# SPEAKER AND STYLE ADAPTATION USING AVERAGE VOICE MODEL FOR STYLE CONTROL IN HMM-BASED SPEECH SYNTHESIS

*Makoto Tachibana, Shinsuke Izawa, Takashi Nose, Takao Kobayashi*

Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology, Yokohama, 226-8502 Japan
Email: {makoto.tachibana,shinsuke.izawa,takashi.nose,takao.kobayashi}@ip.titech.ac.jp

## ABSTRACT

We propose a technique for synthesizing speech with desired style expressivity of an arbitrary target speaker's voice. In an MLLR-based speaker adaptation technique for multiple regression hidden semi-Markov model (MRHSMM), the quality of synthesized speech crucially depends on the initial MRHSMM trained from a certain source speaker's data and it is not always possible to synthesize natural sounding speech with a given target speaker's voice. To overcome this problem, we perform simultaneous adaptation of speaker and style from an average voice model. Experimental results show that the proposed technique provides more natural sounding speech than the conventional one with speaker adaptation only.

***Index Terms***— expressive speech synthesis, style control, hidden Markov model, speaker adaptation, average voice model

## 1. INTRODUCTION

Recently, there is a demand for speech synthesis system capable of expressing various emotions and speaking styles with diverse speaker characteristics. Although many attempts have been made to synthesize expressive speech [1–3], most of them have not always been successful in diversifying styles as well as speaker characteristics. For this purpose, we have proposed a style control technique based on multiple regression hidden semi-Markov model (MRHSMM) [4] in an HMM-based speech synthesis framework. In the MRHSMM-based style control, the mean parameter of the model is given by multiple regression of a low dimensional vector called style vector. By changing the value of the style vector, we can control the degree of the style expressivity in a simple way. Moreover, we have also proposed an MLLR-based speaker adaptation technique for MRHSMM [5] for the style control of arbitrary speaker's voice with only a small amount of target speaker's data.

In our previous work [5], an initial MRHSMM was trained using a sufficient amount of speech data of a certain source speaker, and adapted to a target speaker's model using a small amount of adaptation data. However, the quality of synthesized speech of the adapted model depends on the given source speaker's model crucially and it is not always possible to synthesize natural sounding speech of arbitrary target speakers. A possible approach to solving this problem is to use several types of initial MRHSMM or speaker-independent MRHSMM trained from multiple speakers. However, it is not easy to prepare a large amount of multiple speakers' data for all styles and add a new target style from the viewpoint of recording cost. Moreover, speech characteristics of each style vary depending on individuals and speaker-independent MRHSMM might not be trained appropriately.

To overcome this problem, we propose an alternative approach to training MRHSMM using an average voice model [6] and simultaneous adaptation of speaker and style. The average voice model is a speaker-independent neutral style HSMM trained from multiple speakers' neutral style speech. We adapt the average voice model to target speaker's styles using a technique for simultaneous adaptation of speaker and style. Then, the initial MRHSMM is estimated by the least square method from the adapted HSMMs. This initial MRHSMM will decrease dependency on the source speaker's characteristics of the initial model.

## 2. MODEL ADAPTATION FOR MRHSMM

### 2.1. Style Control Based on MRHSMM

In the MRHSMM-based style control technique [4], each speech synthesis unit is modeled by an $N$-state HSMM. It is assumed that the output probability density functions (pdfs) $b_i(\boldsymbol{o})$ and state duration pdf $p_i(d)$ at state $i$ are characterized by mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$, and mean $m_i$ and variance $\sigma_i^2$, respectively. In MRHSMM, we further assume that $\boldsymbol{\mu}_i$, $m_i$ are modeled using multiple regression as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \quad m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi} \tag{1}$$

where

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{2}$$

and $\boldsymbol{v}$ is a style vector, $L$ is the dimensionality of the style space. The component $v_k$ of the style vector represents the degree or intensity of a certain style in speech. In addition, $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are regression matrices of dimension $M \times (L+1)$ and $1 \times (L+1)$ respectively, and $M$ is the dimensionality of $\boldsymbol{\mu}_i$.

