Style Estimation of Speech Based on Multiple Regression Hidden Semi-Markov Model

Takashi Nose, Yoichi Kato, Takao Kobayashi

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

Abstract

This paper presents a technique for estimating the degree or intensity of emotional expressions and speaking styles appeared in speech. The key idea is based on a style control technique for speech synthesis using multiple regression hidden semi-Markov model (MRHSMM), and the proposed technique can be viewed as the inverse process of the style control. We derive an algorithm for estimating predictor variables of MRHSMM each of which represents a sort of emotion intensity or speaking style variability appeared in acoustic features based on an ML criterion. We also show preliminary experimental results to demonstrate an ability of the proposed technique for synthetic and acted speech samples with emotional expressions and speaking styles.

Index Terms: emotional speech, speaking style, style modeling, style estimation, multiple regression HSMM

1. Introduction

In recent human-computer interaction (HCI) systems which include speech recognition and speech synthesis, modeling of speech with emotional expressions and speaking styles is becoming one of important issues, because there often appear various emotional expressions and speaking styles in actual human speech communication and such paralinguistic/nonlinguistic information could enhance the performance of HCI systems. 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) [1]. Moreover we have proposed several techniques which enable us to control styles in synthetic speech [2, 3].

In emotional speech recognition research area, although a large variety of techniques have been proposed (e.g., see [4]-[7]), most of the techniques focus on only classification of emotions. In contrast, it is often important to know the degree or intensity of the emotion such as “a little irritated” or “very irritated” as well as to detect user’s emotional state in some applications such as triage in call centers.

In this paper, we propose a technique for estimating a set of values each of which represents the degree of a specific style expressivity. The key idea of the proposed technique is based on the style control technique using multiple regression hidden semi-Markov model (MRHSMM) [3], which is an extension of MRHMM-based style control technique of [8]. In this approach, the spectrum and fundamental frequency (F0) of speech are simultaneously modeled using hidden semi-Markov model (HSMM) [9] which can model duration distributions more appropriately than HMM. The mean vector of output and state duration distributions of MRHSMM is given by a function of a parameter vector, called a *style vector*, in speech synthesis units. More specifically, the style vector represents a point in a space, called a *style space*, where each coordinate represents a certain style, such as happiness or sadness, and the mean vector is modeled by using multiple regression of the style vector.

We first review the training algorithm of MRHSMM when training speech samples and corresponding style vectors are given. Then we derive an algorithm for estimating the style vector for given input speech. The proposed technique can be viewed as the inverse process of the style control for speech synthesis based on MRHSMM. In this paper, we focus on derivation of the algorithm, and show preliminary experimental results of style estimation for synthetic and acted emotional speech.

2. Style Modeling Based on MRHSMM

We briefly review the MRHSMM-based style modeling described in [3]. HSMM has output and state duration probability distributions at each state [9]. 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 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, and its components $\{v_k\}$ are predictor variables, called *style components*, each of which represents the degree or intensity of a certain style in speech. $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are regression matrices of dimension $M \times (L+1)$ and $1 \times (L+1)$. 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 [10] 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

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

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

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

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

where K is the total number of observation sequences, $T^{(k)}$ is the number of frames of the k-th observation sequence $\boldsymbol{O}^{(k)}$, $\boldsymbol{o}_s^{(k)}$ is observation vector at time s in $\boldsymbol{O}^{(k)}$, 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 $\boldsymbol{O}^{(k)}$ described in [3].

3. Estimation of Style Vector

Here we consider a problem of estimating a style vector $\boldsymbol{v}$ for an input observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ given the trained MRHSMM λ whose parameters $\boldsymbol{H}_{b_i}, \boldsymbol{\Sigma}_i, \boldsymbol{H}_{p_i}$, and σ_i^2 are fixed. We can rewrite (3) in the following form.

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi} = \boldsymbol{h}_0^{(b_i)} + \boldsymbol{A}_{b_i}\overline{\boldsymbol{v}} \tag{12}$$

where $\overline{\boldsymbol{v}}$ denotes the style vector to be estimated, and

$$\boldsymbol{H}_{b_i} = [\boldsymbol{h}_0^{(b_i)}, \cdots, \boldsymbol{h}_L^{(b_i)}] \tag{13}$$

$$\boldsymbol{A}_{b_i} = [\boldsymbol{h}_1^{(b_i)}, \cdots, \boldsymbol{h}_L^{(b_i)}] \tag{14}$$

$$\boldsymbol{v} = [v_1, \cdots, v_L]^\top \tag{15}$$

The optimal style vector $\boldsymbol{v}_{\max}$ for the input observation $\boldsymbol{O}$ is defined in ML sense as

$$\boldsymbol{v}_{\max} = \arg\max_{\boldsymbol{v}} P(\boldsymbol{O}|\lambda, \boldsymbol{v}). \tag{16}$$

