Analysis and Synthesis of F_0 Contours of Thai Utterances Based on the Command-Response Model

Hiroya Fujisaki[*], **Sumio Ohno**[†] and **Sudaporn Luksaneeyanawin**[‡]

[*] University of Tokyo, Japan
fujisaki@alum.mit.edu

[†] Tokyo University of Technology, Japan
ohno@cc.teu.ac.jp

[‡] Chulalongkorn University, Thailand
sudaporn.1@chula.ac.th

ABSTRACT

Characteristics of F_0 contours representing the five tones of Thai vary widely due to various factors, but the variations can be quantitatively explained and predicted if we have a precise formulation of the underlying process. This paper presents such a formulation in terms of the command-response model, and shows that the model can generate very close approximations to observed F_0 contours by positing systematic patterns of positive and negative tone commands for individual tone types, except for the mid tone that has no tone command. The timing of these commands is shown to vary systematically with the speech rate. These findings are then used as constraints/rules for tone generation in speech synthesis, and the validity of such constraints is confirmed by perceptual evaluation of synthetic speech by native speakers.

1 INTRODUCTION

Thai is a tone language in which a syllable possesses five tone types, representing real but different morphemes in many cases. Therefore the tone is one of the most important factors in the study of Thai speech. Traditionally, the five tones are labeled: mid (T0), low (T1), falling (T2), high (T3), and rising (T4)[1]. Table 1 shows a minimal list of the five tones, and Figure 1 shows the F_0 contours corresponding to the five tones in isolated utterances.

Table 1: Examples of Thai words that are segmentally identical but differ in tone.

word	tone	meaning
/naa0/	mid (T0)	"rice field"
/naa1/	low (T1)	a nick name
/naa2/	falling (T2)	"face"
/naa3/	high (T3)	"uncle/aunt"
/naa4/	rising (T4)	"thick"

Although these labels indicate F_0 contour shapes of isolated utterances qualitatively, they do not necessarily apply to F_0 contour shapes in connected utterances because of tonal coarticulation/assimilation and phrasing. In order to describe the dynamic characteristics of the F_0 contours quantitatively, one needs a mathematical model based on the physiological/physical properties of the control mechanisms of vocal fold vibration. Such a model (henceforth 'the command-response model') has been presented by Fujisaki and his co-workers and has been successfully applied to F_0 contours of speech of various languages including Standard Chinese [2-4], whose tonal structure is rather similar but somewhat simpler than that of Thai. In fact, several attempts have been made to modify it to apply to Thai [5-7]. The present study is conducted with an aim to test the applicability of the model to F_0 contours of Thai utterances, and to demonstrate its use in analysis and synthesis of tonal features of Thai.

2 A MODEL FOR THE F_0 CONTOUR GENERATION OF THAI UTTERANCES

Careful observation of F_0 contours of isolated tones of Thai suggests that the mechanism of laryngeal control for Thai tones is essentially the same, at least qualitatively, as that for the tones of Standard Chinese, and requires tone commands of both positive and negative polarities, with the exception of the mid tone, which is characterized by the lack of tone commands.

Figure 2 shows the model for F_0 contour generation of Thai utterances. The phrase commands (impulses) generate the overall contour of an utterance, and the tone commands (square pulses) generate the local contours corresponding to the five tones. While T1 and T3 are generated by a single tone command (negative in T1 and positive in T3), T2 and T4 are generated by a pair of tone commands (positive-negative in T2 and negative-positive in T4). On the other hand, T0 has no tone commands and hence no tone components. These commands are applied to the respective control mechanisms which produce phrase and tone components.

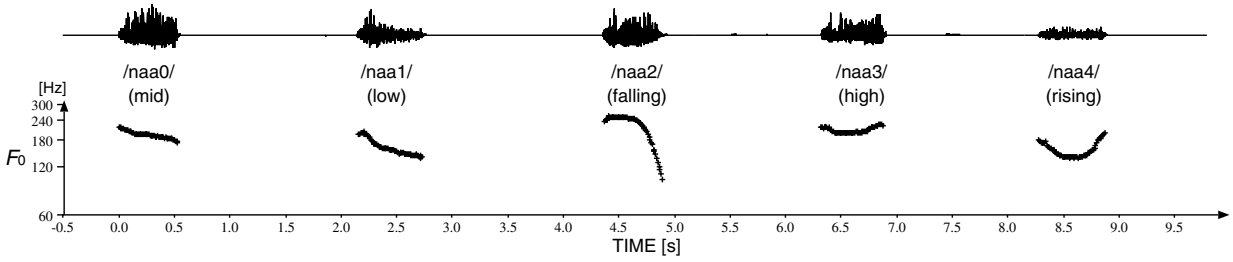

Figure 1: F_0 contours of the five Thai tones produced in isolation by a female speaker.

