SPEAKER-INDEPENDENT STYLE CONVERSION FOR HMM-BASED EXPRESSIVE SPEECH SYNTHESIS

Hiroki Kanagawa[1], *Takashi Nose*[1], *Takao Kobayashi*[1],

[1]Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

kanagawa.h.aa@m.titech.ac.jp, {takashi.nose, takao.kobayashi}@ip.titech.ac.jp

ABSTRACT

This paper proposes a technique for creating target speaker's expressive-style model from the target speaker's neutral style speech in HMM-based speech synthesis. The technique is based on the style adaptation using linear transforms where speaker-independent transformation matrices are estimated in advance using pairs of neutral- and target-style speech data of multiple speakers. By applying the obtained transformation matrices to a new speaker's neutral-style model, we can convert the style expressivity of the acoustic model to the target style without preparing any target-style speech of the speaker. In addition, we introduce a speaker adaptive training (SAT) framework into the transform estimation to reduce the acoustic difference among speakers. We subjectively evaluate the performance of the style conversion in terms of the naturalness, speaker similarity, and style reproducibility.

Index Terms— HMM-based expressive speech synthesis, style conversion, style adaptation, linear transform, speaker adaptive training

1. INTRODUCTION

In the emotional speech synthesis research area, there are many studies for generating expressive speech from neutral-style speech by converting the global/local prosodic characteristics such as average fundamental frequency (F0) and speaking rate using heuristic rules [1]. These rule-based approaches are effective in some typical styles such as happy and sad. However, the conversion performance highly depends on the target style, and the style expressivity is not always satisfactory in some styles.

Recently, several techniques have been proposed to improve the conversion performance of the prosodic features [2, 3]. In [2], some linguistic features were integrated using CART to map the prosody distributions between neutral and emotional speech. In [3], an F0 segment selection approach was proposed where the F0 conversion was expressed as a search problem under contextual constraints. However, the evaluations of these techniques were conducted only under a speaker-dependent condition and the speaker-independent case was not discussed where a target speaker's expressive speech is not available.

In this paper, we propose a novel technique for converting an arbitrary speaker's neutral-style acoustic model to the target style without using any expressive speech of the target speaker. The technique is based on HMM-based speech synthesis with style adaptation [4] and average voice model [5]. In the style adaptation, the transforms estimated between neutral- and target-style speech of the same speaker can be viewed as style conversion functions. However, the obtained transforms are speaker-dependent, and the performance is not always satisfactory when we apply the transforms to the neutral-style model of another speaker. To alleviate this problem, we extend the idea of the style adaptation of a certain speaker to the speaker-independent case by estimating the linear transforms with multiple speaker's data of neutral and target styles. We also introduce the speaker adaptive training (SAT) [6, 7] into the transform estimation to improve the conversion performance.

2. STYLE CONVERSION OF ACOUSTIC MODEL USING SPEAKER-INDEPENDENT LINEAR TRANSFORMS

An outline of the proposed style conversion is shown in Fig. 1. The style conversion has two phases, the first is to obtain speaker-independent linear transforms for the style conversion, and the second is to convert the style expressivity of the target speaker's neutral-style model by applying the obtained transforms.

In the first phase, we prepare training data of multiple speakers, where each speaker utters neutral- and target-style speech. We train an average voice model using the neutral-style speech and estimate transforms from the neutral-style to the target-style using multiple speakers' data of target-style in a style adaptation framework [4]. In this study, we use hidden semi-Markov model [8] and choose a structural maximum a posteriori linear regression (SMAPLR) [9, 10] as an algorithm of the model adaptation. Since the linear transforms are estimated using multiple speakers' data of the neutral and the target styles, this process can be viewed as the speaker-independent linear transformation.

