PROSODY GENERATION USING FRAME-BASED GAUSSIAN PROCESS REGRESSION AND CLASSIFICATION FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Tomoki Koriyama, Takao Kobayashi

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

ABSTRACT

This paper proposes novel models of F0 contours and phone durations using Gaussian process regression and classification (GPR and GPC) for statistical parametric speech synthesis. Although the use of frame-based GPR has shown the effectiveness of spectral feature modeling in previous studies, the application of GPR to prosodic features, i.e., F0 and phone duration, was not investigated sufficiently because the kernel function was designed for phonetic information only. In this paper, therefore, we propose a kernel function available for multiple units such as syllables, moras, and accent phrases. The proposed kernel function is based on temporal acoustic events like the beginning of accent phrase and the relative position between the target frame and the event is utilized for the kernel function. Experimental results of objective and subjective tests show that the GPR/GPC-based F0 and duration modeling improves the prediction accuracy of acoustic features compared with HMM-based speech synthesis.

***Index Terms*—** Statistical parametric speech synthesis, prosody, Gaussian process regression, Gaussian process classification, kernel function

1. INTRODUCTION

Statistical parametric speech synthesis developed in the last decade has shown great success. Especially, hidden Markov model (HMM) based speech synthesis [1] is a dominant approach to the statistical parametric speech synthesis, and has enabled not only generating smoothly changing speech parameters but also speaker and style adaptation because of its flexibility in terms model adaptation. However, the naturalness of synthetic speech generated from the HMM-based technique is not sufficient, and several novel approaches to statistical parametric speech synthesis have been proposed in recent years [2–6].

The use of Gaussian process regression (GPR) is one of such approaches to statistical parametric speech synthesis. Fernandez et al. [4] attempted to model F0 features of speech segments by GPR where the input features are the intermediate values of deep belief networks (DBNs) converted from phone contexts, and showed the effectiveness of using GPR. Another one is modeling of spectral features based on frame-based GPR [5], in which spectral parameters are generated directly for speech synthesis using GPR with frame-level context. It has been shown that the spectral distortion of GPR-based technique becomes smaller than the conventional HMM-based one. The improvement of synthetic spectral features is mainly contributed by a non-parametric nature of GPR, that is, the speech parameters are generated utilizing the characteristics of raw speech data as it is without parameterizing it with a limited number of model parameters. However, the frame-based GPR has applied to only the spectral features, and modeling and generation of prosodic features, such as fundamental frequency (F0) and duration, which affect crucially on naturalness of the synthetic speech, remain to be solved.

In this paper, we propose techniques for modeling and generating prosodic speech parameters, i.e., F0 and duration, based on Gaussian processes. We incorporate a new kernel function for frame-level context so as to cope with prosodic information such as accents. The new kernel function is based on the temporal acoustic events, i.e., the beginning of vocalic phones, the end of accent phrase, and the beginning of the sentence. Furthermore, we construct a speech synthesis system all based on Gaussian process regression/classification (GPR/GPC) and evaluate the performance of the system through objective and subjective tests.

2. FRAME-BASED ACOUSTIC FEATURE MODELING USING GPR

Let $\mathbf{y}_T$ and $\mathbf{y}_N$ be the acoustic feature sequences of test and training data respectively. The predictive distribution of synthetic parameter sequence $\mathbf{y}_T$ using a Gaussian process [5] is obtained by

$$p(\mathbf{y}_T|\mathbf{y}_N) = \int p(\mathbf{y}_T|\mathbf{f}_T)p(\mathbf{f}_T|\mathbf{y}_N)d\mathbf{f}_T \tag{1}$$

$$p(\mathbf{f}_T|\mathbf{y}_N) = \int p(\mathbf{f}_T|\mathbf{f}_N)p(\mathbf{f}_N|\mathbf{y}_N)d\mathbf{f}_N \tag{2}$$

where $\mathbf{f}_T$ and $\mathbf{f}_N$ are latent function variables given by

$$p(\mathbf{f}_N, \mathbf{f}_T) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}_N\\ \mathbf{f}_T\end{bmatrix}; \mathbf{0}, \begin{bmatrix}\mathbf{K}_N^{\mathrm{PIC}} & \mathbf{K}_{NT}^{\mathrm{PIC}}\\ \mathbf{K}_{TN}^{\mathrm{PIC}} & \mathbf{K}_T^{\mathrm{PIC}}\end{bmatrix}\right). \tag{3}$$

