Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

Tone modeling using Gaussian process latent variable model for statistical speech synthesis

Decha Moungsri, Tomoki Koriyama, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Japan
moungsri.d.aa@m.titech.ac.jp, {koriyama,takao.kobayashi}@ip.titech.ac.jp

Abstract

In continuous speech of Thai language, tone pronunciation is affected by several factors. One of significant factors is stress that causes a diversity of F0 contours of tone, and affects syllable durations. Our previous studies have shown that a stressed/unstressed syllable context improves tone modeling accuracy. However, the stress in Thai language is generally unknown for a given input text and it has a wide variety of degrees of stress. Thus the simple stressed/unstressed context is insufficient to represent variation of stress. In this study, we introduce an unsupervised dimensional reduction technique, variational GP-LVM, to represent a diversity of stress. The stress-related information, F0 contour and duration, is projected onto a latent space which has lower dimensionality than the original to represent the variation of stress. Then, we use the latent variable as a stress-related context in GPR-based speech synthesis framework that enables us to determine the similarity of contextual factors continuously using a kernel function. We examine two approaches to data projection: single-space projection and separated-space projection. Objective and subjective evaluation results show that the proposed technique achieves an improvement in tone modeling.

Index Terms: stress, tone modeling, GPR-based speech synthesis, latent variable model, Bayesian GP-LVM

1. Introduction

In tonal languages, tone plays an important role in distinguishing the meanings of syllable with the same phone sequence. Hence, tone modeling is a crucial point in the problem of reproducing natural-sounding synthetic speech. From this point of view, various techniques have been introduced to tone modeling for Thai language. Since the tone is very sensitive in perception, a tone-separated tree structure was proposed to remove tone-type dependency of context in tree-based context clustering for HMM-based speech synthesis framework [1]. However, tone-type context is not enough for F0 modeling since there is a large diversity of F0 contours in each tone. To handle the diversity, tone geometrical feature that represents the shape of tone in a syllable was incorporated as an additional context for the tree-based clustering [2]. Another technique to model the diversity of tone is the use of T-Tilt model [3], a modified Tilt model that has been successfully used for F0 modeling in accentual languages [4]. To overcome the problem of inconsistency between tone labels and speech data that were recorded from non-professional speakers, a quantized F0 context technique was proposed to represent F0 contours by quantized symbols [5]. In the continuous speech, the F0 contour of individual syllables are affected by co-articulation. To avoid the effect of co-articulation in model training, the tone nuclei that are less affected by the adjacent syllables were utilized [6]. A further technique to alleviate the influence of adjacent syllables is using F0 contour of vowel part instead of entire syllable in model training of tone recognition [7], which is based on the fact that vowel part receives low influence from adjacent syllables and contains the main feature of the tone.

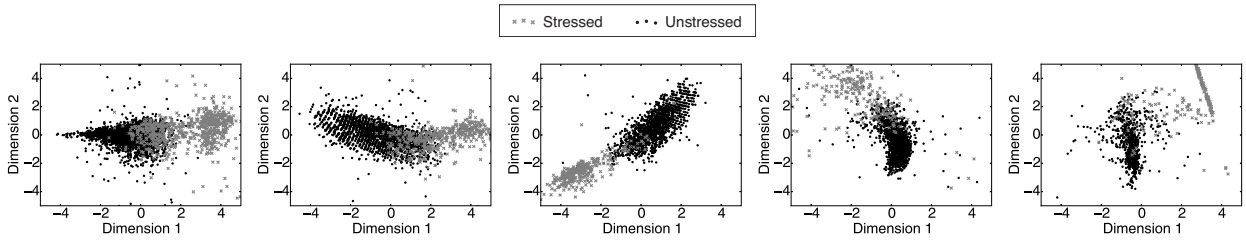
Although tone modeling has been improved by above mentioned techniques, the tone modeling is still imperfect. Specifically, there exists incorrect tone in synthetic speech that causes misunderstanding of meaning and also affects naturalness of synthetic speech. In Thai, the studies of tone and intonation have shown that tones are influenced by various factors [8–10]. One of important factors is stress that affects naturalness and meaning in word and phrase levels. In our previous work, we used a manually annotated stressed/unstressed syllable context in the tree-based context clustering for speech synthesis [11]. Since the manual labeling approach is time-consuming, we also introduced an unsupervised labeling technique for stress labeling by using prosody features such as F0 contour and duration to automatically classify syllables into stress-related classes [12]. It has been shown that the proposed stress-related context can give an improvement in tone correctness and naturalness of synthetic speech. However, the simple stressed/unstressed syllable labeling is not sufficient to represent variety and continuity of characteristics of stress. Moreover, there are syllables where characteristics are unclear between stressed and unstressed ones in real utterances.