When the training data and corresponding style vectors are given, the parameters of MRHSMM, i.e. $\boldsymbol{H}_{b_i}$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{H}_{p_i}$, and $\sigma_i^2$ can be estimated using EM algorithm [4].

### 2.2. Model Adaptation for MRHSMM

In the MLLR-based model adaptation technique for MRHSMM [5], the mean vector of the output pdf of the target speaker's model is assumed to be given by an affine transformation of that of the initial model as follows:

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{b}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{\mu}_i \tag{3}$$

where $\boldsymbol{\mu}_i$ and $\hat{\boldsymbol{\mu}}_i$ are the mean vectors of the initial model and target speakers' model. $\boldsymbol{A}_{b_i}$ is transformation matrix and $\boldsymbol{b}_{b_i}$ is a bias vector. In MRHSMM, since $\boldsymbol{\mu}_i$ and $\hat{\boldsymbol{\mu}}_i$ are given by multiple regression

of the style vector as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \quad \hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi}, \tag{4}$$

Eq. (3) becomes

$$\hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi} = \boldsymbol{b}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}\boldsymbol{\xi}. \tag{5}$$

We assume that the bias term $\boldsymbol{b}_{b_i}$ is also given by multiple regression of the style vector as $\boldsymbol{b}_{b_i} = \boldsymbol{B}_{b_i}\boldsymbol{\xi}$, Eq. (5) is rewritten as

$$\hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi} = (\boldsymbol{B}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i})\,\boldsymbol{\xi}. \tag{6}$$

Consequently, the linear transformation for the output pdf is given by

$$\hat{\boldsymbol{H}}_{b_i} = \boldsymbol{B}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}. \tag{7}$$

Similarly, the linear transformation for the state duration pdf is given by

$$\hat{\boldsymbol{H}}_{p_i} = \boldsymbol{B}_{p_i} + \boldsymbol{A}_{p_i}\boldsymbol{H}_{p_i}. \tag{8}$$

Estimation formulas of these transformation matrices can be found in [5].

## 3. MRHSMM TRAINING FROM AVERAGE VOICE MODEL

### 3.1. Overview of the Proposed Training Method

An outline of the conventional and proposed methods is shown in Fig. 1. In our previous work [5], an initial MRHSMM is trained using a sufficient amount of speech data of a source speaker, and adapted to a target speaker's model using a small amount of adaptation data. On the other hand, the proposed method utilizes the average voice model trained by multiple speakers' neutral style speech data. We adapt the average voice model to target speaker's styles using a technique for simultaneous adaptation of speaker and style. Although it would be possible to synthesize target speaker's specific style speech from the speaker- and style-adapted HSMM, we further train MRHSMM using the adapted HSMMs to enable us to control the degree of the style expressivity, More specifically, the initial MRHSMM is estimated using the least square method from each speaker- and style-adapted HSMM.By using this initial MRHSMM, we can decrease dependency on the source speaker's characteristics of the initial MRHSMM. Then, in the same manner as the conventional method, the initial MRHSMM is adapted to the target speaker's MRHSMM by MLLR-based adaptation technique. Furthermore, the adapted model is modified using ML estimation.

### 3.2. Simultaneous Adaptation of Speaker and Style

For neutral reading style speech, we have already shown that the speech synthesis method using an average voice model and speaker adaptation technique is effective when only a small amount of target speaker's data is available [6]. Moreover, in a similar way to the speaker adaptation technique, style adaptation technique is capable for converting neutral style speech into another style [7]. In this study, we adapt not only speaker's characteristics but also characteristics of each style simultaneously and obtain speaker- and style-adapted HSMM in each style.