In the SMAPLR, the mean parameters of output and state-duration pdfs, $\boldsymbol{\mu} \in \mathcal{R}^{D\times 1}$ and m, of the neutral-style average voice model λ are transformed as

$$\tilde{\boldsymbol{\mu}} = \boldsymbol{J}\boldsymbol{\xi} \tag{1}$$

$$\tilde{m} = \boldsymbol{E}\boldsymbol{\eta} \tag{2}$$

where $\Lambda = (\boldsymbol{J}, \boldsymbol{E})$ represents a set of transformation matrices $\boldsymbol{J} \in \mathcal{R}^{D\times(D+1)}$ and $\boldsymbol{E} \in \mathcal{R}^{1\times 2}$ corresponding to output and state-duration probability density functions (pdfs), respectively. D is the dimensionality of the feature vector, $\boldsymbol{\xi} = [\boldsymbol{\mu}^\top \ 1]^\top \in \mathcal{R}^{(D+1)\times 1}$, and $\boldsymbol{\eta} = [m \ 1]^\top \in \mathcal{R}^{2\times 1}$. The optimal $\hat{\Lambda}$ is estimated using MAP criterion as

$$\hat{\Lambda} = \underset{\Lambda}{\operatorname{argmax}}\, P\left(\Lambda|\boldsymbol{O}, \lambda\right) = \underset{\Lambda}{\operatorname{argmax}}\, P\left(\Lambda\right) P\left(\boldsymbol{O}|\lambda, \Lambda\right) \tag{3}$$

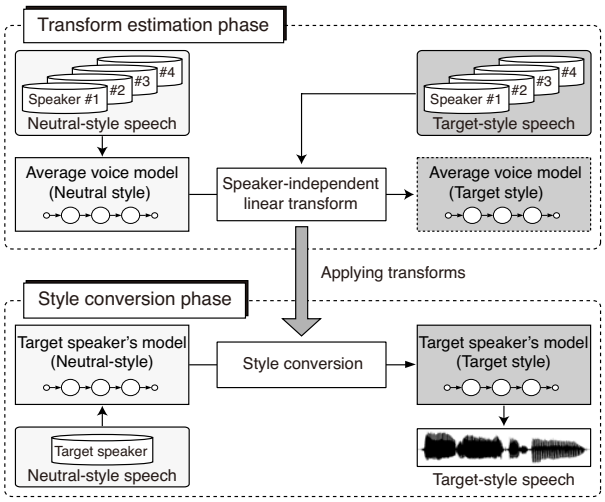


Fig. 1. An outline of speaker-independent style conversion.

where $\boldsymbol{O}$ is observation data for the adaptation, $P(\Lambda)$ is a prior pdf in MAP estimation and is given by

$$P(\Lambda) \propto |\boldsymbol{\Omega}|^{-(D+1)/2} |\boldsymbol{\Psi}|^{-D/2} \tau_p^{-1} |\boldsymbol{\psi}|^{-1/2} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{J}-\boldsymbol{G}\right)^\top \boldsymbol{\Omega}^{-1}\left(\boldsymbol{J}-\boldsymbol{G}\right)\boldsymbol{\Psi}^{-1}\right\} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{E}-\boldsymbol{H}\right)^\top \tau_p^{-1}\left(\boldsymbol{E}-\boldsymbol{H}\right)\boldsymbol{\psi}^{-1}\right\} \quad (4)$$

where $\boldsymbol{\Psi} \in \mathcal{R}^{(D+1)\times(D+1)}$, $\boldsymbol{G} \in \mathcal{R}^{D\times(D+1)}$, $\boldsymbol{\psi} \in \mathcal{R}^{2\times 2}$, and $\boldsymbol{H} \in \mathcal{R}^{1\times 2}$ is hyper parameters for the prior distribution. In this study, we use $\boldsymbol{\Omega} = \tau_b \boldsymbol{I}_{D+1}$, $\boldsymbol{\Psi} = \boldsymbol{I}_{D+1}$, and $\boldsymbol{\psi} = \boldsymbol{I}_2$, which is the same as the setting in the previous study [10]. τ_b and τ_p are positive constant values to control the effect of the prior distribution in MAP estimation, and $\boldsymbol{G}$ and $\boldsymbol{H}$ are transformation matrices in the parent node of $\boldsymbol{J}$ and $\boldsymbol{E}$, respectively.

In the second phase, we apply the speaker-independent transforms to the target speaker's neutral-style model and obtain the target-style model of the target speaker. Finally, we generate synthetic speech from the converted model using the ordinary parameter generation method of the HMM-based speech synthesis. It is noted that the proposed technique does not require any target-style speech of the target speaker in generating the synthetic speech of the style, which is an advantage of the technique compared to the conventional style adaptation approach.

