Duration Prediction Using Multi-Level Model for GPR-Based 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

10.21437/Interspeech.2015-369

Abstract

This paper introduces frame-based Gaussian process regression (GPR) into phone/syllable duration modeling for Thai speech synthesis. The GPR model is designed for predicting frame-level acoustic features using corresponding frame information, which includes relative position in each unit of utterance structure and linguistic information such as tone type and part of speech. Although the GPR-based prediction can be applied to a phone duration model, the use of phone duration model only is not always sufficient to generate natural sounding speech. Specifically, in some languages including Thai, syllable durations affect the perception of sentence structure. In this paper, we propose a duration prediction technique using a multi-level model which includes syllable and phone levels for prediction. In the technique, first, syllable durations are predicted, and then they are used as additional contexts in phone-level model to generate phone duration for synthesizing. Objective and subjective evaluation results show that GPR-based modeling with multi-level model for duration prediction outperforms the conventional HMM-based speech synthesis.

Index Terms: Duration prediction, GPR-based speech synthesis, multi-level modeling, tonal language

1. Introduction

Statistical modeling using hidden Markov model (HMM) has been widely used for speech synthesis, in which spectrum, pitch, and state duration are modeled simultaneously [1]. In the HMM-based framework, the distributions of spectral parameters, pitch parameters, and state duration are estimated at each HMM state and then clustered independently by using a decision-tree technique. However, this approach has two major problems. First, acoustic features are modeled by assuming stationarity of speech signal within a discrete state, nevertheless actual features change even in the same state. Another problem comes from the use of decision tree-based context clustering that reduces the number of states to a small number of leaf nodes and causes loss in contextual diversity. Statistical parametric speech synthesis based on Gaussian process regression (GPR) has been introduced to overcome the limitation of HMM-based speech synthesis [2]. In this technique, frame level information, including relative position in phone unit, is used as an input variable, and the speech parameters are represented by the sum of latent variables and Gaussian noise. To reduce the computational cost of GPR-based speech synthesis, a partially independent conditional (PIC) [3] approximation is incorporated into GPR [4]. It has been shown that the GPR-based approach can achieve better performance than the HMM-based speech synthesis [5, 6].

To generate natural-sounding speech, accurate duration prediction is a crucial issue for determining phone durations appropriately in the speech segment to be synthesized. For this purpose, only phone-level duration model is not enough to synthesize continuous speech correctly, because various prosodic features are based on longer unit such as stressed syllable. For example, in Thai language, pronunciation of stressed/unstressed syllable highly depends on duration [7]. In the conventional HMM-based technique [8], duration has been modeled at phone-unit level. Therefore syllable duration of synthetic speech is determined implicitly by summation of phone durations in a syllable. This results in less natural-sounding synthetic speech with incorrect stress in word pronunciation.

In this context, various techniques have been proposed to improve duration using longer unit than phone. In a speaking rate-dependent hierarchical prosodic model (SP-HPM) [9], various relationships among prosodic-acoustic features of speech signal, linguistic features of associated text, and prosodic tags representing the prosodic structure of speech were incorporated into speaking rate modeling. In [10], maximizing joint probability of state and longer units were proposed to integrate syllable and phrase level speech models with the state-based model for generating better prosody.

In this paper, we introduce the GPR-based framework into speech synthesis of a tonal language, specifically, Thai language. We incorporate frame-based acoustic feature modeling using GPR in a similar way as described in [6]. The different point is that we apply Thai speech contexts and structures into temporal event for the frame context. More importantly, we propose a duration modeling using a multi-level model which consists of phone and syllable levels. First, we train duration model at syllable unit level, then generate syllable duration. After that, we use predicted syllable duration as an event context in phone-level model training for synthesizing speech. We examine the proposed technique and evaluate its performance through objective and subjective evaluation tests.

