# A GENERATIVE MODEL OF FUNDAMENTAL FREQUENCY CONTOURS FOR POLYSYLLABIC WORDS OF THAI TONES

*Pusadee SERESANGTAKUL and Tomio TAKARA*

Department of Information Engineering,
University of the Ryukyus
1 Senbaru, Nishihara, Okinawa, 903-0213, JAPAN
pusadee@iip.ie.u-ryukyu.ac.jp, takara@ie.u-ryukyu.ac.jp

## ABSTRACT

Thai speech synthesis by rule has been developed using cepstral parameters. 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 disyllabic words were analyzed. Based on the analysis of Thai polysyllabic words using this model, rules are derived to synthesize Thai disyllabic words, which we then applied. We performed listening tests to evaluate intelligibility of the model for Thai tone generation. The correct rates were 95% and 99% for no-meaning words and meaning words, respectively. The generative model of $F_0$ contours for Thai words was shown to be effective.

## 1. INTRODUCTION

One of the major proposals of speech synthesis system is to make a system that has high intelligibility and naturalness. 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 effect that tone has on linguistic meaning is shown in the following examples: the mid คา /k$^h$aa/ "to get stuck", the low ข่า /k$^h$àa/ "galangal, a kind of spice", the falling ฆ่า /k$^h$âa/ "to kill" the high ค้า /k$^h$áa/ "trade" and the rising ขา /k$^h$ǎa/ "leg".

The tone is correlated to fundamental frequency ($F_0$). Seresangtakul et al. studied and applied Fujisaki's model to Thai language[2]. 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 Fujisaki's parameters for tonal language. Based on this model, we analyzed disyllabic words and defined rules to synthesize speech sound of such words. The rules were then applied to synthesize Thai disyllabic words. Finally listening tests were performed to evaluate the intelligibility of the model for Thai disyllabic words.

## 2. THAI SPEECH SYNTHESIS SYSTEM

Thai speech synthesis by rule has been developed using cepstral parameters [2]. In the system, synthetic sound is produced from demisyllable parameters, which are composed of the cepstral parameters, the pitch period and the voiced/unvoiced parameters. The parameters are prepared by the cepstral method using linear interpolation. The cepstrum[3] is defined as the inverse Fourier transform of the short time logarithmic amplitude spectrum of the speech signal. The cepstrum is separated into high-quefrency and low-quefrency parts by using a cepstral window with liftering. The fundamental period of the speech signal is extracted from the peak at the high-quefrency part. The Fourier spectrum of the low-quefrency part appears as a vocal tract parameter. Voiced and unvoiced sounds are distinguished by using the threshold parameter at the low frequency part.

In the synthesis part, the synthetic sound is produced using a Log Magnitude Approximation (LMA) [4] filter as the synthesis filter, for which cepstral coefficients are used to characterize the speech sound. The LMA filter is controlled by cepstrum parameters as vocal tract parameters, and driven by pitch impulse series for voiced sounds and by white noise for unvoiced sounds. The pitch of the speech is controlled by the impulse series of the pitch period.

## 3. 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 such as Chinese, German, and Swedish[5-7].

The model has been extended to apply for the $F_0$ contour of Thai[2]. 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:

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

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

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

Where $G_{pi}(t)$ represents the impulse response function of the phrase control mechanism and $G_{t,jk}(t)$ 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, $A_{pi}$ and $A_{t,jk}$ 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. $\alpha_i$ and $\beta_{jk}$ 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.

## 4. ANALYSIS OF $F_0$ CONTOUR OF THAI DISYLLABIC WORD

### 4.1. Speech Material

In this study, a Thai speech corpus was created. Ten native Thai speakers, 3 males and 7 females, participated in the study.

We prepared 25 words of 2 tone sequences. These words were uttered by the speakers. The recording was done in a soundproof room using digital audiotape (DAT) at a sampling rate of 48 kHz. It was down sampled to 10 kHz. In the recording process, all speakers uttered each word at a normal speech rate. Each word was uttered 3 times. The best unit out of three was chosen to analyze and get the $F_0$ contour using the analysis system.

### 4.2. 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 ($F_{0i}$) was transformed to logarithmic $F_0$ in Z-score ($Z_i$)[8], which is a function of mean ($\overline{F_0}$) and standard deviation (SD) as show in expression 2.

$$Z_i = \frac{F_{0i} - \bar{F}_0}{SD} \quad (2)$$

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. Figure 1 shows 25 average $F_0$ contours, where the horizontal axis is duration in frame and the vertical axis is $F_0$ in Hertz scale.

In this work, we model the $F_0$ contours of Thai tones by using the generative model of $F_0$ contours (Fujisaki's model). Therefore to get Fujisaki's parameters for Thai disyllabic 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 between 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-mid

tones 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.