3. NORMALIZING SPEAKER DIFFERENCE IN TRANSFORM ESTIMATION

3.1. CMLLR-based SAT for style conversion

From a preliminary experimental result, we found that the naturalness of the synthetic speech was degraded when the style characteristics of respective speakers in the transform estimation were much different from each other. To alleviate the problem, we employ the idea of speaker adaptive training (SAT) [6, 7] that is a well-known speaker normalization technique for the model training in ASR and TTS. In the ordinary SAT for speech synthesis, the parameters of an average voice model are refined using the transforms from the average voice model to respective speakers' model. By contrast, the transform set Λ is refined in the proposed SAT for style conversion.

First, we create the target-style average voice model λ' by applying the speaker-independent linear transforms to the neutral-style average voice model. Then, we estimate speaker-dependent transform sets $\theta^{(f)} = (\boldsymbol{W}^{(f)}, \boldsymbol{X}^{(f)})$ from λ' to the target-style data of each speaker f where $1 \leq f \leq F$ and F is the number of speakers of the average voice models. $\boldsymbol{W}^{(f)} \in \mathcal{R}^{D\times(D+1)}$ and $\boldsymbol{X}^{(f)} \in \mathcal{R}^{1\times 2}$ are transformation matrices of output and state-duration pdfs for speaker f.

To simplify the derivation of estimation formulas based on SAT, we use feature-space linear transformation based on constrained maximum likelihood linear regression (CMLLR) [11] as an adaptation algorithm to respective speakers. In the CMLLR, the t-th frame observation $\boldsymbol{o}_t^{(f)} \in \mathcal{R}^{D\times 1}$ and duration d of speaker f are transformed to $\tilde{\boldsymbol{o}}_t^{(f)}$ and $\tilde{d}$ as

$$\tilde{\boldsymbol{o}}_t^{(f)} = \boldsymbol{W}^{(f)} \boldsymbol{\zeta}_t^{(f)} \quad (5)$$

$$\tilde{d}^{(f)} = \boldsymbol{X}^{(f)} \boldsymbol{\phi}^{(f)} \quad (6)$$

where

$$\boldsymbol{\zeta}_t^{(f)} = \left[\boldsymbol{o}_t^{(f)\top}\ 1\right]^\top \quad (7)$$

$$\boldsymbol{\phi}^{(f)} = \left[d^{(f)}\ 1\right]^\top \quad (8)$$

Then, the output and state-duration pdfs of speaker f are given by

$$b_i\left(\boldsymbol{o}_t^{(f)}\right) = \frac{1}{\sqrt{(2\pi)^D |\boldsymbol{\Sigma}_i|}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{W}_i^{(f)}\boldsymbol{\zeta}_t^{(f)} - \boldsymbol{J}_i\boldsymbol{\xi}_i\right)^\top \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{W}_i^{(f)}\boldsymbol{\zeta}_t^{(f)} - \boldsymbol{J}_i\boldsymbol{\xi}_i\right)\right\} \quad (9)$$

$$p_i\left(d^{(f)}\right) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2}\left(\boldsymbol{X}_i^{(f)}\boldsymbol{\phi}^{(f)} - \boldsymbol{E}_i\boldsymbol{\eta}_i\right)^2\right\} \quad (10)$$

where $\boldsymbol{\Sigma}_i \in \mathcal{R}^{D\times D}$ and σ_i are variance parameters of output and state-duration pdfs of the target-style average voice model, respectively.

3.2. Estimation of speaker-normalized linear transforms

When the optimal speaker transform set $\hat{\Theta} = \left\{\hat{\theta}^{(1)}, \cdots, \hat{\theta}^{(F)}\right\}$ is given by the CMLLR described in Sect. 3.1, a refined style transform set $\hat{\Lambda}$ is estimated using a SAT framework as

$$\hat{\Lambda} = \underset{\Lambda}{\mathrm{argmax}}\, P(\Lambda) \prod_{f=1}^{F} P\left(\boldsymbol{O}^{(f)} | \lambda, \hat{\theta}^{(f)}, \Lambda\right) \quad (11)$$

