APPENDIX D

MANUSCRIPTS SUBMITTED FOR PUBLICATIONS

1) Potisuk, S., "The Effects of Stress on F_0 Contours of Thai Tones in Continuous Speech."

2) Potisuk, S., "Automatic Classification of Thai Tones in Continuous Speech."

3) Potisuk, S., "Prosody Generation in a Thai Text-to-speech System."

THE EFFECTS OF STRESS ON F_0 CONTOURS OF THAI TONES IN CONNECTED SPEECH

Siripong Potisuk

Department of Electrical and Computer Engineering
Academic Division, Chulachomklao Royal Military Academy
Nakon-nayok, 26001 THAILAND

Address for Editorial Correspondence

Siripong Potisuk, Ph.D.
CRMA P.O. Box 16
Chulachomklao Royal Military Academy
Nakon-nayok, 26001 THAILAND
Tel: 66-037-393484
Fax: 66-037-393484
E-mail: srppts@hotmail.com

ABSTRACT

An experiment was conducted to investigate changes in the fundamental frequency (F_0) contours of Thai tones in connected speech. Thai has five tones: mid, low, falling, high, and rising. It is hypothesized that Thai tones in connected speech are affected by many linguistic factors, such as syllable structure, tonal coarticulation, stress, and intonation. The experiment is designed to isolate stress from other confounding factors. Stimuli consisted of 25 pairs of ambiguous target sentences with disambiguating context. Target syllable was embedded at the beginning of the utterance, and thus eliminating perseverative tonal coarticulation. One member of each pair contained a 2-syllable noun-verb sequence exhibiting a – – stress pattern, the other member a 2-syllable noun compound exhibiting a ~ – stress pattern. Acoustic analysis revealed that F_0 contours of stressed syllables more closely approximate F_0 contours in citation forms than those of unstressed syllables. The degree of approximation is primarily determined by syllable structure. In contrast, F_0 contours of unstressed syllables undergo a more complex process. The average height of all five tones can be classified into three tonal registers: low, mid, and high. The low register comprises the low and the rising tones, the mid register the mid tone, and the high register the falling and the high tones. Based on shape, the falling and high tones are distinguished within the high register. The low and rising tones within the low register. Therefore, a five-way contrast among all five tones appears to be maintained in both stressed and unstressed syllables.

INTRODUCTION

The impetus for this research arose during an investigation on automatic tone classification in connected Thai speech by computer. Since F_0 is the primary acoustic correlate of tone, this study will focus on F_0 realization of Thai tones in connected speech. Tone is a distinctive feature of any tone languages, like Thai, and the differences in lexical tones can be acoustically described in terms of distinct patterns of F_0 contours. While tones in isolation are relatively easy to acoustically describe and to automatically classify because of their rather definite F_0 contour shapes in the tone space, the acoustic manifestation of tones in continuous speech is much more difficult to assess and quantify. It is hypothesized that F_0 contours of Thai tones are influenced by many linguistic factors: syllable structure, tonal coarticulation, stress and intonation. Thus, it is of primary interest to be able to quantify the effects of those confounding factors on F_0 contours of Thai tones. Such information is no doubt essential to the formulation of an algorithm that will automatically assign Thai syllables to their appropriate tonal categories.

Of all the linguistic factors affecting F_0 realization of Thai tones, we are interested in stress and tonal coarticulation. The aim of this study is to investigate changes in F_0 contours of the five Thai tones in connected speech as a function of stress by using a more systematically controlled experiment that isolates stress from other confounding factors. The focus will primarily be on the effects of stress occurring in disyllabic noun compounds. Compounds in Thai are very important not only because of their high frequency of occurrences, but because they provide us with a window to see how prosody may potentially be used by listeners to resolve ambiguities in Thai. The study will attempt to answer questions concerning the effects of stress on individual tones and the contrastive relationship of lexical tones in both stressed and unstressed syllables. Findings will be interpreted in terms of their relevance to the description of sentence prosody in Thai. Implementation issues regarding automatic tone classification will also be addressed.

EXPERIMENT

The Effects of Stress on F_0 Contours of Thai Tones in Continuous Speech

Stress in Thai has been investigated in the past by many linguists, and many stress placement rules postulated. In terms of pitch, those researchers generally agreed on the phonetic realization of stress in "linker syllables" [1]. However, the effects of stress on F_0 contours of unstressed non-linker syllables remain a subject of much controversy up to the present. The disagreement revolves around the issue of whether or not lexical tones of unstressed syllables undergo tone neutralization, i.e., whether or not F_0 contours of all or some of the five tones lose their identities in both height and shape. Of those earlier studies, only a few have presented acoustic-phonetic information on the realization of stress in terms of F_0.

Hiranburana [2,3] presented instrumental findings on stress at the word level. Changes in F_0 contours were shown to vary depending on degree of stress. Her results were based on observations of the pattern of changes in F_0 contours of unstressed syllables obtained from non-final syllables of polysyllabic words, monosyllabic grammatical words, the first syllable of institutionalized compounds, and the reduplicator of the completely reduplicative forms. She concluded that the F_0 contours of the five lexical tones are neutralized to three level tones: high, mid, and modified low.

Gandour [4] argued against tone neutralization in fast casual speech by presenting acoustical measurements of F_0 contours of the initial syllable in pairs of disyllabic noun compounds distinguished minimally or near-minimally by the lexical tone of the initial syllable. His findings indicated no changes in contour tones of unstressed syllables. He concluded that the five lexical tones of unstressed syllables maintain their basic canonical shapes as in citation forms despite being shorter in duration, and the five-way contrast is intact.

Luksaneeyanawin [5] extended the phonetic analysis of stress beyond the word level. Based on acoustic and auditory analyses of passages read by two speakers, her descriptions of F_0 contours of unstressed non-linker syllables were generally in agreement with Hiranburana's, except those of the rising tone. It appears that she also favors the existence of tone neutralization in unstressed non-linker syllables.

The findings in all three aforementioned studies are unfortunately very difficult to interpret because none of them isolated stress from other confounding factors affecting the realization of F_0 contours, such as tonal coarticulation, declination and intonation. Therefore, their results might not be due to the manifestation of stress alone, but to other factors as well.

Method

Subjects

Five native speakers of Thai were selected for this experiment. All subjects were monodialectal speakers of standard Thai from the Bangkok Metropolitan area only. They are free of any speech or hearing disorders by self-report based on a screening interview and as later judged by the investigator during the recording session. Subjects were also chosen based on the following criteria: age, education, gender and geographical profile. All subjects were naive with respect to the purpose of the experiment.

Materials

Stimuli consisted of 25 pairs of ambiguous target sentences. The two members of each pair contained six segmentally identical syllables including two target syllables. The first member (a) contained a 2-syllables noun-verb sequence exhibiting a – – stress pattern, the second member (b) a 2-syllable noun compound exhibiting a ~ – stress pattern. The diacritic ~ represents an unstressed syllable, – a stressed syllable. To minimize tonal coarticulation effects, the two target syllables were embedded at the beginning of the sentence, hence only anticipatory coarticulation on the first syllable is present while carryover coarticulation is eliminated. Thai tones are more greatly influenced by carryover than anticipatory coarticulation [6]. The tones of the two target syllables were also varied to represent all possible two-tone combinations of five Thai tones so that anticipatory coarticulation in all contexts is considered. Of 25 two-tone combinations, only four were fully voiced throughout (MH, MR, LF, and FH); the other 21 two-tone combinations had intervening voiceless obstruents. To maximize the speaker's likelihood of being able to naturally produce the utterance according to its intended meaning, each utterance was preceded by a few sentences of

disambiguating context. A list of the target sentences with their disambiguating contexts is included in the appendix.

Recording Procedure

Speakers were asked to read a target sentence along with a few sentences of disambiguating context typed in Thai script on a 12.7 X 20.32-cm card. Cards were presented in random order and speakers were not told which of the sentences in the paragraph was the target sentence. They were also instructed to produce the sentences at a conversational speaking rate, i.e., at a rate they considered representative of their conversational speech. A random order of presentation and a sufficient pause provided between items were intended to minimize changes in speaking rate and learning or list-reading effects, thus maximizing the likelihood of speakers being able to produce natural sounding utterances. To avoid start and end effects, extra cards were placed at the top and bottom of the deck.

Recordings were made in a soundproof booth using a Mascot ECM-627 unidirectional condenser microphone and a Technics RS-TR210 tape recorder. Speakers were seated and wore a custom-made headband that maintained the microphone at a distance of 20 cm from the lips. For each speaker, the total corpus contained 250 utterances (2 members X 25 tonal combinations X 5 repetitions). There were two recording sessions separated by one week to minimize the possibility of speaker's exaggerating the contrast between the two members [7]. The (a) members of all pairs were assigned to the first recording session, the (b) members to the second session. Before the recording session began, the speakers were allowed to familiarize themselves with the target sentences. During the session, speakers were asked to reread any sentences that the investigator deemed "off-target" until an acceptable version was produced. Each session lasted about 45 minutes.

Measurement Procedure

The tape-recorded stimuli were low-pass filtered at 10 kHz and digitized at a sampling rate of 20 kHz by means of a 16-bit A/D converter with a 5-V dynamic range using the KAY CSL (Computerized Speech Lab) Model 4300 installed on an IBM compatible Pentium III/667 MHz microcomputer. Cursors were positioned on a spectrographic display (8 kHz frequency range, 300 Hz bandwidth) to mark the beginning and end of the target sentence. Total duration of the target sentence was

measured from the release burst of the consonant at the beginning of the sentence to the cessation of the second and higher formants at the end of the sentence. Measurement precision was 4 ms, which simply reflected the resolution of the CSL window when the entire utterance was on the screen.

F_0 was computed directly from the waveform using a CSL algorithm that employs a time domain approach to pitch analysis (modified autocorrelation with center clipping) with nonoverlapping variable framelength. For a particular speaker, framelength was determined by his/her pitch range to ensure that there were at least two complete cycles within a frame. A typical frame length was 20 to 25 ms for male speakers, 15 to 20 ms for female speakers. F_0 analysis sometimes failed to extract a contour from an audio waveform with lengthy stretches of aperiodicity. Other waveforms were contaminated by extraneous background noise or voicing overlap. About 3% of utterances produced by all five speakers were eliminated from the corpus, which resulted in a total of 1210 utterances that were retained for subsequent analysis.

In this study, only the acoustic features of the first syllable of each two-tone combination were of primary interest. Its onset and offset were determined from a simultaneous display of a wide-band (300 Hz) spectrogram with a scale from 0-8 kHz, energy contour, F_0 contour, and audio waveform. Tonal onset was defined as the first F_0 value after voiceless obstruents that coincided with vertical striations in the second and higher formants, or as the first F_0 value of a nasal or liquid. Tonal offset was defined as the last F_0 value preceding the abrupt cessation of second and higher formants of the vowel, or as the last F_0 value preceding the sudden onset of a nasal or liquid based on auditory impression.

Statistical Analysis

F_0 contours of individual syllables were equalized for duration on a percentage scale. Since inter-speaker comparisons were not of interest in this study, F_0 contours were not normalized on a z score scale [8]. Only for display purposes were F_0 contours smoothed by curve fitting.

In view of the perceptual dimensions underlying Thai tones [9, 10], statistical analysis was restricted to F_0 height and shape. To evaluate changes in F_0 height and shape of the stressed and unstressed syllables of each tone, mean and standard deviation of raw F_0 trajectories were computed by pooling across all tokens of all

sequences of all five speakers for that particular tone. The mean and standard deviation were used to assess overall changes in F_0 height and shape, respectively. The standard deviation was further transformed into coefficient of variation to allow meaningful comparisons across all tones. Coefficient of variation is defined as the ratio of the standard deviation to the mean expressed as a percentage. Since raw F_0 values are always positive, coefficient of variation is an appropriate measure of the relative variability with respect to the mean.

Results

Stress Effects on Individual Tones

The means and standard deviations of stressed and unstressed syllables of all 25 two-tone combinations are presented in Table 1, and their corresponding mean F_0 contours are shown in Figure 1.

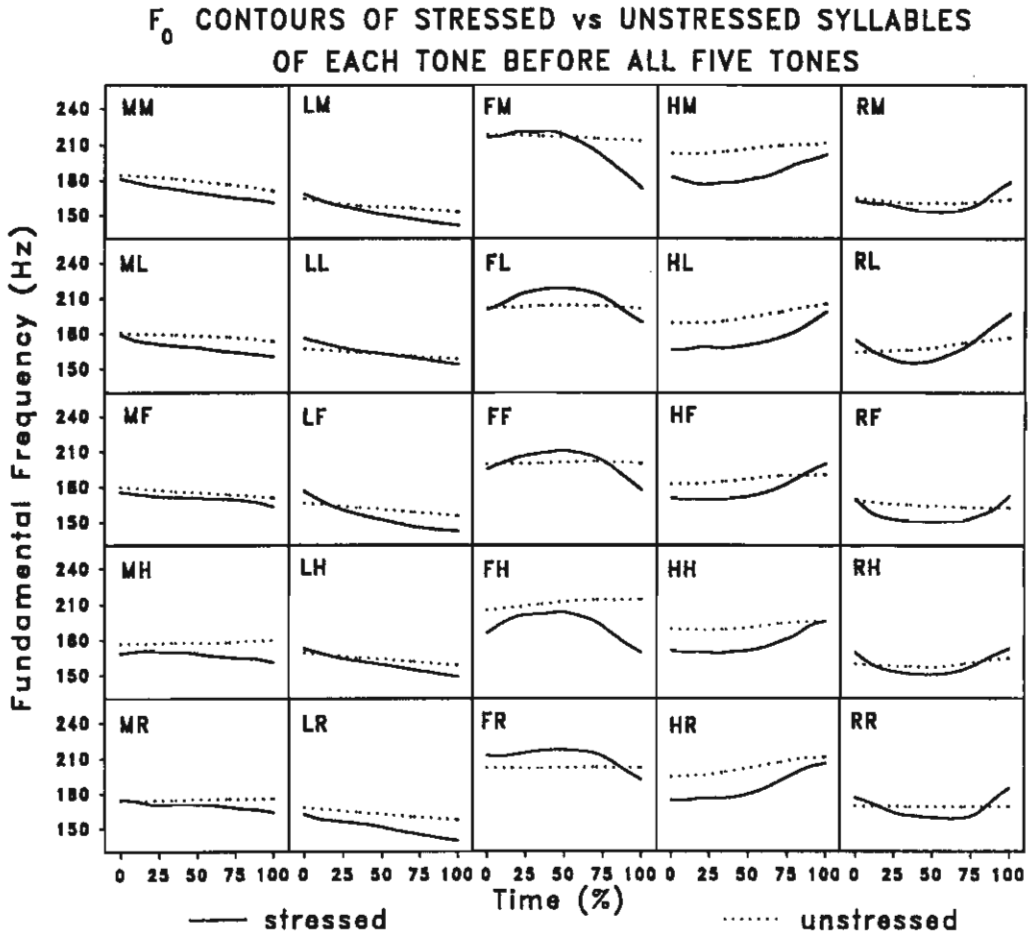

Fig. 1 Mean F_0 contours of each of the 25 two-tone sequences for both stressed and unstressed syllables.

Table 1

The mean and standard deviation in Hz of each tone in all 25 two-tone combinations for both stressed and unstressed syllables

Tone	Sequence	Unstressed syllable		Stressed syllable	
		M	SD	M	SD
Mid	MM	178.70	4.2015	169.61	5.9036
	ML	177.75	1.9424	167.84	4.6796
	MF	175.21	2.6439	170.24	2.7572
	MH	177.88	0.8967	167.21	2.8790
	MR	174.75	0.4719	169.75	2.7126
Low	LM	157.00	3.0314	151.41	7.2660
	LL	162.39	2.481	163.15	6.1326
	LF	190.72	3.2189	153.74	9.6238
	LH	163.62	3.1215	159.31	6.6696
	LR	162.25	3.1435	150.35	6.3734
Falling	FM	215.73	1.6502	208.25	14.606
	FL	203.2	0.8178	210.13	8.2688
	FF	200.47	0.7144	201.90	8.7898
	FH	210.96	2.9327	193.7	9.9143
	FR	202.02	0.2921	211.55	6.8270
High	HM	205.97	3.1677	184.07	7.9676
	HL	195.08	5.5785	174.44	8.9659
	HF	186.56	2.9353	177.15	9.5601
	HH	191.58	2.8525	176.12	8.7743
	HR	201.94	5.9971	184.39	11.003
Rising	RM	161.08	1.3424	158.75	6.5621
	RL	169.06	4.0278	167.26	12.011
	RF	164.19	1.8365	155.30	5.8922
	RH	159.00	2.3062	157.01	6.7814
	RR	168.41	0.14804	165.43	7.3035

Pooled across all tokens of all sequences of each individual tone, F_0 contours of stressed syllables more closely approximate F_0 contours in citation forms [11, 12] than those of unstressed syllables in both average F_0 height and shape. The degree of approximation is primarily determined by syllable structure and the interaction between adjacent tones. In stressed syllables preceding a major phrase boundary, i.e., the first syllable in the noun-verb sequence, F_0 contours of the so-called static tone (mid, low, and high) remained virtually unchanged while those of the so-called dynamic tones (falling and rising) undergo a slight modification and exhibit less extreme F_0 offsets (see Figure 2a). The falling tone does not fall as far as it does in citation forms; the rising tone does not rise as far. In contrast, F_0 contours of

unstressed syllables in noun compounds (see Figure 2b) differ from those of either citation forms or stressed syllables occurring prepausally.

By comparison to F_0 contours in stressed syllables (see Figure 2a), the average height of all five tones in unstressed syllables is raised. With respect to shape, the mid, low, and falling tones exhibit relatively level to slightly falling contours, whereas the high and rising tones show a sharp rise in the terminal portions of their contours.

