DURATION PREDICTION USING MULTIPLE GAUSSIAN PROCESS EXPERTS FOR GPR-BASED 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

This paper proposes an alternative multi-level approach to duration prediction for improving prosody generation in statistical parametric speech synthesis using multiple Gaussian process experts. We use two duration models at different levels, specifically, syllable and phone. First, we individually train syllable- and phone-level duration models. Then, the predictive distributions of syllable and phone duration models are combined by product of Gaussians. The means of combined predictive distributions are used as predicted durations for synthetic speech. We show objective and subjective evaluation results for the proposed technique by comparing with the conventional ones when the techniques are applied to Gaussian process regression (GPR)-based speech synthesis.

Index Terms— Multi-level model, Duration prediction, GPR-based speech synthesis, Product of Guassians, Multiple Gaussian process experts

1. INTRODUCTION

Gaussian process regression (GPR)-based speech synthesis [1] has been successfully developed to overcome the limitations of hidden Markov model (HMM)-based speech synthesis [2]. In the GPR-based technique, frame-level acoustic features and linguistic information are defined as output and input variables of a Gaussian process regression, respectively. Speech parameters are generated by means of inference from new given input variables. The main goal of speech synthesis is to generate natural sounding and intelligible speech. Duration is one of the most important prosodic features which affects naturalness and meaning of synthetic speech. Single phone duration modeling has been successfully applied to duration prediction of given text in the GPR-based framework [3]. However, predicted durations are not perfect because a single phone-level model is insufficient to capture prosodic features in longer units. For example, in Thai language case, stress in the syllable layer is a crucial factor that affects tone contour, syllable duration, and sentence structure [4, 5].

To incorporate the characteristics of multiple layers into prosody generation, various techniques have been proposed to combine multiple models of different layers. In [6], longer unit models were integrated with a state-level model in speech parameter generation by maximizing joint probability. Speaking rate-dependent hierarchical prosodic model (SP-HPM) [7] utilized a hierarchical structure including prosodic-acoustic features, linguistic information, and prosody structure for speaking rate modeling. In [8], a product of experts framework was proposed, which jointly trains multiple acoustic models for speech synthesis. In our previous work [9], we proposed two-stage duration modeling which utilized a syllable-level model for predicting syllable durations and using the result as an additional context for a phone-level model in GPR-based duration prediction. Although the two-stage model has shown significant improvement in duration prediction accuracy, it is still imperfect since the syllable-level model has not been used explicitly in generating duration for speech synthesis.

In this paper, we propose an alternative technique for duration prediction using multiple Gaussian process (GP) experts in the GPR-based speech synthesis. First, we individually train phone and syllable duration models. In duration prediction, we express syllable duration by the sum of phone durations. Then, the predictive distributions of phone and syllable models are combined by product of Gaussians. The predicted duration can be obtained by calculating model parameters of the combined model. We show performance evaluation results of the proposed technique by objective and subjective tests.

2. GPR-BASED SPEECH SYNTHESIS

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]^\intercal$, and $\mathbf{y} = [y_1, y_2, ..., y_N]^\intercal$ be the matrix forms of frame-level contexts, and acoustic features of training data, respectively. The frame-level[1] context $\mathbf{x}_i$ contains temporal events $x_{i,k}$ as follows:

$$\begin{aligned} \mathbf{x}_i &= (x_{i,1}, x_{i,2}, \ldots, x_{i,K}) \\ x_{i,k} &= (\mathrm{p}_{i,k}, c_{i,k}) \end{aligned} \tag{1}$$

where each temporal event $x_{i,k}$ is composed of linguistic information $c_{i,k}$ and relative position $\mathrm{p}_{i,k}$ in speech units. In GPR-based speech synthesis, $\mathbf{y}$ is assumed to be sampled from a Gaussian process that can be expressed by