The auxiliary functions of output and state-duration pdfs for the EM algorithm are given by

$$\mathcal{Q}_b\left(\Lambda;\bar{\Lambda}\right) = \sum_{r=1}^{R}\sum_{f=1}^{F}\sum_{t=1}^{T^{(f)}}\sum_{d=1}^{t}\gamma_r^d(t) \times \sum_{s=t-d+1}^{t}\ln b_r\left(\boldsymbol{o}_s^{(f)}\right) + \ln P(\Lambda) \quad (12)$$

$$\mathcal{Q}_p\left(\Lambda;\bar{\Lambda}\right) = \sum_{r=1}^{R}\sum_{f=1}^{F}\sum_{t=1}^{T^{(f)}}\sum_{d=1}^{t}\gamma_r^d(t)\ln p_r(d) + \ln P(\Lambda) \quad (13)$$

where transforms are tied across R pdfs. By differentiating the auxiliary functions $\mathcal{Q}_b$ and $\mathcal{Q}_p$ with respect to $\boldsymbol{J}$ and $\boldsymbol{E}$, and equating the results to zero, we have

$$\hat{\boldsymbol{j}}_l = \left(\sum_{r=1}^{R}\sum_{f=1}^{F}\sum_{t=1}^{T^{(f)}}\sum_{d=1}^{t}\frac{\gamma_r^d(t)}{\Sigma_r(l)}\sum_{s=t-d+1}^{t}\hat{\boldsymbol{w}}_l^{(f)}\boldsymbol{\zeta}_s^{(f)}\boldsymbol{\xi}_r^{\top} + \tau_b^{-1}\boldsymbol{g}_l\right) \times \left(\sum_{r=1}^{R}\sum_{f=1}^{F}\sum_{t=1}^{T^{(f)}}\sum_{d=1}^{t}\frac{d\gamma_r^d(t)}{\Sigma_r(l)}\boldsymbol{\xi}_r\boldsymbol{\xi}_r^{\top} + \tau_b^{-1}\boldsymbol{I}_{D+1}\right)^{-1} \quad (14)$$

$$\hat{\boldsymbol{E}} = \left(\sum_{r=1}^{R}\sum_{f=1}^{F}\sum_{t=1}^{T^{(f)}}\sum_{d=1}^{t}\frac{\gamma_r^d(t)}{\sigma_r^2}\hat{\boldsymbol{X}}^{(f)}\boldsymbol{\phi}_d^{(f)}\boldsymbol{\eta}_r^{\top} + \tau_p^{-1}\boldsymbol{H}\right) \times \left(\sum_{r=1}^{R}\sum_{f=1}^{F}\sum_{t=1}^{T^{(f)}}\sum_{d=1}^{t}\frac{\gamma_r^d(t)}{\sigma_r^2}\boldsymbol{\eta}_r\boldsymbol{\eta}_r^{\top} + \tau_p^{-1}\boldsymbol{I}_2\right)^{-1} \quad (15)$$

where $\boldsymbol{j}_l \in \mathcal{R}^{1\times(D+1)}$ is the l-th row vector of $\boldsymbol{J}$, and $\boldsymbol{g}_l \in \mathcal{R}^{1\times(D+1)}$ is l-th row vector of $\boldsymbol{G}$. The estimation process of transformation matrices using SAT is summarized as follows:

1. Estimate the initial transform set Λ for style conversion.
2. Obtain the target-style average voice model λ' by applying Λ to the neutral-style average voice model λ.
3. Estimate the speaker transform set Θ using λ' and the target-style data of respective speakers.
4. Update Λ given Θ using Eqs. (14) and (15).
5. Repeat step 2 to 4 until Λ and Θ converge.

