Study on Pitch Contour of Thai Polysyllabic Tone Sequences Using a Generative Model

Pusadee Seresangtakul and Tomio Takara***

*Department of Computer Science, Khon Kaen University
Khon Kaen, Thailand
**Department of Information Engineering, University of the Ryukyus
Okinawa, Japan
pusadee@kku.ac.th, takara@ie.u-ryukyu.ac.jp

Abstract

Thai speech synthesis by rule has been developed. In order to synthesize F_0 contours of Thai tones, the generative model of F_0 contours (Fujisaki's model) for tonal languages is applied. Along with our method, the pitch contours of Thai polysyllabic words were analyzed. Rules are derived and applied to synthesize Thai polysyllabic tone sequences. We performed listening tests to evaluate intelligibility of the model for Thai tone generation. The average intelligibility scores were 98.8%and 96.6% for disyllabic and trisyllabic words, respectively. The generative model of F_0 contours for Thai words was shown to be effective. Furthermore, we derived rules to synthesize suprasegmental F_0 contours using the trisyllabic words' parameters. We performed listening tests to evaluate the intelligibility score and naturalness of synthesized speech. As a result, all phrases/sentences were completely identified. The MOSs (Mean Opinion Score) was 3.50 while the original and analysis/synthesis samples were 4.82 and 3.59, respectively.

1. Introduction

For a tonal language like Thai, tone is an important part of speech. The tone is indicated by contrasting variations in contour of fundamental frequency (F_0) at the syllabic level. Words with the same phoneme sequences may have different meanings if they have different tones. Therefore, tone is one of the most important factors in the speech research field in order to make a system which has high intelligibility and naturalness. Thai has 5 lexical tones traditionally named: mid (M), low (L), falling (F), high (H) and rising (R). Abramson studied and divided the tones into 2 groups: static tones (the high, the mid, the low); and dynamic tones (the falling and the rising) [1].

The tone is correlated to fundamental frequency (F_0). Seresangtakul et al. studied and applied Fujisaki's model to Thai language [2,3]. In the study, Thai monosyllabic words were analyzed and the parameters to synthesize Thai monosyllabic words were obtained. It is clearly effective in the case of isolated syllables. But in the case of polysyllabic words, we can not simply connect 2 tones together because there is a phenomenon that the speech sound is altered in its phonetic manifestation depending on influences from adjacent sounds. Therefore, we attempt to obtain deeper insight into the nature of their interactions and to describe them in terms of the generative model's parameters for tonal language. Based on the model, we analyzed polysyllabic words, and we defined rules to synthesize the F_0 contours of such words. In order to evaluate the intelligibility of the model for Thai polysyllabic words, listening tests were performed. Moreover, we derived rules to synthesize the F_0 contour of phrases and short sentences. Finally, listening tests were performed to evaluate the naturalness of the model for Thai tones in phrases and sentences.

2. A generative model of F_0 contours for tonal languages

The generative model of F_0 contours (Fujisaki's model) is a mathematical model for a quantitative analysis and linguistic interpretation of the F_0 contour characteristics [5]. The model was first proposed for accent of Japanese and successful in many languages [4-6].

The model has been extended to apply for the F_0 contour of Thai [3]. In the original model, the F_0 contour generally contains a smooth rise-fall pattern in the vicinity of the accent components. The F_0 contour is treated as a linear superposition of a global phrase and local accent components on a logarithmic scale. The phrase command produces the base line component while the accent command produces the accent component of an F_0 contour. For Thai, the model to generate pitch contour will consist of the phrase and tone control mechanisms. In Japanese, the F_0 realization of local pitch accents results only in a rise-fall pattern in the F_0 contour. In contrast for Thai, local F_0 variations due to tones result in a combination of both rise-fall and fall-rise patterns.