$$\mathbf{y} \sim \mathcal{N}(0, \mathbf{K}_N + \sigma^2 \mathbf{I}) \tag{2}$$

where $\mathbf{K}_N$ is a covariance matrix of training data. Let $\mathbf{X}_T$ and $\mathbf{y}_T$ be matrix forms for test data. Then, the predictive distribution of $\mathbf{y}_T$ is given by

$$p(\mathbf{y}_T|\mathbf{y}, \mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\mathbf{y}_T; \boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T) \tag{3}$$

$$\boldsymbol{\mu}_T = \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2 \mathbf{I}]^{-1}\mathbf{y} \tag{4}$$

$$\boldsymbol{\Sigma}_T = \mathbf{K}_T + \sigma^2 \mathbf{I} - \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2 \mathbf{I}]^{-1}\mathbf{K}_{NT} \tag{5}$$

where $\mathbf{K}_T$ and $\mathbf{K}_{NT}$ are covariance matrices of test data and between training and test data. In the covariance matrices, the correlation between $\mathbf{x}_m$ and $\mathbf{x}_n$ can be obtained by the kernel function $\kappa(\mathbf{x}_m, \mathbf{x}_n)$

[1]For duration model, phone- or syllable-level context is used.

as follows:

$$\kappa(\mathbf{x}_m, \mathbf{x}_n) = \sum_{k=1}^{K} \theta_k^2 \kappa_k(x_{m,k}, x_{n,k}) + \delta_{mn}\theta_{floor}^2 \quad (6)$$

where θ_k^2 and θ_{floor}^2 are kernel parameters. The function $\kappa_k(\cdot)$ is defined for the k-th temporal event.

When synthesizing speech, we generate a speech parameter sequence using the predictive distribution [3, 10]. Since GPR is a nonparametric model, the predictive mean is obtained directly from acoustic features of training data $\mathbf{y}$, which would result in natural sounding speech parameter sequences.

3. DURATION PREDICTION BY MULTIPLE GP EXPERTS

Thai syllable has four components, initial consonant, vowel, final-consonant, and tone [11]. Final consonant can be absent in some syllables. In studies of Thai language, prosody is often described in syllable unit where a position of stressed/unstressed syllable affects perception in sentence structure [4, 12]. Stress in Thai has two main acoustic features, F0 contour and duration, where the duration is the most dominant [5]. Furthermore, it is shown that whether a syllable is stressed/unstressed influences the durations of vowel and final consonant in that syllable [13]. From this viewpoint, we propose an alternative multi-level model for duration prediction by using syllable- and phone-level models.

In the proposed technique, we use multiple Gaussian process experts in a similar way as [14] which maximize the likelihood of product of multiple predictive distributions: syllable duration $p(\mathbf{d}_T^s|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s)$ and phone duration $p(\mathbf{d}_T^p|\mathbf{d}^p, \mathbf{X}^p, \mathbf{X}_T^p)$ where $\mathbf{X}^s$ and $\mathbf{X}^p$ are input variables of syllable and phone duration models, respectively. Matrix forms of syllable durations $\mathbf{d}^s$ and phone durations $\mathbf{d}^p$ are output variables of syllable and phone duration models, respectively. The product of distribution is expressed as

$$p(\mathbf{d}_T^p|\mathbf{d}^s, \mathbf{d}^p, \mathbf{X}, \mathbf{X}_T) = \frac{1}{Z} p(\mathbf{d}_T^s|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s) \cdot p(\mathbf{d}_T^p|\mathbf{d}^p, \mathbf{X}^p, \mathbf{X}_T^p) \quad (7)$$

$$\mathbf{d}_T^s = [d_1^s, d_2^s, ..., d_n^s]^\intercal \quad (8)$$

$$\mathbf{d}_T^p = [d_{1,1}^p, d_{1,2}^p, d_{2,1}^p, ..., d_{n,m(n)}^s]^\intercal \quad (9)$$

where $\mathbf{d}_T^s$ and $\mathbf{d}_T^p$ are matrix forms of syllable and phone durations of test data, respectively, and Z is a normalization term. Syllable duration d_i^s is determined by the sum of phone durations $d_{i,j}^p$ within the syllable as follows:

$$d_i^s = \sum_{j=1}^{m(i)} d_{i,j}^p \quad (10)$$

where $m(i)$ is the number of phones in i-th syllable, whose value is 2 or 3. Then, the relationship between syllable duration and phone duration can be written in a matrix form using a transformation matrix $\mathbf{W}$ as follows:

$$\mathbf{d}_T^s = \mathbf{W}\mathbf{d}_T^p. \quad (11)$$

For example, suppose that a sentence has 3 syllables and respective syllables have 3, 2, and 3 phones. Then, the equation is expressed as

$$\begin{matrix} \mathbf{d}_T^s \\ \\ \begin{bmatrix} d_1^s \\ d_2^s \\ d_3^s \end{bmatrix} \end{matrix} = \begin{matrix} \mathbf{W} \\ \\ \left[\begin{array}{ccc|cc|ccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}\right] \end{matrix} \begin{matrix} \mathbf{d}_T^p \\ \\ \left[\begin{array}{c} d_{1,1}^p \\ d_{1,2}^p \\ d_{1,3}^p \\ \hline d_{2,1}^p \\ d_{2,2}^p \\ \hline d_{3,1}^p \\ d_{3,2}^p \\ d_{3,3}^p \end{array}\right] \end{matrix}. \quad (12)$$

Since the predictive distribution of syllable duration is Gaussian, it can be reformulated in terms of phone duration in the same way as the formulation of trajectory HMM framework [15] as follows:

$$p(\mathbf{d}_T^s|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s) = \mathcal{N}(\mathbf{W}\mathbf{d}_T^p; \boldsymbol{\mu}_T^s, \boldsymbol{\Sigma}_T^s) \quad (13)$$

$$p(\mathbf{d}_T^p|\mathbf{d}^s, \mathbf{X}^s, \mathbf{X}_T^s) = \mathcal{N}(\mathbf{d}_T^p; \mathbf{P}\mathbf{r}, \mathbf{P}) \quad (14)$$

$$\mathbf{P} = (\mathbf{W}^\intercal {\boldsymbol{\Sigma}_T^s}^{-1}\mathbf{W})^{-1} \quad (15)$$

$$\mathbf{r} = \mathbf{W}^\intercal {\boldsymbol{\Sigma}_T^s}^{-1}\boldsymbol{\mu}_T^s. \quad (16)$$

Since both predictive distributions are Gaussian, Eq. (7) can be rewritten by Gaussian as follows:

$$\begin{aligned} p(\mathbf{d}_T^p|\mathbf{d}^s, \mathbf{d}^p, \mathbf{X}, \mathbf{X}_T) &= \frac{1}{Z'}\mathcal{N}(\mathbf{d}_T^p; \mathbf{P}\mathbf{r}, \mathbf{P}) \cdot \mathcal{N}^p(\mathbf{d}_T^p; \boldsymbol{\mu}_T^p, \boldsymbol{\Sigma}_T^p) \\ &= \mathcal{N}(\mathbf{d}_T^p; \boldsymbol{\mu}_D, \boldsymbol{\Sigma}_D). \end{aligned} \quad (17)$$

Then, the mean and covariance of the predictive distribution are given by

$$\boldsymbol{\mu}_D = \boldsymbol{\Sigma}_D(\mathbf{r} + {\boldsymbol{\Sigma}_T^p}^{-1}\boldsymbol{\mu}_T^p) \quad (18)$$

$$\boldsymbol{\Sigma}_D^{-1} = \mathbf{P}^{-1} + {\boldsymbol{\Sigma}_T^p}^{-1}. \quad (19)$$

