論文 / 著書情報
Article / Book Information

Title	Statistical nonparametric speech synthesis using sparse Gaussian processes
Authors	Tomoki Koriyama, Takashi Nose, Takao Kobayashi
Citation	Proc. INTERSPEECH 2013, Vol. , No. , pp. 1072-1076,
Pub. date	2013, 8
Copyright	(c) 2013 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Statistical Nonparametric Speech Synthesis Using Sparse Gaussian Processes

Tomoki Koriyama, Takashi Nose, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology, Japan

koriyama.t.aa@m.titech.ac.jp, {takashi.nose, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper proposes a statistical nonparametric speech synthesis technique based on a sparse Gaussian process regression (GPR). In our previous study, we proposed GPR-based speech synthesis where each frame of synthesis units is modeled by a regression of Gaussian processes. Preliminary experiments of synthesizing several phones including both vowels and consonants showed a potential of the technique. In this paper, the previous work is extended to full-sentence speech synthesis using sparse GPs and context modification. Specifically, cluster-based sparse Gaussian processes such as local GPs and partially independent conditional (PIC) approximation are examined as a computationally feasible approach. Moreover, frame-level context is extended to include not only a position context from a current phone but also adjacent phones to generate smoothly changing speech parameters. Objective and subjective evaluation results show that the proposed technique outperforms the HMM-based speech synthesis with minimum generation error training.

Index Terms: statistical speech synthesis, Gaussian process regression, non-parametric Bayesian model

1. Introduction

In corpus-based statistical speech synthesis research, HMM-based speech synthesis [1] has been widely studied [2]. One of the reasons is that various speaker and style characteristics are well modeled and reproduced in the synthetic speech using a relatively small amount of training data by the HMM-based speech synthesis. Although the technique can generate stable synthetic speech with smooth speech parameter trajectories using dynamic acoustic features, it is difficult to precisely model acoustic characteristics within a state because they are modeled by a single distribution and a limited number of dynamic features.

In our previous study [3], we proposed frame-level acoustic feature modeling based on Gaussian process regression (GPR) [4] as an alternative approach to parametric speech synthesis. This approach uses a regression model that transforms frame-level linguistic features to corresponding frame-level acoustic features. Since GPR is a non-parametric Bayesian model, the number of parameters increases with the size of training data while keeping the robustness to over-fitting problem. In the previous study, we defined frame-level context and its associated kernel for GPR-based speech synthesis. From preliminary experimental results for isolated phones, we confirmed the potential of this approach.

In this paper, we perform full-sentence speech synthesis and evaluate the effectiveness of the proposed technique. To model trajectories of utterances, it is necessary to express their smoothly changing characteristics. However, the approach proposed in [3] is insufficient because the frame context at the boundary between adjacent phones is not continuously changing and, therefore, this causes unsmoothness of the synthetic speech. Furthermore, since utterances have diverse acoustic characteristics, a large amount and various kinds of training data is required. However it is difficult for GPR to utilize a large amount of training data because the computational complexity of GPR increases with the cube of the number of training data.

To overcome these problems, we propose a novel technique by introducing approaches developed in Gaussian processes (GPs) for machine learning. In the proposed technique, sparse Gaussian processes [5,6], which performs approximation of the covariance matrix of a GP, is incorporated to reduce the computational complexity. In addition, for the problem of the unsmoothness of the acoustic features, an extended frame context including multiple adjacent phones is proposed and convolution kernel [7] is employed as a kernel of the extended context.

2. GPR-based speech synthesis

In the previous study [3], we proposed GPR-based acoustic modeling. We define a training data set $\mathcal{D} = \{(\mathbf{x}_n, y_n)|n = 1, \ldots, N\}$ consisting of N frames and a test data set $\mathcal{D}_T = \{(\mathbf{x}_t, y_t)|t = 1, \ldots, T\}$ consisting of T frames. $\mathbf{x}_i$ represents an input feature vector obtained from linguistic information of the i-th frame, and y_i is the i-th frame's variable of output acoustic feature. Although an acoustic feature is generally a multi-dimensional vector, we here assume that all dimensions are independent and each dimension can be modeled separately. We give y_i using a function $f(\cdot)$ and noise ϵ by