2. GPR-based Thai speech synthesis

2.1. Gaussian process for regression

Let x_n and y_n be input and output variables, respectively. In GPR, y_n is given by

$$y_n = f(x_n) + \epsilon \tag{1}$$

where $f(x_n)$ is a noise-free latent function and ϵ is Gaussian noise. 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 form of input and output variables, and that of latent function values of training data, respectively. We define $\mathbf{X}_T$, $\mathbf{y}_T$, and $\mathbf{f}_T$ as matrix forms for test data. When $f(\cdot)$ is drawn from a Gaussian process, the joint distribution on the function values, $\mathbf{f}$ and $\mathbf{f}_T$, of the training and

test data is given by

$$p(\mathbf{f}, \mathbf{f}_T|\mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\begin{bmatrix}\mathbf{f}\\ \mathbf{f}_T\end{bmatrix}; 0, \mathbf{K}_{N+T}) \quad (2)$$

$$\mathbf{K}_{N+T} = \begin{bmatrix}\mathbf{K}_N & \mathbf{K}_{NT}\\ \mathbf{K}_{TN} & \mathbf{K}_T\end{bmatrix} \quad (3)$$

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

$$p(\mathbf{y}, \mathbf{y}_T|\mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\begin{bmatrix}\mathbf{y}\\ \mathbf{y}_T\end{bmatrix}; 0, \mathbf{K}_{N+T} + \sigma^2\mathbf{I}) \quad (4)$$

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

$$p(\mathbf{y}_T|\mathbf{y}, \mathbf{X}, \mathbf{X}_T) = \mathcal{N}(\mathbf{y}_T; \mu_T, \Sigma_T) \quad (5)$$

$$\mu_T = \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2\mathbf{I}]^{-1}\mathbf{y} \quad (6)$$

$$\Sigma_T = \mathbf{K}_T + \sigma^2\mathbf{I} - \mathbf{K}_{TN}[\mathbf{K}_N + \sigma^2\mathbf{I}]^{-1}\mathbf{K}_{NT} \quad (7)$$

2.2. Frame context for Thai language

The frame context is composed of temporal event and the relative position context as described in [6] :

$$\begin{aligned} x_n &= (x_{n,1}, \ldots, x_{n,K}), \qquad x_{n,k} = (\mathrm{p}_{n,k}, c_{n,k}) \\ \mathrm{p}_{n,k} &= (\mathrm{p}_{n,k}^{(-1)}, \mathrm{p}_{n,k}^{(0)}, \mathrm{p}_{n,k}^{(+1)}), \quad c_{n,k} = (c_{n,k}^{(-1)}, c_{n,k}^{(0)}, c_{n,k}^{(+1)}) \end{aligned} \quad (8)$$

where x_n is an array of partial frame context having K temporal events. $\mathrm{p}_{n,k}$ and $c_{n,k}$ are the relative position and the temporal event vectors, respectively. The temporal events are defined by linguistic information of four layers which are phone, syllable, word, and utterance. In Thai phonological system, Thai sound is often described in terms of syllable unit that is composed of initial consonant, vowel, final consonant, and tone. Thai has five tones that are mid (tone 0), low (tone 1), falling (tone 2), high (tone 3), and raising (tone 4). The phone layer contains phonetic features of Thai language based on the conventional HMM-based Thai speech synthesis [8]. The phonetic features are listed in Table 1. Each phonetic feature is represented by binary variables (+1 or -1). The syllable layer contains tone type feature represented by one-hot vector in which the k-th element of the vector equals one if tone type is tone k, and other elements are zero. For example, tone 0 is represented by [1,0,0,0,0]. The word layer contains part of speech feature which is obtained from the part of speech tag set of T-Sync-1 corpus [11]. We use one-hot vector for representing part of speech feature in the same way as the tone type. The temporal events for the frame context are summarized in Table 2.

2.3. Kernel function

Kernel function for the frame context is defined as follows.

$$\kappa(x_m, x_n) = \sum_{k=1}^{K} \theta_{r,k}^2 \kappa_k(x_{m,k}, x_{n,k}) + \delta_{mn}\theta_{floor}^2 \quad (9)$$

$$\begin{aligned} \kappa_k(x_{m,k}, x_{n,k}) = \sum_{u=-1}^{+1}\sum_{v=-1}^{+1} [&w(\mathrm{p}_{m,k}^{(u)})w(\mathrm{p}_{n,k}^{(v)}) \\ &\cdot \kappa_p(\mathrm{p}_{m,k}^{(u)}, \mathrm{p}_{n,k}^{(v)})\kappa_c(c_{m,k}^{(u)}, c_{n,k}^{(v)})] \end{aligned} \quad (10)$$