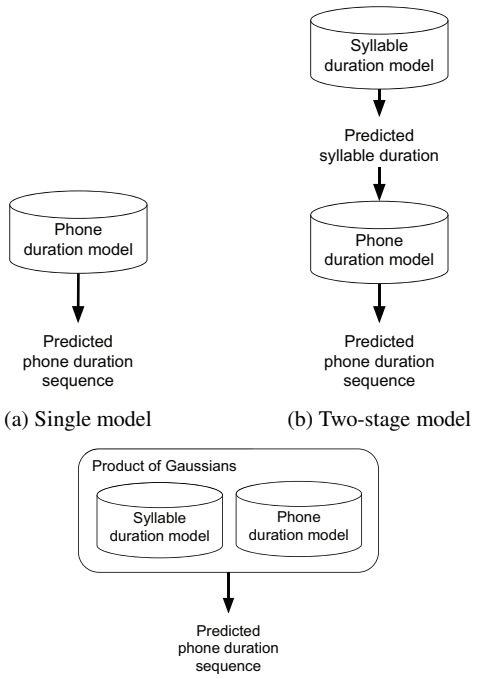
Finally, we use the mean $\boldsymbol{\mu}_D$ as the synthetic phone duration sequence.

4. EXPERIMENTS

We conducted experiments to evaluate the performance of the proposed technique. In the experiments, we compared three techniques: *single model*, multi-level model by *two-stage* prediction, and *multiple Gaussian process (GP) experts* model. Figure 1 summarizes these prediction approaches. The *single model* is the conventional GPR-based approach [3] that uses a single phone model for duration prediction. The *two-stage model* is our previous approach that was proposed in [9]. In the two-stage model, we used syllable duration as an additional context for phone duration prediction. The syllable duration context for test data can be predicted by the syllable duration model. The proposed technique, multiple GP experts model, combines phone and syllable duration models for phone duration prediction as described in Section 3.

4.1. Experimental condition

The speech database used in the experiments was a set of phonetically balanced sentences of Thai speech database, T-Sync-1 developed by NECTEC ([16]). The sentences were uttered by one professional female speaker with clear articulation in the reading style of standard Thai accent. We used 450 utterances and 50 utterances for training and evaluation, respectively. The training data contained



(a) Single model (b) Two-stage model

(c) Multiple Gaussian process experts (Proposed technique)

Fig. 1. Comparison of prediction models.

13733 syllables. The test set for evaluation was not included in the training data.

We used the conventional phone-level context of GPR-based Thai speech synthesis [9] for training and prediction. The phone-based contextual factors are summarized in Table 1. The context is composed of linguistic information in phone, syllable, word, and utterance layers including their relative position in different scale units. The phone-level context was used for training and prediction in each phone-level model of all techniques. For syllable duration model, we used syllable-level context which is composed of linguistic information in syllable, word, and utterance layers. In the syllable layer, the context includes phonetic features of phones in syllable and tone. The syllable-based contextual factors are summarized in Table 2. We used the kernel function described in [9] to calculate the distance of each temporal event.

We used speech signals sampled at a rate of 16kHz. Spectral features, aperiodicity, and F0 were extracted by STRAIGHT [17] 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. In GP model training, we employed partially independent conditional (PIC) approximation [18], and the kernel function parameters were optimized by EM-based method [19].

4.2. Objective evaluation results

In the objective evaluation, we measured duration distortion between synthetic speech and the original one in phone and syllable units. The RMS errors of phone and syllable durations are shown in Figs. 2 and 3, respectively. The two-stage model and proposed technique had smaller distortion than the single model. In phone duration distortion, the two-stage model had lower RMS error than the proposed technique. However, the proposed technique achieved lower RMS error than the two-stage model in syllable duration distortion.

Figure 4 shows an example of syllable duration errors in a test sentence where each bar represents the difference between the predicted syllable duration and the original. It can be seen that the proposed technique provided smaller errors than the other techniques in almost all syllables.

Table 1. Phone-level context based on temporal events for Thai GPR-based speech synthesis.

Unit:	phone
Type:	beginning of each phonetic feature
Scale:	phone-normalized
Unit:	syllable
Type:	{beginning, end} of tone type
	{beginning, end} of syllable duration[a]
Scale:	{syllable, word}-normalized
Unit:	word
Type:	{beginning, end} of part of speech
Scale:	{syllable, word}-normalized
Unit:	utterance
Type:	{beginning, end} of utterance
Scale:	{syllable, word, utterance}-normalized

[a]The temporal context is used only in phone-level duration model of the *two-stage* approach

Table 2. Syllable-level context based on temporal events for Thai GPR-based speech synthesis.