In this paper, we propose an alternative approach to represent stress by using a latent variable model, specifically Bayesian Gaussian process latent variable model (Bayesian GP-LVM) [13]. The advantage of using the latent variable model is that we can represent stress-related features in a low dimensional space in which the similarity between syllables can be easily observed. We use the latent variable as a stress-related context in Gaussian-process-regression(GPR)-based speech synthesis framework [14–17] that enable us to use continuous contextual factors as input features of a kernel function. We perform objective and subjective tests to evaluate the effectiveness of the newly added context.

2. Tone modeling by latent variable model

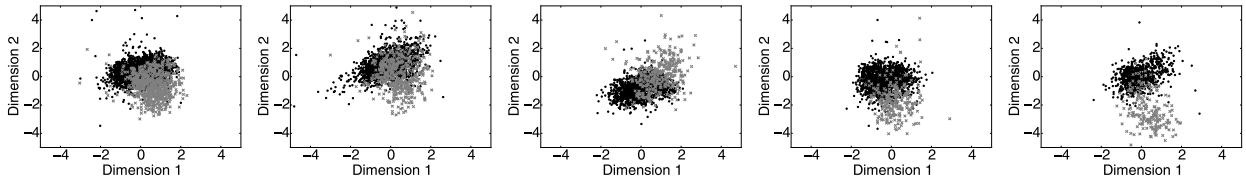
Thai language has five tones in which every syllable is pronounced with one of five tones: mid(0), low(1), falling(2), high(3), and rising(4). Typically, the tones can be distinguished by their F0 contours [18]. However, the F0 contours of tones vary in the continuous speech due to the fact that many factors affect pronunciation [19]. The studies of Thai intonation [8,9,19,20] classified syllables into stressed and unstressed ones by prosodic features where the stressed one has very similar F0 contour to the typical tone and long duration, otherwise it is the unstressed one. Although syllables can be classified into a bi-

 doi: 10.21437/SpeechProsody.2016-208



(a) Tone 0 (b) Tone 1 (c) Tone 2 (d) Tone 3 (e) Tone 4

Figure 1: Visualization of data points in latent space by keeping most two dominant dimension from separated-space projection.



(a) Tone 0 (b) Tone 1 (c) Tone 2 (d) Tone 3 (e) Tone 4

Figure 2: Visualization of data points in latent space by keeping most two dominant dimension from single-space projection.

nary class of stressed or untressed, the actual prosodic features are diverse and such simple classification is not enough to represent the diversity of stress in Thai. In this study, the key idea is to represent stress as a continuous factor by using a latent variable model, specifically, Bayesian Gaussian process latent variable model [13]. Motivation of using the latent variable model is that stress-related features can be easily understood and classified in a lower dimensional space.

