# Tone Modeling Using Stress Information for HMM-Based Thai Speech Synthesis

*Decha Moungsri*[1], *Tomoki Koriyama*[1], *Takashi Nose*[2], *Takao Kobayashi*[1]

[1]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan
[2]Graduate School of Engineering, Tohoku University, Japan

moungsri.d.aa@m.titech.ac.jp, koriyama@ip.titech.ac.jp,
tnose@m.tohoku.ac.jp, takao.kobayashi@ip.titech.ac.jp

## Abstract

This paper describes a modeling technique of Thai tones for HMM-based speech synthesis. Tones are important prosodic features for tonal languages including Thai because the phonetically same words but with different tones give different meanings. Although there have been several approaches to improving tone correctness of synthetic speech by considering tone types, another significant factor, stress, was not used explicitly for prosody modeling. We incorporate stress/unstress information into the framework of the HMM-based speech synthesis. Objective and subjective evaluation results show that the use of stress information improves the performance in Thai tone modeling.

**Index Terms**: HMM-based speech synthesis, tone correctness, stress, context clustering

## 1. Introduction

In the speech synthesis of tonal languages including Chinese, Vietnamese, and Thai, tone correctness of the synthetic speech is a crucial point, because different tones give different meanings even if the phonetic information of words is the same. In this context, various techniques have been examined to improve tone correctness for Thai speech synthesis. A tone-separated tree structure was introduced for HMM-based Thai speech synthesis to reduce the tone-dependent effects in the context clustering process [1]. Moreover, to capture a variety of shapes of fundamental frequency (F0) contours effectively, phrase-intonation and tone-geometrical features derived from Fujisaki-model were used as the contexts in the HMM-based speech synthesis [2]. Another approach to modeling tone in Thai is the use of Tilt model [3]. The Tilt model was extended for modeling F0 contours in tonal languages and the modified Tilt model for Thai is called T-Tilt model [4]. To enhance T-Tilt model, an optimized T-Tilt model by expanding the Tilt curve over the whole syllable was also proposed [5]. If tone information is determined from transcriptions of speech data in the modeling, the quality of synthetic tones much depends on the accuracy of tone labeling. For this problem, a tone modeling technique using a quantized F0 context was proposed to reduce the tone distortion caused by inconsistent tonal labeling, in which quantized F0 symbols were utilized as the context for constructing the decision trees [6].

Although, these techniques improve the naturalness and tone intelligibility of synthetic speech, tone correctness is not perfect and the tones of some syllables in continuous synthetic speech are perceived to be obviously incorrect and unnatural. To alleviate this problem, we focus on one of other factors, which affects prosodic features and has not been considered explicitly as a context in the conventional HMM-based Thai speech synthesis. In [7], it is reported that stressed syllables have largely different characteristics from unstressed ones. Stressed syllables have typical F0 contours and long durations. In contrast, unstressed syllables are diverse. For example, the effect of speaking rate on the F0 contours of unstressed syllables is more extensive, both in terms of height and slope, than that of stressed syllables [8]. As described in [9], coarticulation and intonation affect tonal assimilation and declination. The coarticulatory effect will cause the change in F0 contour shape of a syllable to be assimilated with neighboring syllables.

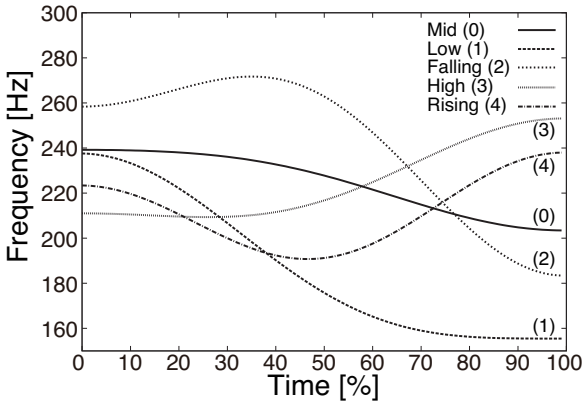



Figure 1: Typical F0 contours of Thai tones. Each contour represents the average of extracted F0 contours from the database TSynC-1.

In this paper, we examine stressed/unstressed effects on Thai tones, and propose a modeling technique using stress information to improve Thai tone correctness of HMM-based synthetic speech. We classify stressed and unstressed syllables manually by using the stress properties described in [7, 10], and incorporate the stress information into the context labeling. We compare the performance of the proposed and conventional methods through objective and subjective evaluations and show the effectiveness of the proposed method.

