A SPEAKER ADAPTATION TECHNIQUE FOR MRHSMM-BASED STYLE CONTROL OF SYNTHETIC SPEECH

Takashi Nose, Yoichi Kato, Takao Kobayashi

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

ABSTRACT

This paper describes a speaker adaptation technique for style control based on multiple regression hidden semi-Markov model (MRHSMM). In the MRHSMM-based style control technique, when available training data is very small, the resultant model would produce unnatural sounding speech. To overcome this problem, we propose a model adaptation technique for MRHSMM, which is similar to the MLLR adaptation technique used in speech recognition and speech synthesis. We formulate the model adaptation problem for MRHSMM based on a linear transformation framework and derive re-estimation formulas for transformation matrices in ML sense. We also describe the results of subjective evaluation tests.

Index Terms— Expressive speech synthesis, Style control, Hidden Markov model, Speaker adaptation, MLLR

1. INTRODUCTION

In recent years, demand for synthetic speech with more variability and expressivity has been increasing. In fact, many attempts have been made to synthesize expressive speech [1, 2]. One of the most essential issues is to give various speaking styles and emotional expressions to synthetic speech. In this context, we have shown that the speaking styles and/or emotional expressions, referred to as *styles*, can be modeled in an HMM-based speech synthesis framework [3] and an intermediate style can be generated using model interpolation [4]. Furthermore, to change the style and its intensity in an intuitive way, we have proposed a style control technique based on multiple regression hidden semi-Markov model (MRHSMM) [5].

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*, in which each component represents the degree or intensity of a specific style. By varying the style vector, we can control the expressivity of styles in synthetic speech. However, in the MRHSMM-based style control, a sufficient amount of training data, preferably about thirty minutes or more, is necessary for each style to train the model appropriately. In other words, when only a small amount of training data is available for each style, the resultant model would produce unnatural sounding speech. For the realization of style control with arbitrary speakers, an alternative approach to model training with less amount of training data is desirable, because it is unrealistic to prepare a large amount of data for all styles of every speaker.

For this purpose, we propose an MRHSMM-based adaptation technique for style control with a small amount of adaptation data. This technique is similar to the well-known MLLR adaptation used in speech recognition [6] and speech synthesis [7]. We train an initial MRHSMM-based model with a sufficient amount of speech data of a source speaker, and adapt it to a target speaker's model using a small amount of adaptation data.

In this paper, we first formulate the model adaptation problem for MRHSMM based on a linear transformation framework, then derive re-estimation formulas for transformation matrices in ML sense using the EM algorithm. We also describe the results of subjective evaluation tests.

2. STYLE CONTROL BASED ON MRHSMM

In the MRHSMM-based style control technique [5], each speech synthesis unit is modeled using a context-dependent MRHSMM. In MRHSMM, the output and state duration probability density functions (pdfs) at state i are given by Gaussian densities as

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{1}$$

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2) \tag{2}$$

where $\boldsymbol{o}$, $\boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are, respectively, observation vector, mean vector, and covariance matrix of the output pdf, and d, m_i, and σ_i^2 are state duration, mean, and variance of the state duration pdf, respectively. We assume that $\boldsymbol{\mu}_i$ and m_i are modeled using multiple regression as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i} \boldsymbol{\xi} \tag{3}$$

$$m_i = \boldsymbol{H}_{p_i} \boldsymbol{\xi} \tag{4}$$

where

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{5}$$

and $\boldsymbol{v}$ is the 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 $\boldsymbol{\mu}_i$. Then the output and duration pdfs $b_i(\boldsymbol{o})$ and $p_i(d)$ are given by

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{H}_{b_i} \boldsymbol{\xi}, \boldsymbol{\Sigma}_i) \tag{6}$$

$$p_i(d) = \mathcal{N}(d; \boldsymbol{H}_{p_i} \boldsymbol{\xi}, \sigma_i^2). \tag{7}$$

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 σ_i^2 can be estimated using the least square method and the EM algorithm [5, 8]. In the speech synthesis phase, the mean parameters of each synthesis unit, $\boldsymbol{\mu}_i$ and m_i are modified based on (3) and (4) with an arbitrarily given desired style vector $\boldsymbol{v}$. Then synthetic speech is generated using the HMM-based speech synthesis framework.