$$y_i = f(\mathbf{x}_i) + \epsilon \tag{1}$$

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$, $\mathbf{y} = [y_1, \ldots, y_N]^\top$, and $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]^\top$ be matrix forms of training data. We define $\mathbf{X}_T$, $\mathbf{y}_T$, and $\mathbf{f}_T$ as matrix forms of test data in the same way as the training data. When $f(\cdot)$ is a Gaussian process, the joint distribution on training and test variables $\mathbf{f}$ and $\mathbf{f}_T$ is given by

$$p(\mathbf{f}, \mathbf{f}_T) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{N+T}) \tag{2}$$

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

where $\mathbf{K}_N$ and $\mathbf{K}_T$ are Gram matrices of training and test frames, respectively. Gram matrices $\mathbf{K}_{NT}$ and $\mathbf{K}_{TN}$ consist of covariances between training and test frames and $\mathbf{K}_{NT} = \mathbf{K}_{TN}^\top$. We assume that the noise ϵ has a variance σ^2. Then the joint distribution on training and synthetic acoustic features is given by

$$p(\mathbf{y}, \mathbf{y}_T) = \mathcal{N}(\mathbf{0}, \mathbf{K}_{N+T} + \sigma^2\mathbf{I}) \tag{4}$$

The predictive distribution of synthetic acoustic features is obtained by

$$p(\mathbf{y}_T|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_T, \boldsymbol{\Sigma}_T) \quad (5)$$

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

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

There are some choices for synthesizing speech from the predictive distribution, such as using the mean sequence for synthetic parameter or generating random sequence from the distribution. We here use the mean sequence of the distribution. For this purpose, a parameter vector $\boldsymbol{\alpha}$ is computed in training procedure, which is given by

$$\boldsymbol{\alpha} = (\mathbf{K}_N + \sigma^2\mathbf{I})^{-1}\mathbf{y} \quad (8)$$

The number of parameters in $\boldsymbol{\alpha}$ is N, which corresponds to the number of training frames. In the synthesis step, the mean sequence is calculated by

$$\boldsymbol{\mu}_T = \mathbf{K}_{TN}\boldsymbol{\alpha} \quad (9)$$

In order to achieve GP modeling, we proposed a kernel function [3] that can represent the relationships among the input features. The input variable of the kernel function is referred to as frame context, which comprises phone context and position context in the phone.

$$\mathbf{x}_n = (c_n, p_n) \quad (10)$$

The phone context p_n is a binary-valued vector of distinctive phonetic features [8] of the preceding, current, and succeeding phones. For the position context c_n, the relative frame position in the phone is used, where the position scale is normalized so that the beginning and end of the phone are 0 and 1, respectively. The kernel function is defined as a product of the kernels for the phone context c and for the position context p:

$$k(\mathbf{x}_m, \mathbf{x}_n) = k_c(c_m, c_n)k_p(p_m, p_n) \quad (11)$$

This kernel gives high covariance between the frames of similar phonetic features and close positions in their phones. Based on the results of the previous study, here we use squared exponential (SE) kernel and linear kernel for the kernels for phone context and position context, respectively.

3. Sparse Gaussian processes

In the training procedure, the matrix inversion needs $\mathcal{O}(N^3)$ calculations to obtain the parameter $\boldsymbol{\alpha}$ in Eq. (8). The value of N is generally at least hundreds of thousands[1]. Therefore the GPR computational complexity is not realistic. In this paper, we introduce two kinds of approximation methods: local GPs [5,6] and partially independent conditional (PIC) approximation [6]. These methods enable feasible computation by approximating matrices to be sparse. Although there are various kinds of approximation methods, e.g., subset of data (SoD) [4,9] and fully independent training conditional (FITC) approximation [9], we choose the local GPs and PIC because they are effective methods to model locally changing features like short-time changing acoustic features within phone segments.