When the phrase commands are assumed to be impulses, they are applied to the phrase control mechanism to generate the phrase components. Further, the tone commands in both positive and negative polarities are applied to the tone control mechanisms to produce local contours corresponding to the tone components. The F_0 contour can be expressed by:

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

$$G_{\mathrm{p}i}(t) = \begin{cases} (\alpha_i^2 t)\exp(-\alpha_i t), & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

$$G_{\mathrm{t},jk}(t) = \begin{cases} [1 - (1 + \beta_{jk} t)\exp(-\beta_{jk} t)], & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

Where $G_{\mathrm{p}i}(t)$ represents the impulse response function of the phrase control mechanism and $G_{\mathrm{t},jk}(t)$ represents the step response function of the tone control mechanism, respectively. The symbols in these equations indicate the following: $F_{\min}$ is the smallest F_0 value in the F_0 contour of interest, $A_{\mathrm{p}i}$ and $A_{\mathrm{t},jk}$ are the amplitudes of the i^{th} phrases and of the j^{th} tone command. T_{0i} is timing of the i^{th} phrase command; T_{1jk} and T_{2jk} are the onset and offset of the k^{th} component of the j^{th} tone command. α_i and β_{jk} are time constant parameters. *I, J,* and *K*(j) are the number of phrases, tones, and components of the j^{th} tone contained in the utterance, respectively.

3. Analysis of F_0 contour of Thai polysyllabic words

3.1. Speech material

In order to create the speech corpus, we prepared 25 words of two-tone sequences and 125 words of three-tone sequences. The 25 sample words were combinations of syllables in the first set: {/niaŋ/, /nòoŋ/, /nûaŋ/, /nɔ́ɔŋ/, /nɔ̌ɔŋ/} together with syllables in the second set: {/yɔɔ/, /yὲɛ/, /wâa/, /lɔ́ɔ/, /lǐi/}. The average durations of disyllabic words in all samples were 0.42 and 0.46 [s] for the first and the second syllable, respectively.

The 125 sample words of 3 tone sequences were combinations of syllables in the first set: {/niaŋ/,/nòoŋ/,/nûaŋ/, /nɔ́ɔŋ/, /nɔ̌ɔŋ/}, together with syllable in the second set: {/yɔɔ/, /yὲɛ/, /wâa/, /lɔ́ɔ/, /lǐi/} and the third set: {/naaŋ/, /lɔ̀ɔn/, /lâam/, /náam/, /lăan/}. All sample words begin with nasal, lateral or semivowel consonants, and the vowels for all 3 syllables were long vowels. The average durations of trisyllabic words in all samples were 0.40, 0.39 and 0.48 [s] for the first, the second and the third syllable, respectively.

3.2. Method

Based on the idea that F_0 contour is significant in tone information, the typical F_0 contours were gotten by averaging F_0 contours of the same sequence of all speakers. Since duration and pitch range between male and female are different, to get the average F_0 patterns normalization processes were performed as follows: First, time normalization was done. The duration of each F_0 contour was obtained by time ratio in percentage of syllable duration across all corresponding syllables in all tone sequences. The normalization was considered syllable-by-syllable. Next, frequency normalization was done. To avoid the difference in pitch range among the speakers, the fundamental frequency in Hertz scale was transformed to logarithmic F_0 in Z-score [7], which is a function of mean and SD. Next, we averaged the F_0 contours of all speakers in Z-score. Finally, the average F_0 contour in Z scale was transformed to Hertz scale by referring to the mean and the standard deviation of the speaker, whose voices were analyzed and used in the Thai speech synthesis system.

In this work, we model the F_0 contours of Thai tones by using the generative model of F_0 contours. Therefore to get the model's parameters for polysyllabic words, a curve fitting method was used to fit the average pitch contour by minimizing the least square error between the average F_0 contour and that of the model on logarithmic scale.

Because the model is based on superposition of phrase and tone components, the phrase component was gotten separately from the tone component. In our work, we hypothesize that mid tone is neutral. We used non-linear curve fitting to get the phrase parameters of the mid tone sequences and set them as the phrase commands of the other patterns to get tone commands from them. The fitting was done under the condition that there is no overlap between 2 tone commands. The initial parameters of a tone can be externally set. The minimized mean square error is obtained through the steepest descent method.