3. SPEAKER ADAPTATION FOR MRHSMM-BASED STYLE CONTROL

3.1. Model Adaptation for MRHSMM

Suppose that we have an MRHSMM-based model of a source speaker and wish to convert it to a target speaker's model. Here, we assume

that the mean vector of the output pdf of the target speaker's model is given by an affine transformation of that of the source speaker's model as follows:

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{b}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{\mu}_i \tag{8}$$

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

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \quad \hat{\boldsymbol{\mu}}_i = \hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi} \tag{9}$$

(8) becomes

$$\hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi} = \boldsymbol{b}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}\boldsymbol{\xi}. \tag{10}$$

If we further assume that the bias term $\boldsymbol{b}_{b_i}$ is also given by multiple regression of the style vector as

$$\boldsymbol{b}_{b_i} = \boldsymbol{B}_{b_i}\boldsymbol{\xi} \tag{11}$$

then we can rewrite (10) as

$$\begin{aligned}\hat{\boldsymbol{H}}_{b_i}\boldsymbol{\xi} =& \boldsymbol{B}_{b_i}\boldsymbol{\xi} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}\boldsymbol{\xi} \\ =& (\boldsymbol{B}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i})\,\boldsymbol{\xi}.\end{aligned} \tag{12}$$

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

$$\hat{\boldsymbol{H}}_{b_i} = \boldsymbol{B}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i}. \tag{13}$$

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

$$\hat{\boldsymbol{H}}_{p_i} = \boldsymbol{B}_{p_i} + \boldsymbol{A}_{p_i}\boldsymbol{H}_{p_i}. \tag{14}$$

3.2. Estimation of Transformation Matrix

Using a similar manner to MLLR [6], the transformation matrices for MRHSMM can be obtained in ML sense using the EM algorithm.

3.2.1. Estimation of Transformation Matrix for Output Pdf

We rewrite (13) as

$$\begin{aligned}\hat{\boldsymbol{H}}_{b_i} &= \boldsymbol{B}_{b_i} + \boldsymbol{A}_{b_i}\boldsymbol{H}_{b_i} \\ &= [\boldsymbol{B}_{b_i} \quad \boldsymbol{A}_{b_i}]\begin{bmatrix}\boldsymbol{I} \\ \boldsymbol{H}_{b_i}\end{bmatrix} = \boldsymbol{W}_{b_i}\boldsymbol{H}'_{b_i}\end{aligned} \tag{15}$$

where $\boldsymbol{H}_{b_i}$ and $\hat{\boldsymbol{H}}_{b_i}$ are the regression matrices for output pdfs of the source and target speakers. $\boldsymbol{A}_{b_i}, \boldsymbol{B}_{b_i}, \boldsymbol{W}_{b_i}$, and $\boldsymbol{H}'_{b_i}$ are matrices of dimension $M \times M$, $M \times (L+1)$, $M \times (M+L+1)$, and $(M+L+1) \times (L+1)$, respectively. From (9) and (15), we have

$$\hat{\boldsymbol{\mu}}_i = \boldsymbol{W}_{b_i}\boldsymbol{H}'_{b_i}\boldsymbol{\xi}. \tag{16}$$

When the adaptation data $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ and corresponding style vectors $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ are given, the auxiliary function for the output pdf of the target speaker is defined by

$$\begin{aligned}&Q_{b_i}(\lambda, \overline{\boldsymbol{W}}_{b_i}) \\ &\quad = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\log b_i(\boldsymbol{o}_s^{(k)}|\overline{\boldsymbol{W}}_{b_i}, \boldsymbol{\xi}^{(k)})\end{aligned} \tag{17}$$

where T_k is the number of frames of the k-th observation sequence $\boldsymbol{O}^{(k)}$, $\boldsymbol{o}_s^{(k)}$ is the observation vector at time s in $\boldsymbol{O}^{(k)}$, and $\gamma_t^d(i)$ is the probability of being in the state i at the period of time from $t-d+1$ to t given $\boldsymbol{O}^{(k)}$. By differentiating the auxiliary function with respect to $\overline{\boldsymbol{W}}_{b_i}$ and equating to zero, we obtain

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot\boldsymbol{\Sigma}_i^{-1}\overline{\boldsymbol{W}}_{b_i}\boldsymbol{H}'_{b_i}\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}\boldsymbol{H}'^{\top}_{b_i} \\ &\qquad = \sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{\Sigma}_i^{-1}\boldsymbol{o}_s^{(k)}\boldsymbol{\xi}^{(k)\top}\boldsymbol{H}'^{\top}_{b_i}.\end{aligned} \tag{18}$$

It is noted that this formula is similar to that for MLLR.

