A Technique for Estimating Intensity of Emotional Expressions and Speaking Styles in Speech Based on Multiple-Regression HSMM

Takashi NOSE(a) and Takao KOBAYASHI(b), Members

SUMMARY In this paper, we propose a technique for estimating the degree or intensity of emotional expressions and speaking styles appearing in speech. The key idea is based on a style control technique for speech synthesis using a multiple regression hidden semi-Markov model (MRHSMM), and the proposed technique can be viewed as the inverse of the style control. In the proposed technique, the acoustic features of spectrum, power, fundamental frequency, and duration are simultaneously modeled using the MRHSMM. We derive an algorithm for estimating explanatory variables of the MRHSSM, each of which represents the degree or intensity of emotional expressions and speaking styles appearing in acoustic features of speech, based on a maximum likelihood criterion. We show experimental results to demonstrate the ability of the proposed technique using two types of speech data, simulated emotional speech and spontaneous speech with different speaking styles. It is found that the estimated values have correlation with human perception.

key words: emotion recognition, emotional expression, speaking style, intensity of style, hidden semi-Markov model (HSMM), multiple-regression HSMM (MRHSMM)

1. Introduction

For the realization of more advanced and user-friendly human-computer interaction (HCI) systems which include speech recognition and speech synthesis, modeling of speech with emotional expressions and speaking styles is one of the important issues. This is because various emotional expressions and speaking styles often appear in actual human speech communication, and such paralinguistic/nonlinguistic information could enhance the performance of HCI systems. If a computer can estimate a user’s emotional state or speaking style from his/her speech, the interaction becomes more human-like because the computer can choose a more appropriate response that matches the user’s state.

In this context, a large number of studies on emotional speech and its classification have been conducted[1],[2]. As for automatic emotion recognition, there are approaches based on Gaussian mixture model (GMM) [3], support vector machine (SVM) [4], hidden Markov Model (HMM) [5],[6], and their hybrids [7],[8]. However, most of the techniques focus only on the classification of emotional states. In contrast, we believe it is also important to detect the variability within a certain emotion (e.g., “a little sad” or “very irritated”) as well as the emotion categories. This is evidenced by the fact that we often soften or emphasize such emotional expressions flexibly depending on the situation in actual human speech communication.

Several studies have analyzed the degree or intensity of emotions in the acoustic features of speech. Hirose et al. [9] investigated how prosodic features change depending on emotional levels using simulated emotional speech toward expressive speech synthesis. Similar analyses of prosodic features for professional announcers or actors were also found in [10],[11]. However, these approaches are based on heuristics, and it is still not easy to estimate objectively the degree or intensity of emotional expressions without any heuristics. To overcome this problem, Song et al. proposed a technique for classifying basic emotions into three levels based on the HMM [12]. In the technique, training data with emotions are labeled in advance according to their emotional types and intensity levels. Then, the sub-HMMs are trained using the training data and corresponding labels. In the classification process, the input speech is classified into emotional categories using an HMM classifier, and its intensity is classified into three levels using trained sub-HMMs. However, this technique is based on the classification of input speech into discrete emotional states or levels, and there are no explicit parameters which represent the intensity of emotional expressions in speech.