In the simultaneous adaptation of speaker and style, the mean vector of average voice model and covariance matrix of output pdf $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and mean and variance of the state duration pdf $m_i$,$\sigma_i^2$ are linearly transformed as follows:

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{\zeta}\boldsymbol{\mu}_i - \boldsymbol{\epsilon}, \qquad \hat{\boldsymbol{\Sigma}}_i = \boldsymbol{\zeta}\boldsymbol{\Sigma}_i\boldsymbol{\zeta}^\top \tag{9}$$

$$\hat{m}_i = \chi m_i - \nu, \qquad \hat{\sigma}_i^2 = \chi\sigma_i^2\chi \tag{10}$$

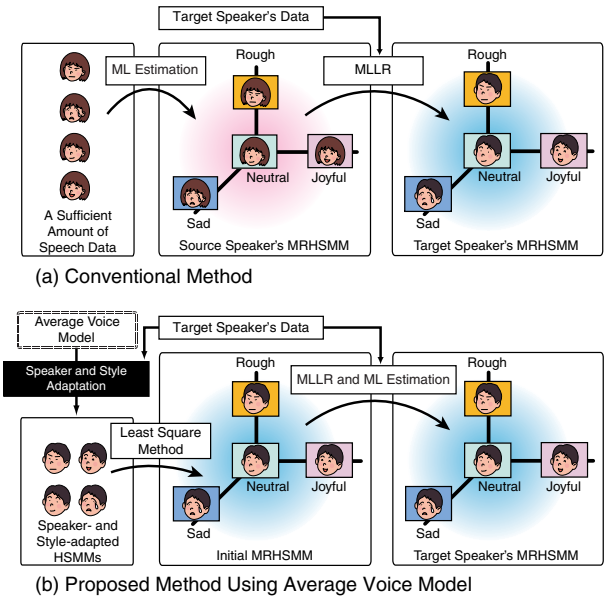


**Fig. 1**. An outline of conventional and proposed methods.

where $\boldsymbol{\zeta}$, $\boldsymbol{\epsilon}$ are transformation matrix and bias vector for output pdf and $\chi$, $\nu$ are transformation coefficient and bias term, respectively. We use HSMM-based constrained structural maximum a posteriori linear regression (CSMAPLR) algorithm [8] as the linear transform algorithm.

### 3.3. Calculation of Initial MRHSMM by Least Square Method

The regression matrices of MRHSMM are calculated from the parameter of each style-adapted HSMM obtained in **3.2** by the least square method. Suppose that speech database contains $S$ styles and each style mean vector and corresponding style vector is given by $\boldsymbol{\mu}_i^{(s)}(1 \le s \le S)$, $\boldsymbol{\xi}^{(s)}$, respectively. We choose $\boldsymbol{H}_{b_i}$ that minimizes

$$E = \sum_{s=1}^{S} \left\| \boldsymbol{\mu}_i^{(s)} - \boldsymbol{H}_{b_i}\boldsymbol{\xi}^{(s)} \right\|^2 \tag{11}$$

as the regression matrices of the initial MRHSMM. By differentiating $E$ with $\boldsymbol{H}_{b_i}$ and setting the result zero, we have

$$\overline{\boldsymbol{H}}_{b_i} = \left( \sum_{s=1}^{S} \boldsymbol{\mu}_i^{(s)}\boldsymbol{\xi}^{(s)\top} \right) \left( \sum_{s=1}^{S} \boldsymbol{\xi}^{(s)}\boldsymbol{\xi}^{(s)\top} \right)^{-1}. \tag{12}$$

In addition, the regression matrices for the state duration pdf $\overline{\boldsymbol{H}}_{p_i}$ can be estimated in a similar way. In this study, $\overline{\boldsymbol{H}}_{b_i}$, $\overline{\boldsymbol{H}}_{p_i}$ are used as the initial model for MLLR-based model adaptation.