2.1. Bayesian Gaussian process latent variable model

The Bayesian Gaussian process latent variable model is a dimensionality reduction technique that relies on Gaussian process latent variable model (GP-LVM) [21]. The GP-LVM is considered to be GP regression model in which inputs Z are fully unobserved, and treated as latent variables. Then, the inputs Z are recovered from the outputs Y based on maximum likelihood (ML) criterion. In Bayesian GP-LVM, the GP-LVM is trained by the variational inference framework [13] that additionally places a Gaussian prior on the latent space, $p(Z) = \prod_{i=1}^{n} \mathcal{N}(z_i|0, I)$. The marginal likelihood of data is given by

$$p(Y) = \int p(Y|Z)p(Z)dZ. \tag{1}$$

The posterior of latent variables $p(Z|Y)$ is approximated by a variational distribution $q(Z)$ given by

$$q(Z) = \prod_{n=1}^{N} \mathcal{N}(z_i|\mu_i, S_i). \tag{2}$$

Then, a variational lower bound $\mathcal{F}$ is derived as

$$\mathcal{F} \leq \log p(Y) \tag{3}$$

$$\mathcal{F} = \langle \log p(Y|Z) \rangle_{q(Z)} - KL(q(Z)\|p(Z)) \tag{4}$$

where $\langle \cdot \rangle_{q(Z)}$ is the expectation with respect to $q(Z)$. The variational parameters μ_i, and S_i are calculated by maximizing the lower bound.

2.2. Projection of Stress-related features

We propose here two approaches to project stress-related features onto a low dimensional space: *separated-space* and *single-space* projections. In the separated-space projection, syllables of respective tones are separately trained and projected onto different latent spaces. In the single-space projection, every syllable is trained all together and projected onto a single latent space. We use two most dominant factors of stress [9], duration and log F0 contour in syllable-unit as the features for the latent variable model training. For log F0 contours, we interpolate F0s in the unvoiced regions and transform them by using discrete cosine transform (DCT).

Visualization examples of data points in latent spaces by plotting the most two dominant dimensions are shown in Figures 1 and 2. In these examples, the data contains 10771 syllables: 3786, 2461, 1892, 1687, and 945 syllables of tones 0, 1, 2, 3, and 4, respectively. Stressed and unstressed syllables were labeled manually. The Bayesian GP-LVM was trained by using the squared exponential kernel and 100 inducing points. We used the same manner as [13] to choose the most dominant dimensions. The figures are plotted separately based on the tone type.

Figure 1 shows the visualization of separated-space projection. In the separated space projection, the observed features are the 0-9th DCT coefficients of log F0 contour and duration, and dimensionality of latent space is 3. It can be observed that some stressed and unstressed syllables are distributed in the same region, especially for Tones 0, 1, and 3. This is because these tones are static ones that have similar shapes of F0 contour independently of being stressed or unstressed. Tone 2 and Tone 4 are dynamic tones in which stressed syllable has high F0 movement and unstressed one has flat F0 contour, and thus they can be distinguished easily.

Figure 2 shows the visualization of single-space projection in which all data were projected onto the same space. In the single-space projection, the observed features are the 0-19th DCT coefficients of log F0 contour and duration, and dimensionality of latent space is 10. In Figure 2, it can be seen that unstressed syllables concentrate in the center of the figures because unstressed syllables of each tone have similar shapes that have flat F0 shape and short duration. Stressed syllables of each tone are located in different region, which depends on similarity of tones.

In the latent space, the similarity of syllables' stress can be determined by using the distance between latent variables. For example, in Figure 1 (a) that shows the visualization of Tone 0, the left most data point is the least stressed syllable (flat F0

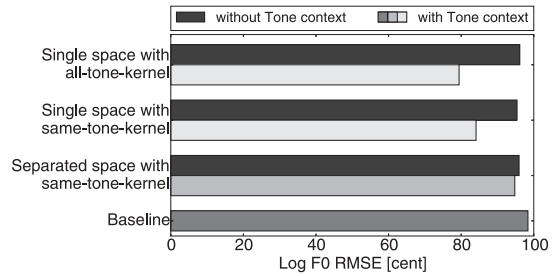


Figure 3: Log F0 distortions between original and synthetic speech.