Table 1: Thai phonetic feature

Initial consonant phonetic features
Cluster-aspirated, Trill, Palatal, Stop, Lateral, Initial, Aspirated, Semivowel, Cluster-Avelolar, Stop-Exploded, Alveolar, Unaspirated, Cluster-unaspirated, Nasal, Cluster-with-l, Velar, Cluster-Labial, Cluster-with-w, Cluster-with-r, Cluster-Voiced, Fricative, Cluster-Fricative, Stop-Voiced, Labial, Cluster, Cluster-Velar, Glottal, Voiced, Unvoiced

Vowel phonetic features
Long, Low, Middle, Dipthong, Vowel, Front, Short, Central, Back, High, Voiced

Final consonant phonetic features
Palatal, Stop, Aspirated, Semivowel, Final, Stop-Unexploded, Alveolar, Unaspirated, Nasal, Velar, Fricative, Labial, Voiced, Unvoiced

Table 2: Temporal context for Thai GPR-based speech synthesis. The scales marked by * are not used for the duration model.

Unit:	phone
Type:	{beginning, end} of each phonetic feature
Scale:	phone-normalized, time*
Unit:	syllable
Type:	{beginning, end} of tone type
Scale:	{syllable, word}-normalized, time*
Unit:	word
Type:	{beginning, end} of part of speech
Scale:	{syllable, word}-normalized, time*
Unit:	utterance
Type:	{beginning, end} of utterance
Scale:	{syllable, word, utterance}-normalized, time*

where $w(\cdot)$, $\kappa_p(\cdot)$, and $\kappa_c(\cdot)$ are a weight function, position kernel, and event feature kernel, respectively. $\theta_{r,k}^2$ and θ_{floor}^2 are kernel parameters. We define the event feature kernel for the binary value and one hot vector as follows.

$$\kappa_c(c_{m,k}^{(u)}, c_{n,k}^{(v)}) = c_{m,k}^{(u)} \cdot c_{n,k}^{(v)} \quad (11)$$

3. Duration prediction using multi-level model

A key idea of the proposed technique is to use syllable duration as an additional context in phone duration modeling. However, the syllable duration corresponding to input text is generally unknown. To overcome the problem, we use a syllable-level model to predict the syllable duration first and use it as a context of input text for phone-level prediction. The squared exponential (SE) kernel is used to measure similarity between two syllable duration contexts in phone-level kernel as follows:

$$\kappa_c(c_{m,k}^{(u)}, c_{n,k}^{(v)}) = \exp\left(-\frac{(c_{m,k}^{(u)} - c_{n,k}^{(v)})^2}{l^2}\right) \quad (12)$$

where $c_{m,k}^{(u)}$ and $c_{n,k}^{(v)}$ are syllable duration contexts. l denotes a length-scale hyper-parameter. By doing this, we aim to model syllable-level duration explicitly in each syllable unit, and to avoid inconsistent phone durations in a syllable.

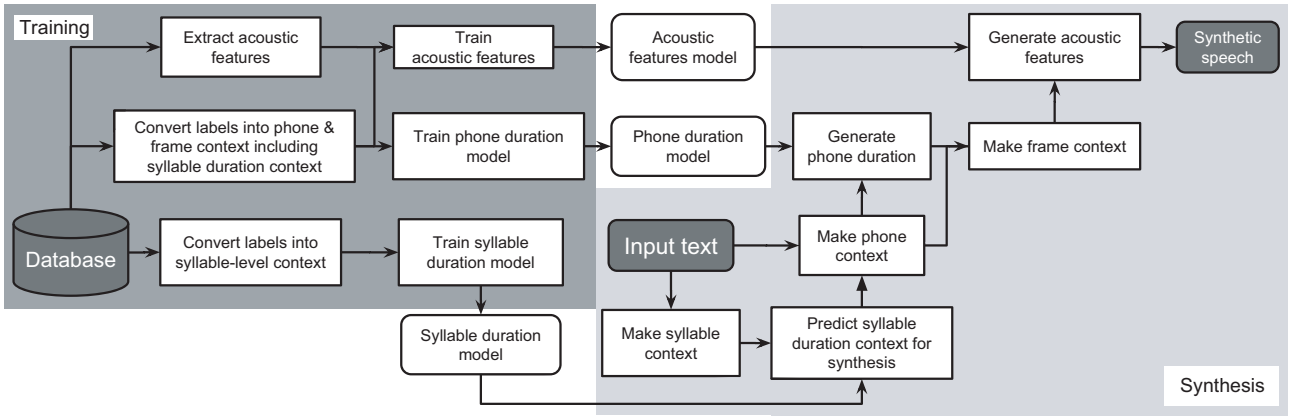


Figure 1: Block diagram of GPR-based speech synthesis system with multi-level model for duration prediction.