### 3.4. Model Modification Using ML estimation

By using ML estimation, we further modify the regression matrix $\hat{\boldsymbol{H}}_{b_i}$ as follows:

$$\boldsymbol{H}_{b_i} = \frac{\tau_{out}\hat{\boldsymbol{H}}_{b_i} + \Gamma_{out}(i)\boldsymbol{H}_{b_i}^{ML}}{\tau_{out} + \Gamma_{out}(i)} \tag{13}$$

**Table 1**. Evaluated MRHSMMs.

| Model | Initial Model | Adaptation Data |
|---|---|---|
| A | FTY (450 sent.×4 styles) | 50 sent.×4 styles |
| B | MMI (450 sent.×4 styles) | 50 sent.×4 styles |
| C | MJI (450 sent.×4 styles) | 50 sent.×4 styles |
| D | target speaker (50 sent.×4 styles) | no adaptation |
| E | target speaker (450 sent.×4 styles) | no adaptation |
| F (proposed) | average voice model (450 sent.×9 persons) | 50 sent.×4 styles |

where

$$\Gamma_{out}(i) = \sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i) \cdot d. \quad (14)$$

$\hat{\boldsymbol{H}}_{b_i}$ is the regression matrix transformed by MLLR-based adaptation in Eq. (7), and $\boldsymbol{H}_{b_i}^{ML}$ is the regression matrix estimated from the adaptation data in ML sense using EM algorithm [4]. $\tau_{out}$ is a positive parameter used to control the modification weight, $K$ is the total number of the observation sequences, $T^{(n)}$ is the number of frames of the the $n$-th observation sequence $\boldsymbol{O}^{(n)}$, and $\gamma_t^d(i)$ is the probability of being in state $i$ at period of time from $t-d+1$ to $t$ given $\boldsymbol{O}^{(n)}$. When enough adaptation data is available at state $i$, the regression matrix $\boldsymbol{H}_{b_i}$ approaches to the $\boldsymbol{H}_{b_i}^{ML}$. $\boldsymbol{H}_{p_i}$ is modified in a similar manner.

The effect of this modification is similar to that of a combined approach based on maximum a posteriori (MAP) adaptation [9] for HSMM [10].

## 4. EXPERIMENTS

### 4.1. Experimental Conditions

We used four styles of read speech — neutral, sad, joyful, and rough (impolite) styles. Speech database contains 503 phonetically balanced ATR Japanese sentences uttered by two male and one female professional narrators, MMI, MJI and FTY, respectively, in each style, and is the same one used in our previous study [5]. The average voice model was trained using five male speakers and four female speakers' utterances taken from the ATR Japanese speech database (Set B). The training data were 450 sentences for each speaker, 4050 sentences in total. In the training stage of the average voice models, The shared-decision-tree-based context clustering (STC) algorithm and the speaker adaptive training (SAT) [11] were applied.

Speech signals were sampled at a rate of 16kHz and windowed by a 25-ms Blackman window with a 5-ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. The feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. We used 5-state left-to-right MRHSMM with diagonal covariance. A three-dimensional style space [4] was used as shown in Fig. 1, and style vectors of training and adaptation data were set as (0,0,0) for the neutral style, (1,0,0), (0,1,0), and (0,0,1) for the sad, rough, and joyful styles, respectively.

Table 1 shows the evaluated MRHSMMs. In the table, "A," "B," and "C" are models obtained by the conventional speaker adaptation method from a source speaker's model, "D" and "E" are speaker-dependent MRHSMMs of the target speaker, "F" is the model obtained by the proposed method using average voice model, respectively. In models A, B, C, and F, the adaptation data were target speaker's 50 sentences in each style, 200 sentences in total and the adaptation was performed using MLLR-based adaptation and ML estimation.