3.1. Local GPs

Local GPs is a method to reduce the computation amount by dividing all of data into local blocks and model each block separately. That is, the covariance matrix $\mathbf{K}_{N+T}$ is approximated by block diagonal one,

$$\mathbf{K}_{N+T} \approx \mathbf{K}_{N+T}^{\mathrm{LOCAL}} = \mathrm{diag}\left[\mathbf{K}_{B_1}, \mathbf{K}_{B_2}, \ldots, \mathbf{K}_{B_S}\right] \quad (12)$$

When all the training frames are divided into S blocks and each block has at most B training frames, the computational cost results in $\mathcal{O}(SB^3)$. By fixing B, the computational complexity increases linearly with the number of training data N.

In order to use the local GPs, it is necessary not only to determine the block of the training frames but also that of synthesis frames from their linguistic features. In this study, we utilize decision-tree-based context clustering, which is effectively used in HMM-based speech modeling. When constructing the decision tree, we stop the node splitting if a node has less than B frames. We perform phone-unit clustering instead of state-level or stream-level clustering because state and stream information is unknown in the synthesis step.

The local GPs and the HMM-based speech synthesis both use the decision tree clustering of context dependent HMMs. In the HMM-based method, the observation samples of each cluster are collected and converted to a limited number of distributions. In contrast, in GPR with local GPs, the covariances of the samples in the same cluster are calculated and GPR training of each cluster yields at most B parameters.

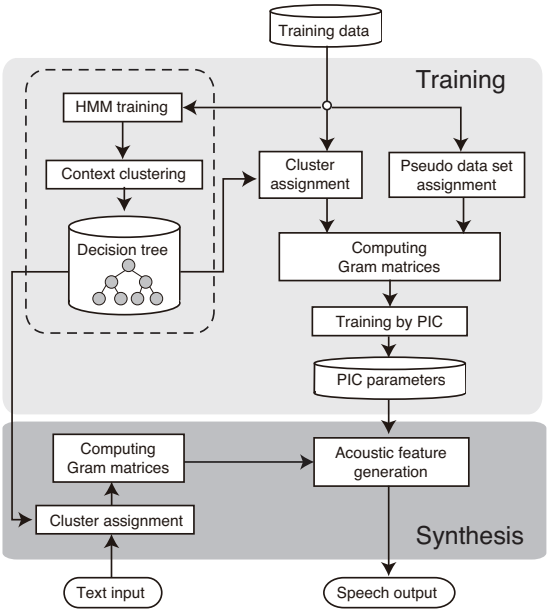
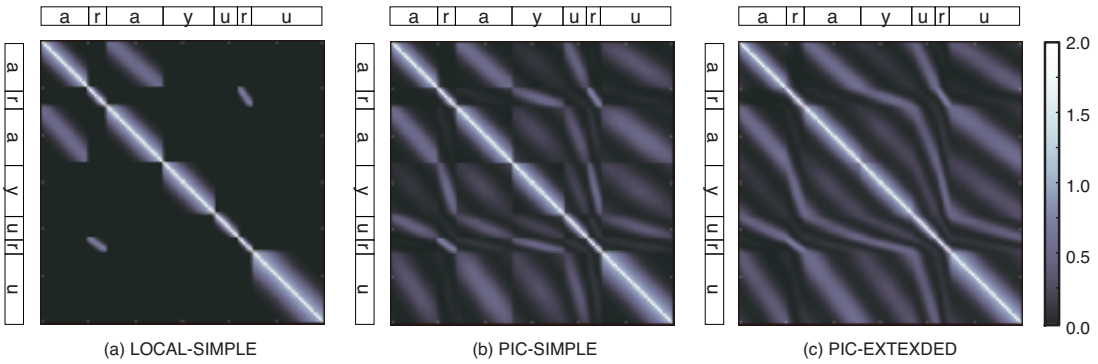


Figure 1: An overview of training and synthesis using PIC approximation.

3.2. Partially independent conditional (PIC) approximation

Although the local GPs can model internally changing features effectively within the blocks, the covariances between different blocks are completely ignored. On the other hand, a partially independent conditional (PIC) approximation estimates the covariances between different blocks using *pseudo-data set*. A pseudo-data set is a small amount of data set with a size of $M \ll N$, and the pseudo-data are expected to be distributed similarly to the training data. PIC divides frames into blocks,

[1] If we have 10 minutes speech data with 5ms shift, N is 120,000.