In general, it is not always able to estimate the transformation matrices for all pdfs because the amount of available adaptation data is limited. We utilize the decision tree constructed in the training phase for tying the transformation parameters. By tying $\overline{\boldsymbol{W}}_{b_i}$ in each node of the decision tree, the adaptation is possible for the states which have no corresponding adaptation data. When the transformation matrix is tied across R pdfs, (18) becomes

$$\begin{aligned}&\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(r)\cdot d\cdot\boldsymbol{\Sigma}_r^{-1}\overline{\boldsymbol{W}}_{b}\boldsymbol{H}'_{b_r}\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}\boldsymbol{H}'^{\top}_{b_r} \\ &\qquad = \sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\gamma_t^d(r)\sum_{s=t-d+1}^{t}\boldsymbol{\Sigma}_r^{-1}\boldsymbol{o}_s^{(k)}\boldsymbol{\xi}^{(k)\top}\boldsymbol{H}'^{\top}_{b_r}.\end{aligned} \tag{19}$$

This re-estimation formula can be solved in a similar manner to that of MLLR and then we obtain the transformation matrices for output pdfs.

3.2.2. Estimation of Transformation Matrix for Duration Pdf

The re-estimation formula of the transformation matrix for the state duration pdf is derived in the same fashion as that for the output pdf. From (14), the linear transformation for the state duration pdf is given by

$$\begin{aligned}\hat{\boldsymbol{H}}_{p_i} &= \boldsymbol{B}_{p_i} + \boldsymbol{A}_{p_i}\boldsymbol{H}_{p_i} \\ &= [\boldsymbol{B}_{p_i} \quad \boldsymbol{A}_{p_i}]\begin{bmatrix}\boldsymbol{I} \\ \boldsymbol{H}_{p_i}\end{bmatrix} = \boldsymbol{W}_{p_i}\boldsymbol{H}'_{p_i}\end{aligned} \tag{20}$$

where $\boldsymbol{H}_{p_i}$ and $\hat{\boldsymbol{H}}_{p_i}$ are the regression matrices for duration pdfs of the source and target speakers. $\boldsymbol{A}_{p_i}, \boldsymbol{B}_{p_i}, \boldsymbol{W}_{p_i}$, and $\boldsymbol{H}'_{p_i}$ are matrices of dimension 1×1, $1 \times (L+1)$, $1 \times (L+2)$, and $(L+2) \times (L+1)$, respectively. The re-estimation formula of the transformation matrices for the state duration pdf is given by

$$\begin{aligned}\overline{\boldsymbol{W}}_p = &\left(\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\frac{\gamma_t^d(r)}{\sigma_r^2}\cdot d\cdot\boldsymbol{\xi}^{(k)\top}\boldsymbol{H}'^{\top}_{p_r}\right)\cdot \\ &\left(\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{d=1}^{t}\frac{\gamma_t^d(r)}{\sigma_r^2}\cdot\boldsymbol{H}'_{p_r}\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}\boldsymbol{H}'^{\top}_{p_r}\right)^{-1}.\end{aligned} \tag{21}$$

4. EXPERIMENTS

4.1. Experimental Conditions

We used four styles of read speech — neutral, sad, joyful, and rough (or irritated/impolite) styles. Speech database contains 503 phonetically balanced ATR Japanese sentences uttered by male and female

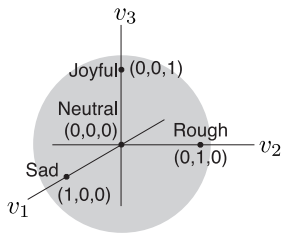

Fig. 1. Style space.

professional narrators, MMI and FTY, respectively, in each style, and is the same one used in our previous study [5, 9].

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. The MRHSMM-based model was trained for each speaker, which will be called speaker-dependent MRHSMM, using 450 sentences in each style, 1800 sentences in total. Then we set each speaker-dependent MRHSMM as the the source speaker's model, and adapted it to the target speaker's model using 50 sentences in each style, 200 sentences in total. We examined the adaptation from FTY to MMI and from MMI to FTY. A three-dimensional style space [5] 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. The transformation matrices were block diagonal which consisted of three blocks for static, delta, and delta-delta parameters. Variance parameters were not adapted. We also trained HSMM-based style-dependent model [9] using 450 sentences in each style of the target speaker.

Subjects were seven males in all tests. For each subject, ten test sentences were chosen at random from 53 test sentences which were contained in neither the training data nor adaptation data.