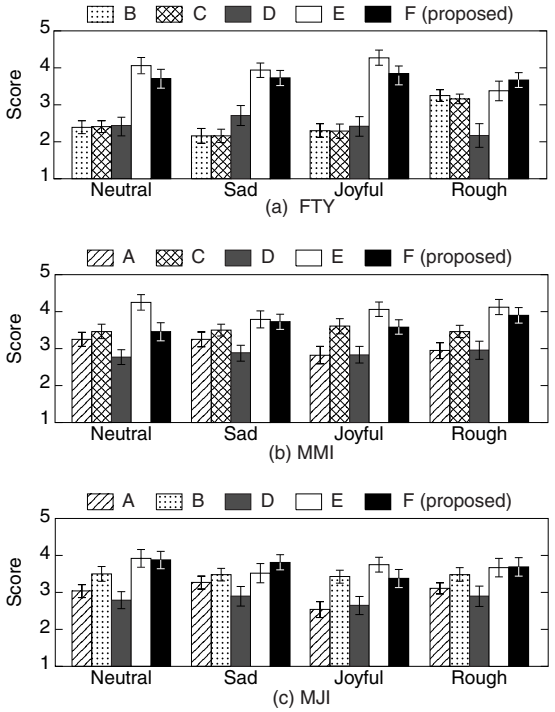


**Fig. 2**. Evaluation of adaptation performance for target speakers (a) FTY, (b) MMI, and (c) MJI.

Subjects were eight persons in all tests. For each subject, eight test sentences were chosen at random from 53 test sentences that were contained in neither the training data nor adaptation data [1].

### 4.2. Subjective Evaluation of Reproducibility

We first evaluated reproducibility of styles and speaker's characteristics by a Comparison Category Rating (CCR) tests. The scale for the CCR test was 5 for very similar and 1 for very dissimilar to reference speech. The reference speech was a target speaker's real utterance with a mel-cepstral vocoder. Test samples were generated from the MRHSMMs with the same style vector used for training in each style. Figure 2 shows the scores with 95% confidence interval of the test. From the result, we can see that the reproducibility of the synthesized speech from adapted models depends on the initial model when the adaptation was performed from a specific speaker's model (A, B, and C). Especially, the similarity between the synthetic speech and the reference speech was greatly decreased when the adaptation was performed from the male speaker's model to the female speaker FTY. On the other hand, the score of the proposed technique (model F) is stable for all target speakers and styles. Since the correct classification rates of the subjective classification tests for the synthetic speech generated from adapted MRHSMMs using the conventional method and speaker-dependent MRHSMMs were more than 80% [5], that of the proposed method would be comparable.

### 4.3. Subjective Evaluation with Changing Style Vector

We next evaluated the naturalness of the synthesized speech of the proposed technique when changing the intensity of each style. For

[1] Several speech samples used in the test are available at http://www.kbys.ip.titech.ac.jp/research/demo/.