(a) LOCAL-SIMPLE (b) PIC-SIMPLE (c) PIC-EXTEXDED

Figure 2: Example of covariance matrices of a Japanese phrase segment "a r a y u r u" using (a) the local GPs and the conventional single frame context, (b) the PIC and the conventional single frame context, and (c) the PIC and the extended frame contexts.

and the covariance matrix of training data is given by

$$\mathbf{K}_N^{\mathrm{PIC}} = \begin{bmatrix} \mathbf{K}_{B_1} & \mathbf{Q}_{B_1B_2} & \cdots & \mathbf{Q}_{B_1B_S} \\ \mathbf{Q}_{B_2B_1} & \mathbf{K}_{B_2} & & \mathbf{Q}_{B_2B_S} \\ \vdots & & \ddots & \vdots \\ \mathbf{Q}_{B_SB_1} & \mathbf{Q}_{B_SB_2} & \cdots & \mathbf{K}_{B_S} \end{bmatrix} \tag{13}$$

where $\mathbf{Q}_{B_iB_j}$ is a matrix given by

$$\mathbf{Q}_{B_iB_j} = \mathbf{K}_{B_iM}\mathbf{K}_M^{-1}\mathbf{K}_{MB_j} \tag{14}$$

$\mathbf{K}_M$ is a self covariance matrix of pseudo data set and $\mathbf{K}_{B_iM}$ and $\mathbf{K}_{MB_j}$ are Gram matrices whose elements are kernel values between the samples of the clustered block and the pseudo data set. The approximation avoids direct inter-blocks Gram matrices calculation by means of pseudo-data set.

When a new input value $\mathbf{x}_*$ is assigned into cluster B_s, the corresponding mean for the new input value is given by

$$\mu_* = \mathbf{K}_{*M}(\mathbf{w} - \mathbf{w}_{B_s}) + \mathbf{K}_{*B_s}\mathbf{p}_{B_s} \tag{15}$$

The first and second terms of the right-hand side of Eq. (15) correspond to global and local characteristics. $\mathbf{w}$, $\mathbf{w}_{B_s}$ and $\mathbf{p}_{B_s}$ are PIC model parameters calculated by

$$\mathbf{w} = \textstyle\sum_{s=1}^{S} \mathbf{w}_s \tag{16}$$

$$\mathbf{w}_s = \mathbf{K}_M^{-1}\mathbf{K}_{MB_s}\mathbf{p}_s \tag{17}$$

$$[\mathbf{p}_1^\top \cdots \mathbf{p}_S^\top]^\top = [\mathbf{K}_N^{\mathrm{PIC}} + \sigma^2\mathbf{I}]^{-1}\mathbf{y} \tag{18}$$

When the maximum block size is B, the number of block is S, and the number of frames of pseudo-data set is M, the computational cost results in $\mathcal{O}(SB^3 + SM^3)$. Methods of determining the blocks and the pseudo-data set are needed to use PIC. The blocks of frames are determined in the same way as the local GPs. We adopt random selection from the training data to select of pseudo-data set.

An overview of speech synthesis using the PIC approximation is shown in Fig. 1. In the training procedure, first the decision tree of contexts are constructed using context-dependent HMMs. Then the pseudo-data set is chosen from training data, and the cluster for each training data frame is assigned by the decision tree. After that, Gram matrices are computed. At the end of the training procedure, the PIC parameters $\{\mathbf{p}_s\}$, $\{\mathbf{w}_s\}$, and $\mathbf{w}$ in Eqs. (16)–(18) are calculated. When synthesizing, the cluster for each frame of a text input is also determined by the decision tree. Next, Gram matrices between synthesis and training frames are computed and the acoustic features of the frames are generated from the Gram matrices and the trained PIC parameters. Finally, a speech utterance is synthesized by using the generated acoustic feature trajectory.

4. Extension of frame context using adjacent phones

Even though PIC can express the covariances between different blocks, the conventional frame context is insufficient for synthesizing natural-sounding speech. A problem of the frame context is the discontinuity of covariances at the boundary of adjacent phones because the context includes only the position information about the current phone. For example, the context of the first frame of a current phone and that of the last frame of the preceding phone are entirely different. The discontinuity in covariance causes synthetic speech to be unsmooth.