## 2. Tone and stress in Thai

Tone represents a change in the pitch of a syllable during its pronunciation. In Thai, every syllable is pronounced in one of five tones: mid (0), low (1), falling (2), high (3), or rising (4).



10.21437/SpeechProsody.2014-200

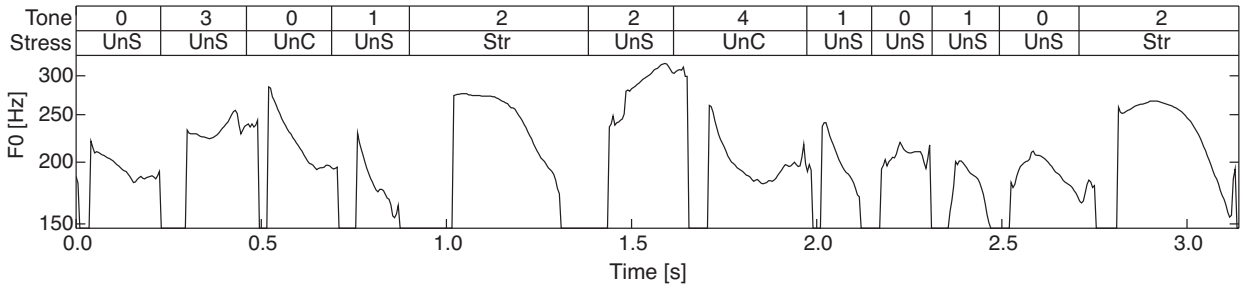
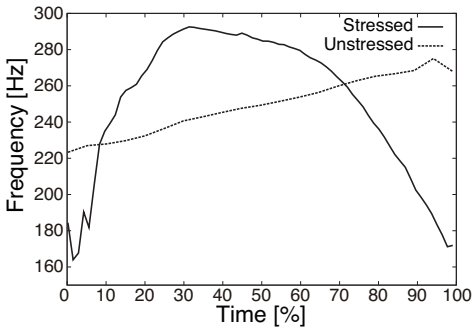
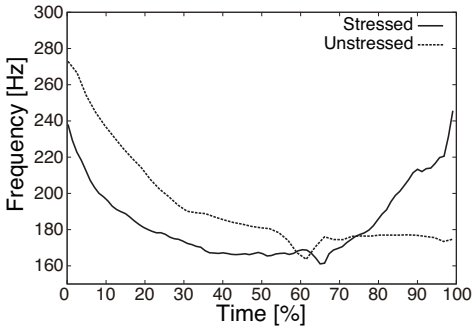



Figure 3: Stressed and unstressed syllables in natural speech (Str : Stressed syllable, UnS : Unstressed syllable, and UnC : Unclear syllable).




(a) Falling tone (2)

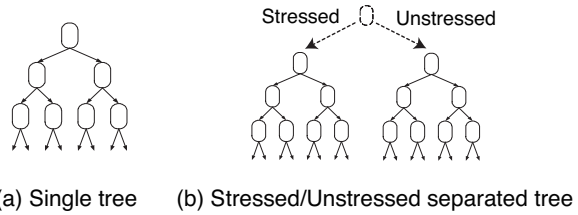



(b) Rising tone (4)

Figure 2: Example of F0 contours in (a) falling tone (2) and (b) rising tone (4) syllables.

The tone must be spoken correctly for the intended meaning of a word to be understood. The identification of a Thai tone relies on the shape of the F0 contour. Figure 1 shows the typical F0 contours of five different tones of isolated syllables. However every F0 contour shape does not always look like the typical one. The shapes depend on stress information of syllables [7]. Figure 2 shows an example of F0 contour shapes of falling tone (2) and rising tone (4) in stressed and unstressed syllables which were extracted from speech samples included in Thai speech database TSynC-1 [11].

The stressed F0 contour shapes are similar to the typical ones, whereas the shapes of unstressed syllables tend to be flat and have less movement of contour, especially in falling tone (2) and rising tone (4). Furthermore, the actual unstressed F0 contours are diverse. Several interacting factors affect F0




(a) Single tree (b) Stressed/Unstressed separated tree

Figure 4: Decision trees for context clustering: (a) single tree structure and (b) stressed/unstressed separated tree structure.

realization of tones, e.g., syllable structure, declination, tonal assimilation, stress, and speaking rate [7–10].