Unit:	syllable
Type:	beginning of each initial-consonant's phonetic feature
	beginning of each vowel's phonetic feature
	beginning of each final-consonant's phonetic feature
	beginning of tone type
Scale:	{syllable, word}-normalized
Unit:	word
Type:	{beginning, end} of part of speech
Scale:	{syllable, word}-normalized
Unit:	utterance
Type:	{beginning, end} of utterance
Scale:	{syllable, word, utterance}-normalized

4.3. Subjective evaluation results

We conducted MOS and forced choice preference tests to evaluate the perceptual quality in the naturalness of predicted duration. Participants were ten Thai native speakers. Each person evaluated ten speech samples that are randomly selected from 50 test samples. In the MOS test, the participants evaluated each sample on a five-point scale from 1 to 5 according to their satisfaction in the naturalness of syllable and phone duration. The definition of the rating was 1:bad, 2:poor, 3:fair, 4:good, and 5:excellent. Participants could repeat playback as many times as they required for evaluation. Figure 5 shows the resultant scores with 95% confidence intervals. It is shown that the proposed technique achieved higher score than the single model. Moreover, the two-stage model got slightly higher score than the proposed technique, but the difference is statistically insignificant.

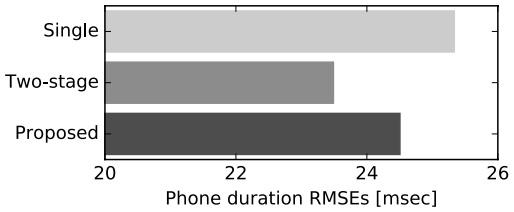


Fig. 2. Phone duration distortion

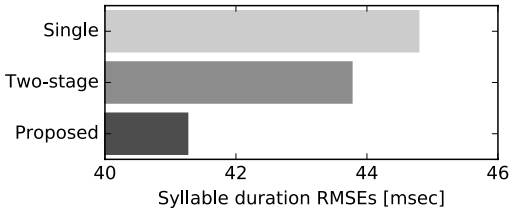


Fig. 3. Syllable duration distortion

In the forced choice preference test, the participants were asked to choose more natural one in terms of phone and syllable durations for each pair of speech samples. The participants could repeat playback as many times as they required in the same way as the MOS test. Figure 6 shows the result of forced choice preference test. It is seen that the participants preferred the proposed technique than the single-level model. When comparing the proposed technique to the two-stage model, we see that the two-stage received more preference even if the proposed technique gave lower syllable duration distortion. One reason might be that the perception of stress intensity is highly dependent on the durations of vowel and final-consonant than that of an entire syllable. This means that even though the duration of initial consonant is very long, the participants may not perceive it as stressed syllable if vowel and final-consonant durations are short. Therefore, the accuracy of phone durations is more significant in the perception of naturalness than syllable durations.

5. CONCLUSION

We have proposed an alternative technique of multi-level model for GPR-based duration prediction. In the proposed technique, we firstly train phone and syllable duration models independently. In duration prediction, we explicitly express syllable duration as the sum of phone durations. Then, the predictive distributions of syllable and phone durations are combined by product of Gaussians. The objective evaluation results showed that the proposed technique gave smaller distortion than the two-stage and single model techniques in syllable duration distortion. The subjective evaluation results showed that the proposed technique is comparable with the two-stage model. In future work, we will conduct experiments with a larger number of syllables since Thai syllables are quite complex and the current amount of syllable data might be insufficient.

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 JP15H02724.

