INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Unsupervised stress information labeling using Gaussian process latent variable model for statistical speech synthesis

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

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

moungsri.d.aa@m.titech.ac.jp, {koriyama,takao.kobayashi}@ip.titech.ac.jp

Abstract

In Thai language, stress is an important prosodic feature that not only affects naturalness but also has a crucial role in meaning of phrase-level utterance. It is seen that a speech synthesis model that is trained with lack of stress and phrase-level information causes incorrect tones and ambiguity in meaning of synthetic speech. Our previous work has shown that manually annotated stress information improves naturalness of synthetic speech. However, a high time consumption is a drawback of the manual annotation. In this paper, we utilize an unsupervised learning technique called Bayesian Gaussian process latent variable model (Bayesian GP-LVM) to automatically put stress annotation on the given training data. Stress related features are projected onto a latent space in which syllables are easier classified into stressed/unstressed classes. We use the stressed/unstressed information as an additional context in GPR-based speech synthesis. Experimental results show that the proposed technique improves naturalness of synthetic speech as well as accuracy of stressed/unstressed classification. Moreover, the proposed technique enables us to avoid ambiguity in meaning of synthetic speech by providing intended stress position into context label sequence to be synthesized.

Index Terms: GPR-based speech synthesis, stress, Thai language, prosody, latent variable model, Bayesian GP-LVM

1. Introduction

In speech synthesis, a main goal is to generate speech which is natural-sounding and also clearly has the intended meaning. Prosody is an important feature that has a great influence on naturalness and meaning of speech. To generate natural-sounding speech, various techniques have been used to model prosodic features. In tonal languages, tone is a major factor used for distinguishing lexical or grammatical meaning of speech. Since Thai is a tonal language and tone is very sensitive in perception, a tone-separated tree structure was proposed to remove tone-dependency on the context in tree-based context clustering for HMM-based speech synthesis [1]. Moreover, pitch contour varies diversely in continuous speech, and thus only tone-type context is not sufficient in F0 modeling. To model diversity of F0 contour in each tone, tone geometrical features that represent the shape of F0 contour were proposed in F0 generation for speech synthesis [2]. Another technique for modeling prosodic feature in Thai is a modified version of Tilt model called T-Tilt [3,4] which was successfully used for representing prosody in accentual languages. Since co-articulation affects F0 contour shape but tone nuclei are less affected by adjacent syllables, a tone nucleus model was used in F0 modeling and generation [5]. Furthermore due to the fact that vowel part of a syllable receives small effect from neighboring syllables and contains the main prosodic feature of syllable, an F0 modeling using only vowel part instead of entire syllable was proposed to reduce complexity and improve accuracy in tone recognition [6].

In addition to tone, stress is another important factor in Thai language which affects naturalness and meaning of sentence. The use of stress information can improve accuracy of tone recognition [7]. In our previous work, we showed that manually annotated stress information can reduce F0 and duration distortions in the HMM-based speech synthesis [8]. To alleviate the problem of high cost of manual labeling, we also proposed an unsupervised labeling technique for classifying syllables into stress-related classes based on F0 movement and syllable duration [9]. However, problems remain; some tones have low F0 movement in both stressed and unstressed cases, and error in F0 extraction may cause a high F0 variance in a syllable.

In this paper, to overcome the problems, we propose a new unsupervised labeling technique for stress annotation. We utilize a dimensionality reduction technique, called Bayesian Gaussian process latent variable model (Bayesian GP-LVM), to project prosodic features onto latent space in which similarity of prosodic features can be easily measured by using distance between latent variables. In our previous work, the latent variables of Bayesian GP-LVM were directly used as additional context [10]. In contrast, the proposed technique clusters the latent variables into simple stressed/unstressed classes and uses the obtained class information as the context. This enables us to give intended stress position into label sequence. We examine stressed/unstressed classification performance to evaluate the effectiveness of the use of latent variables. Then we use Gaussian process regression (GPR)-based speech synthesis [11], which can generate more natural-sounding speech than the HMM-based one [12], and assess the performance of newly added context though objective and subjective tests.