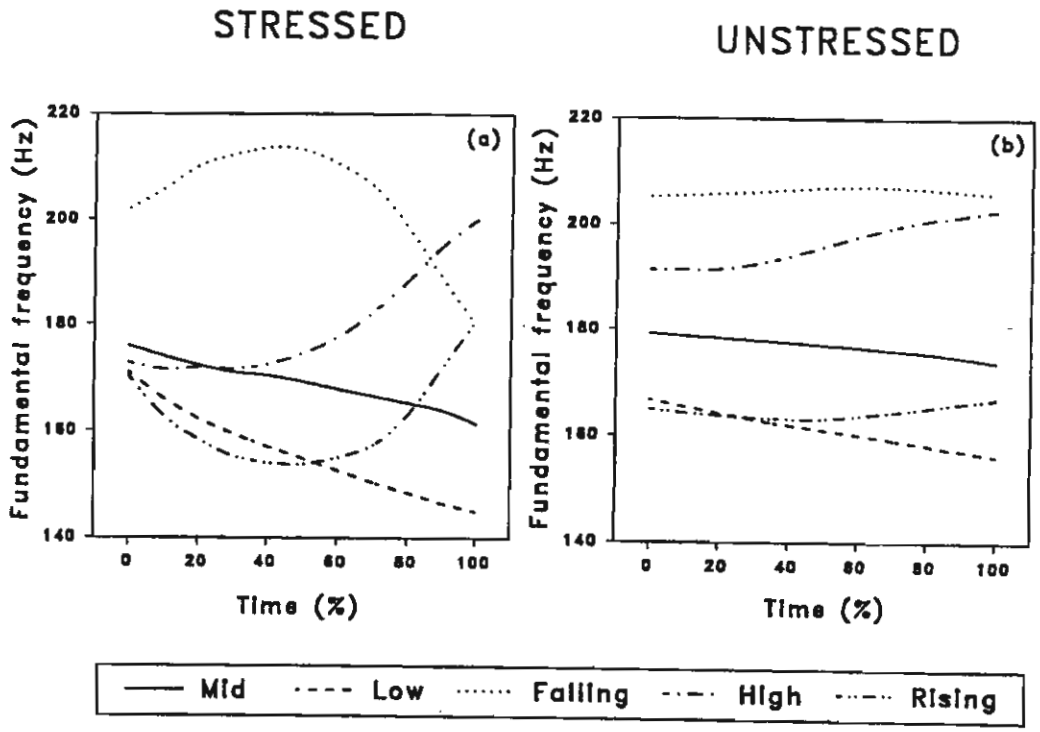

Fig. 2 Overall mean F_0 contours of all five tones
for both stressed and unstressed syllables.

The overall mean, standard deviation, and coefficient of variation of each tone in stressed and unstressed syllables are given in Table 2. Comparisons in average F_0 height and coefficient of variation between stressed and unstressed syllables are also shown in Figure 3a and 3b, respectively.

With respect to average F_0 (see Figure 3a), all five tones of unstressed syllables appear to have higher average F_0 than those of stressed syllables. Dynamic tones (falling and rising) have less of a difference in average F_0 than static tones (mid, low, and high). With respect to coefficient of variation (see Figure 3b), the opposite seems to be the case. All five tones of unstressed syllables appear to have smaller coefficient of variation than those of stressed syllables (see Figure 3a). Dynamic tones have more of a difference in coefficient of variation than static tones. Furthermore, the difference in coefficient of variation appears to be more dramatic than the difference in average F_0.

Stress Effects on the Pattern of Contrast among the Five Tones

The contrastive relationship among the five tones is maintained in both stressed and unstressed syllables (see Figure 2). In stressed syllables, as in citation forms, a five-way contrast is maintained in terms of both average F_0 height and shape despite less extreme F_0 offsets of the falling and the rising tones. In unstressed syllables (see Figure 4a), the five lexical tones can be divided into three subgroups with respect to average F_0 : 1) falling and high, 2) low and rising, and 3) mid. The falling and high tones appear to be higher than the mid, low, or rising tones; the low and rising tones appear to be lower than the mid, falling, or high tones; the mid tone, in turn, is intermediate between these other two subgroups. Hence, it appears that a three-level tonal register is maintained: low, mid, and high. The low register corresponds to the subgroup with the low and rising tones, the mid register to the subgroup with the mid tone, and the high register to the subgroup with the falling and high tones. Also, in unstressed syllables, the two tones within each of the high and low tonal registers appear to be distinguished on the basis of coefficient of variation (see Figure 4b). Within the high register, the high tone exhibits greater variability than the falling; within the low register, the low tone shows greater variability than the rising. These differences in coefficient of variation correspond to differences in shape between the falling and high and between the low and rising tones (see Figure 2b). The high and rising tones exhibit rising contours in their terminal portions, whereas the falling and low tones do not. Therefore, despite differences in height and shape of the tones in unstressed and stressed syllables, a five-way tonal distinction appears to be maintained but in a different tonal space (compare Figures 2a and 2b).

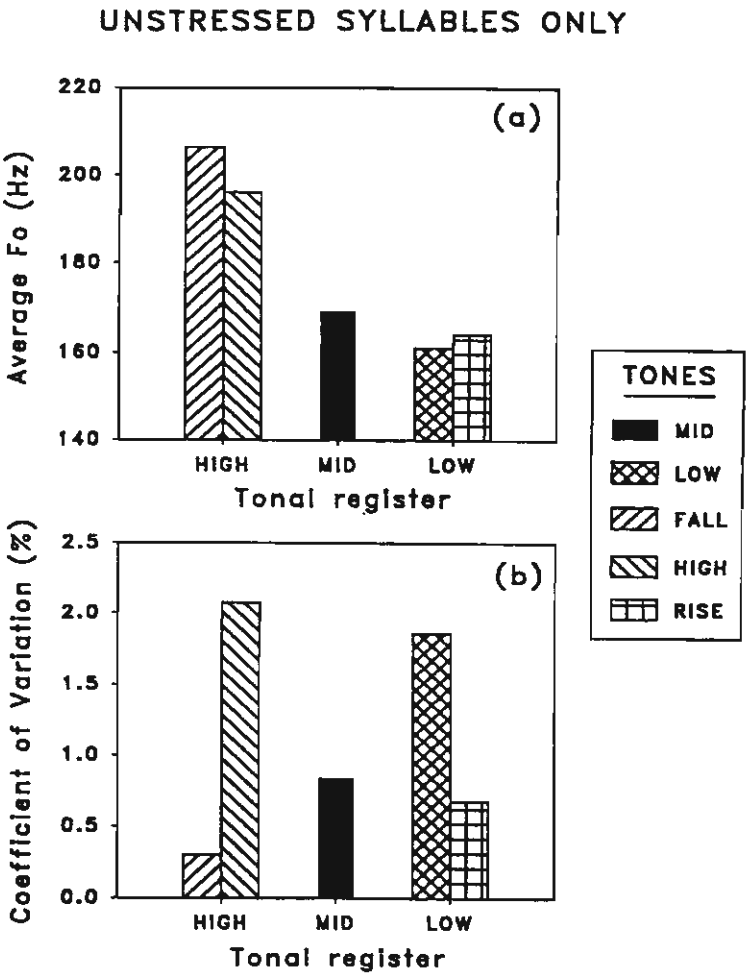

Fig. 4 Comparison of (a) average F_0 heights (b) coefficients of variation among all five tones of unstressed syllables grouped into three tonal registers.

Discussion

Tonal Contrasts in Stressed and Unstressed Syllables in Thai

The major finding of this study is that tonal contrast among all five Thai tones are preserved in both stressed and unstressed syllables in a context of disyllabic noun compound versus a noun-verb sequence despite changes in both average F_0 height and shape of tonal contours. The observed changes in F_0 contours are clearly different from those observed in linker syllables where tone neutralization occurs. The function of the stress rule observed in this study appears to be to distinguish between compounds, noun or verb, and other syntactic phrases. It signals the difference between a major phrase boundary and internal words of a compound, and thus the

effects of stress at the sentence level. We, of course, are aware that a more complicated picture may emerge when extending the study of stress to other sentence positions. In the present study, other factors affecting F_0 contours such as tonal coarticulation, and declination were kept to a minimum. As a result, tonal contrasts in unstressed syllables may or may not be preserved when all factors are taken into account.

Concerning the phonetic realization of F_0 contours of unstressed syllables in noun compounds in Thai, our findings are generally in agreement with earlier studies [2, 3, 4, 5]. However, some discrepancies regarding tone neutralization remain. First, Hiranburana suggested that in unstressed syllables, the five-way contrast is reduced to three and furthermore, that this three-way contrast is maintained on the basis of F_0 height alone. Falling and high tones are neutralized; low and rising tones are neutralized. Though we agree that three tonal registers are maintained based on F_0 height, a contrast is still maintained within the high and low registers based on shape. Falling and high tones are distinguished in the high register; low and rising tones in the low register. Secondly, we agree with Gandour that tone neutralization does not occur. But Gandour reported no changes in the shape of contour tones whereas our findings show a dramatic change in their shapes. Finally, our findings are consistent with Luksaneeyanawin's descriptions of F_0 contours of all five tones in unstressed syllables except that of the rising tones. She stated, "the unstressed rising tones is always realized with a rising contour no matter how much the syllable is reduced in duration..." Our results do not show a rise in certain contexts due to anticipatory coarticulation, i.e., RF and RR (figure 1). The aforementioned discrepancies, we believe, can be attributed to variations in speaking rate from study to study. Gandour used carrier sentences to solicit speech while Luksaneeyanawin based her studies on read passages. Speaking rate in both studies can be classified as low to moderate. Hiranburana used an Allegretto or moderately fast style of speech. However, it is unclear how her speech samples were solicited. The average speaking rate in this study is 4.65 syllables per second which is considered moderate. For low to moderate rate, Gandour's and Luksaneeyanawin's schema is the likely scenario while Hiranburana's scheme prevails for the moderately fast rate. Our findings herein are compatible with moderate speaking rate.

F_0 Correlate of Stress

Our data show that unstressed syllables in a disyllabic noun compound is produced with a higher pitch than when it is stressed in a noun-verb sequence. This finding runs contrary to what is usually found in other languages of the world. Other things being equal, stressed syllables are usually higher in pitch than unstressed syllables. Our data is insufficient to draw a firm conclusion. Nevertheless, we offer the following interpretation for such a phenomenon. In this study, bisyllablic noun compounds are made up of a noun and a verb. When the noun-verb sequence is intended by speakers, both syllables will be produced with a stress by virtue of being content words. Hence, a syntactic break signaling a phrase boundary occurs between them by a lowering of pitch in stressed syllable occurring prepausally. When a compound is intended, speakers destress the first syllable by raising its pitch to maximize the perceptual contrast between compound and phrase boundaries. Thus, pitch raising in unstressed syllables appears to be motivated primarily for the listener's benefit. This phenomenon, we believe, is a manifestation of the tendency for sound patterns in languages of the world to act in accordance with the principle of "sufficient perceptual separation" [13]. In Thai, unstressed syllables are raised in pitch so as make it easier for the listener to distinguish one type of syntactic constituent from another.

REFERENCES

[1] P. Bee, "Restricted phonology in certain Thai linker-syllables," in *Studies in Tai Linguistics in honor of William J. Gedney*, J. Harris and J. Chamberlain, Eds. Bangkok: Central Institute of English Language, 1975, pp. 17-32.

[2] S. Hiranburana, *The role of accent in Thai grammar.* Ph.D. dissertation, School of Oriental and African studies, University of London, 1971.

[3] S. Hiranburana, "Changes in the pitch contours of unaccented syllables in spoken Thai," in *Tai phonetics and phonology*, J. Harris and R. Noss, Eds. Bangkok: Central Institute of English Language, 1972, pp. 23-27.

[4] J. Gandour, "On the representation of tone in Siamese," in *Studies in Tai Linguistics in honor of William J. Gedney*, J. Harris & J. Chamberlain (Eds.). Bangkok: Central Institute of English Language, 1975, pp.170-195.

[5] S. Luksaneeyanawin, *Intonation in Thai,* Ph.D. dissertation, University of Edinburgh, 1983.

[6] S. Potisuk, J. T. Gandour, and M. P. Harper, "Contextual Variations in Trisyllabic Sequences of Thai Tones," *Phonetica*, vol. 54, pp. 22-42. 1997.

[7] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *Journal of Acoustical Society of America*, vol. 90-6, pp. 2956-2970, 1991.

[8] P. Rose, "Considerations in the normalisation of the fundamental frequency of linguistic tone," *Speech Communication,* vol. 6, pp. 343-351, 1987.

[9] J. T. Gandour and R. Harshman, "Cross-language differences in tone perception: A multidimensional scaling investigation," *Language and Speech*, vol. 21, pp. 1-33, 1978.

[10] J. T. Gandour, "Tone perception in Far Eastern languages," *Journal of Phonetics,* vol. 11, pp. 149-175, 1983.

[11] A. S. Abramson, "The vowels and tones of standard Thai: acoustical measurements and experiments," *International Journal of American Linguistics*, vol.28-2, Part III (Publication No.20), 1962. [Bloomington, IN: Indiana University Research Center in Anthropology, Folklore, and Linguistics].

[12] J. T. Gandour, S. Potisuk, and S. Dechongkit, "Inter- and intraspeaker variability in fundamental frequency of Thai tones," *Speech Communication,* vol. 10, pp. 355-372, 1991.

[13] P. Ladefoged, *A Course in Phonetics* (3rd ed.). New York: Harcourt Brace Jovanovich, 1993.

ACKNOWLEDGEMENTS

This material is based upon work supported by Thailand Research Fund under Grant No. RSA/03/2541. The first author would like to extend his gratitude to the Academic division of Chulachomklao Royal Military Academy, the Royal Thai Army for the opportunity to conduct this research. Reprint requests should be sent to: Siripong Potisuk, Ph.D., Department of Electrical Engineering, Academic Division, Chulachomklao Royal Military Academy, Nakon-nayok, 26001 THAILAND.

APPENDIX

Stimuli for the Experiment

1. a) เมืองอยู่ไกล คนไม่ค่อยมาเมืองเลย
   / kʰon mâj kʰɔ̂ːj **maa mɯaŋ** ləəj /
   'The city is too far away. People don't usually come to the city.'
   b) เมืองเลยอยู่ไกล คนไม่ค่อยมาเมืองเลย
   / kʰon mâj kʰɔ̂ːj **maa mɯaŋ**ləəj /
   'The city of Loei is too far away. People don't usually come to Loei.'

2. a) เมืองไม่น่าอยู่ คนไม่ค่อยอยู่เมืองเลย
   / kʰon mâj kʰɔ̂ːj **jùu mɯaŋ** ləəj /
   'The city is not conducive to living. People don't usually live in the city.'
   b) เมืองเลยไม่น่าอยู่ คนไม่ค่อยอยู่เมืองเลย
   / kʰon mâj kʰɔ̂ːj **jùu mɯaŋ**ləəj /
   'The city of Loei is not conducive to living. People don't usually live in Loei.'

3. a) ในเมืองอันตราย คนไม่ค่อยเข้าเมืองเลย
   / kʰon mâj kʰɔ̂ːj **kʰâw mɯaŋ** ləəj /
   'It is dangerous to be in the city. People don't usually come into the city.'
   b) เมืองเลยอันตราย คนไม่ค่อยเข้าเมืองเลย
   / kʰon mâj kʰɔ̂ːj **kʰâw mɯaŋ**ləəj /
   'It is dangerous to be in the city of Loei. People don't usually come into Loei.'

4. a) ถึงเมืองจะแห้งแล้ง คนไม่ค่อยทิ้งเมืองเลย
   / kʰon mâj kʰɔ̂ːj **tʰiŋ́ mɯaŋ** ləəj /
   'Even though the city is hot and dry . People don't usually abandon the city.'
   b) ถึงเมืองเลยจะแห้งแล้ง คนไม่ค่อยทิ้งเมืองเลย
   / kʰon mâj kʰɔ̂ːj **tʰiŋ́ mɯaŋ**ləəj /
   'Even though the city of Loei is hot and dry. People don't usually abandon Loei.'

5. a) เมืองสกปรกจัง คนไม่ค่อยสนเมืองเลย
   / kʰon mâj kʰɔ̂ːj **sŏn mɯaŋ** ləəj /
   'The city is quite dirty. People don't care much about the city.'
   b) เมืองเลยสกปรกจัง คนไม่ค่อยสนเมืองเลย
   / kʰon mâj kʰɔ̂ːj **sŏn mɯaŋ**ləəj /
   'The city of Loei is quite dirty. People don't care much about Loei.'

6. a) คนจนจะมีปัญญาไปเริ่มกิจการอะไรได้ เขาว่าคนจนขาดทุน

   / k^{h}áw wâa k^{h}**oncon k^{h}àat** t^{h}un /

   'The poor are quite at a disadvantage when it comes to starting a business. It's often said that the poor lack funds.'

   b) อัตราภาษีเก่าไม่ยุติธรรมเลย เขาว่าคนจนขาดทุน

   / k^{h}áw wâa k^{h}**oncon k^{h}àatt^{h}**un /

   'The tax rate is really unfair. It's often said that the poor are at a loss.'

7. a) พวกลูกจ้างหยุดงานเพราะกลัวไม่ได้รับเงินค่าแรง เขาว่าคนจ่ายขาดทุน

   / k^{h}áw wâa k^{h}**oncàaj k^{h}àat** t^{h}un /

   'The workers went on strike because they are afraid of not getting paid. There's a rumor that the payer has no money.'

   b) คนถูกหวยงวดนี้กันเพียบเลย เขาว่าคนจ่ายขาดทุน

   / k^{h}áw wâa k^{h}**oncàaj k^{h}àatt^{h}**un /

   'A lot of people strike it rich with this week's lottery. There's a rumor that the payer loses.'

8. a) งานก่อสร้างต้องหยุดชะงักลงกลางครัน เขาว่าคนจ้างขาดทุน

   / k^{h}áw wâa k^{h}**oncâaŋ k^{h}àat** t^{h}un /

   'The construction is suddenly halted. There's rumor that the contractor runs out of funds.'

   b) หลังจากเจรจาตกลงกันแล้ว ดูเหมือนพวกลูกจ้างจะได้กำไร เขาว่าคนจ้างขาดทุน

   / k^{h}áw wâa k^{h}**oncâaŋ k^{h}àatt^{h}**un /

   'After the negotiation, it seems that the workers profit from the deal. The employer is at a loss'

9. a) เศรษฐกิจยังไม่กระเตื้องขึ้นเลย เขาว่าคนซื้อขาดทุน

   / k^{h}áw wâa k^{h}**onsɯ́ɯ k^{h}àat** t^{h}un /

   'The economy is still not recovered. It's often said that consumers don't have the buying power.'

   b) การค้าแบบผูกขาดไม่ดีหรอก เขาว่าคนซื้อขาดทุน

   / k^{h}áw wâa k^{h}**onsɯ́ɯ k^{h}àatt^{h}**un /

   'Monopolized business is no good. It's often said that consumers are at a loss.'

