HMM-Based Thai Speech Synthesis Using Unsupervised Stress Context Labeling

Decha Moungsri, Tomoki Koriyama, and Takao Kobayashi
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan
E-mail: moungsri.d.aa@m.titech.ac.jp, koriyama@ip.titech.ac.jp, takao.kobayashi@ip.titech.ac.jp

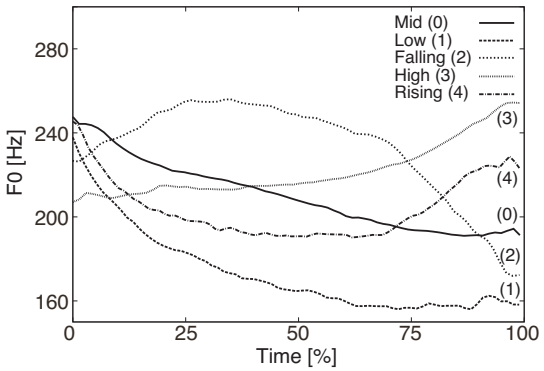
***Abstract*—This paper describes an approach to HMM-based Thai speech synthesis using stress context. It has been shown that context related to stressed/unstressed syllable information (stress context) significantly improves the tone correctness of the synthetic speech, but there is a problem of requiring a manual context labeling process in tone modeling. To reduce costs for the stress context labeling, we propose an unsupervised technique for automatic labeling based on the characteristics of Thai stressed syllables, namely, having high F0 movement and long duration. In the proposed technique, we use log F0 variance and duration of each syllable to classify it into one of stress-related context classes. Objective and subjective evaluation results show that the proposed context labeling gives comparable performance to that conducted carefully by a human in terms of tone naturalness of synthetic speech.**

I. Introduction

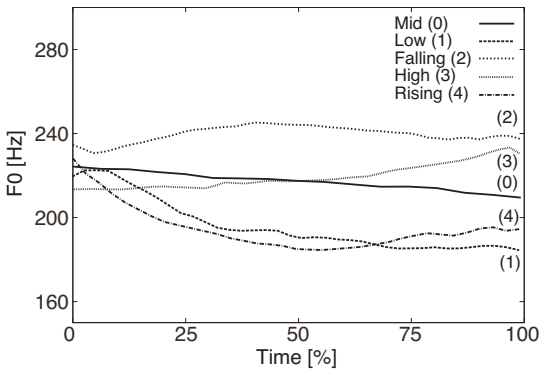
In tonal languages, tone is important for listeners to perceive meaning of a word. Different tones give different meanings, even though the pronunciation of the word is identical. Therefore, it is crucial for speech synthesis to reproduce them appropriately. Various techniques have been investigated to enhance tone correctness and naturalness in HMM-based Thai speech synthesis. For example, a tone-separated tree was introduced to reduce the tone-dependent effects in context clustering [1]. Due to the diversity of fundamental frequency (F0) contour shape, phrase-intonation and tone-geometrical feature were investigated as the context to obtain better tone models [2]. Furthermore, since the quality of synthetic tones highly depend on the accuracy of tone labeling of training data, a quantized F0 context was proposed to reduce the tone distortion caused by inconsistent tonal labeling, in which quantized F0 symbols were used as the context [3].

These methods substantially improve tone correctness and naturalness of synthetic speech. However, synthetic tones still have some incorrectness or unnaturalness in the synthesis of continuous speech. This is caused by the fact that tones in Thai language are influenced by many factors such as stress, speaking rate, tone assimilation, etc. [4–6], and these factors have not been entirely considered as the contexts in the HMM-based speech synthesis framework. To alleviate the problem, we proposed a technique that used stress information as an additional context [7]. In this approach, the stress context was manually labeled and it was shown that the tone correctness and naturalness of resultant synthetic speech were significantly improved. However, there remains a problem that the manual annotation is time-consuming and expensive.

