A Style Control Technique for Speech Synthesis Using Multiple Regression HSMM

Takashi Nose, Junichi Yamagishi, Takao Kobayashi

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

Abstract

This paper presents a technique for controlling intuitively the degree or intensity of speaking styles and emotional expressions of synthetic speech. The conventional style control technique based on multiple regression HMM (MRHMM) has a problem that it is difficult to control phone duration of synthetic speech because HMM has no explicit parameter which models phone duration appropriately. To overcome this problem, we use multiple regression hidden semi-Markov model (MRHSMM) which has explicit state duration distributions to control phone duration. We show that the duration control is important for style control of synthetic speech from the results of subjective tests. We also compare the proposed technique with another control technique based on model interpolation.

Index Terms: HMM-based speech synthesis, speaking style, emotional expression, multiple regression HMM, hidden semi-Markov model.

1. Introduction

For the realization of more advanced human-computer interaction with speech communication, modeling and synthesis of emotional expressions and speaking styles is a crucial problem [1]. There often appears various emotional expressions and speaking styles in actual human speech communication, and people would communicate with others more smoothly using such paralinguistic/nonlinguistic information. We have shown that emotional expressions and/or speaking styles, which will be referred to as *styles*, can be well modeled in a speech synthesis framework based on hidden Markov model (HMM) [2, 3]. Moreover we have proposed style control techniques based on model interpolation [4] and multiple regression HMM (MRHMM) [5] for controlling emotional expressivity and speaking style variability in synthetic speech.

Style control based on model interpolation, called *style interpolation*, is achieved by interpolating model parameters among representative style models [4]. Hence, the system must keep all the representative style models to be used. In contrast, the style control technique based on MRHMM [5] models several styles in a single model simultaneously. Furthermore, in this technique, we can control the style of the synthetic speech intuitively by specifying a desired value corresponding to the degree of expressivity of each style.

However, the MRHMM-based style control technique has a problem that its reproducibility in some styles is lower than that of the style-dependent model [2] which models each style separately. One of the reasons for the problem is that there is no explicit parameter which represents phone duration in MRHMM. To overcome this problem, in this paper, we utilize multiple regression hidden semi-Markov model (MRHSMM) [6] which has explicit state duration distributions for controlling phone duration. We show that the duration control is important in controlling the style of the synthetic speech from the results of subjective tests. We also describe the difference between the style control techniques based on MRHSMM and model interpolation, and then compare the naturalness of synthetic speech between MRHSMM and model interpolation by subjective evaluation.

2. Style Control for Speech Synthesis

2.1. Modeling of Styles Using MRHSMM

In the MRHMM-based style control technique [5], we modeled each speech synthesis unit by using a context-dependent MRHMM, in which mean vectors of the output distributions are given by multiple regression of a set of parameters called *style control vector*, or simply, *style vector*. Here we reformulate this approach by using MRHSMM [6] to take account of explicit duration modeling. HSMM [7] is an extension of HMM and has output and state duration probability distributions at each state. We assume that the i-th state output and duration distributions are given by Gaussian density functions as follows:

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

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

where $\boldsymbol{o}, \boldsymbol{\mu}_i$, and $\boldsymbol{\Sigma}_i$ are observation vector, mean vector, and covariance matrix of output distribution, d, m_i, and σ_i^2 are state duration, mean, and variance of state duration distribution, respectively. In MRHSMM, we further assume that the mean parameters of the output and duration distributions at each state 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, which is a vector on a low dimensional space called *style space*, and 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. Thus we call v_k a *style component*. In addition, $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are $M \times (L+1)$- and $1 \times (L+1)$-dimensional multiple regression matrices, and M is the

dimensionality of $\boldsymbol{\mu}_i$. Then the probability distribution functions at state i are given by

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

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

Based on the EM algorithm, we can derive re-estimation formulas for the parameters of MRHSMM, $\boldsymbol{H}_{b_i}, \boldsymbol{\Sigma}_i, \boldsymbol{H}_{p_i}$, and σ_i^2, in ML sense [6] when training data $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ and corresponding style vectors $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ are given. Then we have

$$\overline{\boldsymbol{H}}_{b_i} = \left(\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i) \left[\sum_{s=t-d+1}^{t} \boldsymbol{o}_s^{(n)} \right] \boldsymbol{\xi}^{(n)\top} \right) \cdot \left(\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i) \cdot d \cdot \boldsymbol{\xi}^{(n)} \boldsymbol{\xi}^{(n)\top} \right)^{-1} \tag{8}$$