4. EXPERIMENTS

4.1. Experimental conditions

We evaluated the performance of the proposed style conversion technique using neutral- and appealing-style speech data. We used parallel speech data of three female professional narrators described in [12]. Furthermore, we recorded additional parallel data of two female professional narrators, and the total number of speakers was five (speaker #1 to #5). Each speaker uttered 176 sentences in both neutral and appealing styles. The appealing-style speech was uttered under a condition where a salesclerk spoke to customers to push some products through mass media commercials. Speech signals were sampled at a rate of 16kHz and the frame shift was 5 ms. We used STRAIGHT analysis [13] for speech feature extraction, and extracted spectral envelope, F0, and aperiodicity features. The spectral envelope was then converted to mel-cepstral coefficients using a recursion formula. The aperiodicity feature was also converted to average values for five frequency sub-bands, i.e., 0–1, 1–2, 2–4, 4–6, and 6–8 kHz. As a result, the feature vector consisted of 39 mel-cepstral coefficients including the zeroth coefficient, log F0, five-band aperiodicity values, and their delta and delta-delta coefficients. The total dimensionality was 138. We used five-state left-to-right HSMM with no skip topology. The output distribution in each state was modeled with a single Gaussian probability density function, and covariance matrices of these models were assumed to be diagonal. In addition, we used SAT algorithm to train the neutral-style average voice model [7].

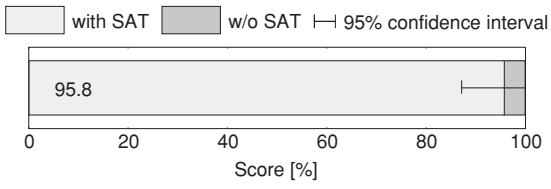
Table 1. Distortions of mel-cepstral, log F0, and duration between original and synthetic speech of the target style.

Target speaker	Method	Mcep [dB]	LogF0 [cent]	Dur [msec]
#1	w/o SAT	6.87	406	29.8
	with SAT	6.62	406	31.4
#2	w/o SAT	6.77	508	28.5
	with SAT	6.29	480	28.7
#3	w/o SAT	7.11	452	31.7
	with SAT	6.72	429	34.3
#4	w/o SAT	7.35	317	44.5
	with SAT	7.05	312	35.8
#5	w/o SAT	7.17	434	34.1
	with SAT	6.84	436	31.1
Average	w/o SAT	7.05	423	33.7
	with SAT	6.70	413	32.2

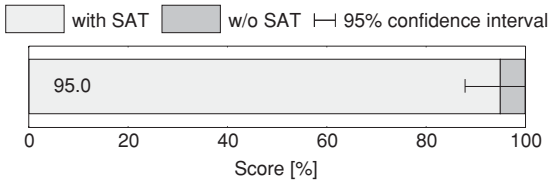
4.2. Objective evaluation

We chose one speaker as a target speaker and used the rest of the speakers to estimate speaker-independent transforms for style conversion, and repeated this evaluation five times by changing the target speaker. Furthermore, we divided 176 utterances into 6 subsets, each of which contains about 30 utterances, and conducted six-fold cross-validation to obtain evaluation results for each target speaker. As the objective measure, we used mel-cepstral distance (Mcep), root mean square errors of log F0 (LogF0) and duration (Dur). In this experiment, we compared these objective measures for the proposed technique with and without SAT.

Table 1 shows the results for respective speakers and their average. From the result, first we see that the conversion with SAT outperformed that without SAT in terms of the total conversion performance. In detail, it is seen that the spectral distortions were consistently reduced by using the SAT. As for the prosody conversion, the transform estimation with SAT also reduced the distortions especially when the distortion values were relatively large, which indicates the SAT is effective for the speakers whose acoustic characteristics are much different from the average voice model.



(a) with spectrum transformation



(b) without spectrum transformation

Fig. 2. Result of the preference test with and without SAT.

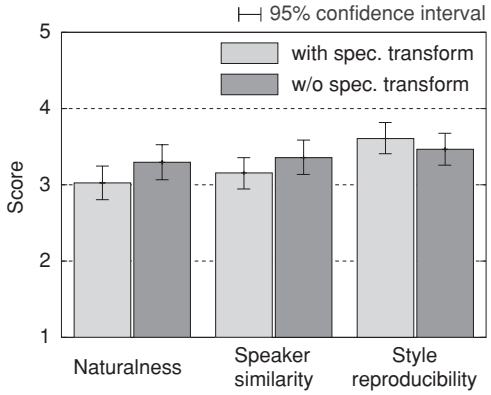


Fig. 3. Mean opinion scores of synthetic speech on naturalness, speaker similarity and style reproducibility with and without spectrum transformation.