10. a) ร้านนั้นดูท่าทางขายดีนะ แต่ยังขยายกิจการไม่ได้ เขาว่าคนขายขาดทุน

    / k^{h}áw wâa k^{h}**onk^{h}ǎaj k^{h}àat** t^{h}un /

    'It seems that the store's business really picks up. But, it can't quite expand. People speculate that the store lacks investment capital.'

    b) ร้านอาหารข้างบ้านปิดไปแล้ว เขาว่าคนขายขาดทุน

    / k^{h}áw wâa k^{h}**onk^{h}ǎaj k^{h}àatt^{h}**un /

    'The restaurant nextdoor was closed. People say that the owner didn't make a profit.'

11. a) ไม่เชื่อกันนี่ บอกแล้วว่าฟืนไหม้ดี

/ bɔ̀ɔk lέɛw wâa **fɯɯn mâj** dii /
'See? You don't believe me. I told you logs burn really well.'

b) ไม่เชื่อกันนี่ บอกแล้วว่าฟืนไม่ดี

/ bɔ̀ɔk lέɛw wâa **fɯɯn mâj** dii /
'See? You don't believe me. I told you the log is no good.'

12. a) ไม่เชื่อกันนี่ บอกแล้วว่าถ่านไหม้ดี

/ bɔ̀ɔk lέɛw wâa **tʰàan mâj** dii /
'See? You don't believe me. I told you charcoal burns really well.'

b) ไม่เชื่อกันนี่ บอกแล้วว่าถ่านไม่ดี

/ bɔ̀ɔk lέɛw wâa **tʰàan mâj** dii /
'See? You don't believe me. I told you the battery is no good.'

13. a) ไม่เชื่อกันนี่ บอกแล้วว่าข้าวไหม้ดี

/ bɔ̀ɔk lέɛw wâa **kʰâaw mâj** dii /
'See? You don't believe me. I told you rice burns really well.'

b) ไม่เชื่อกันนี่ บอกแล้วว่าข้าวไม่ดี

/ bɔ̀ɔk lέɛw wâa **kʰâaw mâj** dii /
'See? You don't believe me. I told you the rice is no good.'

14. a) ไม่เชื่อกันนี่ บอกแล้วว่าไม้ไหม้ดี

/ bɔ̀ɔk lέɛw wâa **máaj mâj** dii /
'See? You don't believe me. I told you wood burns really well.'

b) ไม่เชื่อกันนี่ บอกแล้วว่าไม้ไม่ดี

/ bɔ̀ɔk lέɛw wâa **máaj mâj** dii /
'See? You don't believe me. I told you it's a lousy kind of wood.'

15. a) ไม่เชื่อกันนี่ บอกแล้วว่าหนังไหม้ดี

/ bɔ̀ɔk lέɛw wâa **năŋ mâj** dii /
'See? You don't believe me. I told you cow hide burns really well.'

b) ไม่เชื่อกันนี่ บอกแล้วว่าหนังไม่ดี

/ bɔ̀ɔk lέɛw wâa **năŋ mâj** dii /
'See? You don't believe me. I told you it's a lousy movie.'

16. a) น้องสาวของแนนเป็นเด็กน่ารัก **คนชอบมายอน้องแนน**

/ kʰon cʰɔ̂ɔp maa **jɔɔ nɔ́ɔŋ** nan /
'Nan's sister is a cute kid. Everyone loves to praise Nan's sister.'

b) น้องแนนเป็นเด็กน่ารัก **คนชอบมายอน้องแนน**

/ kʰon cʰɔ̂ɔp maa **jɔɔ nɔ́ɔŋ**nan /
'Nan is a cute kid. Everyone loves to praise Nan.'

17. a) น้องสาวของแนนเป็นเด็กน่ารัก **คนชอบมาแหย่น้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **jɛ̀ɛ nɔ́ɔŋ** nan /

'Nan's sister is a cute kid. Everyone loves to tease Nan's sister.'

b) น้องแนนเป็นเด็กน่ารัก **คนชอบมาแหย่น้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **jɛ̀ɛ nɔ́ɔŋ**nan /

'Nan is a cute kid. Everyone loves to tease Nan.'

18. a) น้องสาวของแนนเป็นเด็กน่ารัก **คนชอบมาเยี่ยมน้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **jiâm nɔ́ɔŋ** nan /

'Nan's sister is a cute kid. Everyone loves to visit Nan's sister.'

b) น้องแนนเป็นเด็กน่ารัก **คนชอบมาเยี่ยมน้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **jiâm nɔ́ɔŋ**nan /

'Nan is a cute kid. Everyone loves to visit Nan.'

19. a) น้องสาวของแนนเป็นเด็กน่ารัก **คนชอบมาล้อน้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **lɔ́ɔ nɔ́ɔŋ** nan /

'Nan's sister is a cute kid. Everyone loves to tease Nan's sister.'

b) น้องแนนเป็นเด็กน่ารัก **คนชอบมาล้อน้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **lɔ́ɔ nɔ́ɔŋ**nan /

'Nan is a cute kid. Everyone loves to tease Nan.'

20. a) น้องสาวของแนนเป็นเด็กน่ารัก **คนชอบมาหลงน้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **lŏŋ nɔ́ɔŋ** nan /

'Nan's sister is a cute kid. Everyone seems crazy about Nan's sister.'

b) น้องแนนเป็นเด็กน่ารัก **คนชอบมาหลงน้องแนน**

/ k^{h}on c^{h}ɔ̂ɔp maa **lŏŋ nɔ́ɔŋ**nan /

'Nan is a cute kid. Everyone seems crazy about Nan.'

21. a) หมอเพิ่งออกไปข้างนอก เดี๋ยวคงกลับ **น่าจะไปรอหมอดู**

/ nâa càa paj **rɔɔ mɔ̌ɔ** duu /

'The doctor has just stepped outside. He'll be back soon. We should wait for the doctor.'

b) หมอดูเพิ่งออกไปข้างนอก เดี๋ยวคงกลับ **น่าจะไปรอหมอดู**

/ nâa càa paj **rɔɔ mɔ̌ɔ**duu /

'The fortuneteller has just stepped outside. He'll be back soon. We should wait for the fortuneteller.'

22. a) หมอสั่งยาไม่ได้เรื่องเลย **น่าจะไปด่าหมอดู**

/ nâa càa paj **dàa mɔ̌ɔ** duu /

'The doctor didn't prescribe high-potent drugs for us. We should let the doctor know how we feel.'

b) หมอดูหลอกเรานี่หว่า น่าจะไปด่าหมอดู

/ nâa càˑ paj **dàa mɔ̌ɔduu** /
'The fortuneteller is quite deceitful. We should give the fortuneteller a piece of our minds.'

23. a) โรงพยาบาลเรายังขาดหมออีกมาก น่าจะไปจ้างหมอดู

/ nâa càˑ paj **câaŋ mɔ̌ɔ** duu /
'Our hospital is still in need of doctors. We should try to hire more doctors.'

b) คดีนี้มืดมนเหลือเกิน น่าจะไปจ้างหมอดู

/ nâa càˑ paj **câaŋ mɔ̌ɔduu** /
'This case is really going nowhere. We should hire a psychic.'

24. a) หมอมีบุญคุณกับพวกเรามาก น่าจะไปเลี้ยงหมอดู

/ nâa càˑ paj **liáŋ mɔ̌ɔ** duu /
'We owe a great deal to the doctor. We should take the doctor out to dinner.'

b) จริง ๆ อย่างที่หมอดูบอกเลยว่าคุณจะได้งาน น่าจะไปเลี้ยงหมอดู

/ nâa càˑ paj **liáŋ mɔ̌ɔduu** /
'The fortuneteller is right on the money about your job prospect. We should take the fortuneteller out to celebrate.'

25. a) ลูกไม่เป็นอะไรมากแล้ว พากลับบ้านเถอะ น่าจะไปขอหมอดู

/ nâa càˑ paj **kʰɔ̌ɔ mɔ̌ɔ** duu /
'Our child is a lot better now. Let's take him home. We should ask the doctor.'

b) หวยใกล้จะออกแล้ว ยังไม่รู้จะซื้อเลขอะไรเลย น่าจะไปขอหมอดู

/ nâa càˑ paj **kʰɔ̌ɔ mɔ̌ɔduu** /
'It's about time for the lottery. I still haven't a clue to what number to play. We should ask the fortuneteller.'

AUTOMATIC CLASSIFICATION OF THAI TONES IN CONTINUOUS SPEECH

Siripong Potisuk

Department of Electrical and Computer Engineering
Academic Division, Chulachomklao Royal Military Academy
Nakon-nayok, 26001 THAILAND

Address for Editorial Correspondence

Siripong Potisuk, Ph.D.
CRMA P.O. Box 16
Chulachomklao Royal Military Academy
Nakon-nayok, 26001 THAILAND
Tel: 66-037-393484
Fax: 66-037-393484
E-mail: srppts@hotmail.com

AUTOMATIC CLASSIFICATION OF THAI TONES IN CONTINUOUS SPEECH

Siripong Potisuk

Department of Electrical and Computer Engineering
Academic Division, Chulachomklao Royal Military Academy
Nakon-nayok, 26001 THAILAND

Abstract—Tone classification is a crucial component of any automatic speech recognition system for tone languages. It is imperative that tonal information be incorporated into the word hypothesization process because patterns of pitch (or tones) contribute to the lexical identification of the individual words. In this paper, we present a novel algorithm for automatically classifying Thai tones in connected speech using an analysis-synthesis method based on an extension of the Fujisaki's model. We have successfully incorporated into the model four major factors affecting the phonetic realization of tones in connected speech: continuity effect due to syllable structure, stress, tonal coarticulation and declination. Also addressed are normalization procedures for achieving speaker-independence. In our preliminary experiment, we were able to achieve 81.7 % classification accuracy.

Index Terms—Analysis-by-synthesis, intonation, lexical tone classification, speech processing, spoken Thai, tonal assimilation.

I. INTRODUCTION

Tone classification is a crucial component of an automatic speech recognition (understanding) system for Thai and other tone languages. Tones, which are indicated by contrastive variations in F_0 at the syllable level, are used to signal differences in lexical meaning. As a result, it is imperative that tonal information be incorporated into the word hypothesization process because patterns of pitch (or tones) contribute to the lexical identification of the individual words.

Phonetically, Thai tones are considered contour tones, which can be specified in terms of gliding pitch movements, rather than in terms of single points within a pitch range. Thai has five tones: mid (M), low (L), falling (F), high (H), and rising (R). The following are examples of five segmentally identical words with different tones and meanings.

คา	/ k^haa /	'to get stuck'
ข่า	/ $k^hàa$ /	'a galangal plant'
ฆ่า	/ $k^hâa$ /	'to kill'
ค้า	/ $k^háa$ /	'to engage in trade'
ขา	/ $k^hǎa$ /	'leg'

Note that the phonemic transcription uses the diacritics /ˋ/, /ˆ/, /˝/, /ˇ/ as tone markers for the low, falling, high, and rising tones, respectively. The mid tone is unmarked.

The primary acoustic correlate of lexical tone is F_0, and differences in tones can be acoustically described in terms of distinct patterns of F_0 contours. Every Thai syllable carries a lexically contrastive F_0 contour. A detailed acoustic study of Thai tones spoken in isolation can be found in [1]. Fig. 1 shows the average F_0 contours for the five Thai tones in isolation from that study.

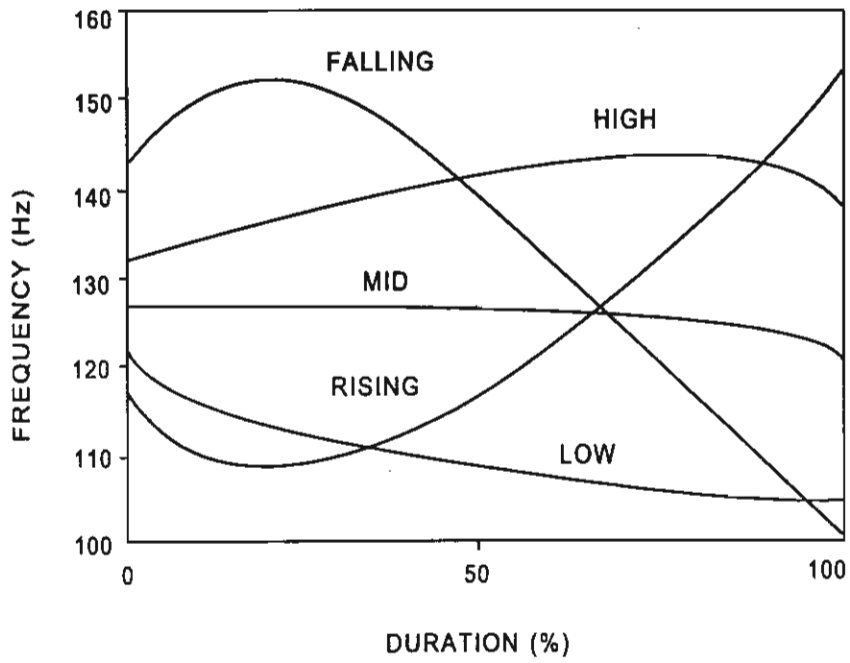

Fig. 1. Average F_0 contours of the five Thai tones in isolation (adapted from Abramson [1]).

From the figure, Thai tones can be classified into two categories: static (mid, low, and high) and dynamic (falling and rising) tones. The dynamic tones are characterized by a large excursion size and a dramatic change in the direction of the F_0 contour; the opposite is true for static tones. Acoustically speaking, it appears that differences in tones can be described in terms of the average F_0 height and the shape of F_0 contours. In fact, in view of the perceptual dimensions underlying Thai tones, F_0 height and movement carry sufficient information for high intelligibility of tones in Thai [2,3].

Automatic tone classification has been investigated by several researchers [4, 5, 6, 7]. Most researches focused on tone recognition of Chinese. The methods used are HMM-based or neural network-based. However, these methods are hypothesized to be less successful for Thai because the tone systems of Chinese and Thai are quite different. Since Thai tone classification has just been vigorously investigated by only a handful of researchers, the best approach for Thai tone recognition still unidentifiable. In this paper, a novel algorithm for automatically classifying Thai tones in continuous speech using an analysis-by-synthesis method is proposed. The analysis-by-synthesis approach to automatic tone classification can be described as follows.

The problem of Thai tones classification in continuous speech can simply be stated as finding the best sequence of tones, $T_1, T_2, \ldots\ldots, T_n$, given an input speech signal. Because the primary acoustic correlate of tone is F_0 and Thai has five distinct F_0 contour patterns, the problem is to find the best possible combination of F_0 contour patterns that closely match the given input F_0 contour. Cast in terms of a pattern recognition system, the general design of a tone classifier involves two major steps: F_0 extraction and pattern matching (classification). Fig.2 illustrates the block diagram of such a design.

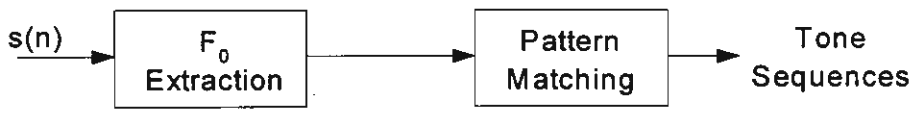

Fig.2 The block diagram of a general tone classifier.

With respect to the pattern matching process, hidden Markov models (HMMs) have proven to be an effective statistical approach to isolated tone recognition [4,5].

However, tone recognition in Thai connected speech using HMMs has never been attempted. We believe that a simple straightforward extension of an HMM isolated tone recognition algorithm is likely to produce unsatisfactory results for connected speech tone classification. This is partly due to the fact that connected speech tone recognition is a more difficult problem than isolated tone recognition. As illustrated in figure 3, there are differences in the F_0 realization of tones in an utterance when each individual word is spoken in isolation (see top panel) and when the whole utterance is naturally spoken in connected speech (see bottom panel). There appear to be interactions among several linguistic factors that affect the F_0 realization of tones in connected speech: continuity effect due to syllable structure, tonal coarticulation, stress, and intonation.

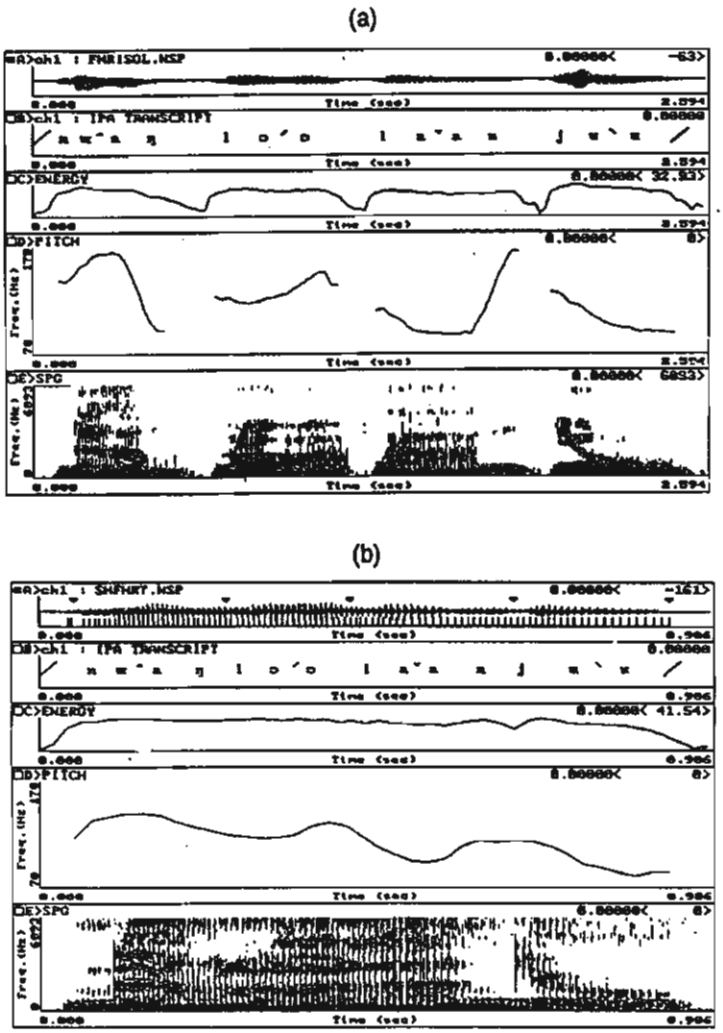

Fig. 3. Differences in the F_0 realization of tones in an utterance when (a) each word is spoken in isolation and (b) when the utterance is naturally spoken.