2. Unsupervised stress information labeling

2.1. Stress in Thai

Stress is a major factor in diversity of prosodic features in syllable-unit [13]. It affects not only naturalness but also meaning of speech. As described in [14], position of stressed syllable has an influence in meaning of phrase. Generally, the position of stressed syllable is unknown and cannot be obtained from text. However, various studies of stress agree that stressed syllables are usually isolated syllable, syllable at the end of phrase, and emphasized syllable or word [13–16]. In terms of acoustic characteristics, it is known that stressed syllables have F0 contours similar to typical F0 contours and long durations, whereas unstressed syllables are otherwise [16]. Additionally, durations

http://dx.doi.org/10.21437/Interspeech.2016-273

of stressed syllables also depend on final consonant [17].

2.2. Bayesian Gaussian process latent variable model

In this paper, we automatically give stress annotation in an unsupervised way. For this purpose, first, we use a dimensionality reduction technique, Bayesian GP-LVM [18], to reduce complexity of stress-related features, specifically, F0 contour and duration. Bayesian GP-LVM is robust to overfitting and we can determine most dominant dimensions of the nonlinear latent space. Secondly, we employ an unsupervised clustering on the latent variables obtained from Bayesian GP-LVM training.

In Bayesian GP-LVM, the output variables Y are observed, and the input variables Z are fully unobserved and treated as latent variables. In the model training, we used the stress-related features as the observed variables Y. In Bayesian GP-LVM training, the marginal likelihood of data is given by

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

Then a variational distribution $q(Z)$ is introduced to approximate the posterior of latent variables $p(Z|Y)$ as follows:

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

where μ_i, and S_i are mean and covariance. 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 obtained by maximizing the lower bound.

To perform stressed/unstressed clustering, we use the means μ_i of variational distribution as features in an unsupervised learning. The stressed/unstressed classes obtained from the clustering are used as an additional context in statistical parametric speech synthesis based on GPR.

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 the 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 GPR-based speech synthesis [19], frame-level context is used as input variables:

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

where x_n is an array of partial frame context having K temporal events. $c_{n,k}$ 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 [20]. 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 similarity between input variables is determined by the kernel function $\kappa(x_m, x_n)$ as follow:

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

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

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.

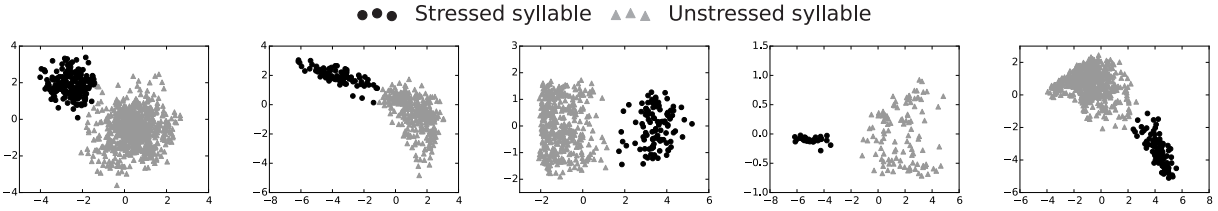
In this paper, the stress information is used as an additional context of a temporal event in a syllable unit. The stress context is represented by binary values: 1 for stressed syllable and 0 for unstressed syllable. The similarity of stress context is calculated by a linear kernel.

4. Evaluation

We first performed an unsupervised learning by using the latent variables to give stress information into context set for speech synthesis. We then performed experiments to measure the improvement by considering stress information in the GPR-based speech synthesis. A set of phonetically balanced sentences of Thai speech database T-Sync-1 from NECTEC [21] was used for training and evaluation. The sentences were uttered with reading style by one professional female speaker with clear articulation and standard Thai accent. Speech signals were sampled at a rate of 16kHz. We used STRAIGHT [22] to extract spectral features, aperiodicity, and F0 with 5-ms frame shift.