## 3. Tone modeling with stressed/unstressed context

### 3.1. Annotation of stress information

As described in [10], stressed syllables have following characteristics:

- Long duation
- F0 contour similar to the prototypes
- High energy

Duration is the predominant feature in the distinction between stressed and unstressed syllables in Thai. The secondary feature is the range of F0 movement [7]. Generally, stressed syllables appear in the end of utterances, isolated phrases, and emphasized words. In other words, the characteristics of unstressed syllables are the opposite ones of stressed syllables.

Figure 3 shows an example of the F0 contour that includes stressed, unstressed, and unclear syllables. The utterance contains 12 syllables, of which the fifth and the twelfth ones are stressed, the third and the seventh ones have unclear stress information, and the others are unstressed.

In this study, we classified stressed and unstressed syllables manually. First, we listened to each syllable individually. If it clearly has the properties of stressed syllables, and it can be classified into only one tone, we annotate a *stressed* label on it. Otherwise, *unstressed* is annotated. However, there are some syllables that are not easy to distinguish, because they are indistinctly uttered by the characteristics between the stressed and unstressed syllables. Currently, we annotate *unstressed* labels for such indistinctly uttered syllables.

Table 1: Average F0 distortion between original and generated speech samples.

| Method | RMS error [cent] | # of leaf nodes |
|---|---|---|
| Conventional | 139.1 | 2532 |
| Single tree | 132.3 | 2470 |
| Separated tree | 132.6 | 2685 |

### 3.2. Context clustering using stress information

We examine two tone modeling methods using stress information. Specifically, we incorporate the stress information into the context clustering that is an essential process in the HMM-based speech synthesis [12]. The first method is adding the stress information to the context set. By incorporating stress information to the context set, the different characteristics are automatically separated during the decision tree clustering. We refer to this method as *single tree*. The second method is to use two trees separated by stressed and unstressed syllables using a manner similar to the tone-separated tree structure proposed in [1]. It is based on the facts that the characteristics of stressed and unstressed syllables are largely different and the frequencies of the stressed and unstressed syllables are imbalanced. In this method, the stress contexts of preceding and succeeding syllables are taken into account based on the results of [1]. We refer to this method as *stressed/unstressed separated tree*. Figure 4 illustrates decision tree structures of the proposed two methods.

## 4. Experiment

### 4.1. Training condition

A set of phonetically balanced sentences of Thai speech database named TSynC-1 from NECTEC [11] was used for training and evaluation. The sentences in the database were uttered by a professional female speaker with clear articulation and standard Thai accent with reading style. A speaker-dependent model was trained using 340 utterances from the database. We used 29 utterances for evaluation, which were not included in the training set.

Speech signals were sampled at a rate of 16kHz. F0 and spectral features were extracted by STRAIGHT [13] with 5-ms frame shift. The feature vector of phone-unit HMM consisted of 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. We used hidden semi-Markov model (HSMM) which has explicit duration distributions. The model topology was 5-state left-to-right context-dependent HSMM. The conventional method uses context clustering as described in [14]. Context clustering of the proposed method was extended from the conventional one by including stress information in the context set.

### 4.2. Objective evaluation result

The proposed and conventional methods were evaluated objectively. The measurements for evaluation were average F0 distortions that were calculated by RMS error between generated and original log F0s.

The results are shown in Table 1. The number of leaf nodes of F0 decision trees for context clustering is also shown in the table. The F0 distortions of the proposed methods were lower than the conventional method. However, there was only slight difference between the single tree and stressed/unstressed separated tree. In addition, the numbers of leaf nodes were not much different for all methods. Figure 7 shows an example of the F0 contours generated by all methods compared to the original speech. The sixth and eleventh syllables are the stressed syllable in the falling tone. The F0 contours of these syllables in Fig. 7 (a), which were generated by the conventional method, are not like the falling tone because they do not fall at the end of the syllable, but they are similar to those of the mid tone. In Figs. 7 (b) and (c), the F0 contours were generated by the proposed methods and they are similar to the original ones. The ends of the F0 contours are falling and were perceived as the falling tone. There is a slight difference in F0 contours between the single tree and stressed/unstressed separated tree structure methods.

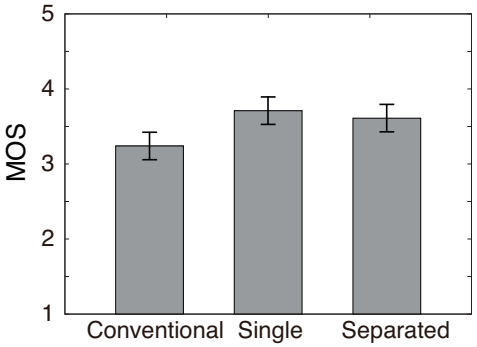



Figure 5: Mean opinion score of naturalness of synthetic tone comparison.

