論文 / 著書情報
Article / Book Information

Title	A Speaker Adaptation Technique For Gaussian Process Regression Based Speech Synthesis Using Feature Space Transform
Authors	Tomoki Koriyama, Syohei Oshio, Takao Kobayashi
Citation	Proc. 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016), Vol. , No. , pp. 5610-5614
Pub. date	2016, 3
Copyright	(c) 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	http://dx.doi.org/10.1109/ICASSP.2016.7472751
Note	This file is author (final) version.

A SPEAKER ADAPTATION TECHNIQUE FOR GAUSSIAN PROCESS REGRESSION BASED SPEECH SYNTHESIS USING FEATURE SPACE TRANSFORM

Tomoki Koriyama, Syohei Oshio, Takao Kobayashi

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

ABSTRACT

In this paper, we propose a speaker adaptation technique for statistical parametric speech synthesis based on Gaussian process regression (GPR). Although it is reported that the GPR-based speech synthesis improves the naturalness of synthetic speech compared with the HMM-based speech synthesis, any speaker adaptation techniques for the GPR-based one have not been established. This is because GPR is a nonparametric model and hence it is impossible to directly apply linear transforms to model parameters. In the proposed technique, we introduce feature-space transform to achieve model adaptation in the framework of GPR-based speech synthesis. Experimental results of objective and subjective tests show that the proposed technique outperforms the conventional HMM-based speaker adaptation framework.

Index Terms— speaker adaptation, statistical parametric speech synthesis, Gaussian process regression, feature-space transform

1. INTRODUCTION

Speaker adaptation is an essential technique for statistical parametric speech synthesis (SPSS) with diverse voices based on hidden Markov model (HMM). It is true that a large amount of speech data of a target speaker promises to give synthetic speech very similar to that speaker, however, it is also true that it is not always possible to prepare a sufficient amount of his/her data. In such a case, speaker adaptation enables us to train a target speaker's model using only a small amount of speech data of the target speaker.

HMM-based speech synthesis [1] has been studied widely and shown that various speaker adaptation techniques, originally developed for automatic speech recognition, still work well in speech synthesis. Maximum likelihood linear regression (MLLR) [2–4] is one of such approaches, in which mean vectors of state output distributions of HMMs are transformed into a target speaker's model using a linear transform. Other approaches, such as constrained MLLR (CMLLR) [5, 6] and CSMAPLR [6], transform not only mean vectors but also covariance matrices.

One of limitations of the HMM-based speech synthesis is that it can generate speech with satisfactory intelligibility but not always high quality in naturalness. In this context, various approaches to SPSS have been proposed as alternatives to the HMM-based speech synthesis in recent years [7–10]. SPSS based on Gaussian process regression (GPR) [10, 11] is one of the alternative approaches. We have shown that GPR-based speech synthesis improves the naturalness of synthetic speech, and moreover, it gives comparable with, or higher performance than the DNN-based one [12].

A part of this work was supported by KAKENHI Grant Number 15H02724.

In this paper, we propose a speaker adaptation technique for GPR-based speech synthesis, which enables the GPR-based speech synthesis to be more flexible and useful as the speaker adaptation for the HMM-based speech synthesis does. However, it is impossible to directly apply linear transform techniques such as MLLR and CMLLR because GPR is a nonparametric model, in other words, there are no mean vectors or covariance matrices to be transformed. DNN-based speech synthesis also has the problem similar to the GPR-based synthesis. To overcome this problem, Fan et al. [13] introduced transform parameters from a hidden layer of neural network to target speaker's feature space. In [14], Wu et al. investigated speaker adaptation in different levels, such as i-vector input, learning hidden unit contribution (LHUC) and feature space transform from an average voice to target speaker's voice. In these studies, it is reported that feature-space transform yields good performance for speaker adaptation.

In the proposed speaker adaptation for GPR-based speech synthesis, we utilize feature-space transform from source speakers to the target speaker, whereas MLLR/CMLLR in the HMM-based framework uses model-space transform from average voice model to the target speaker's one. Then, we define joint distribution among source and target speakers using Gaussian process, and derive optimal transform parameters for speaker adaptation. We show the effectiveness of the proposed technique through objective and subjective evaluations using a small amount of adaptation data.