Analogous to the problem of continuous speech phone recognition in which contextual variations between contiguous phones (i.e., phone coarticulation) must be taken into account, continuous speech tone recognition must also incorporate tonal coarticulation and other linguistic factors into the system. A simple modification of an HMM isolated tone recognizer to recognize tones in continuous speech requires constructing a maximum of 125 (5 previous X 5 current X 5 following tones in a three-tone analysis window) tone models in order to account for both perseverative and anticipatory tonal coarticulation. This model may not conducive to real-time applications even with a parallel implementation. Also, because of subtle changes in F_0 contours due to coarticulatory effects, the usual acoustic features, F_0 and ΔF_0, used in an HMM-based system may not adequately capture the acoustically discriminatory information among coarticulation patterns of tones. For these reasons and because tonal coarticulation appears to be rule-governed, we propose a novel algorithm to classify tones in connected speech using an *analysis-by-synthesis* model.

Analysis-by-synthesis is an abstract model of the speech perception process proposed by Stevens [8]. The basic assumption of the model is that speech perception and production are closely tied. The major claim of the theory is that listeners perceive (analyze) speech by implicitly generating (synthesizing) speech from what they have heard and then comparing the synthesized speech with the auditory stimulus. According to the model, the perceptual process begins with an analysis of auditory features of the speech signal to yield an acoustic description in terms of auditory patterns. A hypothesis (or hypotheses) concerning the distinctive feature representation of the utterance is (are) constructed. This information then becomes the input to a set of generative rules that synthesize candidate patterns. The candidate patterns are subsequently compared with the patterns of the original utterance. The results of this matching process are then sent to a control component that transfers the phonetic description to higher levels of linguistic analysis. This model represents one of many *bottom-up* approaches to speech perception. That is, the model does not incorporate the effects of lexical and other higher-level knowledge into the speech perception process; they are only considered during later stages of understanding.

We adopt this model in the development of a Thai connected speech tone classifier because the model is easily implemented in terms of incorporating linguistic constraints into the model, although there has been little empirical evidence to support its validity. As the name suggests, the model contains two major components: the analysis

and the synthesis module. Roughly speaking, the function of the analysis module is to generate hypothesized tone sequences from the input F_0 contour. The synthesis module, in turn, generates predicted F_0 contours according to the hypothesized tone sequences. These predicted F_0 contours are basically reference templates to be used for pattern matching against the input contour. The synthesis module is based on our extension of Fujisaki's model for synthesizing F_0 contours to tone languages, and linguistic constraints are represented as synthesis rules in the form of the Fujisaki's model parameters. In the next section, we describe a mathematical model for generating F_0 contours based on an extension of the Fujisaki's model to tone languages. Successfully incorporated into the model are the four major linguistic factors affecting the phonetic realization of tones in connected speech: continuity effect due to syllable structure, stress, tonal coarticulation and declination.

II. AN EXTENSION OF FUJISAKI'S MODEL OF F_0 CONTOURS TO TONE LANGUAGES

The Fujisaki's model is a mathematical model for a quantitative analysis and linguistic interpretation of F_0 contour characteristics [9]. The model was first proposed for handling accent in Japanese. Over the years, the model has been successfully extended and used for other languages, such as German and French [10,11]. The model has proven to be a highly effective tool for the analysis and synthesis of F_0 contours in text-to-speech systems in those languages.

A. The Original Model

Fujisaki first observed that an F_0 contour generally contains a smooth rise-fall pattern in the vicinity of the accented Japanese mora. Differences in rise-fall patterns seem to be attributable to the accent type, and these rise-fall patterns appear to be superimposed on a baseline that initially rises and gradually falls toward the end of the phrase or utterance regardless of the accent type. He hypothesized that the observed F_0 contour can be considered as the response of the phonatory system to a set of suprasegmental commands: the phrase (utterance) and the accent command. The phrase command produces the base line component while the accent command produces the accent component of an F_0 contour. From the above observation, he proposed a functional

model for generating an F_0 contour. The model is based on the idea of approximating F_0 contours as the response of a critically damped second-order linear system to excitation commands. The model is considered a superpositional model because it additively superimposes a basic F_0 value (F_{min}), a phrase component, and an accent component together on a logarithmic scale. The logarithmic scale of the frequency scale is based on the biomechanical considerations of the speech apparatus. In short, the output F_0 contour is a linear combination of F_{min}, a phrase component, and an accent component. A block diagram of the model is shown in Fig. 4

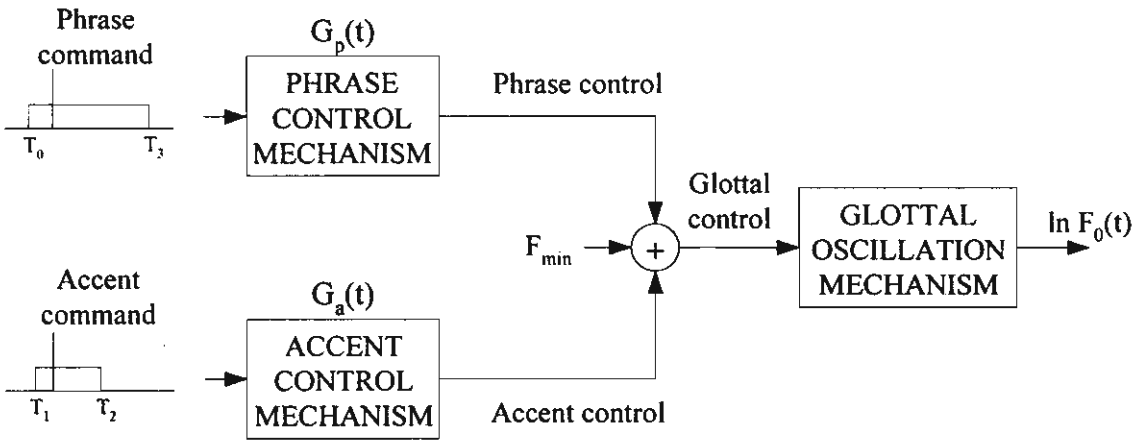

Fig. 4 A block diagram of the Fujisaki's model for synthesizing F_0 contour.

The control mechanisms of the two components are realized as critically damped second-order linear systems responding to rectangular functions. Mathematically speaking, an F_0 contour of an utterance generated from the model has the following functional form:

$$\ln F_0(t) = \ln F_{min} + \sum_{i=1}^{I} A_{pi}[G_{pi}(t-T_{0i}) - G_{pi}(t-T_{3i})] + \sum_{j=1}^{J} A_{aj}[G_{aj}(t-T_{1j}) - G_{aj}(t-T_{2j})]$$

where $G_{pi}(t) = \alpha_i t \exp(-\alpha_i t)\, u(t)$ and

$G_{aj}(t) = [1 - (1+\beta_j t)\exp(-\beta_j t)]\, u(t)$, $u(t)$ = unit step function,

indicate the step response function of the corresponding control mechanism to the phrase and accent command, respectively. F_{min} is the lower limit of F_0 below which vocal fold

vibration cannot be sustained in the glottis of a speaker. A_p's and A_a's are the amplitudes of the phrase and accent commands, respectively. T_{0i} and T_{3i} denote the onset and offset of the ith phrase command; T_{1j} and T_{2j} denote the onset and offset of the jth accent command; The α_i's an β_j's are time constant parameters characterizing a second-order system. I and J are the number of phrases and accented mora, respectively, contained in the utterance. The damping coefficient, which also characterizes a second-order system, is unity in the case of a critically damped system.

From the above description, the phrase component captures the global variation (declination effect) while the accent component captures local variations (accent effect) in the F_0 contour. The model is able to approximate naturally produced F_0 contours very accurately using only a small number of control parameters. These parameters are: the time constant parameters of the phrase and accent control mechanisms and the timing and amplitudes of the phrase and accent commands. The parameters can be empirically obtained by a curve-fitting method (i.e., minimizing the mean square error between the raw F_0 contour and that of the model) on a logarithmic scale. Fujisaki concluded from his experimental results that the time constant parameters and the damping coefficients could be constrained to remain constant without seriously affecting the resulting output F_0 contour.

B. Extending the Model to Thai and Other Tone Languages

To extend the above model to accommodate tone languages requires slight modifications to the model. In Japanese, the F_0 realization of local pitch accents results only in a rise-fall patterns in the F_0 contour. However, in the case of Thai, local F_0 variations due to tones results in a combination of both rise-fall patterns (e.g., a falling tone) and fall-rise patterns (e.g., a rising tone) in the F_0 contour. As a result, a model for tone languages will consist of two components, the phrase (or utterance) and the tone control mechanisms, driven by the phrase and the tone commands. The phrase command and phrase control mechanism are used to capture the declination effect; the tone command and tone control mechanism are used to capture tone types. Instead of a base line, the phrase command will produce a "mid" line. The tone commands in both positive and negative directions with respect to the mid line will produce local contours corresponding to tone types, which are superimposed on the mid line. As before, the model is characterized by time constant parameters and command amplitudes and their

temporal locations. In terms of the damping coefficient, critical damping is assumed for both the phrase and tone control mechanisms. Hence, the damping coefficient is always unity. These parameters are kept constant within the phrase unit. Again, these parameters can be obtained by the curve-fitting method mentioned above. Our extension to Fujisaki's model for tone languages is illustrated in Figure 5.

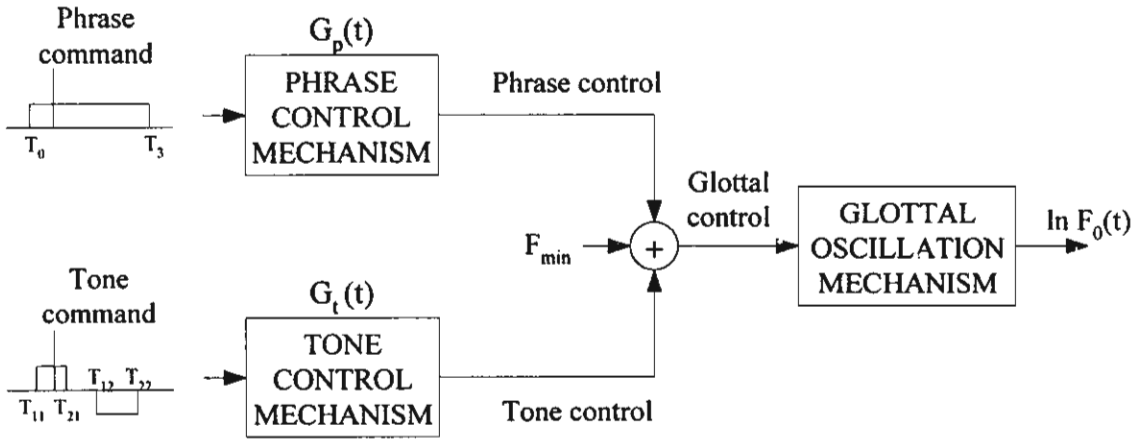

Fig. 5 Our extension of the Fujisaki's model of F_0 contours to tone languages.

Analogous to that of Fujisaki's original model, the tone synthesis model has the following mathematical expression:

$$\ln F_0(t) = \ln F_{\min} + \sum_{i=1}^{I} A_{pi}[G_{pi}(t-T_{0i}) - G_{pi}(t-T_{3i})] + \sum_{j=1}^{J}\sum_{k=1}^{K(j)} A_{t,jk}[G_{t,jk}(t-T_{1jk}) - G_{t,jk}(t-T_{2jk})]$$

where $G_{pi}(t) = \alpha_i t \exp(-\alpha_i t)\, u(t)$ and

$G_{t,jk}(t) = [1 - (1+\beta_{jk}t)\exp(-\beta_{jk} t)]u(t)$, $u(t)$ = unit step function,

indicate the step response function of the corresponding control mechanism to the phrase and tone command, respectively. F_{min} is the smallest F_0 value in the F_0 contour of interest. A_p's and A_t's are the amplitude of the phrase and tone command, respectively. T_{0i} and T_{3i} denote the onset and offset of the ith phrase command. T_{1jk} and T_{2jk} denote the onset and offset of the kth component of the jth tone command. α_i's and β_{jk}'s are time constant parameters characterizing a second-order system. I, J, and $K(j)$ are the number of phrases, tones, and components of the jth tone, respectively, contained in the utterance. It is noted that the logarithmic scale will be replaced by an equivalent-rectangular-bandwidth-rate

(ERB) scale, which is comparable to the logarithmic scale [12] and offers an advantage in that it gives equal prominence to excursions in different pitch registers. This is important in the synthesis of the F_0 contour of male and female speech.

C. Incorporating Linguistic Constraints into the Model

While tones in isolation have rather definite F_0 manifestations in the tone space, they undergo various modifications in connected speech due to syllable structure, stress, interactions from adjacent tones (tonal coarticulation), and declination. These linguistic factors affecting the F_0 realization of tones in connected speech can be easily incorporated into the model in terms of the tone command amplitudes and their temporal locations. In light of the findings in [13] which suggest that coarticulatory effects are physiologically conditioned by the mechanics of the vocal fold vibration, our choice of the Fujisaki's model is quite suited for capturing the effect. Physical phenomena, such as the mass-spring system and, in this case, the mechanical motion of the laryngeal mechanism responsible for pitch control can be mathematically described or modeled by a second-order linear system.

In this research, the values of model parameters which reflect the changes in the F_0 contour due to tonal coarticulation and declination are obtained by the following "training" procedure. Since a three-tone sequence is optimal for capturing coarticulatory effects as suggested by the acoustic experiment in the previous section, the parameters related to the tone command and control mechanism are estimated from each of the 125 possible three-tone sequences. A subset of utterances from the acoustic experiment on tonal coarticulation in [13] totaling 525 utterances (125 utterances X 5 speakers) are used as the set of training utterances. The raw F_0 contours of each utterance were first subject to the usual preprocessing, such as smoothing, normalization, etc.

To account for slight variations in speaking rate within and across speakers, a syllable-by-syllable temporal alignment procedure was used instead of a linear time normalization. Since the findings from the experiment regarding vowel length and stress [14] suggests that temporal variations of an utterance within and across speakers are not due to uniform stretching and shrinking of segments, linear time normalization throughout the whole utterance would be inadequate. Target syllables would not be properly aligned by linear time normalization. As a result, coarticulatory effects would not be consistently measured from utterance to utterance.

The syllable-by-syllable temporal alignment is accomplished as follows. First, an average duration for every syllable in the utterance is obtained by averaging across all corresponding syllable durations in all tonal sequences of all speakers. A ratio expressed in the percent of each average syllable duration to the average total utterance duration is then computed. Finally, a syllable-by-syllable linear time normalization is performed on each utterance based on the ratio for each syllable. The result of the syllable-by-syllable temporal alignment procedure is illustrated in Fig. 6. Three F_0 contours of the same utterance, produced by one speaker at three different speaking rates, were time-normalized without (top panel) and with (bottom panel) temporal alignment.

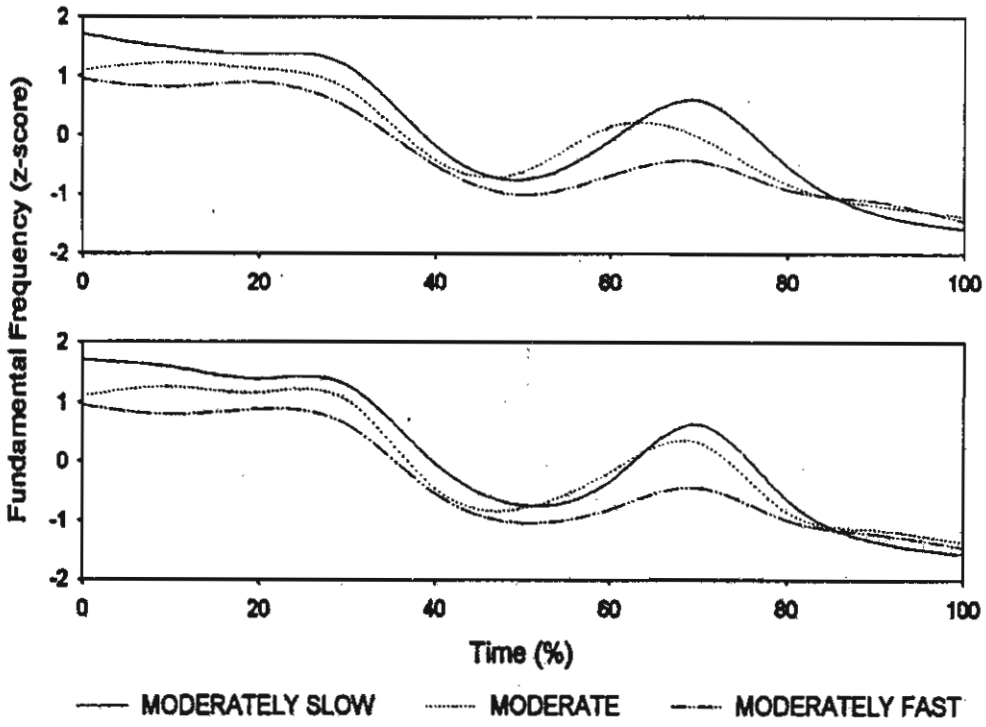

Fig. 6 F_0 Contours, time normalized without (top panel) and with (bottom panel) syllable-by-syllable temporal alignment.

After time-normalization, the resulting F_0 contours are converted to an ERB scale accounting for differences in the excursion size of F_0 movements related to differences in voice range between speakers. At this point, since the model is based on the principle of superposition, the steps of determining the phrase component parameters can be separated from the subsequent determination of the tone component parameters.

To account for the declination effect, an exponential curve is fitted to the F_0 contour. The exponential curve has the same functional form as the response function of the phrase control mechanism of the Fujisaki's model. The parameters of the exponential curve were used to characterize the declination effect of that utterance. The value of the exponential curve was then subtracted from the F_0 contour to eliminate the declination effect. Hence, the resulting difference waveform represents an F_0 contour with only the influence of coarticulatory effects. Before determining the parameters related to the effects, F_0 contours were normalized to a z-score scale to neutralize inter-speaker variability.