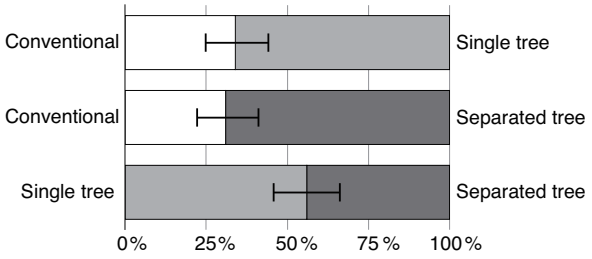
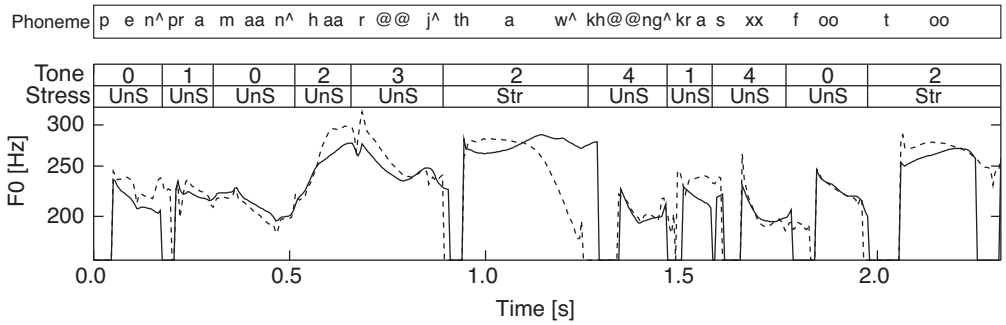



Figure 6: The results of forced choice preference test in tone intelligibility.

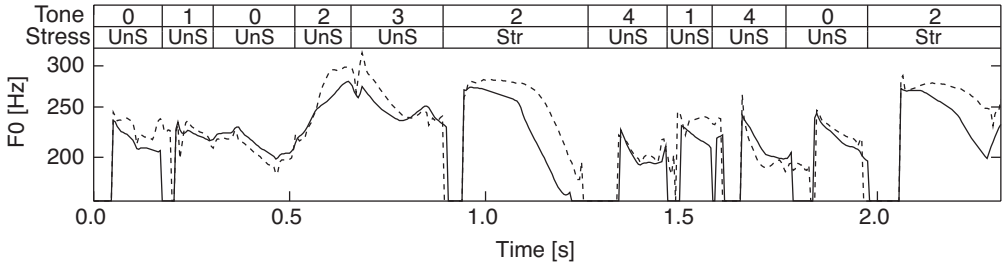
### 4.3. Subjective evaluation result

To ensure the effectiveness of the proposed methods, we evaluated the perceptual quality in terms of naturalness and tone intelligibility. Specifically, we employed mean opinion score (MOS) and forced choice preference tests.

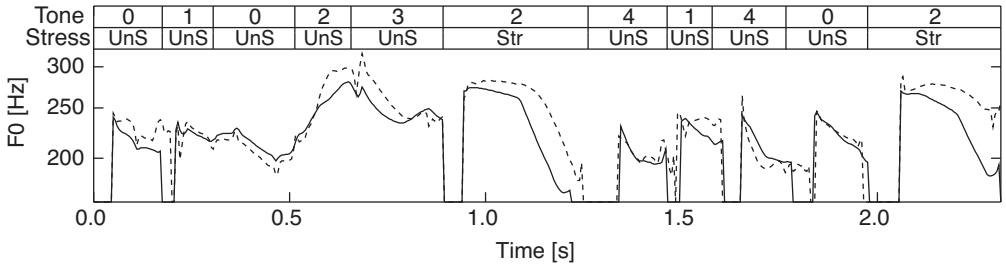
Ten utterances were randomly chosen from the synthetic speech samples used in the objective evaluation test. We assessed the synthetic speech from the proposed two methods and the conventional method. As a result, we compared three types of synthetic speech in the evaluation. Ten Thai native speakers listened to and evaluated the samples. In MOS test, the listeners evaluated each utterance on a five-point scale from 1 to 5 according to their satisfaction with the perceptual naturalness of tones. The definition of the rating was: 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. Listeners could repeat sentences to evaluate as many times as they required for ensuring that they were accurately evaluating. Figure 5 shows the result with




(a) F0 contours comparison between the original one in dashed line and the conventional one in soild line.