$\mathbf{K}_N^{\mathrm{PIC}}$, $\mathbf{K}_{NT}^{\mathrm{PIC}}$, $\mathbf{K}_{TN}^{\mathrm{PIC}}$ and $\mathbf{K}_T^{\mathrm{PIC}}$ are kernel matrices that represent correlations of frames, whose elements are given by kernel function $\kappa(\mathbf{x}_m, \mathbf{x}_n)$ for linguistic frame-level contexts $\mathbf{x}_m$ and $\mathbf{x}_n$, where m and n refer to frame indexes. These matrices may be represented by partially independent conditional (PIC) approximation [7] so that we can avoid infeasible computation of GPR. For the PIC approximation, all frames in training data are clustered into multiple blocks, where the assignment of frames into blocks is conducted on the basis of phone-level decision trees used in HMM-based speech synthesis. In GPR, the speech parameter is supposed to be the sum of the latent variable and Gaussian noise with power σ_ν^2, which is represented by

$$p(\mathbf{y}_T|\mathbf{f}_T) = \prod_{t=1}^{T} \mathcal{N}\left(y_t; f_t, \sigma_\nu^2\right). \tag{4}$$

A part of this work was supported by JSPS Grant-in-Aid for Scien- tific Research 24300071, 25540065, and 25·8776.

3. SPEECH SYNTHESIS SYSTEM

In this study, we consider incorporating the frame-based GPR into the models of prosodic features. We propose voiced/unvoiced, F0, and duration models as follows:

Voiced/unvoiced model: Voiced/unvoiced flags are modeled by Gaussian process classification (GPC) [8], which is regarded as logistic regression of GPR. In GPC using sigmoid function $\varsigma(f_t)$, the output feature y is a binary value of $\{+1, -1\}$ and the distribution of $p(\mathbf{y}_T)$ is given by

$$p(\mathbf{y}_T|\mathbf{f}_T) = \prod_{t=1}^{T} p(y_t|f_t) \tag{5}$$

$$p(y_t = +1|f_t) = \varsigma(f_t). \tag{6}$$

Since we cannot obtain exact inference by GPC, we employ Laplace approximation to the posterior $p(\mathbf{f}_N|\mathbf{y}_N)$ in Eq. (2).

F0 model: Since F0 values are not observed in unvoiced frames, we construct F0 model using only voiced frames.

Duration model: Since the phone durations are phone-level features, we model phone duration by phone-based GPR. We use phone-level context as an input feature of GPR, which is defined similarly to the frame-level context.

Figure 1 shows the outline of the proposed speech synthesis system using GPR and GPC. We train the models of spectral feature, F0, aperiodicity feature, voiced/unvoiced, and duration, individually. When synthesizing, we first generate phone durations and then create the frame-level context from the duration because the frame-based GPR/GPC depends on durations. Using the frame-level context, we determine voiced/unvoiced frames and then generate F0s of the voiced frames, spectral features and aperiodicity features. Finally, we synthesize speech using the generated speech parameters. When generating speech parameters from the predictive distributions obtained by GPR, we employ parameter generation considering GV [9] for spectral feature whereas we use the predictive means without considering GV for the other features.

4. FRAME CONTEXT AND KERNEL FUNCTION FOR PROSODIC INFORMATION

4.1. Conventional kernel function for phonetic information

In the previous work [10], we used the following kernel function for spectral feature modeling:

$$\kappa(\mathbf{x}_m, \mathbf{x}_n) = \sum_{u=-1}^{+1} \sum_{v=-1}^{+1} \Big[w_m^{(u)} w_n^{(v)} \kappa_p(p_m^{(u)}, p_n^{(v)}) \kappa_c(\mathbf{c}_m^{(u)}, \mathbf{c}_n^{(v)}) \Big] + \delta_{mn}\theta_{\text{floor}}^2 \tag{7}$$

where $\kappa_p(\cdot)$ is a position kernel which represents the similarity of frame positions and $\kappa_c(\cdot)$ is a phone context kernel which represents the similarity of phonetic features.The term $\delta_{mn}\theta_{\text{floor}}^2$ is used to keep the kernel function positive definite. The input feature $\mathbf{x}_n$ of n-th frame, which is named frame context, is given by

$$\mathbf{x}_n = (\mathbf{p}_n, \mathbf{c}_n, \mathbf{w}_n), \qquad \mathbf{p}_n = \left(p_n^{(-1)}, p_n^{(0)}, p_n^{(+1)}\right)$$
$$\mathbf{c}_n = \left(c_n^{(-1)}, c_n^{(0)}, c_n^{(+1)}\right), \quad \mathbf{w}_n = \left(w_n^{(-1)}, w_n^{(0)}, w_n^{(+1)}\right). \tag{8}$$

Table 1. Temporal events used for frame context. The scales marked by *are not used for the duration model.