In contrast, this paper presents a technique for estimating a set of continuous values each of which represents the degree or intensity of a specific emotional expression or speaking style based on an HMM framework. The key idea of the proposed technique, called style estimation, is based on the style control technique of synthetic speech using multiple regression hidden semi-Markov model (MRHSMM) [13]. In the MRHSMM-based style modeling, the spectrum, power, fundamental frequency (F0), and duration of speech are simultaneously and explicitly modeled using multi-space probability distribution (MSD) [14] and hidden semi-Markov model (HSM) [15]. Furthermore, the MRHSMM assumes that the mean vectors of output and state-duration probability density functions (pdfs) of each state are given by functions of a parameter vector, called a style vector. More specifically, the style vector represents a point in a space, called a style space, where each coordinate represents a certain style (e.g., a joyful, sad, or impolite style) and the mean vector is modeled by using multiple regression of the style vector [13].
In this paper, we derive an algorithm that estimates the style vector for a given input speech utterance based on a maximum likelihood (ML) criterion [16]. The proposed technique can be viewed as the inverse process of the style control for speech synthesis based on the MRHSMM. Each component of the estimated style vector, called a style component, gives quantitative measure of how much each style affects the acoustic features of speech, including spectral and prosodic information compared to the training data in an ML sense. As a result, we can expect that the estimated values of the style components can be used to detect the degree or intensity of emotions and speaking styles expressed in speech. To confirm that the proposed technique can quantitatively detect the intensity of style expressiveness, we show experimental results of style estimation for three types of simulated emotional speech data — neutral, sad, and joyful styles. We also evaluate the performance of the proposed technique in speaking style classification of spontaneous speech of reading and academic presentation styles.

The rest of this paper is organized as follows. Section 2 gives brief explanations of MRHSMM-based style modeling and training of the MRHSMM using model adaptation. In Sect. 3, we propose an algorithm for estimating the intensity of style for a given speech utterance. Subjective experiments are described in Sect. 4. Section 5 discusses the proposed technique, and Sect. 6 summarizes our findings.

2. Modeling of Multiple Styles and Their Intensity of Speech Based on the MRHSMM

In a previous study on HMM-based style control of synthetic speech [13], [17], we used the MRHSMM to model multiple styles in a single model explicitly taking into account their degree or intensity. For the purpose of style estimation of expressive speech, we use the same MRHSMM-based framework in the modeling of speech units. Here, we briefly review the MRHSMM-based style modeling.

2.1 Speaker-Dependent Style Modeling Based on the Multiple-Regression HSMM

The MRHSMM [18] is an extension of HSMM [15] which has explicit state-duration pdfs. In the MRHSMM, mean parameters of output and state-duration pdfs at each state of the MRHSMM are assumed to be given by Gaussian pdfs, and are represented by multiple regression of a low dimensional vector \( \nu \), called a style vector, as

\[
\mu_i = H_{bi} \xi
\]

\[
m_i = H_{pi} \xi
\]

where

\[
\xi = [1, \nu_1, \nu_2, \ldots, \nu_L]^T = [1, \nu]^T.
\]

\( \mu_i \) and \( m_i \) are the mean parameters of output and state-duration pdfs at the \( i \)-th state of the MRHSMM, respectively. Each component of the style vector \( \nu \) corresponds to the degree or intensity of a certain style. \( H_{bi} \) and \( H_{pi} \) are regression matrices of dimension \( M \times (L + 1) \) and \( 1 \times (L + 1) \), respectively, and \( M \) is the dimensionality of \( \mu_i \).

In the training of the MRHSMM, first we train context-dependent phoneme HSMMs independently without context clustering for the respective styles of a target speaker. Then, the model parameters are tied with a common tree structure for all styles using shared-decision-tree-based context clustering [19]. Using these style-dependent HSMMs, the initial parameters of the speaker-dependent MRHSMM are obtained by least squares estimation (LSE), and re-estimated based on an ML criterion using an expectation maximization (EM) algorithm [20] for the given training data and corresponding style vectors. The details of model training are described in [13].

2.2 Training of the MRHSMM Using Model Adaptation with a Small Amount of Data

When the available training data of a target speaker is limited (e.g., less than a few minutes), it is difficult to estimate parameters of the MRHSMM with high reliability. Although a possible approach to this problem is the speaker-independent one, the cost of collecting training data of many speakers in multiple styles becomes relatively high, and the performance would be unsatisfactory because the expression of emotions and speaking styles varies greatly between individuals. To address this problem, we utilize a model adaptation technique for the MRHSMM with a small amount of speech data [17].