ISBN 1-876346-48-5 © 2003 UAB

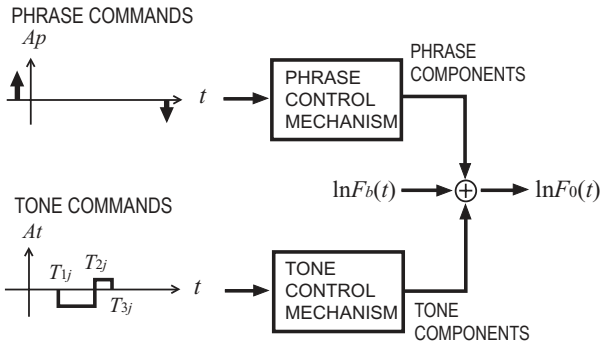

Figure 2: A command-response model for F_0 contour generation of Thai utterances.

These mechanisms are assumed to be critically-dumped second-order linear systems. The phrase components and the tone components are added onto a constant value $\ln F_b$ to produce the final $\ln F_0(t)$. For the rest of the paper, we shall use the word 'F_0 contour' to indicate $\ln F_0(t)$. Physiological and physical evidences supporting the model were presented elsewhere [8].

Thus the F_0 contour as a function of time can be expressed by the following equations:

$$\begin{aligned} \ln F_0(t) &= \ln F_b + \sum_{i=1}^{I} A_{pi} G_p(t - T_{0i}) \\ &\quad + \sum_{j=1}^{J} [A_{t1j}\{G_t(t - T_{1j}) - G_t(t - T_{2j})\} \\ &\quad + A_{t2j}\{G_t(t - T_{2j}) - G_t(t - T_{3j})\}], \end{aligned} \tag{1}$$

$$G_p(t) = \begin{cases} \alpha^2 t \exp(-\alpha t), & \text{for } t \geq 0, \\ 0, & \text{for } t < 0, \end{cases} \tag{2}$$

$$G_t(t) = \begin{cases} \min[1 - (1 + \beta t)\exp(-\beta t), \gamma], & \text{for } t \geq 0, \\ 0, & \text{for } t < 0, \end{cases} \tag{3}$$

where $G_p(t)$ represents the impulse response function of the phrase control mechanism and $G_t(t)$ represents the step response function of the tone control mechanism.

The symbols in Eqs. (1) to (3) indicate

- F_b : baseline value of fundamental frequency,
- I : number of phrase commands,
- J : number of tone command pairs,
- A_{pi} : magnitude of the ith phrase command,
- A_{t1j} : amplitude of the first tone command in the jth command pair,
- A_{t2j} : amplitude of the second tone command in the jth command pair,
- T_{0i} : timing of the ith phrase command,
- T_{1j} : onset of the first tone command in the jth command pair,
- T_{2j} : end of the first tone command and onset of the second tone command in the jth command pair,
- T_{3j} : end of the second tone command in the jth command pair,
- α : natural angular frequency of the phrase control mechanism,
- β : natural angular frequency of the tone control mechanism,
- γ : relative ceiling level of tone components.

The onset of the second tone command is constrained to coincide with the end of the first tone command within a tone command pair. Although Eq. (1) provides a pair of tone commands for every syllable, only one tone command is necessary for tones T1 and T3. Variations in α, β, and γ are found to be quite small across utterances and speakers.

3 ANALYSIS OF F_0 CONTOURS OF THAI UTTERANCES

3.1 SPEECH MATERIAL

The speech material for the present study was recorded at the Centre for Research in Speech and Language at Chulalongkorn University, Bangkok. It consists of two different sets of utterances.

(A) Isolated utterances of the five words of Table 1.

(B) Utterances in which one of the five words is embedded in the carrier sentence /kham0 nii3 phuut2 waa2 __ dooj0 khon0 suuan1 jaj1/ (This word is pronounced as __ by most speakers.).

Each sentence was uttered twice at each of the three speech rates: normal (approx. 3.6 syllables/s), fast (approx. 4.3 syllables/s), and slow (approx. 2.8 syllables/s). The speakers are one female and one male native speakers of Thai.

3.2 ANALYSIS PROCEDURE

The speech signal was digitized at 10 kHz with 16 bit precision. The fundamental frequency was extracted at 10 ms intervals by the modified autocorrelation analysis of the LPC residual. The measured F_0 contour was aligned with the speech waveform whose syllable boundaries as well as onsets of the rhyme were marked by visual inspection of the waveform whenever possible.