To overcome this problem, in this paper, we examine an unsupervised technique for automatic stress context labeling. The key idea of the technique is based on an unsupervised context labeling technique [8, 9] that was originally used for expressive speech synthesis. Differing from the previous study [8, 9], we focus on duration and F0 movement in syllable-unit which have been considered to be the major factors in the stressed and unstressed syllable classification [5]. More specifically, we use additional two-class stress-related context sets in terms of log F0 variance and duration of each syllable. The variance of log F0 represents a degree of F0 movement in a syllable. We investigate how these stress-related context sets are determined optimally and compare the performance of the proposed technique with the manual labeling one through objective and subjective evaluation tests.



(a) F0 contours of stressed syllable.



(b) F0 contours of unstressed syllable.

Fig. 1: Example of F0 contours in (a) stressed and (b) unstressed syllables.

II. Tone and stress in Thai

Tone corresponds to a change of the pitch in a certain pronunciation unit. In Thai, every syllable is pronounced with

978-616-361-823-8

one of five tones: mid (0), low (1), falling (2), high (3), or rising (4). 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. However every F0 contour shape does not always look like the typical one. The shapes depend on stress information of syllables [5]. Figure 1 shows an example of F0 contour shapes of each tone in stressed and unstressed syllables which are the same phones, and were extracted from speech samples included in Thai speech database TSynC-1 [10].

In general, stressed syllables have long duration, F0 contour similar to the typical ones, and high energy [11] but unstressed syllables tend to be flat and have less movement of contour, especially in falling tone (2) and rising tone (4). Indeed, duration is the predominant feature for distinguishing stressed syllables from unstressed ones in Thai. The secondary feature is the degree of F0 movement within a syllable [5]. Stress position in polysyllabic word depend on various factors such as root word, meaning, and part of speech [12]. Generally, stressed syllables appear in the end of utterances, isolated phrases, and emphasized words.

III. Unsupervised stressed/unstressed context labeling

A. Context labeling process

It has been shown that stressed/unstressed context significantly improves the tone correctness and naturalness in the HMM-based Thai speech synthesis [7]. However, to do this, we need stressed/unstressed context labeling for given speech data before model training. To avoid time-consuming and expensive manual labeling process, we utilize an automatic annotation technique that was originally used for expressive speech synthesis [9].

As described in the previous section, two main features of stressed syllables are long duration and large F0 movement. Thus we use two features as the stress-related context, specifically, duration and the variance of log F0 values in each syllable.

We define α as the value for log F0 variance threshold, and β as the value for syllable duration threshold. The context labeling process of log F0 variance based on [9] is summarized as follows :

1) Calculate log F0 variance *V* of original speech for each syllable.
2) Classify the syllable in accordance with the value of *V* into two classes: (1) $V < \alpha$ (small F0 movement), and (2) $V \geq \alpha$ (large F0 movement), where α is a classification threshold.
3) The class index is used as a stress-related context label.

We also apply the same procedure to another stress-related context labeling, i.e., syllable duration context labeling that classifies syllables into two classes of long and short durations using the threshold β.

B. Threshold optimization

We choose optimal thresholds α and β that minimize an objective measure using a similar manner to that described in [9]. As the objective measure, we use root mean square error (RMSE) of log F0 between the original and synthetic speech.

First, we arbitrarily choose thresholds between predetermined lower and upper bounds, $\alpha_s \leq \alpha \leq \alpha_e$ and $\beta_s \leq \beta \leq \beta_e$. Then we classify each syllable into one of the stress-related context classes using the specified thresholds. Next we train HMM synthesis units based on the newly added context labels, and generate F0 contours for all training sentences using the trained HMMs. Finally we calculate the F0 error, and repeat this with changing the thresholds until the F0 error becomes smallest.

IV. Experiments

A. Experimental conditions

We conducted experiments by adding the stress-related context to the conventionally used context set for HMM-based Thai speech synthesis. By using two kinds of stress-related context labels, we can classify syllables into four classes that are stressed syllables (large-variance and long duration), unstressed syllables (small-variance and short duration), and unclear ones (other two classes). We incorporated them into the context clustering that is an essential process in the HMM-based speech synthesis [13].

