	62.โอ้งอ้างไก้ได้	91.ต้องล้ออ้ามแล้ว
5.หน้องกวนต้องดี๊	34.น้องบอกอ๊าตอยู้	63.ออ้างอ้วนได้	92.โอ้งล้อหน้าแล้ว
6.ต้องบอกอ้ามอยู้	35.หน้องบอกอออยู้	64.น้องอ้างตุ๊กได้	93.ออ้ล้อม้ากแล้ว
7.โอ้งบอกไก้อยู้	36.ต้องไล้อ้ามได้	65.หน้องอ้างต้องได้	94.น้องล้ออ๊าตแล้ว

8.ออบอกอ้วนอยู	37.โอ้งไลหน้าได้	66.ตองลออามแล้ว	95.หนองลอออแล้ว
9.น้องบอกตุ๊กอยู	38.ออไลมากได้	67.โอ้งล้อไก่แล้ว	96.ตองหาอามเดี๊ยว
10.หนองบอกตองอยู	39.น้องไลอาตได้	68.ออล้ออ้วนแล้ว	97.โอ้งหาหน้าเดี๊ยว
11.ตองไลอามได้	40. หนองไลออได้	69.น้องล้อตุ๊กแล้ว	98.ออหามากเดี๊ยว
12.โอ้งไลไก่ได้	41.ตองนัดอามแล้ว	70.หนองล้อตองแล้ว	99.น้องหาอาตเดี๊ยว
13.ออไลอ้วนได้	42.โอ้งนัดหน้าแล้ว	71.ตองหาอามเดี๊ยว	100.หนองหาออเดี๊ยว
14.น้องไลตุ๊กได้	43.ออนัดมากแล้ว	72.โอ้งหาไก่เดี๊ยว	101.นานเลนวาวมั่ง
15.หนองไลตองได้	44.,น้องนัดอาตแล้ว	73.ออหาอ้วนเดี๊ยว	102.หมอหนี่หมี่ไหว
16.ตองนัดอามแล้ว	45.หนองนัดออแล้ว	74.น้องหาตุ๊กเดี๊ยว	103.ตอฆาแกะมาก
17.โอ้งนัดไก่แล้ว	46.ตองโออามเดี๊ยว	75.หนองหาตองเดี๊ยว	104.ตอค่าแกะมาก
18.ออนัดอ้วนแล้ว	47.โอ้งโอหน้าเดี๊ยว	76.ตองมองอามดี่	105.โอจับขาอยู
19.น้องนัดตุ๊กแล้ว	48.ออโอมากเดี๊ยว	77.โอ้งมองหน้าดี่	106.โอจับข้าอยู
20.หนองนัดตองแล้ว	49.น้องโออาตเดี๊ยว	78.ออมองมากดี่	107.แมมี่น้ำดี่
21.ตองโออามเดี๊ยว	50.หนองโอออเดี๊ยว	79.น้องมองอาตดี่	108.แมมี่หน้ำดี่
22.โอ้งโอไก่เดี๊ยว	51.ตองมองอามดี่	80.หนองมองออดี่	109.โอขึ้นเขาได้
23.ออโออ้วนเดี๊ยว	52.โอ้งมองไก่ดี่	81.ตองแอบอามอยู	110.โอขึ้นเขาได้
24.น้องโอตุ๊กเดี๊ยว	53.ออมองอ้วนดี่	82.โอ้งแอบหน้าอยู	111.อาเข้ามาได้
25.หนองโอตองเดี๊ยว	54.น้องมองตุ๊กดี่	83.ออแอบมากอยู	112.นานขี่มาเกง
26.ตองกวนอามดี่	55.หนองมองตองดี่	84.น้องแอบอาตอยู	113.นานขี่หมาเกง

27.โอ่งกวนหน่าดี	56.ตองแอบอามอยู่	85.หนองแอบอออยู่	114.หมางับขาแล้ว
28.ออกวนมากดี	57.โอ่งแอบไก่อยู่	86.ตองอ้างอามได้	115.หมางับข้าแล้ว
29.น้องกวนอ๊าตดี	58.ออแอบอ้วนอยู่	87.โอ่งอ้างหน่าได้	

VITA

Nuttavudh Satravaha was born on May 19, 1965 in Samutprakarn, Thailand. He finished his high school from the Demonstration School of Srinakarinvirote University Pratumwan in 1983. In 1987, he received a Bachelor's degree in Computer Engineering from King Mongkut Institute of Technology Ladkrabang (KMITL). He started his career as an engineer for the Telephone Organization of Thailand (TOT) in 1987. From 1989 to 1993, he held a position as a manager of the Computerized Directory Assistance System (CDAS). In June of 1993, he graduated with a Master's degree in Computer Science from Chulalongkorn university. In 1994, he received a scholarship from TOT to study the Master program in telecommunications engineering at Asian Institute of Technology (AIT) and he earned his second master's degree in 1995. After graduating from AIT, he returned to TOT and worked at the Department of Telephone Metropolitan Area1. In 1997, he received a scholarship from TOT to study the Ph.D. program at West Virginia university (WVU), and completed his degree in the Spring of 2002. His major interests include: speech communication, signal processing, pattern recognition, and artificial intelligence.