To account for the coarticulatory effects, the F_0 contour was processed from left to right, tone by tone. The parameter values of the tone component were determined by successive local approximations using an interactive (partly manual) method. That is, the F_0 contour was optimized on a tone-by-tone basis so that the preceding tone commands would not be affected by the following optimization process. Although the results are not expected to be optimal, we have gained an insight into how an automatic procedure can be developed. An automatic algorithm must ensure that a portion of the contour that has already been optimized is not affected by a succeeding tone command. An automatic procedure certainly guarantees the optimality of the estimated parameters. The implementation of such an automatic procedure is the subject of future research

To account for the continuity and stress effect, the temporal locations of the tone commands are adjusted according to the correspondence between abstract rhythmic grouping and the acoustic realization of each grouping (i.e., the relative syllable duration within a grouping). This acoustic realization will depend upon the phonetic structure of the syllables comprising it. These rules will be described in more detail as follows.

Speech is rhythmical not only because of the pattern of sounds and pauses, but also because of the regular recurrence of strongly accented sounds in a series. For example, in a stressed-time language, it is observed that speakers tend to produce stressed syllables at a regularly spaced interval of time while they tend to pause according to the syntax of the utterance 15]. The pause distribution seems to be ruled by syntactic constraints. Speech rhythm is also a psychological correlate of speech timing (an objective instrumental measurement of the duration of segments, syllables, etc.). Thus, in a stressed-time language, stress, pause, and relative syllable duration interact to form speech rhythm. In addition, the phonology and syntax of the language affect the description of speech rhythm as well.

Thai has a stress-timed rhythm [16]. This means that stressed syllables in Thai are perceived to be isochronous (i.e., they recur approximately at equal intervals of time). A phonological unit called foot is used to describe rhythmic groupings within an utterance. A foot is one of many prosodic constituents and is an elementary unit of the prosodic structure in addition to a syllable. A foot is neither a grammatical nor a lexical unit. The domain of a foot extends from a salient (stressed) syllable up to but not including the next salient syllable. A pause is considered a salient syllable, and the beginning of an utterance is always preceded by a pause. It should be noted that a rhythmic pause has a syntactic function, but a disfluency or hesitation pause does not.

In her analysis of Thai rhythm, Luangthongkum [16] posited five-foot structures:

1) | S | = 1-syllable foot,
2) | S W | = 2-syllable foot,
3) | S W W | = 3-syllable foot,
4) | S W W W | = 4-syllable foot,
5) | S W W W W | = 5-syllable foot,

where S and W indicate salient (stressed) and weak (unstressed) syllables, respectively. The 4-syllable and 5-syllable feet are very rare and are omitted from further discussion. Note that foot boundaries are usually inserted in front of the salient syllables.

At an abstract level, Luangthongkum [16] assumed that each rhythmic foot is arbitrarily three units long, regardless of the number of syllables comprising the foot. This suggests that as the number of unstressed syllables in the interval increases, a tendency toward equality of inter-stress intervals causes both the stressed and unstressed syllables to become shorter. Thus, the relative syllable duration for each type of rhythmic foot can be abstractly described as follows:

1) | S | → | 3 |,
2) | S W | → | 2 : 1 |,
3) | S W W | → | 1½ : ¾ ¾ |.

Phonetically, a rhythmic foot is not isochronous. The duration of a foot will differ somewhat depending upon the phonetic structure of the syllables comprising it. Thus, the acoustic realization of a rhythmic foot will be different from the above abstract description. The following is a set of rules proposed by Luangthongkum to predict how syllable duration in each type of foot is realized acoustically. The derived or predicted syllable duration was based on her acoustic analysis of read speech.

| 3 |	→	| 2 |	if the foot is in an utterance-initial position,
	→	| 4 |	if the foot is in an utterance-final position and it does not have a CVS structure.
| 2 : 1 |	→	| 2 : 2 |	if the salient syllable has a CVS structure; or the weak syllable is the first element of a compound that does not have a CVS structure; or both the salient syllable and the weak syllable are function words.
| 1½ : ¾ ¾ |	→	| 1⅔ : 1⅔ 1⅔ |	if the salient syllable has a CVS structure; or it is in an utterance-initial position; or it is a function word and the two weak syllables are two function words or a function word and a linker syllable.

A preliminary study of the extension to the Fujisaki's model for Thai produced very promising results. All of the linguistic factors affecting the F_0 realization of Thai tones in continuous speech have been successfully incorporated into the model: continuity effect, stress, tonal coarticulation, and declination. Figure 7 shows the actual and synthesized F_0 contours of two utterances with the same tone sequence but different segmental makeup: a continuously voiced utterance (top panel) and discontinuously voiced with intervening obstruents utterance (bottom panel) carrying an HLFHR tone sequence. It can be seen that the synthesized F_0 contour closely approximates the actual F_0 contour. Table 1 lists the values of the parameters obtained from the above analysis.

TABLE 1
Fujisaki's model parameters for the above utterances with an HLFHR tone sequence. Note that the last tone (R) has two components.

F_{min} (ERB)	i	T_0 (sec)	T_3 (sec)	A_p	α (sec^{-1})	tone	j	k	T_1 (sec)	T_2 (sec)	A_t	β (sec^{-1})
2.832	1	-.809	1.323	4.318	1.615	H	1	1	.003	.108	.9	20.2
						L	2	1	.135	.378	-1.0	20.2
						F	3	1	.378	.594	.8	20.2
						H	4	1	.783	.891	1.0	20.2
						R	5	1	.945	1.161	-.7	20.2
								2	1.269	1.323	1.0	20.2

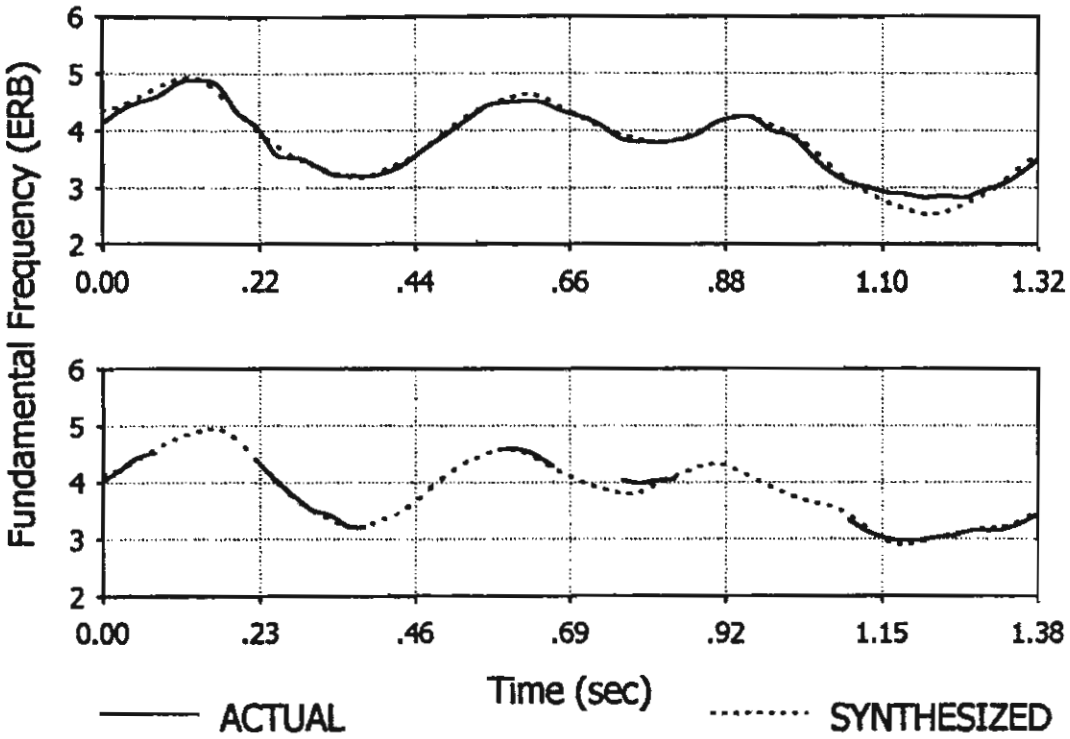

Fig. 7 The actual and synthesized F_0 contours of two utterances with the same tone sequence (HLFHR) but different segmental makeup: a continuously voiced utterance (top panel) and discontinuously voiced with intervening obstruents utterance (bottom panel).

III. THE PROPOSED TONE CLASSIFICATION ALGORITHM

In this section, details of the proposed automatic tone classification algorithm based on the analysis-by-synthesis method are presented. The algorithm takes into account all factors affecting phonetic realization of Thai tones as previously mentioned. Also discussed are important considerations for the normalization procedures to achieve speaker-independence.

The general design of the algorithm involves steps as shown in figure 8. The first three blocks represent the pre-processing of the speech signal to extract relevant information or acoustic features for subsequent classification. These are steps necessary to produce relatively reliable, normalized F_0 contours. The last three blocks represent the tone classification step based on the analysis-by-synthesis method. Each component of the system is described in detail below.

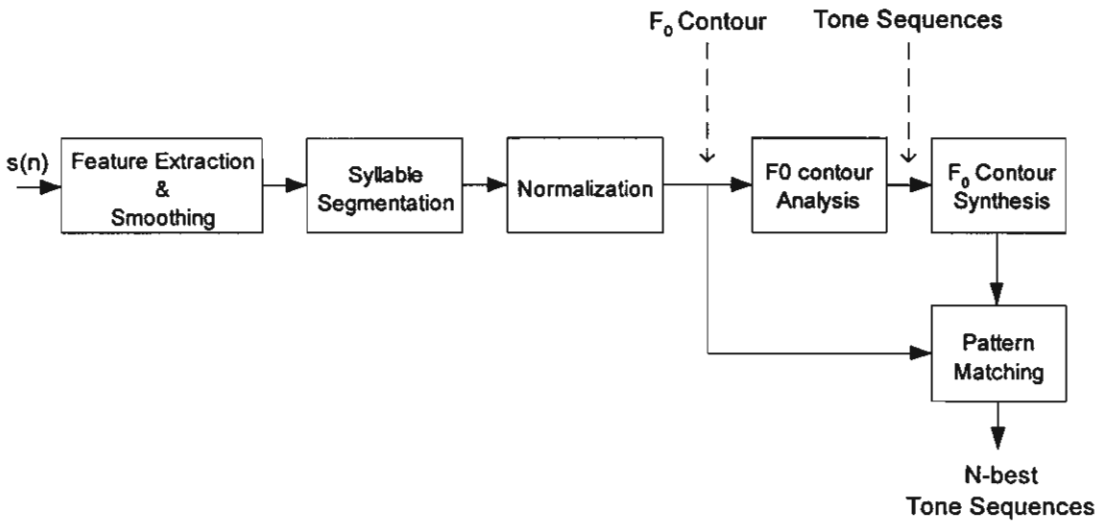

Fig. 8 The block diagram of the proposed tone classifier.

A. Feature Extraction and Smoothing

Two suprasegmental features corresponding to acoustic correlates of prosody (F_0 and intensity) are extracted from the speech input. First, the raw F_0 contour is automatically extracted from the input speech signal using one of several methods to pitch extraction. Our implementation of the tone classifier relies on a CSL pitch extraction algorithm which employs a time domain approach to pitch analysis (modified autocorrelation with center clipping) with nonoverlapping variable frame length. For a particular speaker, frame length will be determined by his/her pitch range to ensure that there were at least two complete cycles within a frame. A typical frame length is 20 to 25 ms for male speakers, 15 to 20 ms for female speakers. To eliminate "drop-outs" during voiced speech segments, spurious pitch values in regions of unvoiced speech segments, and/or "double pulsing" effect, smoothing techniques, such as median filtering and linear interpolation, must be employed. In this experiment, the F_0 contours were smoothed using the linear interpolation technique.

Secondly, the energy (intensity) measure will be used in placed of the amplitude measure of the speech signal since they are closely related. Energy calculation in decibels (dB) will be performed in a nonoverlapping frame-by-frame, pitch asynchronous manner using a Multi-speech algorithm that defines energy as the sum of the square of absolute amplitude values within a frame. Frame length will be kept constant at 20 ms for all speakers. The raw energy value will be converted into dB by computing 20 times the log (base 10) of the square root of the ratio between the energy to the number of

samples in the frame. A smoothing function will be applied to the resulting energy contour.

The energy contours obtained above will be used to crudely identify syllables with CVS structure (i.e., syllables ending with stop consonant, /p/, /t/, and /k/). This is important in determining the rhythmic grouping of the input utterance. Since these coda consonants are glottalized, the syllable ends abruptly and the signal energy decreases very rapidly at the end of the syllable. This rapid energy drop results mainly from the articulatory requirement of the final stop consonant. A syllable ending with a stop consonant will cease abruptly even if the voiced portion preceding the stop consonant has been prolonged. To parameterize this characteristic, a smoothed short-time energy profile $E_S(j)$ is obtained for the voiced portion of the syllable using the above-described procedure. Let $j_{\max}$ denote the frame number in which maximum energy occurs and t_d be the time required for the energy to drop from 90% to 10% of $E_S(j_{\max})$. We can define an energy drop rate as the reciprocal of t_d. That is, $R_D = \frac{1}{t_d}$. It should be noted that the energy drop rate are highly correlated with the syllable duration. The shorter the duration, the faster the energy drops.

B. Syllable Segmentation

Since tones are properties of syllables, it is logical to segment the smoothed and normalized F_0 contour into syllabic units. Syllable boundary information can be provided by an automatic syllable segmentation algorithm based on energy contours and spectral information, or by segmentation information from a phone recognizer unit. In this study, we have developed an automatic procedure for syllable segmentation. Automatic syllable segmentation is a crucial component that provides syllable boundary information necessary for our tone classification system. Traditionally, zero crossing rate and root-mean-square energy (RMSE) of the speech signal are the two most widely used features for locating syllable boundary. In this research, we propose a new segmentation algorithm based on a modified Teager's energy calculation [17]. We present details of the algorithm below.

The most common way of calculating the energy of a speech signal is the root mean square energy (RMSE), which is the square root of the average of the sum of the

squares of the amplitude of the signal samples. Using a window of width W to segment the speech signal into frames, the RMSE of frame n, E_n, is given by:

$$E_n = \left[\frac{1}{W}\sum_{i=1}^{W} s_n^2(i)\right]^{\frac{1}{2}},$$

where $s_n(i)$ denote the i^{th} windowed speech sample in frame number n.

On the other hand, in modeling speech production, Teager developed a new algorithm for computing the energy of a signal. This algorithm has been presented by Kaiser as Teager's Energy Algorithm. Given a signal with the motion of an oscillatory body, its sample is defined as

$$x_i = A\cos(\Omega i + \phi),$$

where A is the amplitude of the oscillation, Ω is the digital frequency, and ϕ is the initial phase. In Teager's Algorithm, the instantaneous energy E_i of the sample x_i is as follows:

$$\begin{aligned} E_i &= x_i^2 - x_{i+1}x_{i-1} \\ &= A^2\sin^2(\Omega) \\ &\approx A^2\Omega^2 \end{aligned}$$

It is noted that the output of Teager's Algorithm is a function of the amplitude of the signal samples, as well as the oscillation frequency. This new energy measure is therefore capable of responding rapidly to the changes in both A and Ω. Thus, it has the ability to track rapid changes as well as the qualitatively different character of various signals.

The fact that the Teager energy algorithm reflects both the amplitude and frequency of a signal suggests that it may be a more suitable measure for different speech events than the RMSE, which reflects only the amplitude of the signal. From the point of view of speech production, the amount of energy used to produce noise-like fricatives should not be an order of magnitude less than that used to produce periodic voiced sounds. Yet, this is the typical difference we often get when using RMSE measure. Fricatives and plosives sounds have very low amplitude, but, unlike most vowels, these sounds have energy distributed in the frequency range above 5 kHz. As a result, Teager's energy measure should be more suitable for the calculation of the energy used in producing those fricatives and plosives.

To apply Teager's energy calculation to the problem of speech segmentation, we observe that the expression for the instantaneous energy can be related to the square of

the samples of the derivative signal. This is equivalent to calculating the RMSE on the derivative of the speech sample x_i. The result is proportional to A^2 and Ω^2 as in Teager's energy calculation. As a result, we propose a new energy calculation based on a modification to Teager's calculation as follows:

1. Calculate the power spectrum of the speech signal;
2. Weight each sample in the power spectrum with the square of the frequency;
3. Take the square root of the sum of the weighted power spectrum.

Based on the above energy calculation, our syllable segmentation algorithm have been evaluated using the speech materials described in the appendix below. To evaluate performance, we visually compare the estimated locations of syllable boundary using the different energy measures (both RMSE and Teager's). Zero crossing rate is also computed and used to aid our visual inspection of the correct boundaries. The detected boundaries are compared with those obtained from manual segmentation via audio playback of the speech signals selected between the detected boundaries.

Preliminary results are encouraging revealing several general properties of this new energy calculation. First, the new measure confirms a higher energy level for fricatives and plosives than that obtained form RMSE measure. Secondly, compared to RMSE, the new measure decreases the energy difference between voiced and voiceless sounds. Lastly, The new measure suppresses the energy level of background noise during silence intervals.

In addition to syllable boundary information obtained above, we also extract the durational patterns of every syllable in the utterance. Based on our automatic syllable segmentation algorithm above, syllable duration is also computed. Note that syllable duration for our purpose is defined as the duration D of the voiced portion of a syllable only. This durational information will be used in discriminating between stress and unstressed syllables in the input utterance. For the purpose of computing the speaking rate, total duration marked by the beginning and end of the utterance is also calculated. The total duration of the target sentence will be measured from the onset of the consonant at the beginning of the sentence to the cessation of the coda consonant (closed syllable) or vowel (open syllable) of the last syllable at the end of the sentence. Speaking rate will then be computed by dividing the total sentence duration by the number of syllables in that sentence. The speaking rate will be used in the normalization process, which will be described next.

C. Normalization

Normalization of the feature parameters is necessary because it will eliminate undesirable time and speaker variations of these parameters. In terms of pitch, for a multiple-speaker system, the normalization process is introduced to neutralize variability from one F_0 contour to the next. Sources of variability include speaker's physiological differences, the kinetics of vocal fold vibration, consonantal perturbations on F_0, and speaking rate. The raw F_0 contour is first converted into an equivalent-rectangular-bandwidth-rate (ERB) scale. This ERB normalization has an effect of neutralizing pitch ranges of different excursion size. To neutralize the declination effect in the F_0 contour, we subtract a time-varying mean F_0 value from the input F_0 contour. A time-varying mean F_0 value is computed by fitting an exponential curve to the overall contour as already discussed. Then, z-score normalization is employed to account for pitch range differences across speakers based on the precomputed mean and standard deviation from all utterances in the training set. This method has the effect of making the first- and second-order moments of the pitch distributions the same.