A block diagram of the MRHSMM training using model adaptation is shown in Fig. 1. First, the speaker-independent HSMM is trained using neutral style speech of many speakers, and then it is adapted to the target speaker’s style-dependent HSMMs using a small amount of adaptation data with respective styles. Constrained structural maximum a posteriori linear regression (CSMAPLR) [21] is used for the model adaptation as in the previous study [17]. Using the speaker- and style-adapted HSMMs, the target speaker’s initial MRHSMM is obtained by the same LSE as used in the speaker-dependent MRHSMM training. Then, the MRHSMM-based maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP)-like adaptation is applied to the initial MRHSMM, and the MRHSMM adapted to the target speaker is obtained. The details of MRHSMM training using model adaptation is described in [17].

3. Estimation Algorithm of the Style Vector Based on the MRHSMM

Given the context-dependent phoneme MRHSMMs, we consider a problem of estimating the style vector, \( \nu \), for an input observation sequence \( O = (o_1, \ldots, o_T) \). An overview of the style estimation procedure is shown in Fig. 1. First, we conduct phoneme recognition for an input utterance
because phonetic information is needed to create a sentence MRHSMM corresponding to the input speech by concatenating the given MRHSMMs. Then, the style vector is estimated for every input utterance using the sentence MRHSMM in ML sense. To derive the re-estimation formula of the EM algorithm, we rewrite Eqs. (1) and (2) in the following form.

\[ \mu_i = H_{bi} \xi = h_{b0}^{(bi)} + A_{bi} v \]  
\[ m_i = H_p \xi = h_{p0}^{(pi)} + A_{pi} v \]  

where

\[ H_{bi} = \begin{bmatrix} h_{b0}^{(bi)} & \cdots & h_{bL}^{(bi)} \end{bmatrix} \]  
\[ A_{bi} = \begin{bmatrix} a_{bi}^{(b)} & \cdots & a_{bi}^{(bL)} \end{bmatrix} \]  
\[ H_{pi} = \begin{bmatrix} h_{p0}^{(pi)} & \cdots & h_{pL}^{(pi)} \end{bmatrix} \]  
\[ A_{pi} = \begin{bmatrix} a_{pi}^{(p)} & \cdots & a_{pi}^{(pL)} \end{bmatrix} \]  
\[ v = [v_1, \cdots, v_L]^T \]  

The optimal style vector, \( v^* \), for the input observation sequence \( O \) is defined in the ML sense as

\[ v^* = \arg\max_v P(O|\lambda, v) \]  

To obtain the ML estimate, we introduce an auxiliary function

\[ Q(\Lambda, \bar{v}) = Q_b(\Lambda, \bar{v}) + Q_p(\Lambda, \bar{v}) \]  

for the EM algorithm. In this equation, \( \Lambda = (\lambda, v) \), and \( Q_b(\Lambda, \bar{v}) \) and \( Q_p(\Lambda, \bar{v}) \) are auxiliary functions of output and state-duration pdfs, respectively. \( Q_b(\Lambda, \bar{v}) \) and \( Q_p(\Lambda, \bar{v}) \) are defined as follows [22].

\[ Q_b(\Lambda, \bar{v}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_i^t(i) \log b_i(o_i|\bar{v}) \]  
\[ Q_p(\Lambda, \bar{v}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_i^t(i) \log p_i(d|\bar{v}) \]  

\[ \log b_i(o_i|\bar{v}) = -\frac{1}{2} \left( a_i - \mathbf{h}_0^{(bi)} - A_{bi} \bar{v} \right)^T \Sigma_i^{-1} \left( a_i - \mathbf{h}_0^{(bi)} - A_{bi} \bar{v} \right) \]  
\[ + \frac{1}{2} \log |\Sigma_i^{-1}| + \text{const.} \]  

\[ \log p_i(d|\bar{v}) = -\frac{1}{2} \left( d - \mathbf{h}_0^{(pi)} - A_{pi} \bar{v} \right)^2 \]  
\[ -\frac{1}{2} \log \sigma_i^2 + \text{const.} \]  