Analysis of the F_0 contour was conducted interactively by assigning first approximations to the phrase and tone commands on the basis of visual inspection and by reducing the difference between the observed F_0 contour and the model-generated F_0 contour by successive approximation until the mean squared difference in $\ln F_0(t)$ between the observed and the model-generated contours is minimized. This allows one to decompose a given F_0 contour into its constituent components, and to estimate their underlying commands.

3.3 EXPERIMENTAL RESULTS

Figure 3 (a) - (e) shows the results of analysis of one sample each of the five utterances of speech material (B), produced by the female speaker at the normal speech rate. Each panel shows, from top to bottom, the speech waveform, measured F_0 values (+ symbols), the model-generated best approximation (solid line), the baseline frequency (dotted line), the phrase commands (impulses), and the tone commands (square pulses). The dashed lines indicate the contributions of phrase components, and the differences between the F_0 contour and the phrase components correspond to the tone components. The results in these panels indicate that the model can generate very good approximations to F_0 contours of all the tone types and supports the initial assumptions made on the tone command patterns of Thai. Equally good approximations were obtained from the analysis of utterances at slow and fast speech rates. Table 2 lists the timing and amplitude of tone commands obtained from the speech material (B) of the female speaker.

ISBN 1-876346-48-5 © 2003 UAB

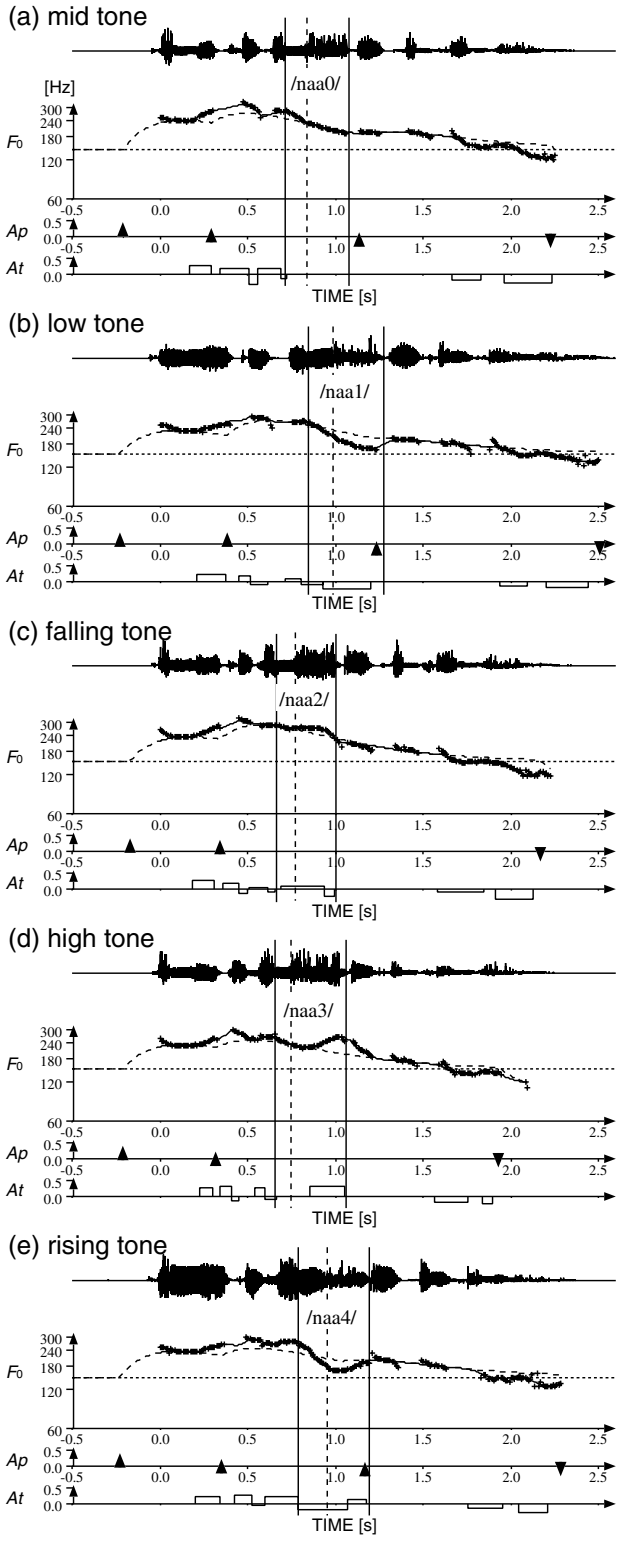

Figure 3: Results of analysis of one sample each of the five utterances in which the target word is embedded in the carrier sentence. The solid vertical lines indicate the syllable boundaries and the dotted vertical line indicates the onset of the rhyme of the target word.