For the duration-related parameters D and R_D, normalization is needed. The speaking rate can be affected by emotional, stylistic and environmental factors, which may change from time to time. For example, the duration of a long syllable can be very short for fast speaking persons. The normalization factors are the precomputed mean from all utterances in the training set.

D. F_0 Contour Analysis

This step is necessary to reduce the number of possible reference templates that have to be generated by the synthesis module, and thus, reduce the amount of time it takes to match against the input F_0 contour. The analysis procedure consists mainly of two steps. First, using the syllable durational patterns, a rhythm grouping among adjacent syllables is determined from the rules given in the previous section. That is, the relative syllable duration for each type of rhythmic foot can be abstractly described together with the corresponding rule for matching the acoustic realization of a rhythmic foot with the abstract description.

Once the rhythmic grouping is determined, the second step involves the peak-and-valley analysis, i.e., the detection of local extrema of the given smoothed, normalized and segmented F_0 contour for that grouping. Local extrema (peaks and valleys) are detected

by using first and second derivatives. The derivative at any point in the contour, except for the first two and last two points, is computed by calculating the linear regression coefficients of a group of five F_0 values consisting of the current point, and its preceding and following two points.

The locations of these extrema coupled with syllable boundary information and the energy drop rate are then used to identify all possible tone labels for the salient syllables in the rhythmic grouping based on some specified rules. For example, between two syllable boundaries, only the falling tone can occur if a maximum occurs, and only the rising or the high tone can occur if a minimum occurs. Also, if a maximum occurs at or in the vicinity of a syllable boundary, the preceding tone can either be a high or a rising tone. If a minimum occurs at or in the vicinity of a syllable boundary, the preceding tone can either be a mid or a low tone, or a sequence of two falling tones. For the rest of the weak or unstressed syllables within the given grouping, only three tonal labels (FH, M, and LR) are assigned depending on the overall temporal pitch variation. The FH label indicates an upward trend, the LR a downward tend, and M a level trend. These labels are derived based on the information obtained from the acoustic experiments described in [18]. They reflect the fact that unstressed syllables suffer tone neutralization, and the contrastive pattern among tones can be divided into roughly three tonal registers.

To deal with syllables with different duration, a time-aligned pitch profile is used [19]. The voiced portion of the syllable is divided evenly into 16 segments. For each segment, a pitch value is obtained from the given F_0 contours using a linear interpolation method. Thus, the pitch profile of each syllable has the same dimension of 16. Given a pitch profile $\{P(1), P(2), \ldots\ldots, P(i), \ldots, P(16)\}$, the overall temporal pitch variation within the profile can be measured using a pitching rising index, I_R, which is defined as

$$I_R = k \cdot \frac{\mathrm{Max}_{i=2}^{15}\{P(i)\} - \mathrm{Min}_{i=2}^{15}\{P(i)\}}{\mathrm{Max}_{i=2}^{15}\{P(i)\} - \mathrm{Min}_{i=2}^{15}\{P(i)\}}$$

$$\text{where } k = \begin{cases} 1 & \arg\mathrm{Max}_{i=2}^{15}\{P(i)\} > \arg\mathrm{Min}_{i=2}^{15}\{P(i)\} \\ -1 & \arg\mathrm{Max}_{i=2}^{15}\{P(i)\} > \arg\mathrm{Min}_{i=2}^{15}\{P(i)\} \end{cases}$$

It is noted that the first and the last segment of the pitch profile ($P(1)$ and $P(16)$) are not used in order to reduce possible errors in the pitch extraction process. The polarity of I_R

indicates the overall temporal trend of pitch movement within the utterance and the magnitude of I_R represents the degree of such variation.

E. F_0 Synthesis

Based on the extension of Fujisaki's model for synthesizing F_0 contours to Thai described in the previous section, the input tone sequences are used to generate predicted F_0 contours. These predicted F_0 contours are basically reference templates to be used for pattern matching against the input contour.

F. Pattern Matching

The classification of input F_0 contours into likely sequences of tones is accomplished in this step by pattern matching against the predicted F_0 contours or reference templates generated by the F_0 model. Pattern matching techniques, such as a simple zero-lag crosscorrelation method or a one-stage dynamic programming search can be used. In both cases, some measure of goodness of fit must be established in order to rank the results so that N-best tone sequences can be obtained. For example, for the zero-lag crosscorrelation method, a correlation coefficient of 0.9 or higher could be used to indicate a relatively good fit. Thus, we can infer that a strong similarity exists between the input and the predicted F_0 contours. For a one-stage dynamic programming search, a distance measure might be more appropriate. In this paper, we used the zero-lag crosscorrelation method.

IV. PERFORMANCE EVALUATION AND DISCUSSION

In order to train and evaluate our computer model, we need additional speech materials. Thirty-five target sentences of 11-15 syllables in length are chosen to closely represent continuous speech. Each target sentence consists of syllables with varying tone sequences. Additional requirement is that some of the sentences comprise voiced sounds throughout in order to increase the level of difficulty in performing the syllable segmentation procedure in our tone classification algorithm. The target sentences described above are listed in the Appendix. They were produced by a set of five speakers. Thus, there was a total of 175 utterances in the test set. Test stimuli were different from the training stimuli used in training the Fujisaki's model.

The classification test was performed on each of the 175 utterances from the test set to obtain the crosscorrelation coefficients between the input contour and each of the predicted contours. All in all, the algorithm misclassified 32 of 175 test utterances. Hence, the classification accuracy for this experiment is approximately 81.7%. In this experiment, the number of N-best output tone sequences is equal to six, i.e., N = 6. The number six was chosen arbitrarily. The reason for outputting N-best tone sequences as inputs to the word hypothesizer is because it is likely that the correct tone sequence could be recovered at that stage by using other linguistic constraints, such as tonal restrictions on the types of syllable structures, etc. The overall performance of the synthesis module was quite reliable in producing F_0 contours. Misclassification mainly occurs with unstressed syllables, especially linker syllables and function words. There are a total of 2,230 syllables in the test stimuli, and only 1822 were correctly classified. This might be due to the fact that unstressed syllables suffer not only from tone neutralization but also from the interaction with adjacent syllables in terms of tonal coarticulation. It is believed that this problem may worsen in the case of polysyllabic words containing linker syllables. However, this problem should not be solved at this stage, but at the stage of word hypothesization where pronunciation dictionary will help rule out ill-formed word.

V. CONCLUSION

A mathematical model for generating F_0 contours for Thai and other tone languages was presented. The model is based on an extension of the Fujisaki's model of F_0 contours. Successfully incorporated into the model are linguistic factors affecting phonetic realization of Thai tones in continuous speech. They are continuity effect due to syllable structure, stress, tonal coarticulation, and declination.

Furthermore, a bottom-up or data-driven approach to automatic classification of Thai tones in connected speech was described. The algorithm is based on the analysis-by-synthesis approach to speech perception, and it is simpler to implement than the left-to-right HMM-based system. Also, we believe that the computational cost of our model is much less than the HMM-based system because it uses fewer parameters.

The present implementation of the algorithm is a continuation of the work done by the principal investigator [20]. Several limitations, such as a lack of automatic segmentation of syllable boundaries, a need to incorporate stress effects into the synthesis module, and a small number of test sentences have been rectified. However, we still are

not quite satisfied with the accuracy of the algorithm, yet the results indicate a step in the right direction toward implementing a connected speech tone recognition system. We believe that the overall performance of the algorithm can be improved through a better training of the model, a better pattern matching method, and a more robust F_0 contour analysis method.

ACKNOWLEDGEMENTS

This material is based upon work supported by Thailand Research Fund under Grant No. RSA/03/2541. The first author would like to extend his gratitude to the Academic division of Chulachomklao Royal Military Academy, the Royal Thai Army for the opportunity to conduct this research. Reprint requests should be sent to: Siripong Potisuk, Ph.D., Department of Electrical and Computer Engineering, Academic Division, Chulachomklao Royal Military Academy, Nakon-nayok, 26001 THAILAND.

REFERENCES

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

[2] J. T. Gandour and R. Harshman, "Cross-language differences in tone perception: A multidimensional scaling investigation," *Language and Speech*, vol. 21, pp. 1-33, 1978.

[3] J. T. Gandour, "Tone perception in Far Eastern languages," *Journal of Phonetics*, vol. 11, pp. 149-175, 1983.

[4] X. Chen, C. Cai, P. Guo, and S. Ying, "A hidden Markov model applied to Chinese four-tone recognition," in *1987 International Conference on Acoustics, Speech and Signal Processing*, Vol. II, May 1987, pp. 787-800.

[5] L., Liu, W. Yang, H. Wang, and Y. Chang. Tone recognition of polysyllabic words in Mandarin Speech," *Computer Speech and Language*, vol. 3, pp. 253-264, 1989.

[6] R. Wu, J. A. Orr, and S-K. Hsu, "Recognition of four tones in Chinese speech by parametric estimation of frequency trajectories," in *1989 2nd Biennial Acoustics, Speech and Signal Processing Central New England Miniconference*, 1989.

[7] T. Lee, P. C. Ching, L. W. Chan, Y. H. Cheng, and Brian Mak, "Tone recognition of isolated Cantonese syllables," *IEEE Trans. Speech and Audio Processing*, vol. 3-3, pp. 204-209, May 1995.

[8] G. H. Yeni-Komshian, "Speech Perception," in *Psycholinguistics*, J. B. Gleason and N. B. Ratner, Eds. Fort Worth: Harcourt Brace, 1993, pp. 89-131.

[9] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The production of Speech*, P. F. MacNeilage, Ed. New York: Springer-Verlag, 1983, pp. 39-55.

[10] B. Möbius, M. Pätzold, and W. Hess, "Analysis and synthesis of German F_0 contours by means of Fujisaki's model," in *Speech Communication*, vol. 13, pp. 53-61, 1993.

[11] G. Bailly, "Integration of rhythmic and syntactic constraints in a model of generation of French prosody," *Speech Communication*, vol. 8, pp. 137-146, 1989.

[12] D. Hermes and J. Van Gestel, "The frequency scale of speech intonation," *Journal of the Acoustical Society of America*, vol. 90, pp. 97-102, 1991.

[13] S. Potisuk, J. T. Gandour, and M. P. Harper, "Contextual Variations in Trisyllabic Sequences of Thai Tones," *Phonetica*, vol. 54, pp. 22-42. 1997.

[14] S. Potisuk, J. T. Gandour, and M. P. Harper, "Vowel and stress in Thai," *Acta Linguistica Hafniensia*, vol. 30, pp. 39-62. 1998.

[15] J. P. Gee and F. Grosjean, "Performance structures: A psycholinguistic and linguistic appraisal," *Cognitive Psychology*, vol. 15, pp. 411-458, 1983.

[16] T. Luangthongkum, *Rhythm in standard Thai,* Ph.D. dissertation, University of Edinburgh, 1977.

[17] J. F. Kaiser, "On a simple algorithm to calculate the energy of the signal," *1990 Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, pp. 381-384. April, 1990.

[18] S. Potisuk, "The effects of stress on F_0 contours of Thai tones in connected speech," Manuscript submitted for publication.

[19] A. Komatsu, A. Ichikawa, K.Nakata, Y. Asakawa, and H. Matsuzaka, "Phoneme recognition in continuous speech," in *1982 Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, pp. 883-886. May, 1982.

[20] S. Potisuk, M. P. Harper, and J. T. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method," *IEEE Transaction on Speech and Audio Processing*, Vol.7, No.1, January 1999, pp.95-102.

APPENDIX

Stimuli for Evaluating Automatic Tone Classification Algorithm

1. หญิงงามอย่างนี้แมวมองไม่มีเมินอย่างแน่นอน
   / jǐŋ ŋaam jàaŋ níi mɛɛwmɔɔŋ mâj mii mɔɔn jàaŋ nêɛnɔɔn /
   'Such a beautiful girl like this is definitely not ignored by the scout.'

2. อย่าลืมว่าเวลามันล่วงเลยมายาวนานแล้ว
   / jàa lɯɯm wâa weelaa man lûaŋ lɔɔj maa jaaw naan lɛ́ɛw /
   'Don't forget that time has passed for so long.'

3. งูเหลือมใหญ่เลื้อยอยู่ในหย่อมหญ้าอย่างเหนื่อยหน่าย
   / ŋuulɯ̌am jàj lɯ́aj jùu naj jɔ̀m jâa jàaŋ nɯ̀aj nàaj /
   'A big python is winding tiredly in the grass.'

4. ลุงหวานขอน้ำหญิงว่างามยิ่งเมื่อยามยิ้มแย้ม
   / luŋwǎan jɔɔ náa jǐŋ wâa ŋaam jîŋ mɯ̂a jaam jím jɛ́ɛm /
   'Uncle Wahn praises aunt Ying for her beauty when she smile.'

5. หล่อนนั่งเหม่อลอยเมื่อโหน่งเล่าว่าหนิงยังไม่ยินยอม
   / lɔ̀ɔn nâŋ mɔ̀ɔlɔ̀ɔj mɯ̂a nòoŋ lâw wâa nǐŋ jaŋ mâj jinjɔɔm /
   'A big python is winding tiredly in the grass.'

6. อย่าหวั่นไหวเมื่องานไม่ลื่นไหลเหมือนเมื่อยังหนุ่ม ๆ อยู่
   / jàa wànwǎj mɯ̂a ŋaan mâj lɯ̂ɯn lǎj mɯ̌an mɯ̂a jaŋ nòm nòm jùu /
   'Don't be discouraged if work doesn't go your way like when you were young.'

7. เหน่งลิ้มลองเนื้อน้อยหน่าหนังเมื่องานเลี้ยงวันวาน
   / nèŋ límlɔɔŋ mɯ́a nɔ́ɔjnàanǎŋ mɯ̂a ŋaan liáŋ wan waan /
   'Neng tasted the meat of a sugar apple at yesterday's party.'

8. นายยิ่งยงยังยุ่ง ๆ อยู่เลยไม่ลางานมาเยี่ยมหลาน ๆ
   / naaj jîŋjoŋ jaŋ jûŋjûŋ jùu lɔɔj mâj laa ŋaan maa jîam lǎan lǎnn /
   'Yingyong is quite busy to take a leave of absence to come visit grandchildren.'

9. หน่อยงุนงงเมื่อนายนนท์เมามายแล้วมาลุ่มล่ามอย่างนี้
   / nɔ̀ɔj ŋunŋoŋ mɯ̂a naaj non mawmaaj lɛ́ɛw maa lûmlaâm jàaŋ níi /
   'Nawj is puzzled that the drunken Non is trying to take advantage of her.'

10. นางนิ่มยืนยันว่าน้องนนนี่นำยาน้ำมันเหลืองมาเล่น

/ naaŋ nîm jɯɯnjan wâa nɔ́ŋ nonniî nam jaanáammanlɯ̌an maa lên /
'Nim insists that Nonni is the one playing with the yellow oil.'

11. หนูยุ้ยล้อเลียนหมอหยองว่างมงายไม่น้อยเลย

/ nǔujúuj lɔ́ɔlian mɔ̌ɔjɔ̌ɔŋ wâa ŋomŋaaj mâj nɔ́ɔj lɔ̀ɔj /
'Yuj is making fun of the fortune-teller, Yong, for being quite superstitious.'

12. ยามนี้ลุงเนื่องย่ำแย่เลยไม่น่ามีเงินเหลือ

/ jaam níi luŋnɯ̂aŋ jâmjêɛ lɔ̀ɔj mâj nâa mii ŋɔn lɯ̌a /
'For the moment, uncle Nuang is in trouble and has no money left.'

13. น้ำแหววิงวอนว่าอย่าลงไม้ลงมือเลย

/ náawɛ̌ɛw wiŋwɔɔn wâa jàa loŋmájloŋmɯɯ lɔ̀ɔj /
'Aunt Waew is pleading for non-violent means.'

14. หม่อมหลวงเนื่องย้ำว่าน้องหมิวยังเยาว์วัยอยู่

/ mɔ̀ɔmlɯ̌aŋ nɯ̂aŋ jám wâa nɔ́ɔŋmǐw jaŋ jaw waj jùu /
'M.L. Nuang emphasizes that Mew is still quite young.'

15. น้องแมวไม่ยอมเล่าว่าลืมแหวนวงนั้นไว้ไหน

/ nɔ́ɔŋmɛɛw mâj joom lâw wâa lɯɯm wɛ̌ɛn woŋ nán waj nǎj /
'Maew refuses to tell where she misplaced that ring.'

16. วันนี้นั้นย่อมไม่เหมือนวันวานอย่างแน่นอนเลย

/ wanníi nán jɔ̂ɔm mâj mɯ̌an wanwaan jàaŋ nêɛnɔɔn lɔ̀ɔj /
'Today is certainly not the same as yesterday.'

17. น้ำหม่องแหย่น้องแหม่มว่าหน้าไม่เหมือนแม่นุ่นเลย

/ náamɔ̀ɔŋ jɛ̀ɛ nɔ́ɔŋmèm wâa nâa mâj mɯ̌an mêɛnûn lɔ̀ɔj /
'Uncle Mong teases Mam for her unresemblance to her mother, Nuun .'

18. น้องแมนงอนแม่เลี้ยงเลยหนีมานั่งนิ่ง ๆ

/ nɔ́ɔŋman jɔɔn mêɛliáŋ lɔ̀ɔj nǐi maa nâŋ níŋníŋ /
'Man was upset at his stepmother and ran from her to sit quietly alone.'

19. หมอเล่าว่าหมู่มั่นเหนื่อยง่ายแม้เวลาวิ่งเหยาะ ๆ

/ mɔ̌ɔ lâw wâa mùu mân nɯ̀aj jâaj méɛ weelaa wîŋ jɔ̀ʔjɔ̀ʔ /
'The doctor says that corporal Mun gets tired easily even when lightly jogging.'

20. ยายเมี้ยนนำเนื้อวัวมาย่างไว้ยำเย็นวันนี้

/ jaajmián nam nɯ́awua maa jâaŋ wáj jam jen wannií /
'Old Mian grilled beef for making tonight's salad.'

21. หลวงลุงนำหมูหยองในย่ามมาโยนไว้ในหลุม

/ lŭaŋluŋ nam mŭujɔ̆ŋ naj jâam maa joon wáj naj lŭm /
'The old monk threw shredded fried pork into the hole.'