4.1. Stressed/unstressed annotation

The training set contained 329 utterances, 10741 syllables in total. The numbers of stressed and unstressed syllables were 1491 and 9250, respectively. We performed Bayesian GP-LVM training by using stress-related features, log F0 contour and duration in syllable-unit, as observed variables. In this paper, we omitted prosodic features in initial consonant part because they did



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

Figure 1: Visualization of stress-related features in latent space by keeping most two dominant dimensions from the projections of syllables with nasal final consonant.

Table 1: Accuracy of stressed/unstressed syllable classification with observed variables. Values represent F1-scores.

Positive class	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4	All
Unstressed	0.975	0.934	0.987	0.943	0.936	0.96
Stressed	0.808	0.542	0.93	0.401	0.641	0.708

Table 2: Accuracy of stressed/unstressed syllable classification with latent variables. Values represent F1-scores.

Positive class	Tone 0	Tone 1	Tone 2	Tone 3	Tone 4	All
Unstressed	0.983	0.972	0.99	0.98	0.989	0.983
Stressed	0.87	0.792	0.982	0.818	0.94	0.88

not have significant differences between stressed and unstressed syllables. We interpolated log F0 contour in the unvoiced region by using a third-order polynomial. Since durations of respective syllables are not equal, the log F0 contour was normalized into 50 samples and its delta and delta-delta were also included in the observed variables. As a result, the observed variables had 150 dimensions of log F0 contour information and 1 dimension of syllable duration. The dimensionality of latent variables was set to 10. Bayesian GP-LVM was trained by using squared exponential kernel as a covariance function and 100 inducing point. The model optimization was conducted by using scaled conjugate gradient method. Since the characteristics of stress depend on tone-type and final consonant type of syllable, we separately trained Bayesian GP-LVMs based on tone-type and final consonant of syllable. We grouped syllables based on their final consonant into three types: non final consonant syllable, nasal final consonant syllable, and non-nasal final consonant syllable. Figure 1 shows the visualization of the observed variables in latent space by projecting syllables that have nasal final consonant.

To evaluate the effectiveness of the latent variable model, we measured accuracy in stressed/unstressed classification. We performed density based clustering using DBSCAN algorithm [23] to cluster the latent variables. Then each cluster was identified to be stressed or unstressed class by observing distribution of a small set of labeled training data in the latent space. We also compared stressed/unstressed classification performance with that obtained by using observed variables. The accuracy was calculated by evaluating consistency with the manual stress annotation. The results are shown in Tables 1 and 2. It can be seen that the use of the latent variables provides higher F1-scores than the observed variables. The accuracy of tone 2 and 4 is higher than other tones because these tones are dynamic ones whose characteristics of stressed syllable are much different from the unstressed ones. The differences between stressed and unstressed static tones, i.e., tones 0, 1 and 3, are not so large as the dynamic ones.

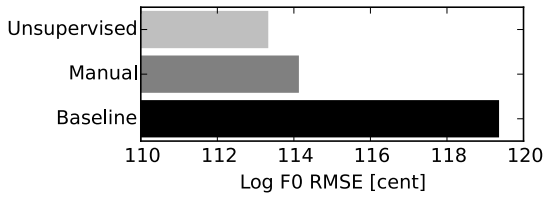


Figure 2: Log F0 distortions between original and synthetic speech.

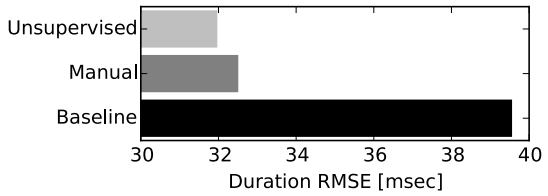


Figure 3: Duration distortions between original and synthetic speech.

4.2. Experimental conditions for objective/subjective evaluation

To clarify the effectiveness of the stress information in speech synthesis, we evaluated synthetic speech with and without using stress information context. The stress information is obtained from the manual labeling and the unsupervised labeling described in 4.1.