$$\overline{\boldsymbol{\Sigma}}_i = \frac{\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i) \cdot \sum_{s=t-d+1}^{t} (\boldsymbol{o}_s^{(n)} - \overline{\boldsymbol{H}}_{b_i}\boldsymbol{\xi}^{(n)})(\boldsymbol{o}_s^{(n)} - \overline{\boldsymbol{H}}_{b_i}\boldsymbol{\xi}^{(n)})^\top}{\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i) \cdot d} \tag{9}$$

$$\overline{\boldsymbol{H}}_{p_i} = \left(\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i) \cdot d \cdot \boldsymbol{\xi}^{(n)\top} \right) \cdot \left(\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i)\, \boldsymbol{\xi}^{(n)} \boldsymbol{\xi}^{(n)\top} \right)^{-1} \tag{10}$$

$$\overline{\sigma}_i^2 = \frac{\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i)\, (d - \overline{\boldsymbol{H}}_{p_i}\boldsymbol{\xi}^{(n)})^2}{\sum_{n=1}^{K} \sum_{t=1}^{T^{(n)}} \sum_{d=1}^{t} \gamma_t^d(i)} \tag{11}$$

where K is the total number of observation sequences, $T^{(n)}$ is the number of frames of the n-th observation sequence $\boldsymbol{O}^{(n)}$, $\boldsymbol{o}_s^{(n)}$ is observation vector at time s in $\boldsymbol{O}^{(n)}$, and $\gamma_t^d(i)$ is a probability of being in the state i at the period of time from $t-d+1$ to t given $\boldsymbol{O}^{(n)}$ and defined by

$$\gamma_t^d(i) = \frac{1}{P(\boldsymbol{O}^{(n)}|\lambda)} \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s^{(n)}) \beta_t(i) \tag{12}$$

where a_{ji} is the state transition probability, and $\alpha_t(i)$ and $\beta_t(i)$ are the forward and backward probabilities given by

$$\alpha_t(i) = \sum_{d=1}^{t} \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s^{(n)}) \tag{13}$$

$$\beta_t(i) = \sum_{d=1}^{T-t} \sum_{\substack{j=1 \\ j \neq i}}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s^{(n)})\, \beta_{t+d}(j) \tag{14}$$

with $\alpha_0(i) = \pi_i$ and $\beta_T(i) = 1$.

2.2. Speech Synthesis with a Desired Style

In speech synthesis stage, for a given style control vector $\boldsymbol{v}$, the mean parameters of each synthesis unit, $\boldsymbol{\mu}_i$ and m_i, are calculated from (3) and (4). Then synthetic speech is generated in the same manner as the speech synthesis framework based on HMM [5]. Consequently, by setting the style vector to a desired point in the style space, we can change the style expressivity of the synthetic speech.

2.3. Comparison with Style Interpolation

The style interpolation technique [4] is another approach to controlling the style of synthetic speech. Here we summarize the difference between the style control techniques based on MRHSMM and model interpolation.

In the style interpolation technique, synthesis units are trained for each style separately and representative style models are prepared. The new model for an intermediate style is obtained by interpolating the corresponding parameters among the representative style models with a desired interpolation ratio. It is easy to add a new style to the system because model training is required only for the representative style to be added.

In contrast, the MRHSMM-based technique models all styles in a single model simultaneously. Thus model retraining is required for all styles when a new style is added to the system. However, in this technique, the style space is defined in which the style vector represents the degree or intensity of each style. Furthermore, control parameters for styles represented by the style vector are used consistently in both the training and synthesis stages, which would lead to more intuitive style control than the style interpolation.

3. Experiments

3.1. Experimental Conditions

We used four types of read speech in neutral, sad, joyful, and rough (irritated) styles. Speech database contains phonetically balanced 503 ATR Japanese sentences uttered by a male professional narrator MMI in each style, and is the same one used in our previous studies [2]-[5].

Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis [8]. 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 and trained the model using 450 sentences in each style, 1800 sentences in total. We also trained the style-dependent model [3] using 450 sentences in each style. The model parameters were tied using shared decision tree context clustering (STC) [5, 9] in both MRHSMM and style-dependent models. This means that the style-dependent model of each style had the same number of distributions as the MRHSMM-based model. The number of distributions of each model was 1701 for spectrum, 1934 for F0, and 764 for state duration. We used the style-dependent models as the representative style models for the style interpolation.

Subjects were eight males in all subjective evaluation tests. For each subject, eight test sentences were chosen at random from 53 test sentences which were not contained in the training data.