### 4.3. Result and discussion

Both phrase and tone component parameters were gotten by using non-linear least square fitting. The values of the parameters varied 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 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 disyllabic tone sequences. Table 1 shows the result of the tone components of each tone in each position.

The beginning times of these commands are found, in case of the phrase commands of disyllabic words, approximately 22 frames before the onset of an utterance. In case of the tone commands, they are found to start at 0 to 6 frames before the onset of the utterance for the mid, the low, the falling, and the rising tones, and 17.7 frames after the onset of the utterance for the high tone. Alpha is approximately 2.9, which is the same value as that of monosyllabic words [2]. An example of synthetic $F_0$ and average $F_0$ contours of the FR type is shown in Figure 2.

## 5. LISTENING TEST

### 5.1. Speech materials.

In order to evaluate the intelligibility, we prepared 2 datasets of all combination of 2 Thai tone sequences. The first dataset was composed of 25 disyllabic words that were generated from the same spectrum of the phoneme /màa màa/ uttered by a native Thai male speaker. The second dataset was synthesized from the phoneme /nàa lɔ̀ɔ/.

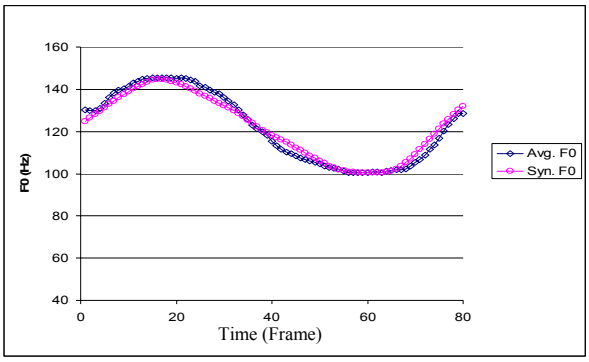


**Figure 2**. An average $F_0$ contour and a synthetic $F_0$ contour of the FR type contour generated using Fujisaki's parameters.