2. GPR-BASED SPEECH SYNTHESIS

Let N and D are the numbers of frames in training data and dimensions of acoustic feature, respectively. In the GPR-based speech synthesis [10, 11], we assume that speech parameter sequences are outputs sampled from a Gaussian process. This assumption can be expressed by following joint distribution:

$$\mathbf{Y}_N \sim \mathcal{MN}\left(\mathbf{O}, \mathbf{K}_N + \sigma^2\mathbf{I}, \mathbf{V}\right) \tag{1}$$

where $\mathbf{Y}$ is $N \times D$ matrix that consists of all speech parameters included in training data given by

$$\mathbf{Y}_N = \left[\mathbf{y}_1, \ldots, \mathbf{y}_N\right]^\top. \tag{2}$$

$\mathcal{MN}(\cdot)$ denotes matrix variate normal distribution and $\mathbf{V} = \mathrm{diag}[v_1, \ldots, v_D]$ is a variance matrix which represents variances of respective dimensions. $\mathbf{K}_N$ is a Gram matrix that represents the relationship of frame-level contexts. All speech parameters are normalized to zero mean.

Using this assumption, we obtain a predictive distribution of

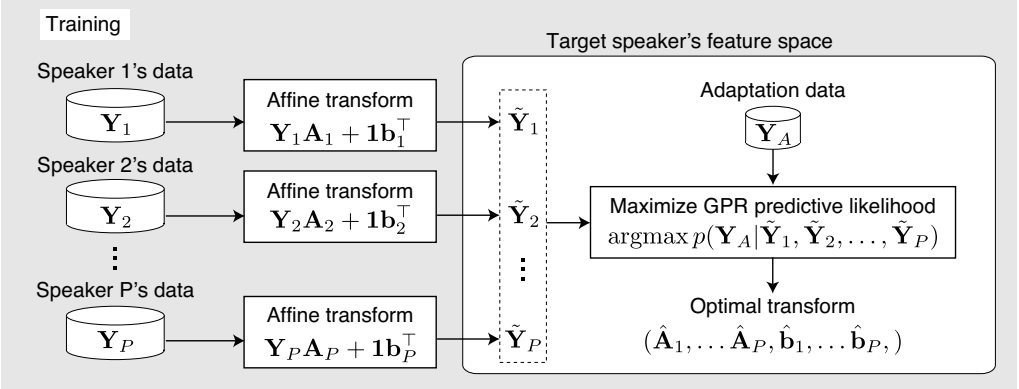

Figure 1: Outline of training of adaptation parameters.

synthetic speech parameters $\mathbf{Y}_T$ as follows:

$$p(\mathbf{Y}_T|\mathbf{Y}_N) = \mathcal{MN}(\mathbf{Y}_T; \mathbf{M}_T, \mathbf{\Sigma}_T, \mathbf{V}) \tag{3}$$

$$\mathbf{M}_T = \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{Y}_N \tag{4}$$

$$\mathbf{\Sigma}_T = \mathbf{K}_T - \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{K}_{NT} + \sigma^2\mathbf{I}_T. \tag{5}$$

Since direct calculation of GPR is computationally unrealizable, we employ partially independent conditional (PIC) approximation [15] as in the previous study [10].

3. SPEAKER ADAPTATION FOR GPR-BASED SYNTHESIS

3.1. Speaker adaptation framework

Suppose that there exists P speakers' speech data for model training. Let $\mathbf{Y}_i$ $(i = 1, \ldots, P)$ be a speech parameter sequence of source speaker i included in training data. In this study, we assume that transformed speech parameters from source speakers, $\tilde{\mathbf{Y}}_i$ $(i = 1, \ldots, P)$, can be regarded as the training data of GPR-based speech synthesis in target speaker's space. This assumption leads to following joint probability distribution of speech parameters of source speakers $(1, \ldots, P)$ and target speaker T.