To obtain ML estimation, we use the EM-algorithm in which the auxiliary function is given by

$$Q(\boldsymbol{v}, \overline{\boldsymbol{v}}) = \sum_{\text{all }\boldsymbol{q}}\sum_{\text{all }\boldsymbol{l}} P(\boldsymbol{q}, \boldsymbol{l}|\boldsymbol{O}, \lambda, \boldsymbol{v}) \log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l}|\lambda, \overline{\boldsymbol{v}}) \tag{17}$$

where $\boldsymbol{q} = \{q_1, q_2, \cdots, q_T\}$ is a possible state sequence, and $\boldsymbol{l} = \{l_1, l_2, \cdots, l_N\}$ is a possible sequence of state duration for observation sequence $\boldsymbol{O}$ and the state sequence $\boldsymbol{q}$. It is noted that $\sum_{i=1}^{N} l_i = T$. The auxiliary function (17) has the similar form to that in [11], and only the output distributions are affected in the re-estimation process for the style vector $\overline{\boldsymbol{v}}$. Hence the auxiliary function (17) can be rewritten as

$$Q_b(\boldsymbol{v}, \overline{\boldsymbol{v}}) = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{t} P(q_t = i, l_i = d|\boldsymbol{O}, \lambda, \boldsymbol{v}) \cdot \log\left(\prod_{s=t-d+1}^{t} b_i(\boldsymbol{o}_s|\overline{\boldsymbol{v}})\right) \tag{18}$$

where

$$b_i(\boldsymbol{o}_s|\overline{\boldsymbol{v}}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{o}_t - \boldsymbol{h}_0^{(b_i)} - \boldsymbol{A}_{b_i}\overline{\boldsymbol{v}})^\top \cdot \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_t - \boldsymbol{h}_0^{(b_i)} - \boldsymbol{A}_{b_i}\overline{\boldsymbol{v}})\right] \tag{19}$$

and note that $P(q_t = i, l_i = d|\boldsymbol{O}, \lambda, \overline{\boldsymbol{v}}) = \gamma_t^d(i)$. By differentiating the auxiliary function (18) with respect to $\overline{\boldsymbol{v}}$ and equating to zero, we have

$$\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot \boldsymbol{A}_{b_i}^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{A}_{b_i}\overline{\boldsymbol{v}} = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{A}_{b_i}^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_s - \boldsymbol{h}_0^{(b_i)}). \tag{20}$$

Consequently, the re-estimation formula of the style vector for output distribution is given by

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\cdot d\cdot \boldsymbol{A}_{b_i}^\top\boldsymbol{\Sigma}_i^{-1}\boldsymbol{A}_{b_i}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^d(i)\sum_{s=t-d+1}^{t}\boldsymbol{A}_{b_i}^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{o}_s - \boldsymbol{h}_0^{(b_i)})\right). \tag{21}$$

Note that (21) is very similar to the re-estimation formula of the mean parameters in speaker adaptive training (SAT) algorithm [12] and that of the gesture parameter in parametric HMM [13].

The obtained values of style components give quantities how much each style affects the acoustic features including spectral and prosodic information compared to those of the training data in ML sense. As a result, we can expect that the estimated values of the style components can be used to detect the degree of emotions and speaking styles expressed on speech.

4. Experiments

4.1. Experimental Conditions

To examine whether the proposed algorithm works well, we conducted preliminary experiments using synthetic and real speech samples. We used four types of acted speech in neutral, sad, joyful and rough (or irritated) styles. Speech database contains phonetically balanced 503 ATR Japanese sentences uttered by male and female professional narrators, MMI and FTY, respectively, in each style, and is the same one used in [1].