A set of phonetically balanced sentences of Thai speech database named TSynC-1 from NECTEC [10] 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, approximately 52 minutes in total, from the database. We used 29 utterances for evaluation, which were not included in the training set.

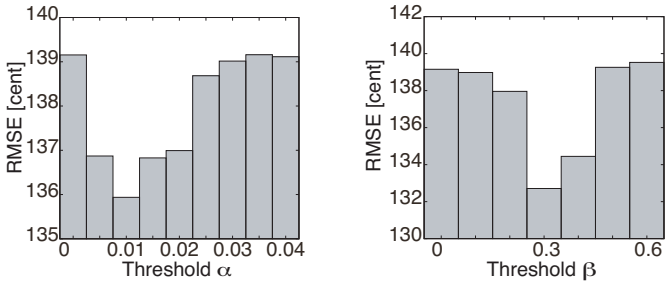
Speech signals were sampled at a rate of 16kHz. F0 and spectral features were extracted by STRAIGHT [14] with 5-ms frame shift. The feature vector of phone-unit HMM consisted of 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. We used hidden semi-Markov model (HSMM) which has explicit duration distributions. The model topology was 5-state left-to-right context-dependent HSMM. The conventional technique used the context clustering as described in [15] and the manual stressed/unstressed labeling case used the one described in [7]. We performed a 5-fold cross-validation test in threshold optimization using the training data.

B. Optimal threshold in model training

We first determined optimal thresholds for the log F0 variance and syllable duration by using the algorithm mentioned in the previous section. In the first iteration, we set $\alpha = \alpha_s$ and $\beta = \beta_s$, respectively, and at the n-th iteration, we set

$$\alpha = \alpha_s + (n-1)\Delta\alpha \qquad (\alpha \leq \alpha_e)$$
$$\beta = \beta_s + (n-1)\Delta\beta \qquad (\beta \leq \beta_e)$$

where $\Delta\alpha$ and $\Delta\beta$ are increments in each iteration. Specifically, in this experiment, α_s and β_s were fixed to zero, and α_e and β_e were set to the maximum values obtained from the



(a) (b)

Fig. 2: RMS error of log F0 with different classification threshold.

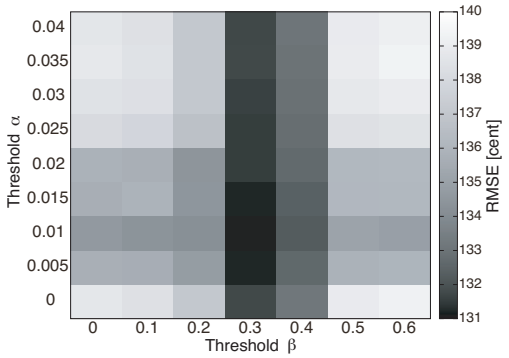


Fig. 3: RMS error of log F0 with different classification thresholds α and β in threshold optimization.

training data. In addition, we set α_e to 0.04 and $\Delta\alpha$ to 0.005 for the log F0 variance context, and β_e to 0.6 sec and $\Delta\beta$ to 0.1 sec for syllable duration context.

Figure 2(a) shows the average RMS errors of log F0 between synthetic and original speech samples with different values of α when only the log F0 variance context was added to the conventional context set. In contrast, Figure 2(b) shows the case when only the syllable duration context was used. It is seen that, for individual context, the optimal thresholds of α and β are 0.01 and 0.3, respectively. The result for the case when both of the stress-related contexts were used is shown in Figure 3. It can be seen that the optimal thresholds for using both contexts are identical to those obtained by individual optimization.

C. Objective evaluation results

We objectively evaluated the effect of the log F0 variance and syllable duration contexts. For the objective measure, we again used the RMS error of log F0 between the original and synthetic speech. Figure 4 shows the result of comparison among the conventional and newly added stress-related context sets. We can see that the syllable duration context reduces the F0 error, and moreover, the use of both the duration and log F0 variance contexts further reduces the F0 error.