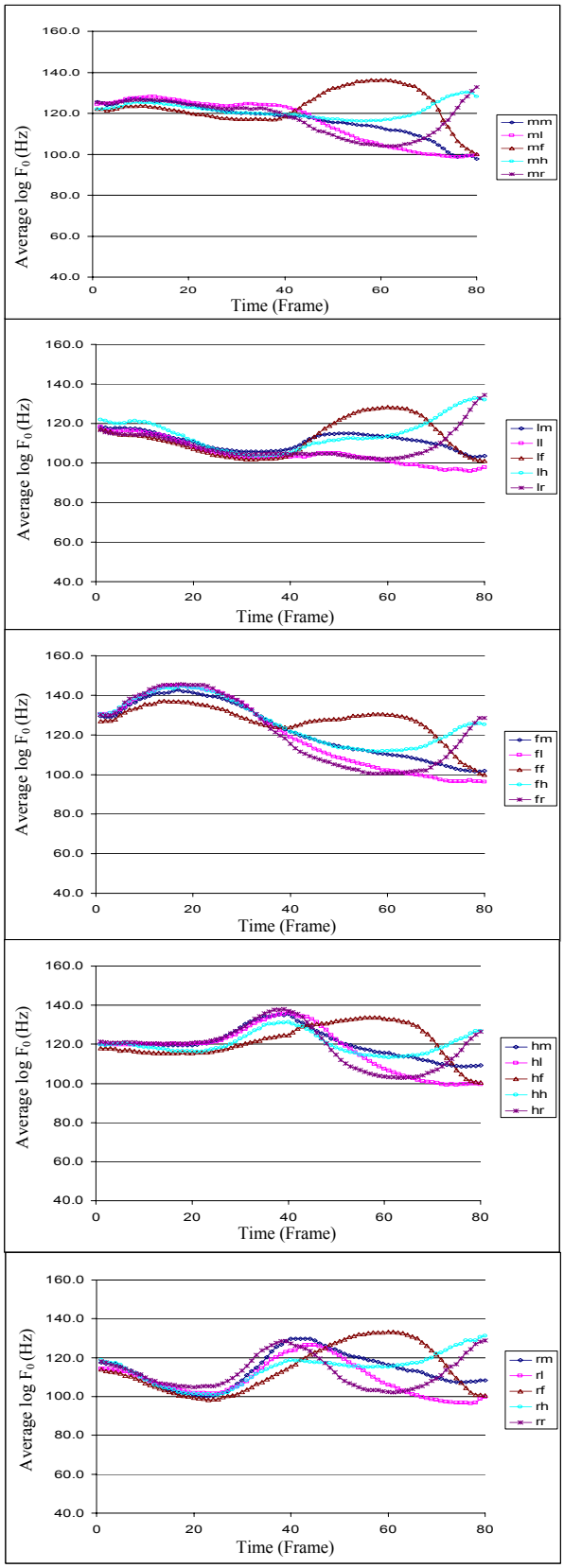


**Figure 1**. The average $F_0$ contours of 2 Thai tone sequences.

More than 70% of the words in the first dataset are no-meaning words and 80% of the words in the second dataset are meaning words. The $F_0$ contours of both datasets were generated from the parameters shown in Table 1. The time onset and offset of the tone parameters were adjusted by a lengthening/shortening factor, which is ratio of the desired length to the model length.

### 5.2. Listening test and result

In the experiment, six native Thai listeners participated in the listening test. All listeners have normal hearing ability and most of them are not familiar with synthetic sounds.

In the test, 2 datasets were presented to the listener. 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 the word that corresponds to the sound that they heard before playing the next word. A summary of the average intelligibility scores, calculated after removing the score of the listeners with maximum and minimum error value, is shown in table 2. We found that the errors came from no-meaning words.

**Table 2**: Listening test result after removing the maximum and minimal error score.

| Word | No of Sample | Error Rate |
|---|---|---|
| /maa maa/ | 25x2 | 5% |
| /naa lɔɔ / | 25x2 | 1% |

## 6. CONCLUSION

In this work, we studied pitch contours of Thai polysyllabic words and applied the generative model of $F_0$ contours (Fujisaki's model) for tonal languages in order to synthesize the $F_0$ contours of Thai tones. Based on the analysis of Thai disyllabic words using this model, the phrase and tone commands parameters for 25 patterns of 2 Thai tone sequences were gotten. The interactions between tones were described in term of Fujisaki's parameters. Ten groups of the parameters were derived to synthesized Thai disyllabic words. To show the intelligibility of the model for synthesizing Thai disyllabic words, listening tests were performed. The results showed that for the proposed method, the intelligibility rates were 95% and 99% for no meaning and meaning words, respectively. Therefore, the generative model of $F_0$ contours for Thai disyllabic words was shown to be effective.

## 7. REFERENCES


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

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

[3] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", Marcel Dekker, inc., New York, USA, 2001.

[4] S. Imai, "Log Magnitude Approximation (LMA) filter" Trans. IECE Japan, J63-A, 12, pp. 886-896, 1980-12 (In Japanese).

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

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

[7] H. Fujisaki and S. Ohno, "The use of generative model of $F_0$ contours for multilingual speech synthesis", Proceedings of ICSLP'98, pp. 714-717, 1998.

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


**Table1:** Fujisaki's parameters for disyllabic words and their standard deviations (in parentheses)

| Number of Syllable | Syllable No. | Parameter | Tone | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Mid** | **Low** | **Falling** | | **High** | **Rising** | |
| | | $k$ | 1 | 1 | 1 | 2 | 1 | 1 | 2 |
| 2 | 1 | $A_{t,1k}$ | 0.030 (0.04) | -0.172 (0.02) | 0.222 (0.05) | -0.155 (0.04) | 0.276 (0.05) | -0.261 (0.05) | 0.324 (0.06) |
| | | $T_{11k}$ | 0.0 (0.00) | -5.6 (0.47) | -3.8 (0.97) | 30.5 (1.08) | 17.7 (1.84) | -5.1 (1.08) | 26.9 (2.51) |
| | | $T_{21k}$ | 39.0 (0.0) | 29.6 (0.9) | 13.4 (2.09) | 37.2 (4.02) | 36.4 (1.67) | 16.4 (1.18) | 36.2 (1.84) |
| | | β | 11.5 | 11.5 | 11.5 | 11.5 | 9.5 | 11.5 | 11.5 |
| | 2 | $A_{t,2k}$ | 0.036 (0.03) | -0.08 (0.03) | 0.315 (0.05) | -0.278 (0.05) | 0.307 (0.03) | -0.188 (0.12) | 0.486 (0.06) |
| | | $T_{12k}$ | 40.0 (0.0) | 40.0 (0.0) | 38.1 (0.31) | 66.8 (0.31) | 53.7 (2.42) | 40.0 (0.0) | 63.7 (1.0) |
| | | $T_{22k}$ | 79.0 (0.0) | 79.0 (0.0) | 55.3 (1.7) | 79.0 (0.0) | 79.0 (0.0) | 52.1 (4.5) | 79.0 (0.0) |
| | | β | 11.5 | 11.5 | 11.5 | 11.5 | 9.5 | 11.5 | 11.5 |