3.3. Result and discussion

Both phrase and tone component parameters were gotten by using non-linear least square fitting. The values of the parameters vary little and the shape of F_0 contour at each position is very similar. Therefore, to reduce the number of rules to synthesize the F_0 contours of disyllabic and trisyllabic words, the parameters were grouped and averaged by tone at each position. That means, there are 10 groups (5 preceding tones and 5 following tones) for 2 tone sequences, and 15 groups (5 preceding, 5 middle, and 5 following tones) for trisyllabic tone sequences. Table 1 and Table 2 show the tone component parameters for disyllabic and trisyllabic words, respectively. The values in parentheses show the SD for each parameter. The beginning times of these commands are found, in case of the phrase commands of both disyllabic and trisyllabic words, approximately 200 [ms] before the

onset of an utterance. Alpha is approximately 2.9. Beta values are 9.5 and 11.5 for the high and the other tones (the mid, low, falling and rising tones), respectively. These are the same values as that of monosyllabic words [2].

Table 1: *Tone-command parameters for disyllabic words.*

Tone	Syl.(j)	A_{tj1}	$T_{1j1\,[S]}$	$T_{2j1\,[S]}$	A_{tj2}	$T_{1j2\,[S]}$	$T_{2j2\,[S]}$
Mid	1st	-0.002 (0.019)	0.0 (0.0)	0.410 (0.0)			
	2nd	0.031 (0.014)	0.420 (0.0)	0.870 (0.0)			
Low	1st	-0.170 (0.01)	-0.049 (0.005)	0.318 (0.024)			
	2nd	-0.072 (0.001)	0.420 (0.0)	0.870 (0.0)			
Falling	1st	0.221 (0.033)	-0.049 (0.005)	0.150 (0.004)	-0.073 (0.023)	0.333 (0.024)	0.390 (0.045)
	2nd	0.298 (0.055)	0.403 (0.016)	0.622 (0.015)	-0.163 (0.055)	0.721 (0.017)	0.870 (0.0)
High	1st	0.267 (0.052)	0.190 (0.006)	0.395 (0.016)			
	2nd	0.301 (0.033)	0.576 (0.021)	0.870 (0.0)			
Rising	1st	-0.288 (0.038)	-0.053 (0.008)	0.166 (0.014)	0.336 (0.038)	0.299 (0.026)	0.377 (0.029)
	2nd	-0.164 (0.068)	0.420 (0.0)	0.531 (0.026)	0.412 (0.056)	0.690 (0.009)	0.870 (0.0)

Table 2: *Tone-command parameters for trisyllabic words.*

Tone	Syl (j)	A_{tj1}	$T_{1j1\,[S]}$	$T_{2j1\,[S]}$	A_{tj2}	$T_{1jk\,[S]}$	$T_{2j2\,[S]}$
Mid	1st	-0.019 (0.053)	0.00 (0.0)	0.360 (0.009)			
	2nd	-0.011 (0.027)	0.400 (0.001)	0.780 (0.022)			
	3rd	-0.054 (0.027)	0.800 (0.0)	1.270 (0.0)			
Low	1st	-0.159 (0.029)	-0.055 (0.002)	0.330 (0.022)			
	2nd	-0.161 (0.029)	0.398 (0.002)	0.710 (0.015)			
	3rd	-0.149 (0.040)	0.800 (0.0)	1.270 (0.0)			
Falling	1st	0.241 (0.031)	-0.023 (0.013)	0.160 (0.013)	-0.107 (0.150)	0.307 (0.012)	0.365 (0.016)
	2nd	0.243 (0.032)	0.378 (0.016)	0.550 (0.017)	-0.085 (0.087)	0.709 (0.013)	0.781 (0.017)
	3rd	0.252 (0.44)	0.773 (0.017)	0.960 (0.010)	-0.312 (0.054)	1.090 (0.021)	1.270 (0.0)
High	1st	0.256 (0.026)	0.181 (0.005)	0.360 (0.004)			
	2nd	0.260 (0.047)	0.577 (0.011)	0.758 (0.011)			
	3rd	0.166 (0.051)	0.967 (0.018)	1.270 (0.0)			
Rising	1st	-0.231 (0.060)	-0.052 (0.007)	0.170 (0.014)	0.296 (0.036)	0.281 (0.017)	0.357 (0.053)
	2nd	-0.217 (0.060)	0.379 (0.015)	0.560 (0.006)	0.294 (0.043)	0.670 (0.011)	0.755 (0.009)
	3rd	-0.263 (0.047)	0.800 (0.012)	0.990 (0.014)	0.425 (0.101)	1.160 (0.020)	1.270 (0.0)