Next, we compared the proposed technique with the conventional one and the manual stressed/unstressed labeling one. Figure 5 shows the RMS log F0 errors for the conventional one [15], the single tree and separated tree structure of manual labeling described in [7], and the proposed one. It is shown that the F0 error of proposed technique is smaller than that

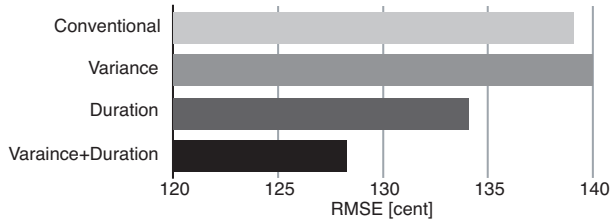


Fig. 4: Effect of using the stress-related context in terms of RMS log F0 error.

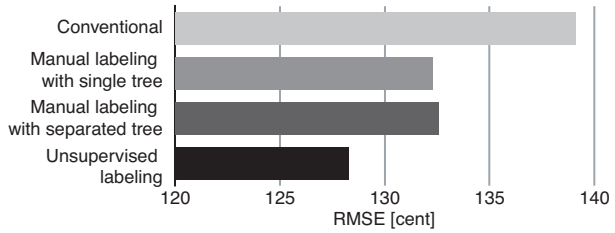


Fig. 5: Comparison among conventional, manual labeling, and unsupervised labeling techniques in terms of RMS log F0 error.

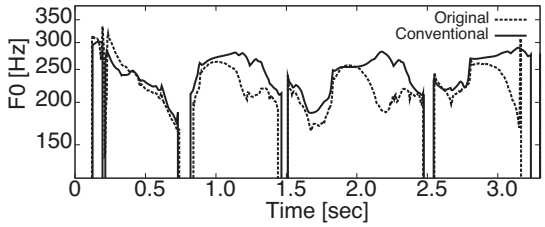
of the manual labeling one. Figure 6 shows an example of F0 contours generated using conventional technique, manual labeling one, and unsupervised labeling one. From the figure, we can see that the F0 contour of the synthetic speech with unsupervised labeling method is similar to the manual labeling one, and closer to the original one than the conventional one.

D. Subjective evaluation results

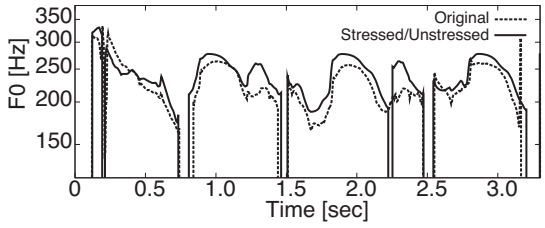
To confirm the tone intelligibility of the unsupervised labeling method, we evaluated the perceptual quality in terms of naturalness and tone intelligibility. Specifically, we used mean opinion score (MOS) and forced choice preference tests. Ten utterances were randomly chosen from the synthetic speech samples used in the objective evaluation test. We assessed the synthetic speech from the proposed technique, the manual labeling, and the conventional technique. For the manual labeling, we selected the single tree one for subjective evaluation because it had achieved the best score as described in [7]. As a result, we compared three types of synthetic speech in the evaluation. Thirteen Thai native speakers listened to and evaluated the synthetic speech samples.

In the 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 playing back sentences to evaluate as many times as they required for ensuring that they were accurately evaluating. Figure 7 shows the resultant scores with 95% confidence intervals. It can be observed that the proposed methods outperformed conventional technique. The score of manual labeling is slightly higher than that of the unsupervised labeling one but there is no significant difference.

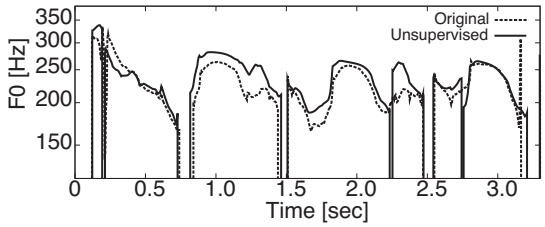
In the forced choice preference test, the listeners were asked to choose more natural-sounding tone from each pair of synthetic speech samples. The listeners could repeat playing back sentences as many times as they required in the same way as the MOS test. The results of the forced choice preference



(a) Conventional



(b) Manual labeling



(c) Unsupervised labeling

Fig. 6: Example of F0 contours compared with original F0 contour.