$$p\begin{pmatrix}\tilde{\mathbf{Y}}_1\\ \vdots\\ \tilde{\mathbf{Y}}_P\\ \mathbf{Y}_T\end{pmatrix} = \mathcal{MN}\left(\begin{bmatrix}\tilde{\mathbf{Y}}_1\\ \vdots\\ \tilde{\mathbf{Y}}_P\\ \mathbf{Y}_T\end{bmatrix}; \mathbf{O}, \mathbf{K}_{N+T} + \sigma^2\mathbf{I}, \mathbf{V}\right) \tag{6}$$

where

$$\mathbf{K}_{N+T} = \begin{bmatrix}\mathbf{K}_N & \mathbf{K}_{NT}\\ \mathbf{K}_{TN} & \mathbf{K}_T\end{bmatrix} \tag{7}$$

$$\mathbf{K}_N = \begin{bmatrix}\mathbf{K}_{11} & \cdots & \mathbf{K}_{1P}\\ \vdots & \ddots & \vdots\\ \mathbf{K}_{P1} & \cdots & \mathbf{K}_{PP}\end{bmatrix} \tag{8}$$

$$\mathbf{K}_{NT} = \begin{bmatrix}\mathbf{K}_{1T} & \cdots & \mathbf{K}_{PT}\end{bmatrix}. \tag{9}$$

$\mathbf{K}_{ij}$ $(i, j = 1, \ldots, P \text{ or } T)$ represents a Gram matrix between speaker i and j. In accordance with the estimation procedure in GPR, the predictive distribution of target speaker's speech parameters is obtained as follows:

$$p(\mathbf{Y}_T|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_P) = \mathcal{MN}(\mathbf{Y}_T; \mathbf{M}_T, \mathbf{\Sigma}_T, \mathbf{V}) \tag{10}$$

$$\mathbf{M}_T = \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma^2\mathbf{I}_N\right)^{-1}\begin{bmatrix}\tilde{\mathbf{Y}}_1\\ \vdots\\ \tilde{\mathbf{Y}}_P\end{bmatrix} \tag{11}$$

$$\mathbf{\Sigma}_T = \mathbf{K}_T - \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{K}_{TN} + \sigma^2\mathbf{I}_T. \tag{12}$$

In this study, we employ a simple affine transform to target speaker's feature space as the model adaptation in GPR-based speech synthesis:

$$\tilde{\mathbf{y}} = \mathbf{A}_i^\top\mathbf{y} + \mathbf{b}_i \tag{13}$$

where $\mathbf{A}_i$ and $\mathbf{b}_i$ are a transform matrix and a bias vector, respectively. Using these definitions, the matrix form of transformed speech parameters is represented by

$$\tilde{\mathbf{Y}}_i = \mathbf{Y}_i\mathbf{A}_i + \mathbf{1}_i\mathbf{b}_i^\top. \tag{14}$$

The proposed adaptation framework is shown in Figs. 1 and 2. In the training phase, we estimate optimal transform $\hat{\mathbf{A}}_i$ and $\hat{\mathbf{b}}_i$ by maximizing predictive likelihood of adaptation data. The estimation method is described in the next section. In the synthesis phase, source speakers' data are transformed using the optimal transform parameters and then the predictive distribution is calculated by GPR. After that, we generate speech parameters by maximizing predictive likelihood in the same way as the speaker dependent GPR-based synthesis [12].

3.2. Estimation of transform parameters

We define

$$\mathbf{L} \triangleq \left[\mathbf{L}_1, \ldots, \mathbf{L}_P\right] \triangleq \mathbf{K}_{TN}\left(\mathbf{K}_N + \sigma^2\mathbf{I}_N\right)^{-1} \tag{15}$$

which appears in (11) and (12). Then the log predictive distribution of speech parameters of adaptation data $\mathbf{Y}_A$ is given by

$$\begin{aligned}\mathcal{L} &= \log p(\mathbf{Y}_A|\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_P)\\ &= -\frac{T_A D}{2}\log(2\pi) - \frac{T_A}{2}\log|\mathbf{V}| - \frac{D}{2}\log|\mathbf{\Sigma}_T|\\ &\quad - \frac{1}{2}\mathrm{Tr}\left[\mathbf{V}^{-1}\left(\mathbf{Y}_A - \sum_{i=1}^{P}\mathbf{L}_i\tilde{\mathbf{Y}}_i\right)^\top\mathbf{\Sigma}_T^{-1}\left(\mathbf{Y}_A - \sum_{i=1}^{P}\mathbf{L}_i\tilde{\mathbf{Y}}_i\right)\right]\end{aligned} \tag{16}$$