Table 3: Context for syllable-unit in syllable-based duration model.

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

3.1. Frame context of syllable-level model

The frame context of the syllable-level model contains temporal events and relative position in the same manner as the phone-level model. The temporal events are defined in three layers that are syllable, word, and utterance. The temporal event for syllable unit is composed of initial-consonant, vowel, final-consonant, and tone. The context of each phonetic component contains a list of its own phonetic features shown in Table 1, each of which is represented by a binary value. Tone type context is represented by one hot vector. The structure of syllable-level context is summarized in Table 3. In word and utterance levels, we use the same contexts as used in the phone-level GPR-based speech synthesis described in section 2.2.

3.2. Speech synthesis system with multi-level model

Figure 1 shows the overview of speech synthesis system with multi-level model for duration prediction. The respective processes of the blocks are as follows:

- Training part
  1. Extract acoustic features including mel-cepstral coefficients, aperiodicity, and F0.
  2. Convert labels into phone-level context including syllable duration context.
  3. Train the models of mel-cepstral coefficients, aperiodicity, F0, and phone-level duration individually.
  4. Convert labels into syllable-level context as described in section 3.1.
  5. Train syllable-level duration model.
- Synthesis part
  1. Make syllable-level context of input text.
  2. Predict syllable duration of input text by using syllable-level duration model.
  3. Make phone-level context of input text with predicted syllable duration.
  4. Generate phone duration by using phone-level model.
  5. Make frame context and synthesize speech.

4. Evaluation

4.1. Experimental condition

We used a set of phonetically balanced sentences of Thai speech database, T-Sync-1 from NECTEC [11], for training and evaluation data. The sentences were uttered by one professional female speaker with clear articulation and standard Thai accent with reading style. The training set contains 450 utterances, approximately 54 minutes in total. We used 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 [12] with 5-ms frame shift. The acoustic feature vector consisted of 0-39th mel-cepstral coefficients, 5-band aperiodicity, log F0, and their delta and delta-delta coefficients. PIC approximation [4] and optimization by EM-based method [13] were performed for each model.

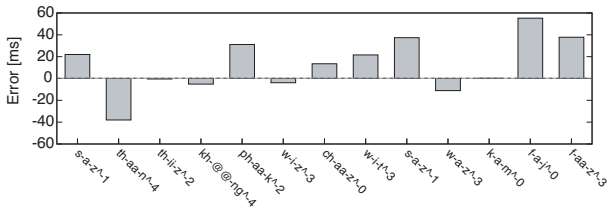
In the conventional HMM-based speech synthesis, we used hidden semi-Markov model (HSMM), which has explicit duration distributions. The model topology was 5-state left-to-right context-dependent HSMM. The context set consisted of the information of phone, syllable, word, and utterance.

4.2. Objective evaluation results

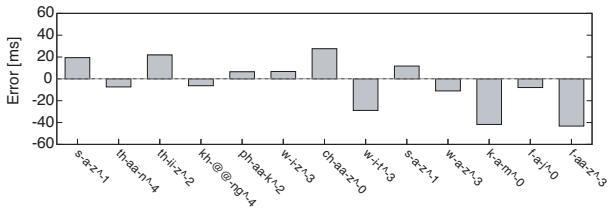
We compare the conventional HMM-based and the proposed GPR-based techniques objectively by using distortions of mel-cepstrum, log F0, and syllable/phone duration between generated speech and original one. The results are shown in Table 4. The mel-cepstrum and log F0 distances of GPR-based synthetic speech became lower than HMM-based synthetic speech significantly. In duration distortion, there was not much difference between HMM and GPR. Comparing duration distortion

Table 4: Objective evaluation result. The values represent mel-cepstral distances, and RMSEs of log F0, phone duration, and syllable duration, respectively.