4.3. Subjective evaluation

In the subjective evaluation tests, we used parameter generation algorithm considering global variance (GV) [14] to improve the perceptual quality of the synthetic speech. Since we cannot use any target-style speech of the target speaker, the target speaker's target style GV model need to be construct without these speech data. Therefore we first constructed an average voice GV model using multiple speaker's neutral-style speech data in the framework of training for context-dependent GV model [15]. Secondly, we estimated speaker-independent transform from the neutral style to the target style using a style adaptation framework. Finally, by applying the estimated transform to target speaker's neutral-style GV model, we obtained target speaker's neutral-style one. The number of participants for subjective evaluation was seven.

First, we compared the naturalness of the synthetic speech with and without SAT through a preference test to examine the effectiveness of SAT-based transform estimation in style conversion. We randomly chose eight samples of synthetic speech with and without SAT for each participant. Then, each participant listened to the speech samples of the two methods in random order and was asked which sample was more natural. We performed the evaluation for two types of synthetic speech. One was synthetic speech where all speech features are converted to the target style. The other was synthetic speech where the spectral feature was not converted and that of the neutral-style was used instead. Fig. 2 shows the result. We can see that the naturalness of the synthetic speech was significantly improved by using SAT.

Next, we examined the performance of the proposed technique with SAT in terms of the naturalness, speaker similarity, and style reproducibility. In this experiment, we evaluated synthetic speech samples with and without spectrum transformation to examine the effective of spectrum transformation in style conversion. For each participant, we randomly chose eight samples of synthetic speech for each method. Then, participants rated the naturalness, speaker similarity, and style reproducibility of test samples using a five-point scale: "1" for bad, "2" for poor, "3" for fair, "4" for good, and "5" for excellent. In the evaluation of speaker similarity, synthetic speech generated from target speaker's neutral-style model was used as the reference. Fig. 3 shows the result. From the result, we see that the proposed technique generated synthetic speech similar to the target speaker and style while keeping the naturalness. It is also found that the spectral conversion slightly improved the style reproducibility but degraded the naturalness.

5. CONCLUSIONS

We have proposed a novel style conversion technique for creating expressive-style model of a certain speaker without using his/her target-style data in an HMM-based speech synthesis framework. The technique estimates speaker-independent linear transforms for the style conversion using multiple speakers' data of neutral and target styles with a SAT framework. The experimental results have shown that the SAT-based speaker normalization in the transform estimation is effective and the performance of the proposed technique with prosody conversion is between fair and good in terms of the naturalness, speaker similarity and style reproducibility. The future work will include the evaluation of the proposed technique using speech data of other styles.

6. ACKNOWLEDGMENTS

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 23700195 and 24300071.

7. REFERENCES

[1] M. Schröder, "Emotional speech synthesis: A review," in *Proc. EUROSPEECH 2001*, 2001, pp. 561–564.

[2] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1145–1154, 2006.

[3] Z. Inanoglu and S. Young, "Data-driven emotion conversion in spoken english," *Speech Communication*, vol. 51, no. 3, pp. 268–283, 2009.

[4] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1092–1099, 2006.

[5] M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "Text-to-speech synthesis with arbitrary speaker's voice from average voice," in *Proc. EUROSPEECH*, 2001, pp. 345–348.

[6] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP-96*, 1996, pp. 1137–1140.

[7] 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.

[8] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, 2007.

[9] O. Shiohan, Y. Myrvoll, and C.H. Lee, "Structural maximum a posteriori linear regression for fast HMM adaptation," *Computer Speech & Language*, vol. 16, no. 3, pp. 5–24, 2002.

[10] 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.

[11] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, pp. 75–98, 1998.

[12] H. Nakajima, N. Miyazaki, A. Yoshida, T. Nakamura, and H. Mizuno, "Creation and analysis of a Japanese speaking style parallel database for expressive speech synthesis," in *Proc. Oriental COCOSDA*, 2010, http://desceco.org/O-COCOSDA2010/proceedings/paper_30.pdf.

[13] 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.

[14] 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.

[15] J. Yamagishi, H. Zen, Y.J. Wu, T. Toda, and K. Tokuda, "The hts-2008 system: Yet another evaluation of the speaker-adaptive hmm-based speech synthesis system in the 2008 blizzard challenge," in *Proc. Blizzard Challenge 2008 Workshop*, 2008.