22. มาลินลืมล้อมหมูไว้ในเล้าเมื่อเย็นวานนี้

/ maalin lɯɯm lɔ̂ɔm mŭu wáj naj láw mɯa jen waannîi /
'Malin forgot to herd her pigs into the pen yesterday evening.'

23. เชิญท่องเที่ยวทั่วถิ่นแคว้นแดนไทยไปกับทัวร์เอื้องหลวง

/ cʰɔɔn tʰɔ̂ŋtʰîaw tʰûa tʰìn kʰwéen deen tʰaj paj kàp tʰua ʔɯ̂aŋlŭaŋ /
'Come visit every inch of Thailand with the Royal Orchid Tour.'

24. ไอทีวีมีเรื่องราวหลากหลายให้ได้ชมกันทุกวัน

/ ʔajtʰiiwii mii rɯ̂aŋraaw làaklăaj hâj dâaj cʰom kan tʰúk wan /
'ITV offers a wide variety of programs for our viewing pleasure everyday.'

25. ทุกคนชื่นชมคนซื่อสัตย์เฉกเช่นชายชื่อชวน

/ tʰúkkʰon cʰɯ̂ɯncʰom kʰon sɯ̂ɯsàt cʰèk cʰên cʰaaj cʰɯ̂ɯ cʰuan /
'Everyone admires an honest person like Mr. Chuan.'

26. แชมพูสมุนไพรช่วยบำรุงผมให้กลับนุ่มเงางามได้

/ cʰɛempuu samŭnpraj cʰûaj bamruŋ pʰŏm hâj klàp ŋawŋaam dâaj /
'Herbal shampoo helps revitalize your hair for soft and silky feel.'

27. รายการเหลียวหลังแลหน้ากล้าเจาะลึกประเด็นข่าวสำคัญ

/ raajkaan lĭawlăŋleenâa klâa cɔ̂ʔ luík pràden kʰàaw sămkʰan /
'The program "Glance Back and Look Ahead" dares to probe important issues.'

28. เชิญแวะมาชิมอาหารหลากรสได้ที่ร้านซุ้มสามสาว

/ cʰɔɔn wɛ́ʔ maa cʰim ʔaahăan làak rój dâaj tʰîi ráan súmsăamsăaw /
'You're welcome to sample many delicious dishes at Sam Sao restaurant.'

29. แม่ เตือนน้องข่ามว่าอย่าเข่นเขี้ยวเคี้ยวฟันเมื่อยามโกรธ

/ mêɛ tɯan nɔ́ɔŋkʰàam wâa jàa kʰènkʰîawkʰiáwfan mɯ̂a jaam kròt /
'Mother warns Kham not to grind her teeth when angry.'

30. กับแกล้มบ้านลุงโกร่งพอจะกล้อมแกล้มไปได้บ้าง

/ kàpklêɛm bâan luŋkròŋ pʰɔɔ càʔ klɔ̂ɔmklêɛm paj dâaj bâaŋ /
'Appetizer at Uncle Krong's house will do.'

31. ไม่ควรด่วนตัดสินใจไขว่คว้าหาคู่ครองเมื่อยังเด็กอยู่

/ mâj kʰwuan dùan tàtsǐncaj kʰwàjkʰwáa hǎa kʰûukʰrɔɔŋ muûa jaŋ dèk jùu /
'Don't decide to get marry at a young age.'

32. อรอนงค์ออกอาการอึดอัดเมื่ออ๊อดเอ่ยปากชวน

/ ʔɔɔnʔanooŋ ʔɔ̀ɔk ʔaakaan ʔɯ̀tʔàt muûa ʔɔ́ɔt ʔəəj paak cʰuan /
'Awn-anong felt uncomfortable when Aut invited her.'

33. น้องต้องตวาดหนูตุ๊ดตู่ว่าเดินต้วมเตี้ยมเป็นเต่าเลย

/ nɔ́ɔŋtɔ̂ŋ tawàat nŭu túttùu wâa dəən tûamtîam pen tàw ləəj /
'Tong snaps at Toot-too for walking so slow like a turtle.'

34. เธอควรเอาพระบนหิ้งมาห้อยเผื่อเวลาเคราะห์หามยามร้าย

/ tʰəə kʰuan ʔaw pʰrá? bon hîŋ maa hɔ̂ɔj pʰùa weelaa kʰrɔ́ʔ hǎam jaam ráaj /
'You should take Buddha images from the shelf and wear in case of bad luck.'

35. ความวัวยังไม่ทันหาย ความควายดันเข้ามาแทรก

/ kʰwaamwua jaŋ mâj tʰan hǎj kʰwaamkʰwaaj dan kʰâw maa sɛ̂ɛk /
'No sooner had one bad thing subsides than the occurrence of the other.'

Cover Page

Paper Information

Title: Prosody Generation in a Thai Text-to-speech System

Authors: Dr. Siripong Potisuk

Affiliation: Department of Electrical & Computer Engineering, Chulachomklao Royal Military Academy

Address: Academic Division, CRMA
Muang District, Nakorn-nayok 26001

Technical Areas: Speech and Audio Processing

Correspondence Author

Name: Siripong Potisuk
E-mail address: srppts@hotmail.com
Phone number: (037) 393-484 (office), (01) 170-3655
Fax number: (037) 393-484
Mailing address: CRMA P.O. Box 16, Muang District, Nakorn-nayok 26001

Prosody Generation in a Thai Text-to-speech System

Abstract

At present, it is generally agreed that prosody generation or synthesis is one of the least developed parts of existing systems for converting text to speech. This paper describes our preliminary work on a prosody-generating aspect of a text-to-speech system for Thai. Specifically, we are interested in modeling prosody by predicting symbolic markers from text (i.e., prosodic phrase boundaries, accent, and intonation boundaries), and then using these markers to generate pitch, energy, and duration patterns for the synthesis module of the system. The first part of this paper describes the prosody annotation process in which the foot structure (i.e., the rhythm of the utterance) is obtained from text. Then, the second part deals with the prosody synthesis process, including the prediction of segmental duration patterns and the generation of fundamental frequency (F_0) and energy contours. The mathematical model used for generating F_0 and energy contours is based on an extension of the Fujisaki's model of F_0 contours to tone languages.

1. Introduction

In general, speech generation or synthesis can be accomplished by one of the following three methods: general-purpose concatenative synthesis, corpus-based synthesis, and phrase splicing.

First of all, for the general-purpose concatenative synthesis, the system translates incoming text into phoneme labels, tone labels (for the case of tone languages), stress and emphasis tags, and phrase break tags. This information is then used to compute a target prosodic pattern (i.e., phoneme duration, and pitch and energy contours). In order to generate an output utterance, signal-processing methods are used to retrieve acoustic units from a stored inventory, modify the units so that they match the target prosody, and glue and smooth them together. Such acoustic units are primarily fragments of speech corresponding to short phoneme sequences such as diphones. As for speech quality and scope, general-purpose concatenative synthesis is able to handle any input sentence but generally produces mediocre quality.

Secondly, corpus-based synthesis, although quite similar to general-purpose concatenative synthesis, uses a stored inventory consisting of a large corpus of labeled speech. And, instead of

modifying the stored speech to match the target prosody, the corpus is searched for speech phoneme sequences whose prosodic patterns match the target prosody. Corpus-based synthesis can produce very high quality, but only if its speech corpus contains the right phoneme sequences with the right prosody for a given input sentence. If the corpus contains the right phonemes but with the wrong prosody, the end result may locally sound quite good, but the utterance as a whole may have a bizarre sing-song quality with confusing accelerations and decelerations.

Finally, for phrase splicing, stored prompts, sentence frames, and stored items used in the slots of these frames, are glued together. And, obviously, phrase splicing methods produce completely natural speech, but can only say the pre-stored phrases or combinations of sentence frames and slot items; naturalness can be a problem if the slot items are not carefully matched to the sentence frames in terms of prosody.

For the purpose of this paper, we are concerned with concatenative speech synthesis only. Concatenative synthesis has the edge on size because of an increasing interest in using speech synthesis on handheld devices. This is true since its quality limitations are less of a problem given that the acoustic capabilities of handheld devices are themselves limited. However, the price we pay is that the cost of generating a corpus or an acoustic unit inventory is significant. Besides making the speech recordings, each recording has to be analyzed microscopically by hand to determine phoneme boundaries, phoneme labels, and other tags.

At present, widespread use of text-to-speech technology is limited by its inability to produce high-quality speech. Thai is, intelligibility and naturalness of synthetic speech is still not quite at the level acceptable by human listeners. It is quite obvious from the above discussion that prosody generation must be of primary concern for improving the quality of synthetic speech.

Prosody is often described as a suprasegmental feature of speech (a term for describing phonological features of those aspects of speech that involve more than single consonants or vowels). Acoustically speaking, prosody can be defined as change in the fundamental frequency (F_0), timing, and amplitude of a speech signal. Speakers control the prosody of an utterance in order to signal linguistic and affective information. Linguistic prosody is used by speakers to signal grammatical information at the syllable, word, or sentence level (e.g., stress, intonation). Affective prosody, on the other hand, is used to convey information that indicates speaker's intentions, attitudes, or emotional states. In addition to linguistic and affective information, prosody can also be used to convey non-linguistic information concerning speakers' personal characteristics such as age, gender, idiosyncrasy, speaking style, and physical condition. Such characteristics may or may not be under the speaker's volitional control. It is part of the intelligibility and naturalness of his/her speech. This paper will deal only with linguistic prosody.

The role of linguistic prosody in spoken language is similar to that of punctuation in written language. Punctuation is used to divide a stream of text into smaller segments such as a phrase, clause, or sentence, and thus, it helps readers interpret the message according to the intentions of the writer. Likewise, prosodic information helps listeners interpret a spoken sentence in the way the speaker intends. The need for punctuation or prosody can be attributed in part to the inherent ambiguity of natural language.

Intuition tells us that intelligibility and naturalness can be attributed to prosody in our everyday use of speech. Some words in a sentence are louder and longer than others. Because function words are acoustically less prominent than the semantically important content words, such as nouns and verbs, we can prosodically distinguish them. Pauses tend to be inserted at certain points in the sentence, and words at the end of the sentence are likely to be lengthened. This suggests the existence of prosodic constituents that are used in the overall prosodic structure or melody of an

utterance. Linguists have posited units such as syllables, prosodic words, phonological phrases, and intonational phrases.

The use of prosody by speakers, in attempting to sound intelligibly and naturally, can be best exemplified by considering its use in ambiguous sentences. When two sentences are segmentally identical, a problem of identifying the correct meaning arises for a listener, especially when the contextual information is not adequate. In such cases, the listener can make use of another type of information, namely prosody. The question arises, from the speaker's point of view, as to how this prosodic information should be encoded, and from the listener's point of view, how this information is decoded or associated with different meanings. At an abstract level, a commonly accepted hypothesis is that there is a direct relationship between the syntactic structure of a sentence and its prosodic structure [1,2]. This hypothesis implies that an ambiguous sentence will have a different prosodic structure for each syntactic structure, and as such it can be used to determine the correct meaning. At the phonetic level, the speaker tends to manipulate the acoustic correlates of prosody, such as F_0, segmental and pause duration, amplitude, and spectrum of the speech signal in order to signal prosody. The listener, in turn, will try to translate the changes in these physical correlates into abstract linguistic concepts in order to arrive at the intended meaning of the utterance.

As in human speech production, it is believed that prosodic information can help improve performance of a text-to-speech system. Prosodic information is particularly helpful in generating synthetic speech because of lexical and structural ambiguities of written forms. Prosodic information could be used by computers to generate phonetically similar, but syntactically different utterances.

In the following sections, a novel method for generating prosody in a text-to-speech system will be described. Two specific issues will be addressed: the prosodic annotation process and the prosody synthesis process. The prosodic annotation process will be abstractly described and demonstrated by using structurally ambiguous sentences involving different types of compounds in Thai. Compounds are a major cause of structural ambiguity in Thai and often create problems because of their high frequency of occurrence [3]. Compounding is the most widespread word formation process in Thai. Structural ambiguities often result from compounds because Thai words lack inflectional and derivational affixes to indicate, for example, subject-verb agreement. Nevertheless, compounds can be prosodically distinguished from syntactic phrases by differences in stress patterns. As for the prosody synthesis process, a mathematical model for generating F_0 contours proposed by Fujisaki for Japanese is chosen. An extension of the Fujisaki's model for tone languages, particularly Thai, will also be given [4]. In addition, the process of generating durational patterns for the synthesis module will also be described.

2. Prosodic annotation

Prosodic annotation or encoding provides to the prosody-synthesis module relevant information that adequately captures the essence of the prosodic structure of the input sentence or text. Prosodic encoding usually involves the process of predicting prosodic labels for the input sentence according to the intended meaning. The labeling criteria provide a mechanism for mapping abstract prosodic labels into a sequence of acoustic correlates of prosody. As a result, prosodically-labeled sentences contain information concerning the correspondence between the phonological and phonetic attributes of the prosodic structure of utterances and their intended meanings. Prosodic labels should be chosen to represent abstract linguistic categories of prosody, such as rhythmic groupings (or phrasing) and prominence. Also, they should be chosen such that they are used consistently within and across human labelers, and they make the automatic labeling process tractable and consistent. An example of a prosodic labeling system for English speech is described next.

Price et al. [5] proposed a labeling system consisting of seven labels, called prosodic break indices. These break indices express the degree of perceived decoupling or separation between every pair of words in an utterance. A boundary within a clitic group (e.g., det-noun, two-word verb, etc.) is indicated by a 0 break index; a normal word boundary by a 1; a boundary marking a minor grouping of words by a 2; an intermediate phrase boundary by a 3; an intonational phrase by a 4; a boundary marking a grouping of intonational phrases by a 5; and a sentence boundary by a 6. In terms of prominence, prominent syllables in an utterance are indicated by *P*1 for a major phrasal prominence; *P*0 for a lesser prominence; *C* for contrastive stress; and *s* for syllables with no prominence. Price demonstrated that these metrics could be used effectively by human labelers to determine how speakers encode prosodic cues for structural ambiguities in structurally ambiguous sentences.

In this paper, we adopted the Price's methods in the development of our prosodic encoding scheme for Thai. However, we made a slight modification to take advantage of our Constraint Dependency parsing framework based on Dependency grammar formalism. The encoding of the prosodic structure is accomplished by annotating each word in the sentence with a prosodic feature called strength. We describe next how the strength features are derived and compare them with Price's break indices.

The strength feature is chosen based on the dependency representation of syntax. In [6], Potisuk, et.al., described how a dependency theory is used for syntactic representation. A dependency grammar expresses the syntactic relations that lexical items can have with each other using governor-dependent relations in a D-tree. According to the congruency model of syntax and prosody [7], a relation of dominance between two adjacent lexical items can be established based on their positions in the D-tree. Figure1 illustrates the four basic configurations of relational marks between adjacent lexical items in a D-tree. ID or independence indicates no direct link between the two items; IT or interdependence indicates the dependence of the two lexical items on the same governor; LD or left dependence indicates the dependence on the following word; RD or right dependence indicates the dependence on the preceding word. It is noted that LD and RD are relational marks between two lexical items at different levels of the D-tree; ID and IT at the same level.

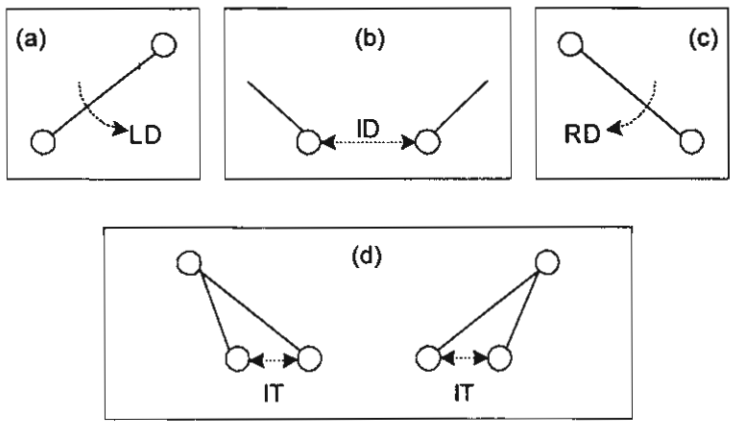

Figure 1. The four basic configurations of the relational marks in the dependency tree: (a) left dependence (LD), b) independence (ID), (c) right dependence (RD), and (d) interdependence (IT).

In addition, we have developed a new set of relational marks called strength dynamics in order to take into account the information about the lexical category of each word in addition to its position in the D-tree. Lexical category information is important because it is related to the stress placement rules in spoken language. Content words are usually stressed; function words are usually unstressed.

There are four levels of strength dynamics: strong dependence (SD), dependence (DE), independence (ID), and strong independence (SI). SD describes a strength dynamic at the word boundary within a clitic group, within a compound, between a content and a function word, or between two function words that are interdependent (i.e., both depend on the same governor). DE describes strength dynamic at minor phrase boundaries, i.e., between a subject noun phrase and a verb phrase, between a verb and an object noun phrase, or between two content words. ID describes strength dynamic at major phrase boundaries (intonational phrases). And, SI describes strength dynamic at the sentence boundary.

Like the break indices used in Price's labeling system, these strength dynamics indicate the degree of connection between the present and the preceding words in an input sentence. They are similar in a sense that both represent the relationship between two adjacent words in a sentence. The stronger the dependency strength is, the smaller the break index. Nonetheless, the strength dynamic has an added benefit in terms of the lexical category information. In addition to the strength feature, a word at the end of a phrase or an utterance will receive the feature *'final'* to indicate that it is affected by the final lengthening effect. Final lengthening is always accompanied by a pause. A word with a *'final'* feature also automatically receives strength of ID or SI.

In addition, we utilize a prosodic encoding scheme that integrates both syntactic and rhythmic constraints. That is, the prosodic structure of an utterance is established by minimizing speech disrhythmy while maintaining the congruency with syntax.

Speech is rhythmical not only because of the pattern of sounds and pauses, but also because of the regular recurrence of strongly accented sounds in a series. For example, in a stressed-time language, it is observed that speakers tend to produce stressed syllables at a regularly spaced interval of time while they tend to pause according to the syntax of the utterance [8]. The pause distribution seems to be ruled by syntactic constraints. Speech rhythm is also a psychological correlate of speech timing (an objective instrumental measurement of the duration of segments, syllables, etc.). Thus, in a stressed-time language, stress, pause, and relative syllable durations interact to form speech rhythm. In addition, the phonology and syntax of the language affect the description of speech rhythm as well.