Table 2: Parameters of tone commands in Tones 1 to 4 obtained from the utterances of the female speaker produced at the normal speech rate. The timing is in ms with respect to rhyme onset.

	T_{1j}	T_{2j}	T_{3j}	A_{t1j}	A_{t2j}
T1	−43	221		−0.20	
T2	−60	178	240	0.19	−0.13
T3	87	260		0.27	
T4	−79	149	275	−0.20	0.16

Figure 4 shows the relationships between the timing of the tone commands and the speech rate. The ordinate indicates the timing of the onset and the offset of tone commands, while the abscissa indicates the length of the rhyme of the target word. Panels (a), (b), (c), and (d) respectively correspond to T1, T2, T3, and T4. Except for T3, the onset of the first tone command is found to occur approximately 40 - 80 ms before the onset of the rhyme regardless of the speech rate, while the offset of tone command is found to vary almost linearly with rhyme duration in all the tones.

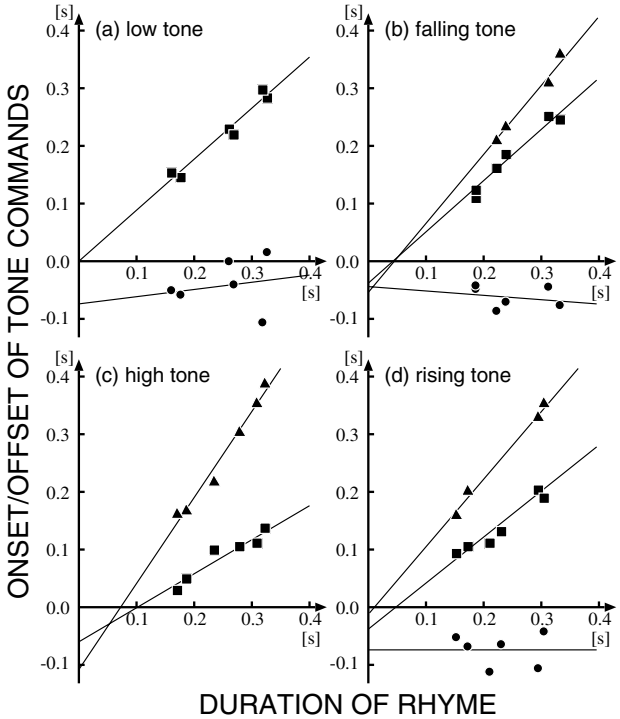

Figure 4: Relationships between the onset/offset of tone commands and the duration of the rhyme.

4 SYNTHESIS OF TONAL FEATURES OF THAI AND ITS PERCEPTUAL EVALUATION

Since the model can generate very close approximations to F_0 contours of natural utterances, it is clear that the model will be useful in speech synthesis and in speech coding. From the point of view of speech synthesis by rule, it is of interest to find out to what extent these parameters can be standardized without degrading the naturalness of prosody of synthetic speech. For this purpose, the following preliminary experiment was conducted.

ISBN 1-876346-48-5 © 2003 UAB

4.1 STIMULI

The following four sets of stimuli were synthesized.

(1) Analysis-resynthesis of one utterance each of the five sentences (i.e., with the original F_0 contours produced at the normal speech rate).

(2) Take one of the utterances in which the target word has T0 (to be referred to as U0) and modify its tone type into T1 to T4 by introducing the tone command(s) following the timing information shown in Table 2.

(3) Modify the timing of all the tone commands in U0 following the timing information indicated by the lines in Fig. 4.

(4) Modify all the tone commands in U0 following the timing information given in Fig. 4 and the amplitude information given in Table 2.

Thus a total of 19 different stimuli were synthesized. These stimuli were arranged in a random order and presented to subjects through a loudspeaker in a quiet room. Each subject made 10 judgments on the identity of the tone of the target word as well as on the naturalness of tone and intonation of each stimulus as a whole utterance (i.e., not just of the target word) on a 5-point opinion scale (5: excellent, 4: good, 3: acceptable, 2: somewhat unnatural, 1: unacceptable).

4.2 SUBJECTS

The subjects were three native speakers of Thai. Two were born in Bangkok, while one was born in Songkhla. Although the tones of the target words were correctly identified by the three subjects, significant differences were found between the Bangkok subjects and the Songkhla subject in the judgment of naturalness. The following analysis was made on the results of the Bangkok subjects.