The training set contained 329 utterances, approximately 50 minutes in total, and 40 utterances were used for evaluation which were not included in the training set. The acoustic feature vector consisted of the 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. The acoustic models were trained by using PIC approximation [24] and EM-based optimization [25]. We used the context set of GPR-based model described in [20] for the baseline context set. In the proposed technique, we included stress information as an additional context in the context set. The manual stress labeling was the same set as we used in our previous work [8]. In the test set, we manually gave stress information.

4.3. Objective evaluation

We used log F0 and duration distortions between synthetic and original speech samples for objective evaluation. The result of log F0 distortion is shown in Figure 2. In the figure, "Baseline" represents the result without using stress information, "Manual" and "Unsupervised" represent the results with stress information by manual labeling and automatic labeling using the proposed technique, respectively. It is seen that the stress context gives smaller log F0 distortion than the baseline. It is noted that

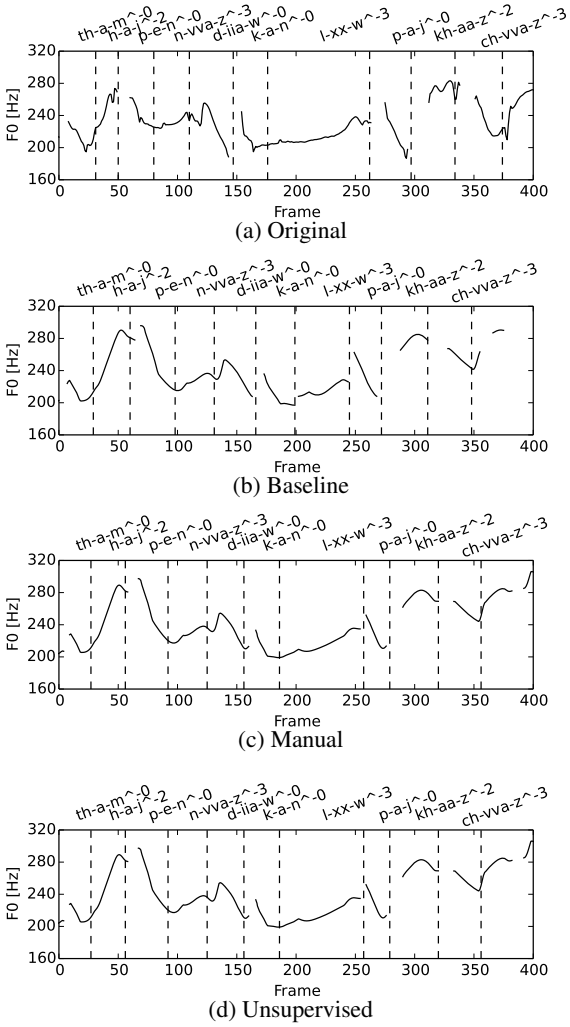


Figure 4: Example of F0 contours and syllable duration compared with original. The sentence means "... bring mixed milk into sterilization ... ". The number suffixed to each syllable indicates its tone type.

there is only small difference between manual and unsupervised labeling cases. Figure 3 shows the duration distortions. It shows the similar result to that of log F0 distortions that the stress context can reduce distortion and there is no significant difference between the manual and unsupervised labeling.

Figure 4 shows an example of F0 contours and syllable durations of original, baseline, manual and unsupervised labeling. It can be seen that the results for manual and unsupervised labeling are closer to the original than the baseline, especially at the seventh syllable (l-xx-w^3). In this example, the baseline produced ambiguous meaning because the incorrectness of the seventh syllable affects the meaning of the sentence. By giving the stress context onto the seventh syllable, we can synthesize speech having unique meaning.

4.4. Subjective evaluation

We conduced the subjective evaluation by mean opinion score (MOS) and forced choice preference tests. Ten Thai native speakers participated the evaluation. Ten synthetic speech samples were randomly selected from the test set of the objective evaluation. Speech samples were evaluated in a five-point scale corresponding to 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 they require for ensuring in quality. Figure 5 shows the result of MOS test with 95% confidence interval. Synthetic speech with stress context received higher score than the baseline. Moreover, there is no significant difference between the manual and unsupervised labeling.