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

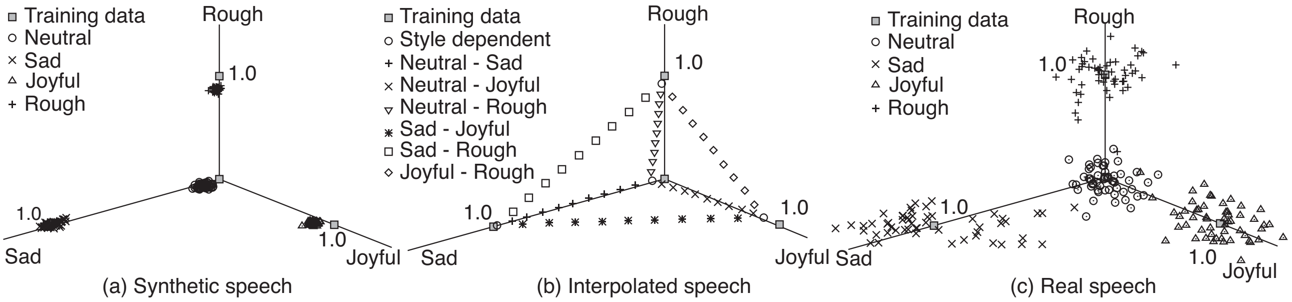


Figure 1: *Estimation results for speaker MMI's (a) synthetic speech samples generated from style-dependent models, (b) synthetic speech samples generated from interpolated models between two representative styles, and (c) real speech samples taken from speech database.*

including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. We used 5-state left-to-right MRHSMMs, which were trained using 450 sentences in each style, and 1800 sentences in total.

A three-dimensional style space was used which was shown to be an appropriate for the purpose of style control in speech synthesis [3]. Neutral style was positioned at the origin of the style space, and all other styles were assumed to be independent with each other. For all training speech data in each style, style vectors $\boldsymbol{v} = (v_1, v_2, v_3)$ were fixed at $(0, 0, 0)$ for neutral, $(1, 0, 0)$ for sad, $(0, 1, 0)$ for joyful, and $(0, 0, 1)$ for rough.

4.2. Estimation of Style Vector Using Context-dependent MRHSMM

We utilized the context-dependent MRHSMMs used in the style control for speech synthesis [3], where contexts, i.e., possible phonetic and linguistic factors that would affect spectral and prosodic features were taken into account. Furthermore, we assumed that the phonetic and linguistic information of input utterances were known before estimation of the style vectors.

We first estimated the style vectors for synthetic speech samples. Speech samples were synthesized using the style-dependent model [1] trained using the same training data as the MRHSMM in each style. Figure 1 (a) shows the distribution of the estimated style vectors for 53 test sentences in each style. These test sentences were not contained in the training data. Each plotted point in the figure corresponds to one sample, and there are 53 points for each style. It is obvious that the synthetic speech generated from the style-dependent model has close acoustic features to the speech used for the model training. This fact agrees with subjective evaluation results of the style-dependent models [1].

Figure 1 (b) shows estimation result for the synthetic speech samples with intermediate styles generated using a style interpolation technique [2]. We chose four style-dependent models described above as the representative style models in interpolation. We obtained style interpolated speech samples for all combinations of two styles out of four representative styles. We chose a test sentence and gradually changed the interpolation weights from (1.0, 0.0) to (0.0, 1.0) at 0.1 intervals. It can be seen from the figure that style interpolated speech has the acoustic features corresponding to those having linearly interpolated style components in the style space. This fact again agrees with subjective evaluation results of the style-interpolated models [2].

We then estimated style vectors for real speech samples. We used 53 utterances taken from the database in each style. These sentences are the same as those for the synthetic speech samples in the previous experiment. The result is shown in Fig. 1 (c). As mentioned in **3**, an estimated style vector can be considered to represent the degree of expressivity on an emotion or a speaking style. Since the intensity of emotional expressions and speaking styles appeared in acoustic features of real speech is not always constant among utterances, the estimated vectors are distributed around the points for the training data.

4.3. Estimation of Style Vector Without Using Context Information

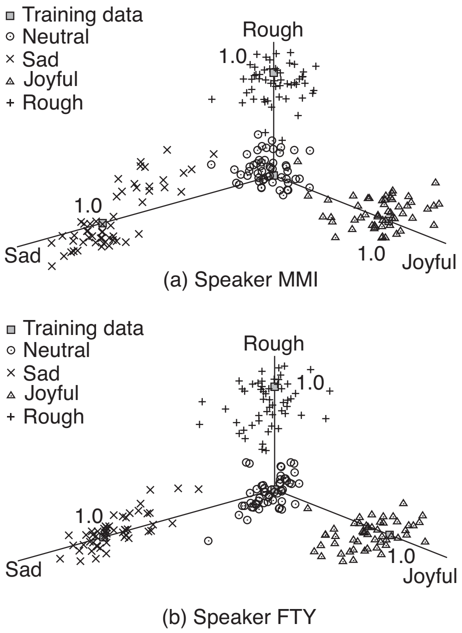
To evaluate the estimation performance under a more realistic condition, we conducted an experiment when the phonetic and linguistic information for input utterances was unknown. Specifically, before we estimated the style vectors for real speech samples, transcriptions of the input utterances were estimated by phoneme recognition. The feature vector used for the recognition consisted of 12 mel-cepstral coefficients and their deltas not including the zeroth coefficient. For each speaker, style-independent triphone HMMs were trained using the same database as the training of MRHSMM. We used phonetic networks based on Japanese phonetic concatenation rules in the recognition. The recognition accuracy for whole test samples was 78.6%.