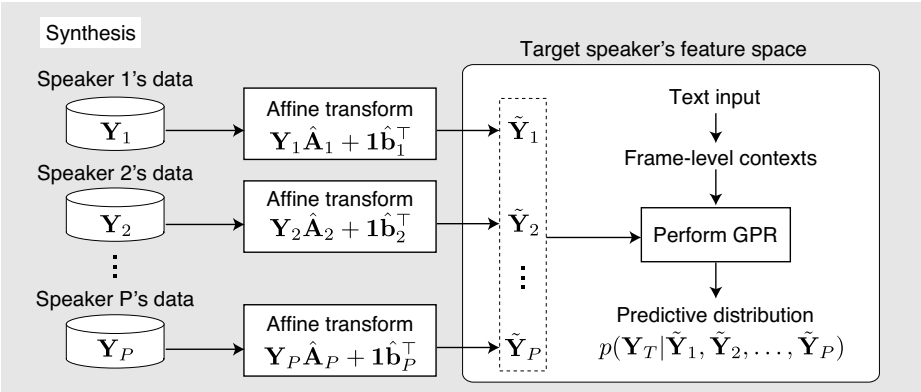

Figure 2: Outline of inference of predictive distribution in synthesis phase.

where T_A represents the number of frames in adaptation data.

Here, we train optimal transform parameters $\hat{\mathbf{A}}_i$ and $\hat{\mathbf{b}}_i$ by maximizing the predictive likelihood of (16). The optimal transform parameters can be obtained by setting following derivatives to zero.

$$\frac{\partial \mathcal{L}}{\partial \mathbf{A}_i} = \mathbf{V}^{-1} \left(\mathbf{Y}_A - \sum_{j=1}^{P} \mathbf{L}_j \tilde{\mathbf{Y}}_j \right)^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{Y}_i \tag{17}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_i} = \mathbf{V}^{-1} \left(\mathbf{Y}_A - \sum_{j=1}^{P} \mathbf{L}_j \tilde{\mathbf{Y}}_j \right)^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{1}_i. \tag{18}$$

This corresponds to solving the following equations:

$$\sum_{j=1}^{P} \mathbf{A}_j \mathbf{Y}_j^\top \mathbf{L}_j^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{Y}_i + \sum_{j=1}^{P} \mathbf{b}_j \mathbf{1}_j^\top \mathbf{L}_j^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{Y}_i = \mathbf{Y}_A^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{Y}_i \tag{19}$$

$$\sum_{j=1}^{P} \mathbf{A}_j \mathbf{Y}_j^\top \mathbf{L}_j^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{1}_i + \sum_{j=1}^{P} \mathbf{b}_j \mathbf{1}_j^\top \mathbf{L}_j^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{1}_i = \mathbf{Y}_A^\top \mathbf{\Sigma}_T^{-1} \mathbf{L}_i \mathbf{1}_i. \tag{20}$$

As a result, we can obtain the optimal parameters analytically.

4. EXPERIMENTS

4.1. Experimental conditions

We used ATR Japanese speech database set B [16] as speech data for evaluation. Source speakers were two males (MHO and MMY) and 450 utterances (about 35 to 40 minutes) for each speaker were used for training. Target speakers were two males (MHT and MSH) and the number of adaptation utterances varied from 5 to 50, which corresponds to about 20 sec to 4 min. The sentences for adaptation were included in the utterances of source speakers. 53 utterances, which were not included in neither source nor adaptation utterances, were used for synthesis. Speech signals were sampled at 16kHz and spectral envelope, aperiodicity, and F0 were extracted using STRAIGHT [17]. We used 0–39th mel-cepstrum, 5-band aperiodicity feature, and log F0, and their delta and delta-delta features.

We used partially independent conditional (PIC) approximation [15] for feasible computation of GPR, where the maximum number of frames in clusters was set to 1024 and pseudo data size was 1024. Phone-level clustering [12] was employed for the PIC approximation. The number of transforms was fixed to one for each speaker and we estimated the transform parameters for each acoustic feature individually.

We compared the proposed framework (GPR-SA) with HMM-based speaker adaptation using CSMAPLR with MAP modification [6]. HMM topology was 5-state hidden semi-Markov model (HSMM) with single mixture and a diagonal covariance matrix. Shared context decision tree clustering (STC) [18] and speaker adaptive training (SAT) [19] were employed to improve the performance. The number of transforms for the HMM-based method was determined according to the amount of adaptation data.