Unit:	phone
Types:	{beginning, end} of {vocalic, high, low, anterior, back, coronal, plosive, affricative, continuant, voiced, nasal, semi-vowel, silent} phoneme
Scale:	phone-normalized scale, time*
Unit:	mora
Types:	{beginning, end} of high/low mora
Scale:	{mora, accent phrase}-normalized scale, time*
Unit:	accent phrase
Types:	{beginning, end} of accent phrase {beginning, end} of accent nuclear mora {beginning, end} of phrase tone mora
Scale:	{mora, accent phrase}-normalized scale, time*
Unit:	breath group
Types:	{beginning, end} of breath group
Scale:	{mora, accent phrase, breath group}-normalized scale, time*
Unit:	sentence
Types:	{beginning, end} of sentence
Scale:	{mora, accent phrase, breath group, sentence}-normalized scale, time*

Here, the superscripts (-1), (0), and $(+1)$ correspond to the preceding, current, and succeeding phones, and $\mathbf{p}_n$, $\mathbf{c}_n$, and $\mathbf{w}_n$ are relative positions in the phone, phone contexts, and weights, respectively. Weights are used to emphasize the effect of phones that are closer to the current frame and are therefore dependent on relative position $\mathbf{p}_n$.

4.2. Frame context based on temporal events

The relative position in phones was solely used as the frame position information in the previous work [10]. However, in order to model F0 contours, it is important to use multiple scales. Various studies on HMM-based speech synthesis have shown that the positions of syllables, moras, and phrases are effective for modeling. In addition, Fujisaki model [11], a generative F0 contour model, utilizes time scale as an input feature. On the basis of this consideration, we define a new frame context as follows:

$$\mathbf{x}_n = (\mathbf{x}_{n,1}, \dots, \mathbf{x}_{n,K}), \qquad \mathbf{x}_{n,k} = (\mathbf{p}_{n,k}, \mathbf{c}_{n,k})$$
$$\mathbf{p}_{n,k} = \left(\mathbf{p}_{n,k}^{(-1)}, \mathbf{p}_{n,k}^{(0)}, \mathbf{p}_{n,k}^{(+1)}\right), \quad \mathbf{c}_{n,k} = \left(c_{n,k}^{(-1)}, c_{n,k}^{(0)}, c_{n,k}^{(+1)}\right). \tag{9}$$

A major difference from the conventional frame context is the use of an array of partial frame context $(\mathbf{x}_{n,1}, \dots, \mathbf{x}_{n,K})$, where $k = (1, \dots, K)$ indexes different temporal events such as the beginning of vowel, the end of accent nuclear mora, and the beginning of sentence.

Table 1 shows the list of temporal events defined for Japanese speech synthesis. The events are defined in multiple units. Moreover, multiple scales used for the relative position vector $\mathbf{p}_{n,k}^{(u)}$ are defined individually for each unit. Phone-normalized and mora-normalized scales represent the lengths of phones and moras are normalized to 1, respectively. By incorporating these scales, we can re-