Table 1. Evaluation of reproducibility of styles for MMI

(a) Adapted MRHSMM

Style and Style Vector	Classification Rate (%)				
	Neutral	Sad	Rough	Joyful	Other
Neutral (0,0,0)	90.0	0.0	4.3	1.4	4.3
Sad (1,0,0)	15.7	80.0	2.9	0.0	1.4
Rough (0,1,0)	7.1	0.0	90.0	0.0	2.9
Joyful (0,0,1)	11.4	0.0	1.4	87.1	0.0

(b) Speaker-dependent MRHSMM

Style and Style Vector	Classification Rate (%)				
	Neutral	Sad	Rough	Joyful	Other
Neutral (0,0,0)	100.0	0.0	0.0	0.0	0.0
Sad (1,0,0)	1.4	97.1	0.0	0.0	1.4
Rough (0,1,0)	4.3	1.4	94.3	0.0	0.0
Joyful (0,0,1)	4.3	0.0	0.0	95.7	0.0

Table 2. Evaluation of reproducibility of styles for FTY

(a) Adapted MRHSMM

Style and Style Vector	Classification Rate (%)				
	Neutral	Sad	Rough	Joyful	Other
Neutral (0,0,0)	85.7	0.0	8.6	1.4	4.3
Sad (1,0,0)	1.4	92.9	4.3	0.0	1.4
Rough (0,1,0)	11.4	7.1	80.0	0.0	1.4
Joyful (0,0,1)	4.3	0.0	4.3	87.1	4.3

(b) Speaker-dependent MRHSMM

Style and Style Vector	Classification Rate (%)				
	Neutral	Sad	Rough	Joyful	Other
Neutral (0,0,0)	98.6	0.0	0.0	1.4	0.0
Sad (1,0,0)	7.1	92.9	0.0	0.0	0.0
Rough (0,1,0)	7.1	20.0	70.0	0.0	2.9
Joyful (0,0,1)	15.7	0.0	0.0	84.3	0.0

4.2. Subjective Evaluation of Reproducibility of Styles

We first did a classification test for the synthetic speech generated from the adapted MRHSMM with the same style vector used for the training data of each style. For comparison, we also did the same test for the synthetic speech generated from the speaker-dependent model of the target speaker. Subjects were asked which style they perceived in the test speech. The available choices for perceived styles were "neutral," "sad," "rough," and "joyful." Speech samples that were not assigned by the subjects to one of these groups were classified as "other." Tables 1 and 2 show the classification rates. It can be seen from the result that the classification rates for the adapted MRHSMM are slightly worse than or comparable to the speaker-dependent MRHSMM in all styles. It should be noted that the adapted MRHSMM was trained using only 50 sentences of the target speaker in each style, whereas the speaker-dependent MRHSMM was trained using 450 sentences.

4.3. Subjective Evaluation of Adaptation Performance

We next conducted a Comparison Category Rating (CCR) to evaluate the adaptation performance for each style. Subjects compared a test sample with a pair of reference samples and rated it. The test samples were generated from the adapted MRHSMM with the same style vector used for training in each style. The reference samples were generated from HSMM-based style-dependent models of the source and target speakers. The rating was done using a 5-point scale, that is, 5 for almost the same as the target speaker, 4 for closer to the target speaker, 3 for close to neither, 2 for closer to the source speaker, and 1 for almost the same as the source speaker. For comparison, we also evaluated synthetic speech generated using an HSMM-based MLLR adaptation technique [9]. The adaptation was done between the HSMM-based style-dependent models of the source and target speakers using 50 sentences for respective styles. Figures 2 and 3 show the result. A confidence interval of 95% is also shown in the figures. The result shows that the performance of MRHSMM-based adaptation is comparable to that of HSMM-based adaptation in all styles for both speakers MMI and FTY. In addition, a larger amount of training data for the source speaker's model of MRHSMM compared to HSMM might have led to slight improvement in the scores. It is noted that the MRHSMM-based technique can control expressivity of styles while the HSMM-based one cannot.

4.4. Subjective Evaluation of Naturalness

Finally, we evaluated the naturalness of the synthetic speech of the proposed technique when controlling the intensity of each style. We generated synthetic speech samples from the adapted MRHSMM by varying the value of the style vector along each axis of the style space. For each style except for the neutral style, we changed the style component corresponding to the target style from 0.5 to 1.5

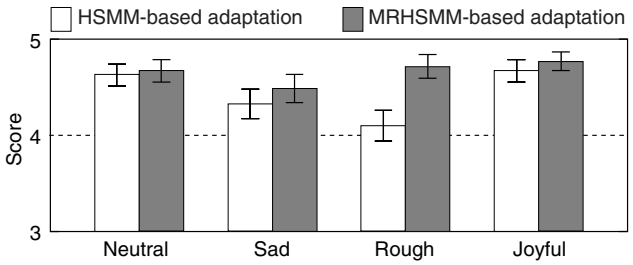

Fig. 2. Evaluation of adaptation performance for each style of MMI.