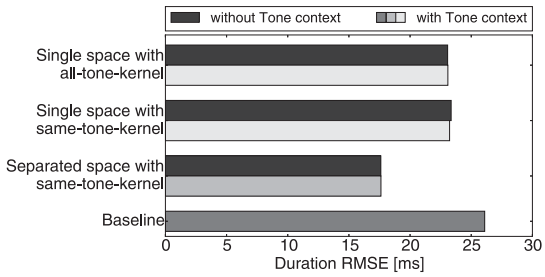


Figure 4: Phone duration distortions between original and synthetic speech.

shape and short duration) and stress intensity becomes higher and diverse on the right side of the figure.

3. GPR-Based speech synthesis

Let $\mathbf{X} = [x_1, \ldots, x_N]^T, \mathbf{y} = [y_1, \ldots, y_N]^T$, and $\mathbf{f} = [f(x_1), \ldots, f(x_N)]^T$ be the matrix representation of input and output variables, and that of latent function values of training data, respectively. In GPR, the relationship between inputs x_n and outputs y_n is given by

$$y_n = f(x_n) + \epsilon. \tag{5}$$

Let $\mathbf{X}_T, \mathbf{y}_T$, and $\mathbf{f}_T$ be denoted for the variables of test data. The joint distribution on the function values of the training and test data is given by

$$p(\mathbf{f}, \mathbf{f}_T | \mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_T \end{bmatrix}; 0, \mathbf{K}_{N+T}) \tag{6}$$

$$\mathbf{K}_{N+T} = \begin{bmatrix} \mathbf{K}_N & \mathbf{K}_{NT} \\ \mathbf{K}_{TN} & \mathbf{K}_T \end{bmatrix} \tag{7}$$

where $\mathbf{K}_N$ and $\mathbf{K}_T$ are covariance matrices of training and test frames, respectively. The joint distribution of $\mathbf{y}$ and $\mathbf{y}_T$ is given by

$$p(\mathbf{y}, \mathbf{y}_T | \mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_T \end{bmatrix}; 0, \mathbf{K}_{N+T} + \sigma^2 \mathbf{I}). \tag{8}$$

The predictive distribution of $\mathbf{y}_T$ is obtained by

$$p(\mathbf{y}_T | \mathbf{y}, \mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\mathbf{y}_T; \mu_T, \Sigma_T) \tag{9}$$

$$\mu_T = \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2 \mathbf{I}]^{-1} \mathbf{y} \tag{10}$$

$$\Sigma_T = \mathbf{K}_T + \sigma^2 \mathbf{I} - \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2 \mathbf{I}]^{-1} \mathbf{K}_{NT}. \tag{11}$$

In the frame-based feature modeling for GPR-based speech synthesis [17], the frame-based context is defined as follows:

$$\begin{aligned} x_n &= (x_{n,1}, \ldots, x_{n,K}), \qquad x_{n,k} = (\mathrm{p}_{n,k}, c_{n,k}) \\ \mathrm{p}_{n,k} &= (\mathrm{p}_{n,k}^{(-1)}, \mathrm{p}_{n,k}^{(0)}, \mathrm{p}_{n,k}^{(+1)}), \quad c_{n,k} = (c_{n,k}^{(-1)}, c_{n,k}^{(0)}, c_{n,k}^{(+1)}) \end{aligned} \tag{12}$$

where x_n is an array of partial frame context having K temporal event. $c_{n,k}$ and $\mathrm{p}_{n,k}$ are the temporal events and the relative position vectors, respectively. In Thai GPR-based speech synthesis, the temporal events are the linguistic information of phone, syllable, word, and utterance units [22]. The relative position vectors are defined individually for each unit. The superscripts (-1), (0), and $(+1)$ denote preceding, current, and succeeding of corresponding units. The kernel function for determining the similarity of frame context is defined as follows:

$$\kappa(x_m, x_n) = \sum_{k=1}^{K} \theta_{r,k}^2 \kappa_k(x_{m,k}, x_{n,k}) + \delta_{mn} \theta_{floor}^2 \tag{13}$$