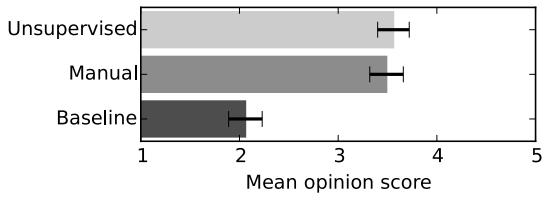


Figure 5: Result of mean opinion score test.

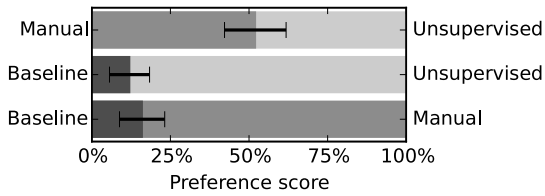


Figure 6: Result of forced choice preference test.

In the forced choice preference test, the participants were asked to choose more natural-sounding and clear meaning of speech from each pair of samples. The participants could repeat playback in the same way as MOS test. Figure 6 shows the result of the preference test. It can be seen that the proposed technique can achieve higher score than the baseline. Additionally, there is no significant difference between manual and unsupervised labeling.

5. Conclusion

In this paper, we presented an unsupervised labeling technique for giving stress information into context set for statistical speech synthesis. Stress is an important factor which affects naturalness and meaning of utterance. We utilized an unsupervised learning technique called Bayesian Gaussian process latent variable model to obtain stress information automatically. We used the prosodic features of stress, log F0 contour and syllable duration, as observed variables of Bayesian GP-LVM. Then we conducted a latent variable model training to project the observed variable into latent space in which we can easily classify syllables into stressed and unstressed ones. The objective and subjective results showed that the proposed technique is comparable to the manual labeling and also outperforms the baseline. In future work, we will utilize the latent variable model into longer speech unit than syllable unit in which the prosodic feature is affected by various factors.

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 KAKENHI Grant Number 15H02724.

7. References

[1] S. Chomphan and T. Kobayashi, "Design of tree-based context clustering for an HMM-based Thai speech synthesis system," in *Proc. SSW*, 2007, pp. 160–165.

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

[5] 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.

[6] J. Chaiwongsai and Y. Miyanaga, "Improved tone model for low complexity tone recognition," in *Proc. SICE*, 2014, pp. 1124–1129.

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

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

[9] D. Moungsri, T. Koriyama, and T. Kobayashi, "HMM-based Thai speech synthesis using unsupervised stress context labeling," in *Proc. APSIPA ASC*, 2014, http://www.apsipa.org/proceedings.htm.

[10] D. Moungsri, T. Koriyama, and T. Kobayashi, "Tone modeling using Gaussian process latent variable model for statistical speech synthesis," in *Proc. Speech Prosody 8*, 2016, pp. 1014–1018, http://www.isca-speech.org/archive/sp2016/.

[11] 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.

[12] T. Koriyama and T. Kobayashi, "A comparison of speech synthesis systems based on GPR, HMM, and DNN with a small amount of training data," in *Proc. INTERSPEECH*, 2015, pp. 3496–3500.

[13] 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.

[14] S. Luksaneeyanawin, "Intonation in Thai," *University of Edinburgh*, 1983.

[15] A. S. Abramson, "Lexical tone and sentence prosody in Thai," 1979, pp. 380–387.

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

[17] S. Potisuk, J. Gandour, and M. P. Harper, "Vowel length and stress in Thai," *Acta linguistica hafniensia*, vol. 30, no. 1, pp. 39–62, 1998.

[18] M. K. Titsias and N. D. Lawrence, "Bayesian Gaussian process latent variable model," in *Proc. AISTATS*, 2010, pp. 844–851.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996, pp. 226–231.

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

[25] 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.