4. Listening Test

4.1. Speech material

In order to evaluate the intelligibility of the model of Thai polysyllabic tones, listening tests were performed. We prepared 5 datasets. The first 4 datasets were composed of all 25 combinations of 2 Thai tone sequences that were generated from the same spectrum of the phonemes /maa maa/, /nɔɔŋ yɛɛ/, /nàa lɔ̀ɔ/, /k^{h}aa law/, respectively. More than 50% of the words in the first and the second datasets are nonsense words and 80% of the words in the third, and the fourth datasets are actual words. The fifth dataset composed of 125 trisyllabic words of 3 Thai tone sequences, of which 65% are actual words. These words were synthesized from the same spectrum of the word /nɔɔŋlɔɔlaam/. The F_0 contours of the first 4 datasets and the fifth dataset were generated from the parameters in table 1 and table 2, respectively. The spectrum parameter of the mid-mid sequence is used for all 25 combinations. The time onset and offset of the tone commands were adjusted as the expression (3)

$$T = B + \frac{D}{D_m} T_m \qquad (3)$$

Where T is an actual time onset or time offset parameter of the generative model. D_m and D are syllable's durations of the model and the synthesized one, respectively. B is the beginning time point of the synthesized syllable. T_m is a relative time onset/offset expressed as following expression (4).

$$T_m = T_{i,jk} - B_m \qquad (4)$$

Where $T_{i,jk}$ are model's time onset or offset shown in table1 and table2. B_m is the beginning time point of the model's syllable, which are equal to 0 [s] and 0.42 [s] for the first and the second syllable of disyllabic word, and equal to 0[s], 0.40[s] and 0.80[s] for the first, the second, and the third syllable of trisyllabic word, respectively.

4.2. Listening test and result

In the experiment, six native Thai speakers participated in the listening test. All listeners have normal hearing ability and most of them are not familiar with synthetic sounds. Five datasets were presented to the listeners and the tests were done for each dataset. The experiment was done in a soundproof room. The listener listened with headphones. Each word was played 2 times with 2-second intervals, which is sufficient for listeners to choose a Thai script word that corresponds to the sound that they heard. A summary of the average intelligibility scores, calculated after removing the score of the listeners with maximum and minimum error values, is shown in Table 3.

The results show that for the proposed method, the average intelligibility rates are 98.8% and 96.6% for disyllabic words (A), (B), (C), (D) and trisyllabic words (E), respectively. We found that all errors came from nonsense words. For the disyllabic words the average intelligibility scores are 98.3% and 99.3% for datasets composed of mostly nonsense disyllabic words (A)(B), and mostly actual disyllabic words (C)(D), respectively. We can see that the higher the ratio of actual words, the

higher the intelligibility scores. As a result, the generative model of F_0 contours for Thai polysyllabic words was shown to be effective.

Table 3: *Listening test result after removing the maximum and minimal error scores.*

	Word	No of Sample	Intelligibility score
A	/maa maa/	25x2	98.0%
B	/nɔɔŋ yɛɛ/	25x2	98.5%
C	/naa lɔɔ/	25x2	99.0%
D	/k^{h}aa law/	25x2	99.5%
E	/nɔɔŋlɔɔlaam/	125	96.6%

5. Suprasegmental F_0 contour generation

In order to synthesized the F_0 contour of phrases and short sentences, the trisyllabic words' parameters were applied by the following rules: The starting, the ending and the other remaining syllables in a sentence were assigned from the first, the third, and the second parameters of the trisyllabic word's parameters show in table 2. The original spectrum parameters were used for each synthesize sequence.