4.3 RESULTS

Table 3 shows the mean (μ) and the standard deviation (σ) of the opinion scores of the two Bangkok subjects for the four sets of stimuli. It is interesting to note that stimulus Set 4 received the highest score, though analysis of variance indicates that the differences in the four mean opinion scores are all insignificant.

Table 4 shows the average opinion score of the two subjects for each tone in each stimulus set. Although some of the differences of scores are significant, the opinion scores are at least close to or above 4 except for T1 of Set 1. In fact, all the tones in Set 4 received scores above 4.0, which may indicate that introduction of timing and amplitude constraints results in higher quality of synthetic speech when they are introduced simultaneously to all syllables within an utterance, rather than only to the target word.

Table 3: Mean and standard deviation of the opinion score of the two Bangkok subjects for each stimulus set.

Stimulus set	(1)	(2)	(3)	(4)
μ	4.41	4.43	4.49	4.55
σ	0.71	0.64	0.64	0.58

Table 4: Average opinion scores for each tone and stimulus set.

Stimulus set	(1)	(2)	(3)	(4)
T0	4.75	4.70	4.85	4.75
T1	3.50	4.60	4.65	4.60
T2	4.95	3.85	3.95	4.70
T3	4.60	4.70	4.95	4.60
T4	4.25	4.30	4.05	4.10

5 SUMMARY AND CONCLUSION

This paper has shown the application of the command-response model to the analysis and synthesis of F_0 contours of Thai. It was shown by our preliminary analysis that the structure of the model developed for F_0 contours of Standard Chinese remains essentially the same for Thai. The commands for Thai tones were found to be partially similar to those for the four tones of Standard Chinese in that their polarities can be both positive and negative, but the mid tone was found to have no tone command. The configurations of the tone commands were found to remain unchanged in connected utterances and to be stable across various speech rates. The results of analysis of connected utterances were used as constraints in speech synthesis. Perceptual evaluation of speech synthesized with these constraints showed little degradation of naturalness of prosody, indicating the usefulness of our findings to speech synthesis by rule.

ACKNOWLEDGEMENTS

This work was supported by the Grant-in-Aid for Scientific Research of Priority Areas (B) No. 12132102 “Analysis, Formulation, and Modeling of Prosody” (Principal Investigator: Hiroya Fujisaki).

REFERENCES

[1] S. Luksaneeyanawin, “Intonation in Thai,” in *Intonation Systems. A Survey of Twenty Languages*, D. Hirst & A. Di Cristo, Eds., pp. 376–394. Cambridge, U. K.: Cambridge University Press, 1998.

[2] H. Fujisaki, P. Hallé and H. Lei, “Application of F_0 contour command-response model to Chinese tones,” *Reports of Autumn Meeting, Acoust. Soc. Jpn.*, vol. 1, pp. 197–198, 1987.

[3] H. Fujisaki, K. Hirose, P. Hallé and H. Lei, “Analysis and modeling of tonal features in polysyllabic words and sentences of the Standard Chinese,” *Proceedings of 1990 Int’l Conf. on Spoken Language Processing*, vol. 1, pp. 841–844, 1990.

[4] C. Wang, H. Fujisaki, R. Tomana and S. Ohno, “Analysis of fundamental frequency contours of Standard Chinese in terms of the command-response model and its application to synthesis by rule of intonation,” *Proceedings of the 6th Int’l Conf. on Spoken Language Processing*, vol. 3, pp. 326–329, 2000.

[5] H. Potisuk, M. Harper and J. Gandour, “Classification of Thai tone sequences in syllable-segmented speech using the Analysis-by-Synthesis method,” *IEEE Transactions on Speech & Audio Processing*, vol. 7, no. 1, pp. 95–102, 1999.

[6] P. Seresangtakul and T. Takara, “Analysis of pitch contour of Thai tone using Fujisaki’s model,” *Proceedings of 2000 IEEE Int’l Conf. on Acoust., Speech, & Signal Processing*, vol. 1, pp. 505–508, 2002.

[7] H. Mixdorff, S. Luksaneeyanawin, H. Fujisaki and P. Charnvivit, “Perception of tone and vowel quantity in Thai,” *Proceedings of the 7th Int’l Conf. on Spoken Language Processing*, vol. 2, pp. 753–756, 2002.

[8] H. Fujisaki, “Modeling in the study of tonal features of speech — with application to multilingual speech systhesis —,” *Proceedings of the Joint Int’l Conf. of SNLP-Oriental COCOSDA 2002*, pp. D1–D10, 2002.