4.2. Objective evaluation

To evaluate the performance of proposed technique, we calculated acoustic feature distortions between original and synthetic speech. Figure 3 shows the average distortions of mel-cepstrum, log F0, and phone duration for the 53 test utterances as a function of the number of adaptation utterances. In the figure, the results of speaker dependent (SD) models, HMM-SD and GPR-SD, trained by 450 utterances are also shown. From the figure, we see that the proposed GPR-SA consistently gave lower distortions than HMM-SA for both speakers. Moreover, it is noted that the proposed technique using only 5 adaptation utterances reduced F0 and phone duration distortions compared with HMM-SD that used 450 utterances of the target speaker.

4.3. Subjective evaluation

To evaluate the perceptual quality of synthetic speech, we performed three subjective tests: MOS and preference test in terms of naturalness and XAB test for speaker similarity[1]. We compared HMM-SA and GPR-SA using 10 adaptation utterances. Synthetic speech samples were generated using global variance (GV) constraint [20] . In the MOS test, the listeners rated the naturalness of synthetic speech on a five-point scale: 5: excellent, 4: good, 3: fair, 2: poor, and 1: bad. In the preference test, participants could choose neutral if there was no preference between two methods. In the XAB test, vocoded speech samples were used as the reference. Seven participants listened to twenty speech samples in the MOS test, whereas eight participants listened to ten speech samples in the other tests. Speech

[1]Some synthetic speech samples used in the subjective evaluation are available at http://www.kbys.ip.titech.ac.jp/demo/gpradapt/koriyama/

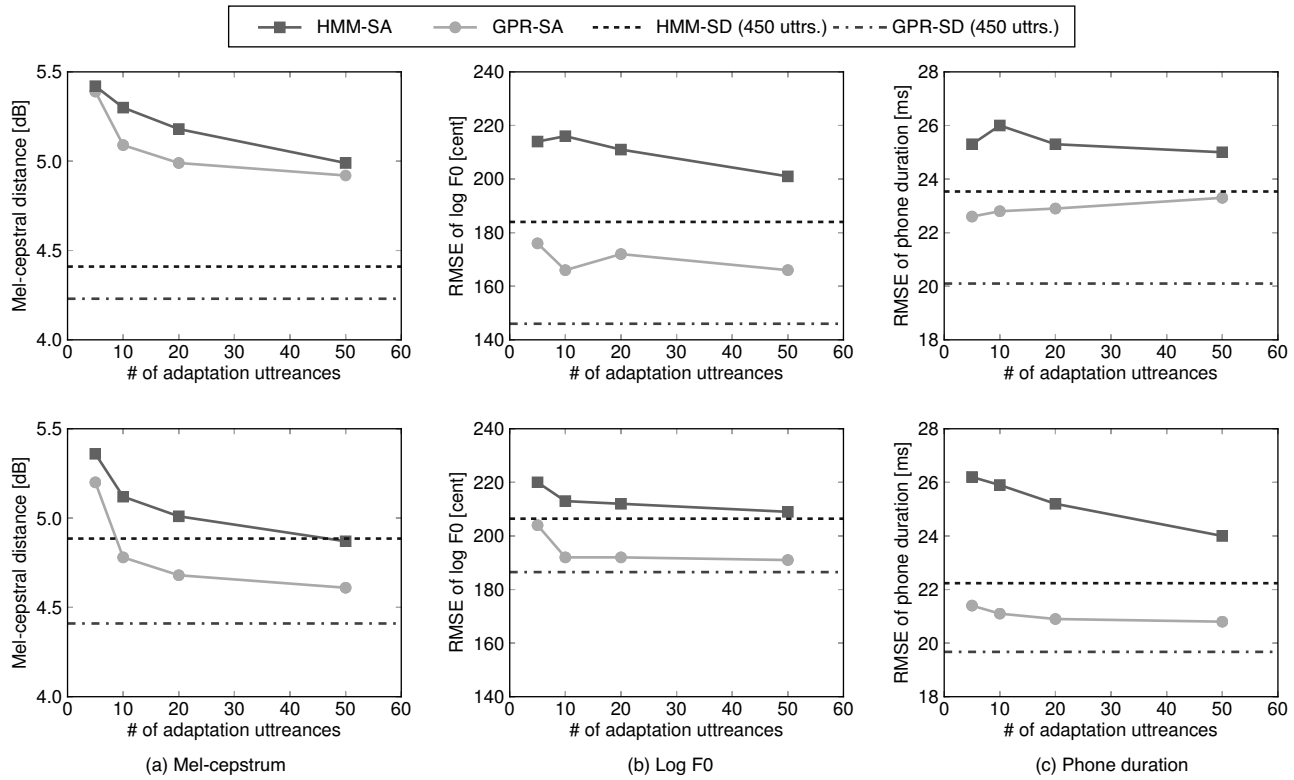

Figure 3: Acoustic feature distortions between original and synthetic speech. Upper and lower rows show the results of speaker MHT and MSH, respectively.