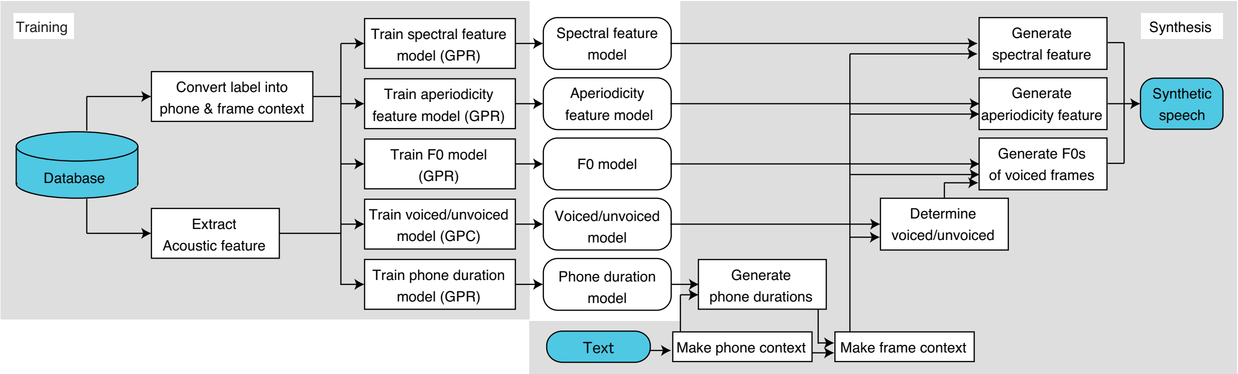

Training
Synthesis
Database
Convert label into phone & frame context
Extract Acoustic feature
Train spectral feature model (GPR)
Train aperiodicity feature model (GPR)
Train F0 model (GPR)
Train voiced/unvoiced model (GPC)
Train phone duration model (GPR)
Spectral feature model
Aperiodicity feature model
F0 model
Voiced/unvoiced model
Phone duration model
Generate phone durations
Text
Make phone context
Make frame context
Determine voiced/unvoiced
Generate spectral feature
Generate aperiodicity feature
Generate F0s of voiced frames
Synthetic speech

Fig. 1. Outline of parametric speech synthesis using GPR/GPC.

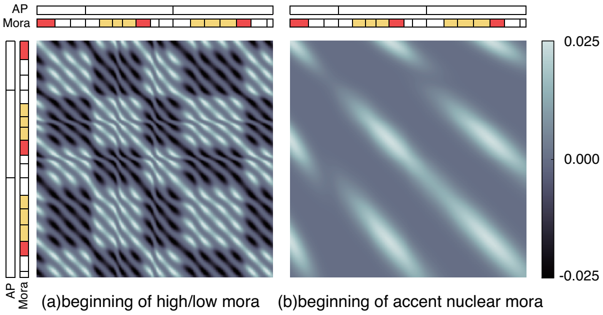


Fig. 2. Example of kernel matrices using events, for the segment that includes 3 accent phrases (APs) and 17 moras. Both red- and yellow-colored moras represent high ones, while red-colored moras are accent nuclear ones.

gard the similarity of mora-normalized position from the beginning of an accent phrase as that of the mora position in the accent phrase.

In Eq. (9), $c^{(u)}_{n,k}$ is an event feature represented by a binary value of $\{+1, -1\}$. The binary values for phone unit events represent distinctive phonetic feature [12] and those for mora unit events correspond to high/low tones.

In this study, the proposed frame context is used in common for all models shown in Fig. 1 including the spectral feature model. However the time scale in Table 1 is not used for the duration model because it is unknown information when phone durations are predicted.

4.3. Kernel function for the proposed frame context

We define a kernel function for the new frame context considering prosodic information as follows.

$$\kappa(\mathbf{x}_m, \mathbf{x}_n) = \sum_{k=1}^{K} \theta^2_{r,k}\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) + \delta_{mn}\theta^2_{\text{floor}} \qquad (10)$$

$$\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k}) = \sum_{u=-1}^{+1}\sum_{v=-1}^{+1}\left[w\left(\mathbf{p}^{(u)}_{m,k}\right) w\left(\mathbf{p}^{(v)}_{n,k}\right) \cdot \kappa_p\left(\mathbf{p}^{(u)}_{m,k}, \mathbf{p}^{(v)}_{n,k}\right)\kappa_c\left(c^{(u)}_{m,k}, c^{(v)}_{n,k}\right)\right] \qquad (11)$$

where $\theta^2_{r,k}$ is a kernel parameter that represents the relevance of partial kernel function $\kappa_k(\mathbf{x}_{m,k}, \mathbf{x}_{n,k})$. $w(\cdot)$ is a weight function used to emphasize the effect of closer events to the current frame, and given by

$$w\left(\mathbf{p}^{(v)}_{n,k}\right) = \exp\left(-\left(\mathbf{p}^{(v)}_{n,k}\right)^{\top}\boldsymbol{\Theta}^{-1}_{w,k}\mathbf{p}^{(v)}_{n,k}\right). \qquad (12)$$