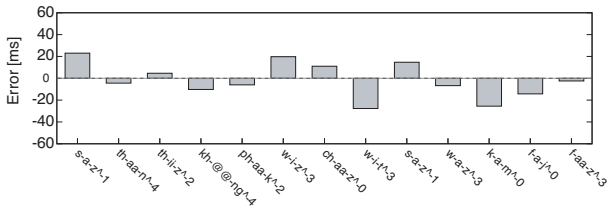
Model	Mcep [dB]	Log F0 [cent]	Duration [ms]	
			phone	syllable
HMM	4.92	126.4	25.8	47.8
GPR	4.46	98.4	26.1	45.8
GPR with multi-level	↑	↑	23.7	43.9



(a) Conventional HMM



(b) GPR



(c) GPR with multi-level model

Figure 2: Comparison of duration prediction distortions in syllable unit level. The sentence is "The place belongs to the department of electrical engineering." in English.

between HMM and GPR with multi-level, we can see that duration distortion was reduced significantly both at phone and syllable levels. Figure 2 shows an example of differences of syllable durations in a sentence generated by the conventional and proposed techniques from original ones. It can be seen that the GPR with multi-level technique generates the closest syllable duration sequence to the original one, and the improvement is significant at the final syllable of the sentence.

4.3. Subjective evaluation results

To confirm the improvement, we also evaluated the perceptual quality in naturalness of synthetic speech. We conducted mean opinion score (MOS) and forced choice preference tests for subjective evaluation. Participants were twenty Thai native speakers. Each participant listened to five speech samples that were randomly selected from the evaluation set. In MOS test, the participants evaluated each sample on a five-point scale from 1 to 5 according to their satisfaction in naturalness of sample. 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 3 shows the resultant scores with 95% confidence intervals. The GPR-based techniques achieved significantly higher scores than the HMM-based technique. In addition, the GPR-based technique with multi-level model gave slightly higher scores than the GPR without using multi-level model.

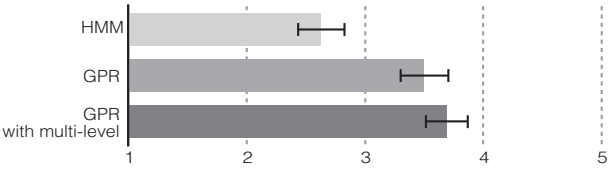


Figure 3: Result of MOS test in subjective evaluation of naturalness.

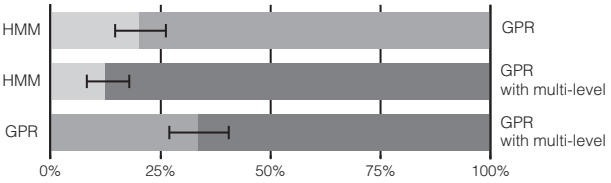


Figure 4: Results of forced choice preference test in subjective evaluation of naturalness.

In forced choice preference test, the participants were asked to choose the more natural-sounding one for each pair of speech samples. The participants could repeat samples as many times as they required in the same way as MOS test. Figure 4 shows the results of forced choice preference test. GPR-based methods obviously outperformed HMM-based technique. Comparing scores between GPR-based systems with and without multi-level model, we can see that the participants preferred the GPR with multi-level model significantly.

5. Conclusions

In this paper, we introduced GPR-based approach into Thai speech synthesis. We designed phone-level context based on linguistic information of Thai. In order to improve duration prediction performance at syllable unit level, we proposed multi-level model duration prediction. This technique uses a two-level model that consists of phone and syllable duration models. First, we train syllable-level model and use it to predict syllable duration for input text. Then, we use predicted syllable duration as an additional context for phone-level model. The results in objective and subjective evaluations showed that the GPR-based method outperformed the HMM-based one. In duration prediction, the GPR with multi-level model gave lower distortion than the GPR without multi-level model and achieved significantly higher score in subjective evaluation.

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] 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.

[2] 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.

[3] E. Snelson and Z. Ghahramani, "Local and global sparse Gaussian process approximations," in *Proc. AISTATS*, 2007, pp. 524–531.

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

[5] 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.

[6] 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.

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

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

[9] 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. on, Audio, Speech, and Lang. Process.*, vol. 22, no. 7, pp. 1158–1171, July 2014.

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

[11] 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.

[12] 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.

[13] 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.