(b) F0 contours comparison between the original one in dashed line and the single tree one in soild line.




(c) F0 contours comparison between the original one in dashed line and the separated tree one in soild line.

Figure 7: Examples of F0 contours generated by (a) the conventional method, (b) the single tree method, and (c) the separated tree method. The sentence means that "it is about 500 times of the photoelectric." The tone numbers represent mid tone (0), low tone (1), falling tone (2), high tone (3), and rising tone (4).

95% confidence intervals. It can be observed that the proposed methods outperformed the conventional method. The scores of both proposed methods are not significantly different and it is consistent with the result of objective evaluation.

In the forced choice preference test, the listeners were asked to choose more natural-sounding tone from each pair of synthetic speech. The listener could repeat sentences as many times as they required in the same way as the MOS test. The results of the forced choice preference test are shown in Fig. 6. It is seen again that the proposed methods outperformed the conventional method. When comparing between the single tree structure method and the stressed/unstressed separated tree structure method, the listeners preferred the single tree method, but the difference is statistically not significant.

## 5. Conclusion

This paper has described a modeling technique of Thai speech synthesis using the stress information of syllables. Although stress is an important factor for tone perception in Thai, stress information has not been included in the context clustering in conventional techniques. This causes generation of incorrect tones. To overcome the problem, we added stress information to context clustering. The objective evaluation showed that the proposed method can reduce the F0 distortion significantly. The subjective tests also yielded results that corresponded to those from the objective tests. Although we have confirmed that the stress information could improve the tone correctness, there still exist unnatural tones in synthetic speech. For improvement of the tone naturalness, syllable-level unit might not be appropriate. Thus, in future work, we will investigate a tone modeling technique based on longer unit such as word or phrase.

## 6. Acknowledgements

We would like to thank Dr. Vataya Chunwijitra of NECTEC, Thailand, for providing us with helpful discussion and the TSynC-1 speech corpora. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071.

## 7. References


[1] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker-independent average-voice-based Thai speech synthesis," *Speech Communication*, vol. 51, no. 4, pp. 330–343, 2009.

[2] S. Chomphan and T. Kobayashi, "Incorporation of phrase intonation to context clustering for average voice models in HMM-based Thai speech synthesis," in *Proc. ICASSP*, 2008, pp. 4637–4640.

[3] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *J. Acoust. Soc. Am.*, vol. 107, p. 1697, 2000.

[4] A. Thangthai, N. Thatphithakkul, C. Wutiwiwatchai, A. Rugchatjaroen, and S. Saychum, "T-Tilt: a modified tilt model for F0 analysis and synthesis in tonal languages," in *Proc. INTERSPEECH*, 2008, pp. 2270–2273.

[5] A. Thangthai, A. Rugchatjaroen, N. Thatphithakkul, A. Chotimongkol, and C. Wutiwiwatchai, "Optimization of T-Tilt F0 modeling," in *Proc. INTERSPEECH*, 2009, pp. 508–511.

[6] V. Chunwijitra, T. Nose, and T. Kobayashi, "A tone-modeling technique using a quantized F0 context to improve tone correctness in average-voice-based speech synthesis," *Speech Communication*, vol. 54, no. 2, pp. 245–255, 2012.

[7] S. Potisuk, J. Gandour, and M. Harper, "Acoustic correlates of stress in Thai," *Phonetica*, vol. 53, no. 4, pp. 200–220, 1996.

[8] J. Gandour, A. Tumtavitikul, and N. Satthamnuwong, "Effects of speaking rate on Thai tones," *Phonetica*, vol. 56, no. 3-4, pp. 123–134, 1999.

[9] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "Tone recognition in Thai continuous speech based on coarticulaion, intonation and stress effects," in *Proc. INTERSPEECH*, 2002.

[10] N. Thubthong, B. Kijsirikul, and S. Luksaneeyanawin, "Stress and tone recognition of polysyllabic words in Thai speech," in *Proc. Int. Conf. Intelligent Technologies*, 2001, pp. 356–364.

[11] C. Hansakunbuntheung, A. Rugchatjaroen, and C. Wutiwiwatchai, "Space reduction of speech corpus based on quality perception for unit selection speech synthesis," *Proc. SNLP*, pp. 127–132, 2005.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH*, 1999, pp. 2347–2350.

[13] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[14] S. Chomphan and T. Kobayashi, "Implementation and evaluation of an HMM-based Thai speech synthesis system," in *Proc. INTERSPEECH*, 2007, pp. 2849–2852.