We define position kernel $\kappa_p(\cdot)$ and event feature kernel $\kappa_c(\cdot)$, which corresponds to phone context kernel, as follows:

$$\kappa_p\left(\mathbf{p}^{(u)}_{m,k}, \mathbf{p}^{(v)}_{m,k}\right) = \exp\left(-\left(\mathbf{p}^{(u)}_{m,k} - \mathbf{p}^{(v)}_{n,k}\right)^{\top}\boldsymbol{\Theta}^{-1}_{b,k}\left(\mathbf{p}^{(u)}_{m,k} - \mathbf{p}^{(v)}_{n,k}\right)\right) \qquad (13)$$

$$\kappa_c\left(c^{(u)}_{m,k}, c^{(v)}_{m,k}\right) = c^{(u)}_{m,k}\cdot c^{(v)}_{m,k}. \qquad (14)$$

Kernel parameters $\theta^2_{r,k}$, θ^2_{floor}, $\boldsymbol{\Theta}^{-1}_{w,k}$, and $\boldsymbol{\Theta}^{-1}_{b,k}$ are automatically optimized by EM-based hyperparameter optimization algorithm [9], and thus we can choose appropriate events and their scales for training data.

Figure 2 shows an example of kernel matrices using events of different units, where the input sequence has 17 moras in 3 accent phrases. We picked up the beginning of high/low mora from the mora-unit events and the beginning of accent nuclear mora from the accent-phrase-unit ones. It is seen from the figure that the accent-phrase-unit event represents the frame relationships with a longer scale than the mora-unit event. We expect that the use of multiple units enables modeling of additional structure of F0 contours effectively.

5. EXPERIMENTS

5.1. Experimental conditions

We used speech data of one male (MMY) and one female (FKS) included in ATR Japanese speech database set B [13] for evaluations. We chose 450 sentences for training and used the remaining 53 sentences as test data. Speech signals were sampled at a rate of 16kHz, and the frame shift used for feature extraction was 5ms. The spectral envelope, F0, and aperiodicity were extracted by STRAIGHT [14], and 0-39th mel-cepstral coefficients, log F0, and five-band aperiodicity feature were used as acoustic feature vectors, where the feature value was normalized to zero mean and unit variance in each dimension. The number of temporal events, kernel function parameters and

Table 2. Accuracy of voiced/unvoiced prediction. Values represent F-scores.

Method	MMY	FKS
HMM	0.952	0.944
GPR/GPC	0.964	0.957

Table 3. RMS errors of log F0 and phone duration between original and synthetic speech.

	Log F0 [cent]		Duration [msec]	
Method	MMY	FKS	MMY	FKS
HMM	220.0	192.6	20.5	22.5
GPR/GPC	181.6	170.7	18.8	20.8

input features were 25, 112, and 205 for the duration model, and 38, 227, and 397 for the other models of frame-level acoustic features, respectively.

Two parameters of PIC approximation [7], the number of pseudo data set frames and the maximum number of frames in each block, were both set to 1024. The parameters of the kernel function were optimized by EM-based method [9], where 50 sentences included in the training data set were used for the optimization and the number of iterations for the EM-based method was five. The optimization was performed individually for each model.

The acoustic feature vector used to train HMM-based decision tree for PIC approximation included delta and delta-delta parameters. HMM topology was 5-state hidden semi-Markov model (HSMM) with single mixture and a diagonal covariance matrix. Context set consisted of the information of phone, mora, accent phrase, breath group, and sentence length. We employed the decision trees of mel-cepstrum, log F0, and phone duration for the voiced/unvoiced, F0, and duration models, respectively.

The HMM-based speech synthesis was used as a conventional technique. The model settings and acoustic feature vector were the same as that used for HMM-based decision tree for PIC approximation. The minimum description length (MDL) criterion [15] was used for constructing decision trees. Note that the HMM-based system used the dynamic features for speech parameter generation whereas the proposed GPR-based system did not.

5.2. Results

To evaluate the accuracy and reproducibility of the proposed method, we calculated acoustic distortions between original and synthetic speech. Table 2 shows the accuracy of voiced/unvoiced prediction by F-scores, and Table 3 shows the RMS errors of log F0 and phone duration between original and synthetic speech, respectively. As can be seen from Tables, the proposed method had higher accuracy and smaller distortions than conventional HMM-based speech synthesis.