We trained triphone MRHSMMs using the same data as the previous experiment and used them for estimating the style vectors with the transcriptions obtained by the phoneme recognition. We used 53 real utterances in each style that were same as the previous experiment. The result is shown in Figs. 2. We can see that the estimated style vectors had similar distribution to that estimated with the phonetic and linguistic information.

4.4. Perceptual Evaluation for Estimated Style Vector

We assessed whether the result of style estimation matched human perception in terms of the degree of expressions. First, for each style, three samples having the largest three values of the style components were chosen as the *strong* intensity group from the 53 samples estimated in **4.3**. Similarly, another three samples having the smallest three style components were chosen as the *weak* intensity group. Finally, another three samples having the style components nearest to the average of the largest and smallest ones were chosen as the *medium* intensity group. As a result there were nine test samples for each style. Six subjects listened to all combinations of two samples chosen randomly from the nine samples except for the combination of the samples in the same group, and were asked which sample had stronger intensity for each style.

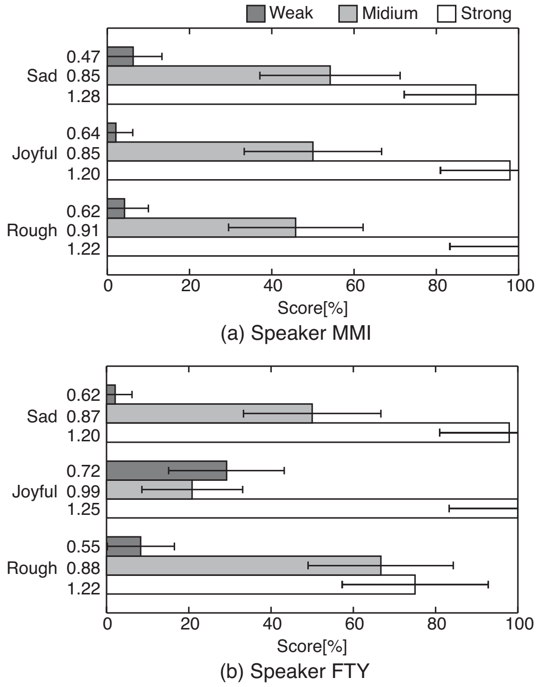
Figure 3 shows the result with a confidence interval of 95%. The mean values of the style components of each group are also shown in the figure for each style. We can see that the style vectors estimated for real speech samples well matched the perceptual scores for speaker MMI. For speaker FTY, although the estimated value did not always match the score, there were sig-



(a) Speaker MMI

(b) Speaker FTY

Figure 2: *Estimation results for real speech samples of speaker MMI and FTY without using context information.*



(a) Speaker MMI

(b) Speaker FTY

Figure 3: *Correspondence between perceptual and estimated intensity for real speech samples of speakers MMI and FTY.*

nificant differences between the strong and weak groups in all styles.

5. Conclusion

In this paper, we have presented a technique for estimating the degree of emotions and speaking styles in speech. The technique is based on the acoustic modeling of both spectral and prosodic information simultaneously using MRHSMM. The predictor variables of MRHSMM represent the degree of styles and are estimated for input speech using ML criterion. Moreover, the proposed technique can be viewed as the inverse process of the MRHSMM-based style control for speech synthesis. We have shown that the proposed technique is promising in estimation of expressivity of emotions and speaking styles.

Future work will be to refine the training algorithm of MRHSMM by incorporating the re-estimation of the style vector and to evaluate the estimation performance of styles for spontaneous speech. We will also evaluate the estimation performance when using MRHSMM-based speaker adaptation with small amount of speech data [14].

6. References

[1] 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.

[2] 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.

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

[4] R. Cowie, E. Dougla-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, Jan. 2001.

[5] L. Bosch, "Emotions, speech and the ASR framework," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, Apr. 2003.

[6] T. Nwe, S. Foo, and L. D. Silva, "Speech emotion recognition using hidden Markov models," *Speech Communication*, vol. 41, no. 4, pp. 603–623, Nov. 2003.

[7] J. Cichosz and K. Slot, "Low-dimensional feature space derivation for emotion recognition," in *Proc. INTERSPEECH 2005-Eurospeech*, Sep. 2005, pp. 477–480.

[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] 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.

[10] 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.

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

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

[13] A. Wilson and A. Bobick, "Parametric hidden Markov models for gesture recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 884–900, Sep. 1999.

[14] 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, pp. 833–836.