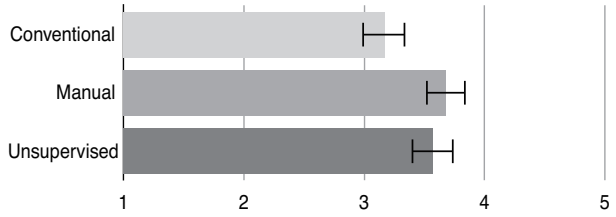


Fig. 7: Result of MOS test in subjective evaluation of naturalness of synthetic tone.

test are shown in Figure 8. The listeners preferred the synthetic speech with the stress-related context rather than the conventional one. For the manual labeling and the unsupervised one, the listeners preferred the manual one but the difference is statistically not significant.

V. Conclusion

This paper has described an unsupervised stress context labeling technique for HMM-based Thai speech synthesis. Stress is an important factor for perception in Thai. To alleviate the problem of time-consuming and expensive manual labeling, we introduced unsupervised context labeling by considering two factors that are the log F0 variance and duration of syllable. The experimental results showed that both factors are necessary for unsupervised labeling to achieve comparable quality as the manual labeling. The proposed unsupervised labeling enabled us to avoid the time-consuming stress context labeling and achieved similar quality with the manual labeling one. However, tone naturalness is still imperfect. Tone in Thai

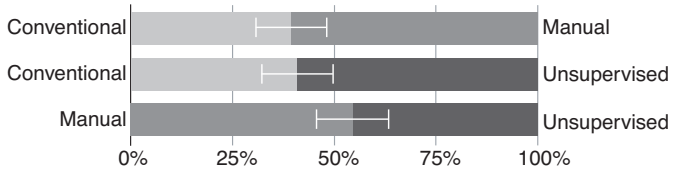


Fig. 8: Results of forced choice preference test in subjective evaluation of tone intelligibility.

is also depended on sentence structure. Thus in future work, we will focus on the tone in word level and larger unit to improve Thai speech synthesis.

Acknowledgment

We would like to thank Dr. Vataya Chunwijitra of NECTEC, Thailand, for her helpful discussion and providing the TSynC-1 speech database. We would also thank to Dr. Takashi Nose for his valuable discussions. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071.

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] 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.

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

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

[6] J. Gandour, S. Potisuk, S. Dechongkit, and S. Ponglorpisit, "Tonal coarticulation in Thai disyllabic utterances: a preliminary study," *Linguistics of the Tibeto-Burman Area*, vol. 15, no. 1, pp. 93–110, 1992.

[7] D. Moungsri, T. Koriyama, T. Nose, and T. Kobayashi, "Tone modeling using stress information for HMM-based Thai speech synthesis," in *Proc. Speech Prosody*, 2014, pp.1057–1061.

[8] Y. Maeno, T. Nose, T. Kobayashi, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "HMM-based emphatic speech synthesis using unsupervised context labeling," in *Proc. INTERSPEECH*, 2011, pp. 1849–1852.

[9] Y. Maeno, T. Nose, T. Kobayashi, T. Koriyama, Y. Ijima, H. Nakajima, H. Mizuno, and O. Yoshioka, "Prosodic variation enhancement using unsupervised context labeling for HMM-based expressive speech synthesis," *Speech Communication*, vol. 57, pp. 144–154, 2014.

[10] 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.

[11] 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.

[12] P. Peyasantiwong, "Stress in Thai," in *Papers from a Conference on Thai Studies in Honor of William J. Gedney. Michigan Papers on South and Southeast Asia, Center for South and Southeast Asian Studies, University of Michigan, Ann Arbor*, 1986, pp. 19–39.

[13] 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.

[14] 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.

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