Then, to evaluate the perceptual naturalness of synthetic speech, we conducted a paired comparison test on naturalness. Ten sentences were randomly chosen from the 53 test sentences for each of the participants and for each speaker. Eight participants were asked which synthetic speech was more natural. Figure 3 shows the result of the paired comparison test. We see that the proposed method outperformed the conventional one, and the difference of the scores was significant at the 5% significance level.

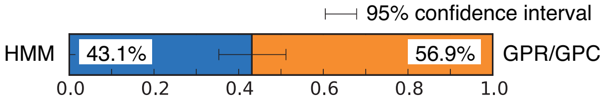
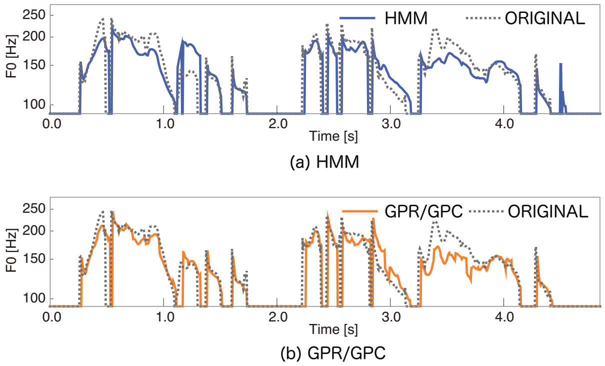


Fig. 3. Result of paired comparison test in naturalness of synthetic speech.



(a) HMM

(b) GPR/GPC

Fig. 4. Examples of generated F0 sequence.

In addition, we show examples of F0 contours generated by the HMM-based and GPR/GPC-based speech synthesis of the same sentence in Fig. 4. It is seen that the F0 contour generated by HMM had larger distortions than that by GPR/GPC around 1.1 sec. Such difference can be considered as the main cause of the decrease of acoustic distortion.

6. DISCUSSION

We showed that the proposed method using GPR and GPC outperformed the HMM-based one throughout the evaluations. However, the difference of perceptual naturalness was relatively small compared with the decrease of acoustic distortions, although the improvement of spectral feature modeling using GPR was significant [5, 9].

A possible reason is phone-based block assignment in the PIC approximation. Although an F0 contour is generally smooth in an accent phrase, the predictive F0 mean sequence is not smooth between at the boundaries of phones whose blocks are different. This results in discontinuous oscillations, which cause unnaturalness, as shown around 3.5 sec of the F0 contour generated by GPR/GPC in Fig. 4. Another possible reason is phone-based decision tree clustering, which does not directly model supra-segmental features such as shapes of F0 contours. That is, the performance of modeling of the F0 contour shape depends on that of clustering rather than GPR.

7. CONCLUSIONS

This paper investigated the effectiveness of the use of Gaussian process regression and classification for voiced/unvoiced, F0, and phone duration models on statistical parametric speech synthesis. In order to enable the prosodic feature modeling, we defined the kernel function based on temporal events. The objective and subjective evaluation showed that the proposed technique is effective for voiced/unvoiced prediction and F0 and duration generation as well as spectral feature modeling. In future work, we will investigate appropriate block assignment in the PIC approximation, such as prosody-based decision trees instead of phone-based ones, to improve the naturalness of synthetic speech. Moreover, we should examine the performance of the proposed approach for the respective acoustic feature on the perceptual quality.

8. 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] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in *Proc. ICASSP*, 2013, pp. 8012–8016.

[3] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. ICASSP*, 2013, pp. 7962–7966.

[4] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "F0 contour prediction with a deep belief network-Gaussian process hybrid model," in *Proc. ICASSP*, 2013, pp. 6885–6889.

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

[6] Y. Qian, Y. Fan, W. Hu, and Frank K. Soong, "On the training aspects of deep neural network (DNN) for parametric TTS synthesis," in *Proc. ICASSP*, 2014, pp. 3829–3833.

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

[8] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT press Cambridge, MA, 2006.

[9] 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. 3862–3866.

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

[11] K. Hirose, H. Fujisaki, and M. Yamaguchi, "Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information," in *Proc. ICASSP*, 1984, pp. 597–600.

[12] T. Fukuda and T. Nitta, "Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition," *IEICE Trans. Inf. & Syst.*, vol. 87, no. 5, pp. 1110–1118, 2004.

[13] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357–363, Aug. 1990.

[14] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[15] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *Acoustical Science and Technology*, vol. 21, no. 2, pp. 79–86, 2000.