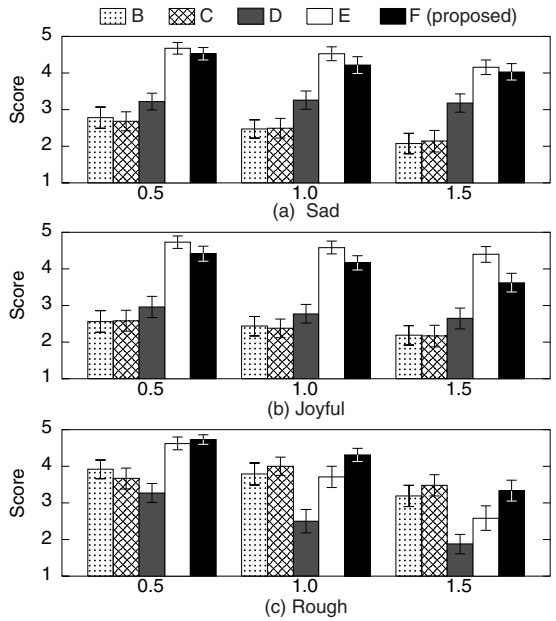


**Fig. 3**. Evaluation of naturalness in style control for FTY.

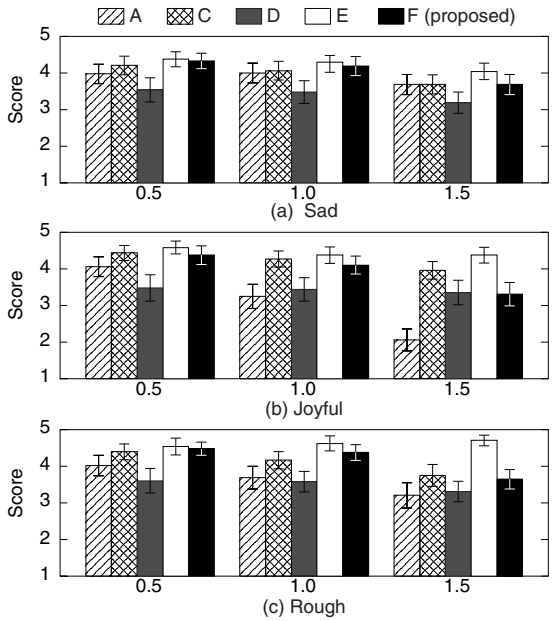


**Fig. 4**. Evaluation of naturalness in style control for MMI.

each style except for the neutral style, we changed the style component corresponding to the target style from 0.5 (weaken the intensity of the target style) to 1.5 (emphasize) with an increment of 0.5 and fixed the other style components to zero. Subjects rated the naturalness of the test samples using a 5-point scale including 5 for good, 3 for acceptable, and 1 for bad. Figures 3 and 4 show the results for the female speaker FTY and that for the male speaker MMI with 95% confidence interval of the test, respectively. Note that we confirmed the result of the male speaker MJI had a similar tendency to that of MMI. We can see that naturalness of the synthesized speech from the proposed technique (model F) was stable and closest to that from speaker-dependent MRHSMM (model E). Moreover, the score of model F exceeded that of speaker-dependent model trained by 50 sentences of the target speaker (model D) in almost all cases. Therefore, the proposed technique would be more effective when only a small amount of target speaker's style data is available.

## 5. CONCLUSION

In this paper, we have proposed a training method for multiple regression hidden semi-Markov model (MRHSMM) using an average voice model and simultaneous adaptation of speaker and style. From the results of subjective evaluation tests, we have shown that the proposed technique provides more natural speech than the conventional one with speaker adaptation only and dependency on the initial model can be decreased. Our future work is further improvement in naturalness of the synthesized speech when the intensity of styles is emphasized.

## 6. ACKNOWLEDGMENTS

A part of this work was supported by Grant-in-Aid for JSPS Fellows (1910295).

## 7. REFERENCES

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

[2] Donna Erickson, "Expressive speech: Production, perceptionand application to sp eech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 317–325, July 2005.

[3] J.F. Pitrelli, R. Bakis, E.M. Eide, R. Fernandez, W. Hamza, and M.A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 4, pp. 1099–1108, July 2006.

[4] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sept. 2007.

[5] T. Nose, Y. Kato, and T. Kobayashi, "A speaker adaptation technique for MRHSMM-based style control of synthetic speech," in *Proc. ICASSP 2007*, Apr. 2007.

[6] 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, Feb. 2007.

[7] 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, Mar. 2005.

[8] Y. Nakano, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Constrained structural maximum a posteriori linear regression for average-voice-based speech synthesis," in *Proc. INTERSPEECH 2006-ICSLP*, Sept. 2006, pp. 2286–2289.

[9] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 294–300, July 1996.

[10] K. Ogata, M. Tachibana, J. Yamagishi, and T. Kobayashi, "Acoustic model training based on linear transformation and map modification for hsmm-based speech synthesis," in *Proc. INTERSPEECH 2006-ICSLP*, Sept. 2006, pp. 1328–1331.

[11] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda, and T. Kobayashi, "A training method of average voice model for HMM-based speech synthesis," *IEICE Trans. Fundamentals*, vol. E86-A, no. 8, pp. 1956–1963, Aug. 2003.