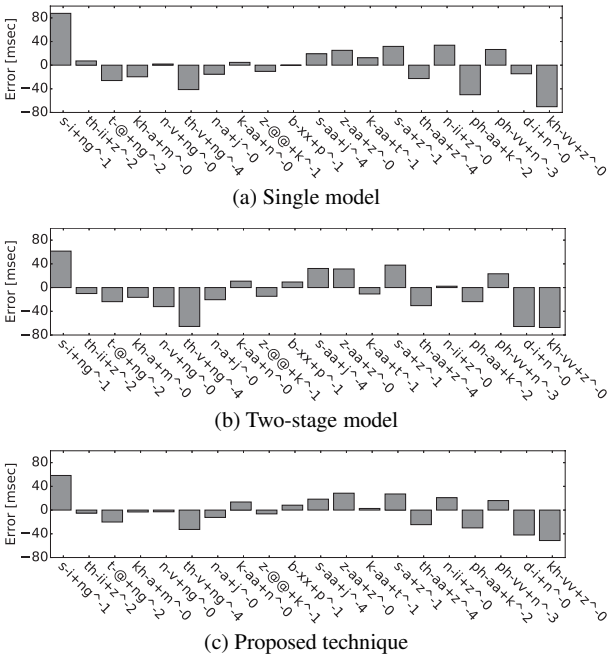


Fig. 4. Comparison of duration prediction errors in syllable unit. The sentence is "... the points of concern in design of antenna at ground station is ..." in English.

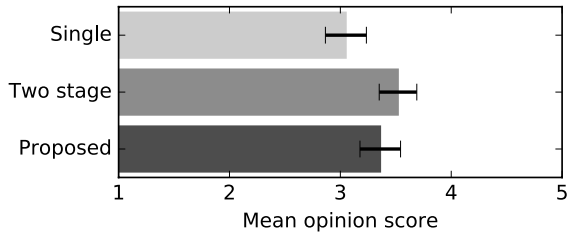


Fig. 5. Result of MOS test in subjective evaluation of naturalness.

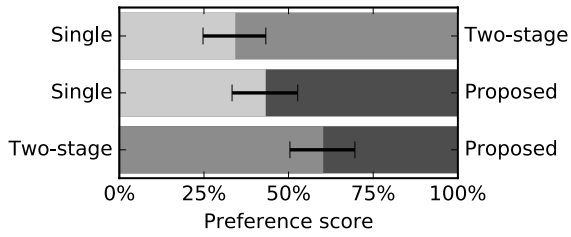


Fig. 6. Result of forced choice preference test in subjective evaluation of naturalness.

7. REFERENCES

[1] 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, pp. 173–183, 2014.

[2] 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.

[3] 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.

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

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

[6] Y. Qian, Z. Wu, B. Gao, and F. K. Soong, "Improved prosody generation by maximizing joint probability of state and longer units," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 6, pp. 1702–1710, 2011.

[7] S.-H. Chen, C.-H. Hsieh, C.-Y. Chiang, H.-C. Hsiao, Y.-R. Wang, Y.-F. Liao, and H.-M. Yu, "Modeling of speaking rate influences on mandarin speech prosody and its application to speaking rate-controlled TTS," *IEEE/ACM Trans., Audio, Speech, and Lang. Process.*, vol. 22, pp. 1158–1171, 2014.

[8] H. Zen, M. J.F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 3, pp. 794–805, 2012.

[9] 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.

[10] 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.

[11] C. Wutiwiwatchai and S. Furui, "Thai speech processing technology: A review," *Speech Commun.*, vol. 49, pp. 8–27, 2007.

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

[13] S. Potisuk, J. Gandour, and M. P. Harper, "Vowel length and stress in Thai," *Acta linguistica hafniensia*, vol. 30, pp. 39–62, 1998.

[14] N. C.V. Pilkington, H. Zen, and M. J.F. Gales, "Gaussian process experts for voice conversion," in *Proc. INTERSPEECH*, 2011, pp. 2772–2775.

[15] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Computer Speech & Language*, vol. 21, pp. 153–173, 2007.

[16] 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.

[17] 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 Commun.*, vol. 27, pp. 187–207, 1999.

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

[19] 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.