Therefore we propose an extended context in order to achieve smoothly changing trajectories of synthetic speech. The point is that a certain frame has not only the information of the current phone but also that of nearby phones. For example, the first frame of a current phone can also be regarded as the next frame of the last frame of the preceding phone. Hence, the extended context $\mathbf{x}_n$ is defined as a set of position and phone contexts of adjacent phones.

$$\mathbf{x}_n = \{(w_n^{(i)}, c_n^{(i)}, p_n^{(i)}) | i \in \{-1, 0, +1\}\} \tag{19}$$

Here, the subscriptions -1, 0, and $+1$ of the variables correspond to the preceding, current, and succeeding phones. $w_n^{(i)}$ represents a weight used to emphasize the effect of closer phones. In this study the following sine window function is used for the weight

$$w_n^{(i)} = \begin{cases} \sin(\pi(p_n^{(i)} + 0.5)/2) & (-0.5 \le p_n^i \le 1.5) \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

The kernel function between the extended contexts is defined using a convolution kernel [7]. The convolution kernel enables us to define kernel function of input variables that has multiple values such as the extended kernel. The proposed extended kernel is given by

$$k(\mathbf{x}_m, \mathbf{x}_n) = \sum_{i\in\{-1,0,+1\}} \sum_{j\in\{-1,0,+1\}} \left[w_m^{(i)} w_n^{(j)} k_c(c_m^{(i)}, c_n^{(j)}) k_p(p_m^{(i)}, p_n^{(j)}) \right] \tag{21}$$

Figure 2 shows an example of covariance matrices. In Fig. 2 (a), since the local GPs are used, we can see many of the elements are zeros because only intra-cluster covariances are calculated. By using PIC in Fig. 2 (b) and (c), inter-clustered covariances are estimated. And in Fig. 2 (c), the extended context gives smooth covariances around the boundaries of adjacent

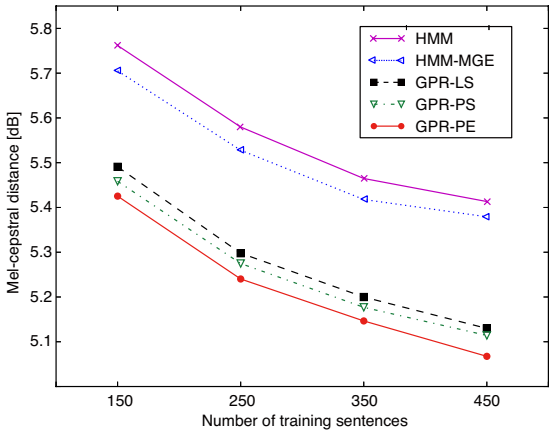


Figure 3: Average spectral distortions between original and synthetic speech as a function of the number of training sentences.

phones.

5. Experiments

5.1. Experimental conditions

We used speech data of a Japanese female voice actress. The speaker uttered 503 phonetically balanced sentences with a reading style. These sentences were taken from ATR Japanese speech database set B [10]. The total length of the 503 speech samples was about 45 minutes. The phone boundary information was annotated manually. Speech signals were sampled at a rate of 16kHz, and the frame shift was 5ms. In this study, we only modeled and generated a spectral feature. The 0-39th mel-cepstral coefficients derived from the spectral envelop extracted by STRAIGHT [11] were used as the spectral features. The maximum number of frames B of the cluster described in Sect. 3 was set to 1000 for the local GPs and PIC, and the number of pseudo data sets M was set to 200. For comparison, we also evaluated the HMM-based speech synthesis with and without minimum generation error (MGE) training [12]. The model topology was 5-state, left-to-right, no-skip hidden semi-Markov model (HSMM). The output distribution in each state was modeled with a single Gaussian pdf, and covariance matrices were assumed to be diagonal. The feature vector included delta and delta-delta dynamic features as well as the static one. Triphones were used for the context set for the HMM training. In the decision-tree-based context clustering for parameter tying, the MDL was used as a stopping criterion [13].