$$\begin{aligned} \kappa_k(x_{m,k}, x_{n,k}) = \sum_{u=-1}^{+1} \sum_{v=-1}^{+1} [w(\mathrm{p}_{m,k}^{(u)}) w(\mathrm{p}_{n,k}^{(v)}) \\ \cdot \kappa_p(\mathrm{p}_{m,k}^{(u)}, \mathrm{p}_{n,k}^{(v)}) \kappa_{c,k}(c_{m,k}^{(u)}, c_{n,k}^{(v)})] \end{aligned} \tag{14}$$

where $w(\cdot)$, $\kappa_p(\cdot)$, and $\kappa_c(\cdot)$ are weight function, position kernel, and event feature kernel, respectively. $\theta_{r,k}^2$ and θ_{floor}^2 are kernel parameters.

In the proposed technique, we incorporate stress information into GPR-based Thai speech synthesis. The different point between the baseline [22] and the proposed techniques is whether the stress-related context is included as an additional temporal event of the syllable-unit context set. The stress-related context is represented by the latent variables described in section 2.2. The squared exponential kernel $\kappa_{c,lat}(\cdot)$ is used to calculate the distance of the stress-related context. Here, we propose two kernels for the stress-related context based on the single-space projection: *all-tone-kernel* defined by (15), and *same-tone-kernel* defined by (16). The difference between these kernels is that the *same-tone-kernel* ignores the calculation between different tones. For the stress-related context that is obtained from the separated-space projection, we calculated the distance of the stress-related context with only the same-tone-kernel:

$$\kappa_{c,lat}(c_{m,k}^{(u)}, c_{n,k}^{(v)}) = \exp\left(-\frac{(c_{m,k}^{(u)} - c_{n,k}^{(v)})^2}{l^2}\right) \tag{15}$$

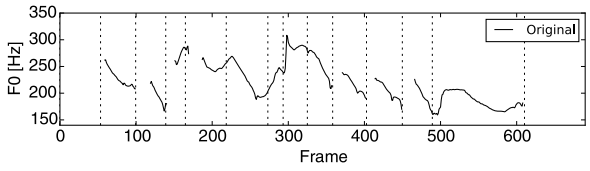
$$\kappa_{c,lat}(c_{m,k}^{(u)}, c_{n,k}^{(v)}) = \delta_{t_m,t_n} \cdot \exp\left(-\frac{(c_{m,k}^{(u)} - c_{n,k}^{(v)})^2}{l^2}\right) \tag{16}$$

where δ_{t_m,t_n} is the Kronecker delta for t_m and t_n.

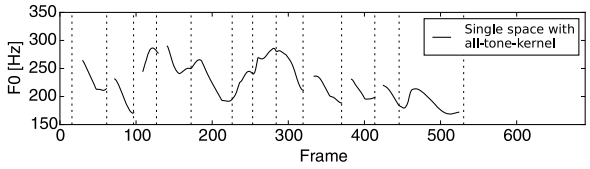
4. Evaluation

4.1. Experimental conditions

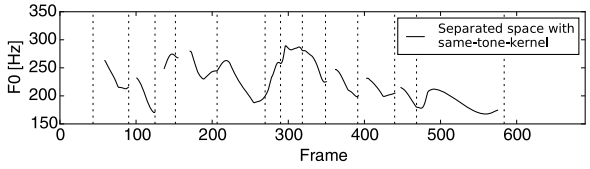
A set of phonetically balanced sentences of Thai speech database T-Sync-1 from NECTEC [23] was used for training and evaluation. The sentences were uttered by one professional female speaker with clear articulation and standard Thai accent with reading style. The training set contained 450 utterances, approximately 54 minutes in total, and 50 utterances for evaluation which were not included in the training set. Speech signals were sampled at a rate of 16kHz. Spectral features, aperiodicity, and F0 were extracted by STRAIGHT [24] with 5-ms frame shift. The acoustic feature vector consisted of the 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. We used the context set of GPR-based model described in [22] for the baseline. Each acoustic model was trained using PIC approximation [25] and EM-based optimization [16]. In the proposed technique, the vector of stress-related context was three most dominant dimensions of latent variables.