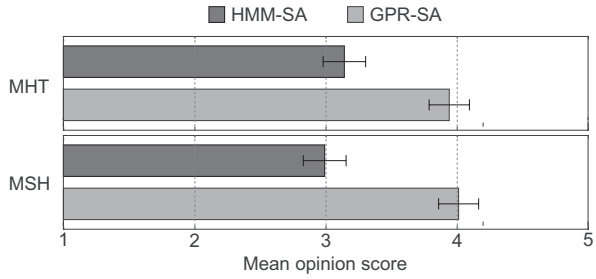

Figure 4: Results of MOS test in terms of naturalness.

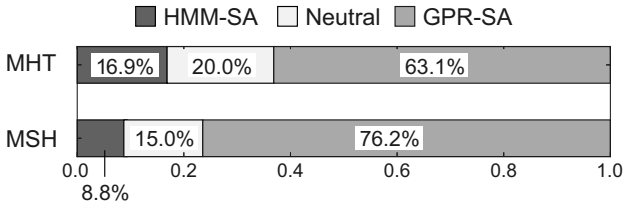

Figure 5: Results of preference test in terms of naturalness.

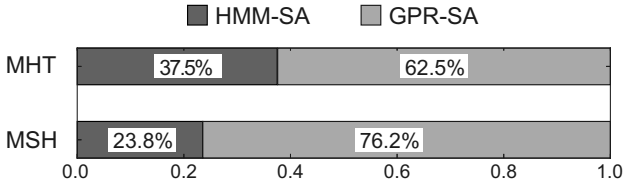

Figure 6: Results of XAB test in terms of speaker similarity.

samples were randomly chosen from the test sentences, for each target speaker. The results are shown in Figs. 4 to 6. It is seen from the figure that the proposed GPR-SA had significantly higher scores than the conventional HMM-SA in both naturalness and speaker similarity for the two target speakers.

5. CONCLUSIONS

In this paper, we have proposed a speaker adaptation technique for GPR-based speech synthesis. In the proposed technique, feature-space transform matrices to target speaker's acoustic feature space are introduced. We derived optimum transformation parameter estimation algorithm in a GPR framework for speech synthesis. Objective and subjective evaluation results showed that the proposed method outperformed the conventional HMM-based speaker adaptation. In future work, we should investigate the effect of speech data for training because we performed experiments under only limited conditions as the first step for establishing speaker adaptation in GPR-based synthesis.

6. 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] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

[3] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," in *Proc. ICASSP*, 2001, vol. 2, pp. 805–808.

[4] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 2, pp. 533–543, 2007.

[5] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech and Audio Process.*, vol. 3, no. 5, pp. 357–366, 1995.

[6] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 66–83, 2009.

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

[8] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. INTERSPEECH*, 2014, pp. 1964–1968.

[9] R. Fernandez, A. Rendel, B. Ramabhadran, and R. Hoory, "Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks," in *Proc. INTERSPEECH*, 2014, pp. 2268–2272.

[10] T. Koriyama, T. Nose, and T. Kobayashi, "Statistical parametric speech synthesis based on Gaussian process regression," *IEEE J-STSP*, vol. 8, no. 2, pp. 173–183, 2014.

[11] 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.

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

[13] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. ICASSP*, 2015, pp. 4475–4479.

[14] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. INTERSPEECH*, 2015, pp. 879–893.

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

[16] 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, 1990.

[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 Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[18] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A context clustering technique for average voice models," *IEICE Trans. Inf. & Syst.*, vol. 86, no. 3, pp. 534–542, 2003.

[19] T. Aanastasakos, "A compact model for speaker-adaptive training," *ICSLP*, vol. 2, 1996.

[20] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.