5.2. Objective evaluation

First, we objectively compared the performance of the conventional and proposed techniques. Mel-cepstral distance between synthetic and original speech was used as an objective distortion measure. 150, 250, 350, and 450 sentences were used as the training data, and 53 sentences were used as the test data. The test data was not included in the training data. We compared two kinds of HMM-based methods and three kinds of proposed GPR-based methods. The results are shown in Fig. 3. In the figure, "HMM" represents the HMM-based method where the model parameters was optimized by the ML criterion. "HMM-MGE" used MGE training for the model parameter optimization. In the proposed GPR-based methods, L and P denote local GPs and PIC for approximation, respectively, and S and E denote conventional single frame context and proposed extended

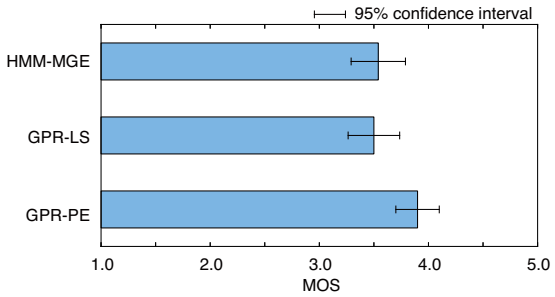


Figure 4: Naturalness of the synthetic speech.

frame context, respectively. From the result, it is seen that both HMM-MGE and GPR-based methods gave smaller distortions than HMM. Moreover, the GPR-based methods gave significantly smaller distortions than HMM-MGE which means that the frame-level regression has a good performance though the proposed methods do not include dynamic acoustic features. By comparing GPR-LS and GPR-PS, we can see that the distortion decreased slightly for all of the training sets. In addition, GPR-PE had slightly higher reproducibility than GPR-PS.

5.3. Subjective evaluation

To examine the total performance of the proposed technique, we evaluated HMM-MGE, GPR-LS, and GPR-PE by a mean opinion score (MOS) test. The number of training sentences was 450. Speech samples were synthesized by STRAIGHT using generated spectral features and original F0 and phone durations. Five participants listened to the synthetic speech samples and rated the naturalness of synthetic speech on a five-point scale, i.e., 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad. For each participant, ten sentences were randomly chosen from the 53 sentences. Figure 4 shows the mean opinion score (MOS). The error bars stand for 95% confidence intervals. When comparing GPR-LS and HMM-MGE, we see that the scores were comparable whereas GPR-LS gave smaller distortion in objective evaluation. This is because the generated acoustic features were not smooth at the phone boundaries and this discontinuity degraded the naturalness. In contrast, GPR-PE, which had continuity on the covariance matrices, gave the highest score of the three methods. There is a significant difference between GPR-PE and HMM-MGE at a 5% significance level ($p = 0.027$).

6. Conclusion

This paper described a technique for the GPR-based speech synthesis. We used block-based sparse GP approximations such as local GPS and PIC for trajectory modeling of utterances with feasible computation. Moreover, for the generation of smooth parameter trajectory, the frame context was extended to include nearby phone information. From the objective and subjective evaluation, the proposed method using the PIC approximation and the extended context achieved better performance than the HMM-based methods. However there are various kinds of parameters and kernel structures that has to be optimized. Therefore, in future work, we have to refine them and this might lead to improvements in the quality

7. Acknowledgments

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 24300071 and 25540065.

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] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] T. Koriyama, T. Nose, and T. Kobayashi, "Frame-level acoustic modeling based on gaussian process regression for statistical nonparametric speech synthesis," in *Proc. ICASSP, in press*, 2013.

[4] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT press Cambridge, MA, 2006.

[5] H. Wackernagel, *Multivariate Geostatistics*. Springer, 2003.

[6] E. Snelson and Z. Ghahramani, "Local and global sparse gaussian process approximations," in *Proceedings of Artificial Intelligence and Statistics*, 2007.

[7] D. Haussler, "Convolution kernels on discrete structures," in *Technical Report UCSC-CRL-99-10*. Dept of Computer Science, University of California at Santa Cruz., 1999.

[8] 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.

[9] J. Quiñonero-Candela and C. E. Rasmussen, "A unifying view of sparse approximate Gaussian process regression," *The Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.

[10] 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.

[11] 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.

[12] Y. J. Wu and R. H. Wang, "Minimum generation error training for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 1, 2006, pp. 889–892.

[13] 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.