In order to evaluate the intelligibility and naturalness of the pitch contour of the suprasegmental level, listening tests were performed. In the listening tests, we synthesized 6 phrases/sentences shown in Table 4. Six native Thai speakers participated in the listening test. In the first test, all listeners listened to the phrases/sentences and wrote what they heard in Thai script. The result showed that all answered correctly. Therefore, the second test was performed to evaluate the naturalness of the synthesized speech. In the second test, we prepared 4 datasets by the following methods: (1) original sound, (2)analysis/synthesis, (3)the sound with original spectrum and F0 contour generated by the generative model, (4) sound synthesized by rules with the generative model. In the listening tests which were performed twice for each participant, all sentences were randomly presented to 6 native Thai speakers 2 times. The listeners were asked to answer the judgement score on a 5 point scale (5-excellent, 4-Good, 3-Fair, 2-Poor, and 1-Bad) by considering the naturalness aspect. The MOS (Mean Opinions Score) became 4.82, 3.59, 3.50, and 2.29 for the above (1), (2), (3), and (4), respectively

6. Conclusion

In this paper, we studied pitch contours of Thai polysyllabic words and applied the generative model of F_0 contours for tonal languages to synthesize the F_0 contours of Thai tones. Based on the analysis of Thai polysyllabic tone sequences using this model, the phrase and tone commands' parameters for 25 patterns of 2 tone sequences and 125 patterns of 3 Thai tone sequences were gotten. The interactions between tones were expressed in term of the generative model's parameters. To show the intelligibility of the model for synthesizing Thai polysyllabic words, listening tests were performed. The results showed that for the proposed method, the average intelligibility rates were 98.8% and 96.6% for disyllabic and trisyllabic words, respectively. All errors came from the nonsense words of the test data. Therefore, the generative model of F_0 contours for Thai polysyllabic words was shown to be effective. Furthermore, we derived rules to synthesize pitch contour in the suprasegmental level (phrase or sentence). The listening tests were performed to evaluate intelligibility and naturalness of synthesized speech by the proposed method. All phrases/sentences were completely identified and the MOS of the proposed method was 3.50 compared to the original sound and analysis/synthesis sound, which were 4.82 and 3.59, respectively.

Table 4: *List of phrase/sentence used in experiment.*

Phoneme	**Thai script**
/bâan-náa/	บ้านน้า
/tâng-tɛ̀ɛ-k^{h}âm-wan-níi/	ตั้งแต่ค่ำวันนี้
/con-t^{h}ɯ̌ɯ-khâm-wan-p^{h}rŭŋ-níi/	จนถึงค่ำวันพรุ่งนี้
/túk-p^{h}âa-?aa-kàat-yaŋ-k^{h}oŋ-yen-yùu/	ทุกภาคอากาศยังคงเย็นอยู่
/?àk-k^{h}a-rà-hâa-wan-nĭi-nə̂ə-c^{h}áa/	อักขรห้าวันหนีเนินช้า
/kaan-tâŋ-tôn-ŋaan-níi-săm-k^{h}an-tîi-sùt/	การตั้งต้นงานนี้สำคัญที่สุด

7. References

[1] A.S. Abramson, "Lexical tone and sentence prosody in Thai", *Proceeding of the ninth International Congress of Phonetics Science*, Copenhagen, Denmark, 380-387, August 1979.

[2] P. Seresangtakul, T. Takara, "Analysis of pitch contours of Thai tone using Fujisaki's model", *Processing of ICASSP 2002*, Orlando, USA, 505-508, May 2002.

[3] P.Seresangtakul and T. Takara, "Analysis and Synthesis of Pitch Contour of Thai Tone Using Fujisaki's Model", *IEICE Trans. Inf.&Syst.*, Vol. E86-D, No. 10, 2223-2230, 2003.

[4] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", *J. Acoustic Society Japan (E)* 5, No.4, 133-142, 1984.

[5] H. Fujisaki, K. Hirose, P. Halle, and H. Lei, "Analysis and modeling of tonal features in polysyllabic words and sentences of the standard Chinese", *Proc. of ICSLP'90*, 841-844, 1990.

[6] H. Fujisaki and S. Ohno, "The use of generative model of F_0 contours for multilingual speech synthesis", *Proc. of ICSP'98*, 714-717, 1998.

[7] P. Rose, "Considerations in the normalization of the fundamental frequency of linguistic tone", *Speech Communication*, Vol. 6, 343-351, 1987.