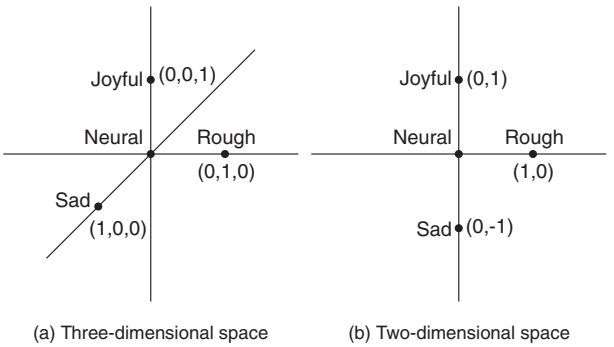


Figure 1: *Style spaces.*

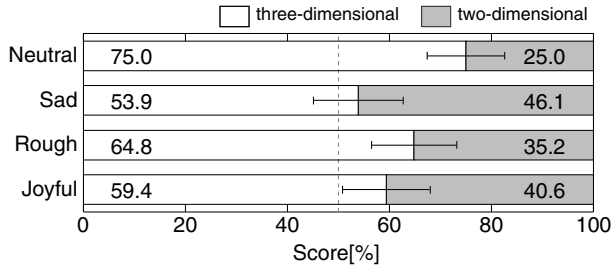


Figure 2: *Comparison of reproducibility of synthetic speech using different style spaces.*

3.2. Evaluation for Reproducibility of Styles

We first examined the choice of style spaces which would affect the reproducibility of style in synthetic speech. Here two different style spaces shown in Fig.1 were used. Neutral style was positioned at the origin in both style spaces. In Fig.1 (a), all styles except for the neutral style were assumed to be independent with each other. In Fig.1 (b), the joyful and sad styles are assumed to be on one axis as used in [5]. Thus the style spaces became three-dimensional and two-dimensional ones. For all training speech data in each style, the style vector was fixed in the style space as shown in Fig.1. Test speech samples were synthesized using the style vectors of target styles which were the same as those for training data. Subjects were presented with a pair of speech samples synthesized using the style spaces of Fig.1 (a) and (b) in random order, and then asked which sample sounded more similar to a reference speech sample. The reference speech samples for the target style were synthesized by a mel-cepstral vocoder. The result is shown in Fig.2 with a confidence interval of 95%. From the result, we can see that the reproducibility of styles using three-dimensional space is better than two-dimensional one for all styles, and the superiority is evident for the neutral style. This is because a style intermediate between the joyful and sad styles is similar to, but not the same as, the neutral style, and this fact leaded to the decrease of reproducibility for the neutral style when using the two-dimensional space. Hence we use the three-dimensional style space of Fig.1 (a) for all of following experiments.

We next assessed the reproducibility of synthetic speech using MRHMM, MRHSMM, and style-dependent HSMM (STD-HSMM). For the cases of MRHMM and MRHSMM, the desired

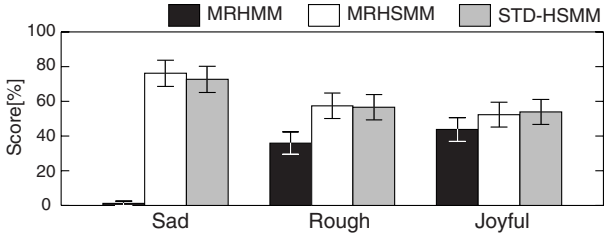


Figure 3: *Comparison of reproducibility of synthetic speech using different style control techniques.*

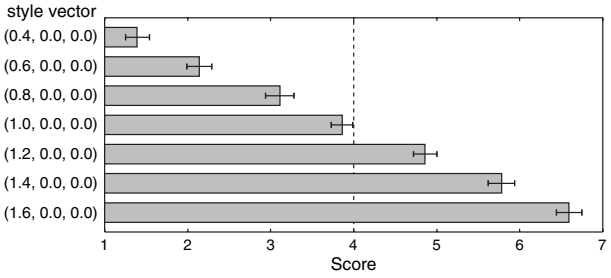


Figure 4: *Evaluation for the degree of expressivity of sad style.*

style vector was set to the same as that for training data in each style, and other conditions were the same as those of [5]. Subjects were presented with a pair of speech samples chosen from those synthesized using MRHMM, MRHSMM, and STD-HSMM in random order, and then asked which sample sounded more similar to a reference speech sample. The reference samples were the same as the previous experiment. We did not evaluate on the neutral style because there were not significant difference between STD-HMM and STD-HSMM in the neutral style [3].

Figure 3 shows the preference score with a confidence interval of 95%. It can be seen from the figure that duration control implemented by MRHSMM improves the reproducibility of styles. Moreover MRHSMM achieves comparable scores to the style-dependent model in all styles.

3.3. Evaluation for Intensity of Styles

We evaluated whether intuitive control of style expressivity in synthetic speech was achieved by the style vector. We generated synthetic speech samples 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 0.4 to 1.6 with an increment of 0.2 and fixed the other style components to zero. Subjects listened to synthesized speech samples chosen randomly from test sentences and rated their style intensity comparing to those of the reference speech samples. The rating was done using a 7-point scale, that is, 7 for very strong, 4 for equal, and 1 for very weak. The reference speech samples were synthesized from the STD-HSMM of the target style. The results are shown in Figs.4–6. These figures also show the average score for each desired style vector with a confidence interval of 95%. From these results, we can see that subjective scores for all styles increase almost in proportion to the value of style components.