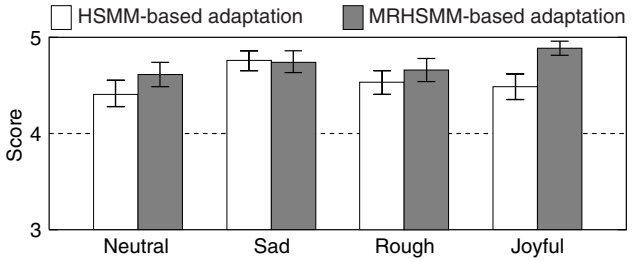

Fig. 3. Evaluation of adaptation performance for each style of FTY.

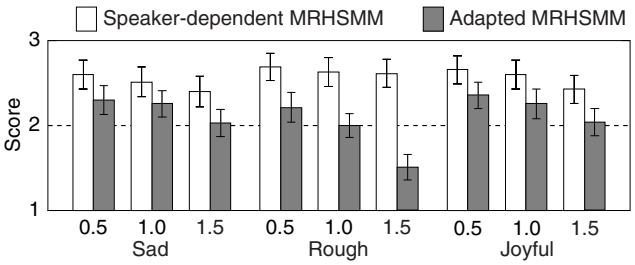

Fig. 4. Evaluation of naturalness in style control for MMI.

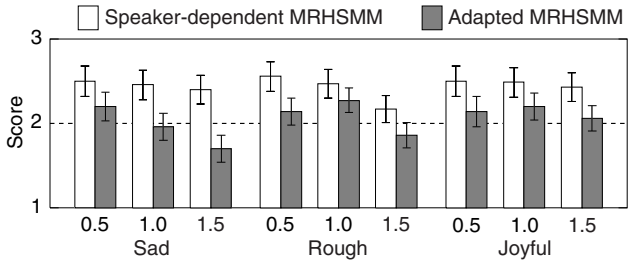

Fig. 5. Evaluation of naturalness in style control for FTY.

with an increment of 0.5 and fixed the other style components to zero. Subjects rated the naturalness of test samples and the rating was done using a 3-point scale, that is, 3 for good, 2 for acceptable, 1 for bad. For comparison, we also evaluated synthetic speech generated from the speaker-dependent MRHSMM trained using 450 sentences. Figures 4 and 5 show the scores with 95% confidence interval of the test. From the result, we can see that the adaptation degrades the naturalness of synthetic speech especially when the style component is larger than 1.0. In general, the naturalness of the synthetic speech using model adaptation depends on the initial model. The average-voice-based approach [10] might improve the dependency of the initial model and generate more natural sounding speech.

5. CONCLUSION

In this paper, we have proposed a technique of speaker adaptation for style control based on multiple regression hidden semi-Markov model (MRHSMM). In MRHSMM-based speaker adaptation, the initial model trained with a sufficient amount of data is adapted with a small amount of data using linear transformation in a similar way to the MLLR adaptation technique. We have formulated the model adaptation problem for MRHSMM based on linear transformation and derived re-estimation formulas for transformation matrices in ML sense. From the results of subjective evaluation tests, we have shown that MRHSMM can be trained using the proposed adaptation technique with only small amount of speech data. Our future work is to apply the average-voice-based approach to the adaptation for MRHSMM to improve the naturalness of synthetic speech.

6. 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. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. INTERSPEECH 2003-EUROSPEECH*, Sept. 2003, pp. 2461–2464.

[4] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[5] T. Nose, J. Yamagishi, and T. Kobayashi, "A style control technique for speech synthesis using multiple regression HSMM," in *Proc. INTERSPEECH 2006-ICSLP*, Sept. 2006, pp. 1324–1327.

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

[7] J. Yamagishi, T. Masuko, and T. Kobayashi, "MLLR adaptation for hidden semi-Markov model based speech synthesis," in *Proc. INTERSPEECH 2004-ICSLP*, 2004, pp. 1213–1216.

[8] K. Miyanaga, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based speech synthesis," in *Proc. INTERSPEECH 2004-ICSLP*, Oct. 2004, pp. 1437–1440.

[9] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Performance evaluation of style adaptation for hidden semi-Markov model based speech synthesis," in *Proc. INTERSPEECH 2003-EUROSPEECH*, Sept. 2003, pp. 2805–2808.

[10] K. Ogata, M. Tachibana, Y. Junichi, and T. Kobayashi, "Acoustic model training based on linear transformation and MAP modification for HSMM-based speech synthesis," in *Proc. INTERSPEECH 2006-ICSLP*, 2006, pp. 1328–1331.