Thai has a stress-timed rhythm [9]. This means that stressed syllables in Thai are perceived to be isochronous (i.e., they recur approximately at equal intervals of time). A phonological unit called foot is used to describe rhythmic groupings within an utterance. A foot is one of many prosodic constituents and is an elementary unit of the prosodic structure in addition to a syllable. A foot is neither a grammatical nor a lexical unit. The domain of a foot extends from a salient (stressed) syllable up to but not including the next salient syllable. A pause is considered a salient syllable, and the beginning of an utterance is always preceded by a pause. It should be noted that a rhythmic pause has a syntactic function, but a disfluency or hesitation pause does not.

In her analysis of Thai rhythm, Luangthongkum [9] posited five-foot structures:

1) | S | = 1-syllable foot,
2) | S W | = 2-syllable foot,
3) | S W W | = 3-syllable foot,
4) | S W W W | = 4-syllable foot,
5) | S W W W W | = 5-syllable foot,

where S and W indicate salient (stressed) and weak (unstressed) syllables, respectively. The 4-syllable and 5-syllable feet are very rare and are omitted from further discussion. Note that foot boundaries are usually inserted in front of the salient syllables.

Based on the discussion above, the strength dynamics assigned earlier can be used to obtain the information about the foot structure using the following rules. Since we only distinguish between two classes of stress, the salient syllable immediately after a weak syllable receives a strength of SD; otherwise, it receives a strength of DE. The weak syllable receives a strength of SD. A word before a pause receives a strength of DE as well as the final feature. A word after a pause receives a strength of SI if it is in the utterance-initial position; otherwise, it receives a strength of ID.

3. Prosody Synthesis

In this section, we describe a method of transforming symbolic prosodic markers or labels into the acoustic correlates of prosody. In other words, those markers are used to generate pitch, energy, and duration patterns.

3.1 Prediction of Durational Patterns

First, we describe the criteria for obtaining duration and pause information from the above strength features (through the derived foot structure). These criteria establish the correspondence between the phonological (strength dynamics) and the phonetic (acoustic correlates) attributes of prosody.

At an abstract level, Luangthongkum [9] assumed that each rhythmic foot is arbitrarily three units long, regardless of the number of syllables comprising the foot. This suggests that as the number of unstressed syllables in the interval increases, a tendency toward equality of inter-stress intervals causes both the stressed and unstressed syllables to become shorter. Thus, the relative syllable durations for each type of rhythmic foot can be abstractly described as follows:

1) | S | → | 3 |,

2) | S W | → | 2 : 1 |,

3) | S W W | → | 1½ : ¾ ¾ |.

Phonetically, a rhythmic foot is not isochronous. The duration of a foot will differ somewhat depending upon the phonetic structure of the syllables comprising it. Thus, the acoustic realization of a rhythmic foot will be different from the above abstract description. The following is a set of rules proposed by Luangthongkum to predict how syllable durations in each type of foot are realized acoustically. The derived or predicted syllable durations were based on her acoustic analysis of read speech.

| 3 | → | 2 | if the foot is in an utterance-initial position,

→ | 4 | if the foot is in an utterance-final position and it does not have a CVS structure.

| 2 : 1 | → | 2 : 2 | if the salient syllable has a CVS structure; or the weak syllable is the first element of a compound that does not have a CVS structure; or both the salient syllable and the weak syllable are function words.

| 1½ : ¾ ¾ | → | 1⅔ : 1⅔ 1⅔ | if the salient syllable has a CVS structure; or it is in an utterance-initial position; or it is a function word and the two weak syllables are two function words or a function word and a linker syllable.

Figure 2 depicts the process of predicting duration patterns from strength dynamics for two sentence hypotheses of an ambiguous sentence, แกงเผ็ดมากไป.

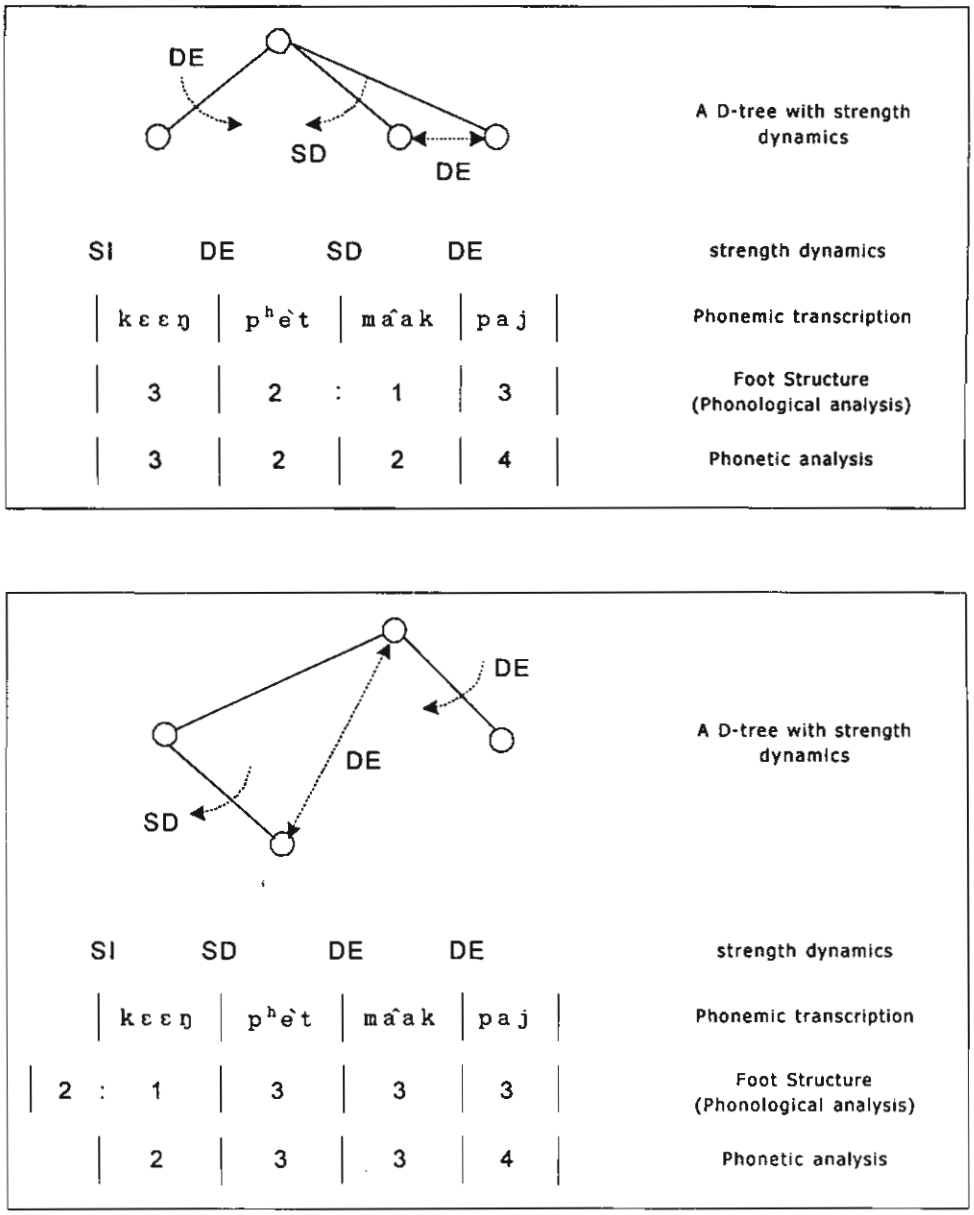

Figure 2. A prediction of duration patterns for two sentence hypotheses of an ambiguous sentence, แกงเผ็ดมากไป. The top panel indicates the first interpretation, 'the curry is too spicy'. The bottom panel indicates the second interpretation, 'There is too much curry.'

3.2 F_0 and Energy Contours Generation

Fujisaki [10] observed that an F_0 contour generally contains a smooth rise-fall pattern in the vicinity of the accented Japanese mora (a unit of syllable quantity). Differences in rise-fall patterns seem to be attributable to the accent type, and these rise-fall patterns appear to be superimposed on a base line that initially rises and gradually falls toward the end of the phrase or utterance regardless of the accent type. He hypothesized that the observed F_0 contour can be considered as the response of the phonatory system to a set of suprasegmental commands: the phrase (utterance) and the accent command. The phrase command produces the base line component while the accent command produces the accent component of an F_0 contour. Hence, Fujisaki proposed a functional model for generating an F_0 contour based on the idea of approximating F_0 contours as the response of a critically damped second-order linear system to excitation commands. The model is considered a superposition model because it additively superimposes a basic F_0 value (F_{min}), a phase component, and an accent component together on a logarithmic scale. The logarithmic frequency scale is based on the biomechanical considerations of the speech apparatus. In short, the output F_0 contour is a linear combination of F_{min}, a phrase component, and an accent component.

In Japanese, the F_0 realization of local pitch accents results only in rise-fall patterns in the F_0 contour. However, in the case of Thai, local F_0 variations due to tones result in a combination of both rise-fall patterns (e.g., F) and fall-rise patterns (e.g., R). As a result, a model for tone languages will consist of two components, the phrase and tone control mechanisms, driven by the phrase and tone commands. The phrase command and phrase control mechanisms are used to capture the declination effect; the tone command and tone control mechanisms are used to capture tone types. Instead of a base line, the phrase command will produce a midline. The tone commands in both positive and negative directions with respect to the midline will produce local contours corresponding to tone types, which are superimposed on the midline. As before, the model is characterized by time constant parameters, and command amplitudes and their temporal locations. Critical damping is assumed for both the phrase and tone control mechanisms; hence, the damping coefficient is always unity. Our extension of Fujisaki's model to tone languages is shown in Fig. 3.

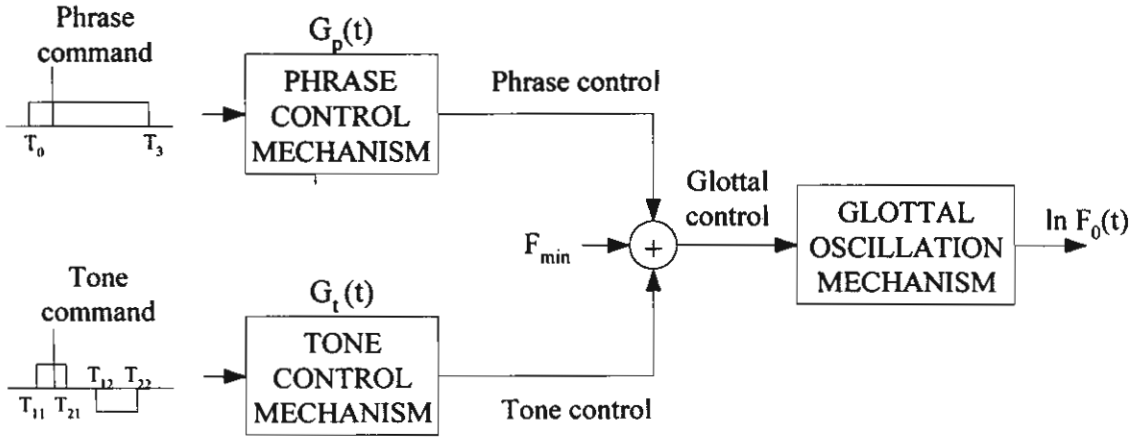

Figure 3. Our extension of Fujisaki's model for F_0 contour generation to tone languages.

Analogous to that of Fujisaki's original model, the tone synthesis model has the following mathematical expression:

$$\ln F_0(t) = \ln F_{min} + \sum_{i=1}^{I} A_{pi}[G_{pi}(t-T_{0i}) - G_{pi}(t-T_{3i})] + \sum_{j=1}^{J}\sum_{k=1}^{K(j)} A_{t,jk}[G_{t,jk}(t-T_{1jk}) - G_{t,jk}(t-T_{2jk})]$$

where $G_{pi}(t) = \alpha_i t \exp(-\alpha_i t)\, u(t)$ and $G_{t,jk}(t) = [1 - (1+\beta_{jk}t)\exp(-\beta_{jk} t)]u(t)$, $u(t)$ = unit step function, indicate the step response function of the corresponding control mechanism to the phrase and tone command, respectively. F_{min} is the smallest F_0 value in the F_0 contour of interest. A_p's and A_t's are the amplitude of the phrase and tone command, respectively. T_{0i} and T_{3i} denote the onset and offset of the ith phrase command. T_{1jk} and T_{2jk} denote the onset and offset of the kth component of the jth tone command. α_i's and β_{jk}'s are time constant parameters characterizing a second-order system. I, J, and $K(j)$ are the number of phrases, tones, and components of the jth tone, respectively, contained in the utterance. It is noted that the logarithmic scale will be replaced by an equivalent-rectangular-bandwidth-rate (ERB) scale, which is comparable to the logarithmic scale [11] and offers an advantage in that it gives equal prominence to excursions in different pitch registers. This is important in the synthesis of the F_0 contour of male and female speech.

While tones in isolation have rather definite F_0 manifestations in the tone space, they undergo various modifications in connected speech due to stress, interactions from adjacent tones (tonal coarticulation), and declination. These linguistic factors affecting the F_0 realization of tones in connected speech can be easily incorporated into the model in terms of the tone command amplitudes and their temporal locations. In this paper, the values of model parameters that reflect changes in the F_0 contour due to stress, tonal coarticulation and declination are manually obtained. In order for the model to be trainable given a corpus of speech, an automatic procedure is being investigated in our laboratory. Figure 4 shows the synthesized F_0 contours of two utterances with the same tone sequence but different segmental makeup.

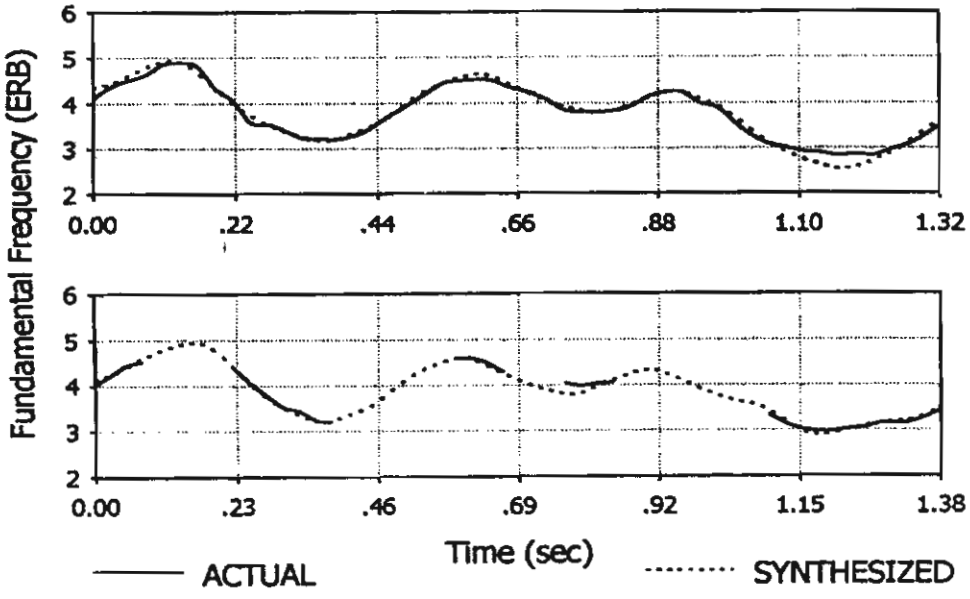

Figure 4. The actual and synthesized F_0 contours based on the extension of the Fujisaki's model for a discontinuously voiced with intervening obstruents utterance (bottom panel) and a continuously voiced utterance (top panel) carrying an HLFHR tone sequence.

4. Summary and Conclusion

In this paper, we have described our preliminary work on the issue of prosody generation in order to improve intelligibility and naturalness of synthetic speech produced by a text-to-speech system. Such improvement will undoubtedly make this type of speech technology more attractive and acceptable to human listeners. Since the algorithm is still in an early stage of development, we are planning to test and evaluate our system using a FESTIVAL text-to-speech system developed at Oregon graduate Institute as a part of the CSLU toolkit.

5. References

[1] E. O. Selkirk, *Phonology and Syntax: The Relation Between Sound and Structure.* MIT press, 1984.

[2] M. Nespor and I. Vogel, *Prosodic Phonology*, J. Koster and H. V. Riemsdijk, Eds. Dordrecht-Holland, 1986.

[3] P. E. Vongvipanond, "Linguistic problems in computer processing of the Thai language," in *1993 Proceedings of the Symposium on Natural Language Processing in Thailand.* Chulalongkorn University, Bangkok, Thailand, 1993, pp. 519-545.

[4] S. Potisuk, M. P. Harper, and J. T. Gandour, "Classification of Thai tone sequences in syllable-segmented speech using the analysis-by-synthesis method," *IEEE Transaction on Speech and Audio Processing*, Vol.7, No.1, January 1999, pp.95-102 .

[5] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong, "The use of prosody in syntactic disambiguation," *Journal of Acoustical Society of America*, vol. 90-6, pp. 2956-2970, 1991.

[6] S. Potisuk and M. P. Harper, "CDG: An alternative formalism for parsing written and spoken Thai." *Proceedings of the Fourth International Symposium on Language and Linguistics: Pan-Asiatic Linguistics*, Vol. 4, pp. 1177-1196, 1996.

[7] G. Bailly, "Integration of rhythmic and syntactic constraints in a model of generation of French prosody," Speech *Communication*, vol. 8, pp. 137-146, 1989.

[8] J. P. Gee and F. Grosjean, "Performance structures: A psycholinguistic and linguistic appraisal," *Cognitive Psychology*, vol. 15, pp. 411-458, 1983.

[9] T. Luangthongkum, *Rhythm in standard Thai,* Ph.D. dissertation, University of Edinburgh, 1977.

[10] H. Fujisaki, "Dynamic characteristics of voice fundamental frequency in speech and singing," in *The production of Speech*, P. F. MacNeilage, Ed. New York: Springer-Verlag, 1983, pp. 39-55.

[11] D. Hermes and J. Van Gestel, "The frequency scale of speech intonation," *Journal of the Acoustical Society of America*, vol. 90, pp. 97-102, 1991.