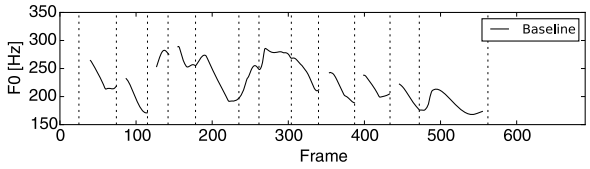
(a) Original



(b) Single space with all-tone-kernel



(c) Separated space with same-tone-kernel



(d) Baseline

Figure 5: Example of F0 contours and syllable duration compared with original. The sentence is / khw-aa-mˆ-0 th-ii-zˆ-1 th-ii-zˆ-2 ch-a-jˆ-3 ng-aa-nˆ-0 m-a-jˆ-2 d-aa-jˆ-2 m-ii-zˆ-0 ph-iia-ngˆ-0 khw-aa-mˆ-0 th-ii-zˆ-1 d-iia-wˆ-0 /, meaning "... utilizing frequency, there is not only one frequency." in English. The number suffixed to each syllable indicates its tone type.

4.2. Objective evaluation

We objectively evaluated the proposed technique in terms of log F0 and duration distortion. Since the stress-related context contains the tone information, we evaluated the proposed technique by including and not including the tone type context. Figure 3 shows the log F0 distortion. The distortion of proposed technique without tone-type context is slightly lower than the baseline. Furthermore, the proposed technique with adding the tone-type context gave significantly smaller distortions than the baseline. Figure 4 shows phone duration distortion comparison. The proposed technique achieved lower distortion than the baseline, especially when using the separated space projection. A comparison of F0 contours and syllable duration is shown in Figure 5. There are small differences between the original F0 contour and synthetic one. In syllable duration distortion, the separated-space with same-tone-kernel gave closer result to the original one than the baseline in the ending part of the utterance.

4.3. Subjective evaluation

In subjective assessment, we evaluated perceptual quality by using mean opinion score (MOS) and forced choice preference tests. The single-space projection with all-tone-kernel, the separated-space projection with same-tone-kernel, and the baseline ones were used in this experiment. Ten Thai native speakers participated the evaluation. For each participant, ten synthetic

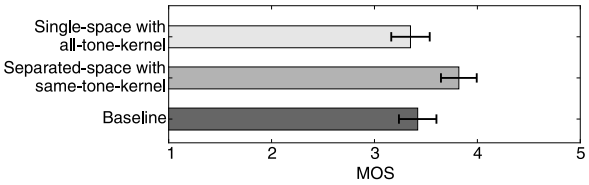


Figure 6: Result of MOS test.

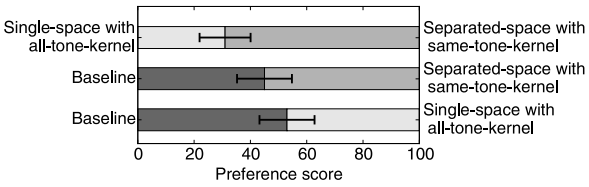


Figure 7: Result of preference test.

speech samples were randomly selected from the same test set as the objective evaluation. In MOS test, each sample was listened and evaluated in a five-point scale corresponding to their perception in naturalness of synthetic speech. The definitions of scores were 1-bad, 2-poor, 3-fair, 4-good, and 5-excellent. The participants could listen to the sample as many time as their require for ensuring in quality. Figure 6 shows the result of MOS test with 95% confidence interval. It can be seen that the separated-space outperformed other cases.

In the forced choice preference test, the participants were asked to choose more natural-sounding speech from each pair of samples. The participants could repeat playback as many time as require in the same way as MOS test. The result of the preference test is shown in Figure 7. The comparison between the baseline and the single-space projection reveals that the difference is small. Comparing the separated-space projection to the single-space projection and baseline, we can see that the separated-space one achieved higher score than the others, especially the single-space one.