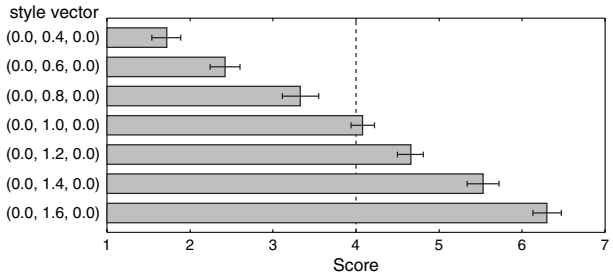


Figure 5: *Evaluation for the degree of expressivity of rough style.*

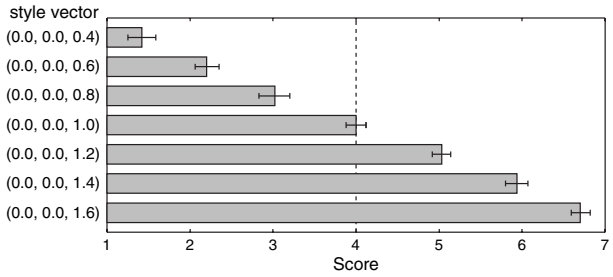


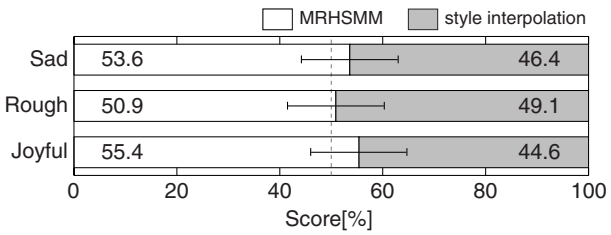
Figure 6: *Evaluation for the degree of expressivity of joyful style.*

3.4. Comparison with Style Interpolation

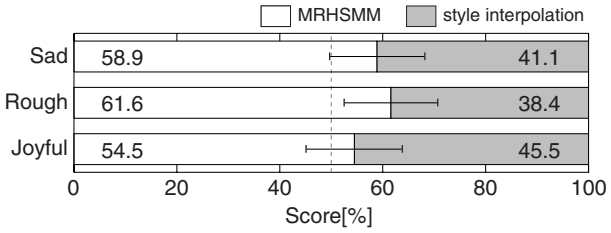
We compared naturalness of synthetic speech between MRHSMM and the style interpolation. Synthetic speech samples were generated for (a) intermediate styles between neutral and one of the other three styles and (b) emphasized styles except for the neutral style. In the style interpolation, the interpolation ratios between neutral and the target style were set to (a) $0.5 : 0.5$ and (b) $-0.5 : 1.5$. The style component of the target style in MRHSMM was set to (a) 0.5 and (b) 1.5, respectively, and the other style components were fixed to zero. This corresponds to doing style interpolation with the interpolation ratio described above. Subjects were presented with a pair of speech samples synthesized using MRHSMM and style interpolation in random order and then asked which sample sounded more natural. The results are shown in Fig.7 with a confidence interval of 95%. It is shown that the style control technique using MRHSMM is comparable to or slightly better than the style interpolation technique in naturalness of the synthetic speech.

4. Conclusion

In this paper, we have proposed a technique for controlling the degree or intensity of speaking styles and emotional expressions in synthetic speech using multiple regression HSMM (MRHSMM) which has explicit duration parameters. We have shown the proposed technique is superior to the conventional technique based on MRHMM, and the duration control is important in reproducing styles of the training speech samples or controlling intensity of speaking styles and emotional expressions from the results of subjective tests. We have also described the difference between the style control techniques based on MRHSMM and model interpolation and compared them by a subjective test. Future work will be to apply MRHSMM to emphasize characteristic of speakers.



(a) interpolation



(b) extrapolation

Figure 7: *Comparison of naturalness of synthetic speech between multiple regression HSMM and model interpolation.*

5. References

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

[2] 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.

[3] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 503–509, Mar. 2005.

[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] 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.

[6] N. Niwase, J. Yamagishi, and T. Kobayashi, "Human walking motion synthesis with desired pace and stride length based on HSMM," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2492–2499, Nov. 2005.

[7] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-markov model based speech synthesis," in *Proc. INTERSPEECH 2004-ICSLP*, Oct. 2004, pp. 1393–1396.

[8] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP-92*, Mar. 1992, pp. 137–140.

[9] 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.