From the results of objective and subjective evaluation, it can be seen that the accuracy of duration has more impact in perception than the log F0 one.

5. Conclusions

We presented a tone modeling technique using a Bayesian Gaussian process latent variable model. In continuous speech of tonal language, tone is affected by stress which causes a variety of shapes of F0 contour and duration. To represent the continuity and diversity of stress, we projected stress-related features onto a low dimensional space by using Bayesian GP-LVM. We proposed two approaches to the projection: single-space projection in which all tone are projected onto the same space, and separated-space projection in which each tone is separately projected. The latent variables were used as additional stress-related contexts for GPR-based speech synthesis. We also proposed kernel functions to determine the similarity of the stress-related context. The objective evaluation revealed that the use of stress-related context can reduce the distortion of F0 and duration. Furthermore, the subjective evaluation showed that the separated-space projection outperformed the baseline.

6. Acknowledgements

We would like to thank Dr. Vataya Chunwijitra of NECTEC, Thailand, for providing the T-Sync-1 speech database. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 15H02724.

7. References

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

[2] 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.

[3] 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.

[4] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *The Journal of the acoustical society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[5] 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.

[6] O. Krityakien, K. Hirose, and N. Minematsu, "Generation of fundamental frequency contours for Thai speech synthesis using tone nucleus model." in *Proc. INTERSPEECH*, 2013, pp. 1037–1041.

[7] J. Chaiwongsai, W. Chiracharit, K. Chamnongthai, Y. Miyanaga, and K. Higuchi, "Tone model enhancement for low complexity tone recognition," in *2013 World Congress on Sustainable Technologies (WCST)*, Dec 2013, pp. 60–65.

[8] 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.

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

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

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

[12] D. Moungsri, T. Koriyama, and T. Kobayashi, "HMM-based Thai speech synthesis using unsupervised stress context labeling," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, Dec 2014, pp. 1–4.

[13] M. K. Titsias and N. D. Lawrence, "Bayesian Gaussian process latent variable model," in *International Conference on Artificial Intelligence and Statistics*, 2010, pp. 844–851.

[14] T. Koriyama, T. Nose, and T. Kobayashi, "Frame-level acoustic modeling based on Gaussian process regression for statistical nonparametric speech synthesis," in *Proc. ICASSP*, 2013, pp. 8007–8011.

[15] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on Gaussian process regression," *IEEE J. Selected Topics in Signal Process.*, vol. 8, no. 2, pp. 173–183, 2014.

[16] T. Koriyama, T. Nose, and T. Kobayashi, "Parametric speech synthesis based on Gaussian process regression using global variance and hyperparameter optimization," in *Proc. ICASSP*, 2014, pp. 3834–3838.

[17] T. Koriyama and T. Kobayashi, "Prosody generation using frame-based Gaussian process regression and classification for statistical parametric speech synthesis," in *Proc. ICASSP*, 2015, pp. 4929–4933.

[18] C. Wutiwiwatchai and S. Furui, "Thai speech processing technology: A review," *Speech communication*, vol. 49, no. 1, pp. 8–27, 2007.

[19] S. Luksaneeyanawin, "Intonation in thai." *University of Edinburgh*, 1983.

[20] S. Hiranburana, "Changes in the pitch contours of unaccented syllables in spoken Thai," *Tai phonetics and phonology*, pp. 23–27, 1972.

[21] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *The Journal of Machine Learning Research*, vol. 6, pp. 1783–1816, 2005.

[22] D. Moungsri, T. Koriyama, and T. Kobayashi, "Duration prediction using multi-level model for GPR-based speech synthesis," in *Proc. INTERSPEECH*, 2015, pp. 1591–1595.

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

[24] 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.

[25] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical nonparametric speech synthesis using sparse Gaussian processes," in *Proc. INTERSPEECH*, 2013, pp. 1072–1076.