where \( \Sigma_i \) is the covariance matrix of the output pdf at state \( i \), and \( d \) and \( \sigma_i^2 \) are the state duration and variance of the \( i \)-th state-duration pdf, respectively. \( \gamma_i^t(i) \) is the probability of being in state \( i \) during the period of time from \( t - d + 1 \) to \( t \) given \( O \) and defined by

\[ \gamma_i^t(i) = \frac{1}{P(O|\lambda)} \sum_{j=1}^{N} \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \beta_s(i) \]  

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

\[ \alpha_t(i) = \sum_{d=1}^{T} \sum_{j=1}^{N} \alpha_{t-d}(j) a_{ji} p_i(d) \prod_{s=t-d+1}^{t} b_i(o_s) \]  
\[ \beta_t(i) = \sum_{j=1}^{N} \sum_{d=1}^{T-t} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(o_s) \beta_{t+d}(j) \]  

with \( \alpha_0(i) = \pi_i \) and \( \beta_T(i) = 1 \), and \( \pi_i \) is the initial state probability of being in state \( i \) at time \( t = 1 \).

By differentiating the auxiliary function \( Q(\Lambda, \bar{v}) \) of Eq. (12) with respect to \( \bar{v} \) and equating the result to zero, we have

\[ \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_i^t(i) \left( d A_p^\top \Sigma_i^{-1} A_{bi} + \frac{1}{\sigma_i^2} A_p^\top A_{pi} \right) \bar{v} \]  
\[ = \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_i^t(i) \left( \sum_{s=t-d+1}^{t} A_{bi}^\top \Sigma_i^{-1} (o_s - \mathbf{h}_0^{(bi)}) + \frac{1}{\sigma_i^2} A_p^\top (d - \mathbf{h}_0^{(pi)}) \right) \]  

Fig. 1 A block diagram of MRHSMM training and style estimation using a speaker-independent model and model adaptation.
Consequently, the re-estimation formula of the style vector for the input observation is given by

\[
\hat{\nu} = \left[ \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{T} \gamma_i^t(i) \left( \sum_{i}^{T} A_k^r \Sigma_{h_i}^{-1} A_h + \frac{1}{\sigma_i^2} A_k^T A_h \right) \right]^{-1} \\
\times \left[ \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{d=1}^{T} \gamma_i^t(i) \left( \sum_{i}^{T} A_k^r \Sigma_{h_i}^{-1}(o_i - h_{d1}^{(b)}) + \frac{1}{\sigma_i^2} A_k^T(d - h_{d1}^{(b)}) \right) \right].
\]

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

4. Experiments

4.1 Speech Database

In the following experiments, we evaluated the performance of the proposed technique using two types of speech data, simulated emotional speech and spontaneous speech with different speaking styles. For emotional speech, we used speech data uttered by a professional narrator and non-professional speakers with three types of emotions — neutral, sad, and joyful styles. 500 sentences taken from the 503 phonetically balanced sentences of the ATR Japanese speech database were uttered in each style by a male professional narrator MMI, who had some experience in speaking with simulated styles, uttered 100 sentences of subsets A and I in each style. For spontaneous speech data, we used the corpus of spontaneous Japanese (CSJ) database [23]. We used speech data uttered by two non-professional speakers, a male speaker (speaker ID: 423) and a female speaker (speaker ID: 514) with two types of speaking styles — reading and academic presentation styles. From the database of each speaker and style, we extracted 100 utterances segmented by silences with 200 ms or longer and having no fillers. To train the speaker-independent model, we used speech data of 100 speakers (50 males and 50 females) included in the Japanese Newspaper Article Sentences (JNAS) database [24]. The training data were 100 sentences for each speaker, 10,000 sentences in total. The 100 sentences were not included in the above ATR 503 sentences.

4.2 Experimental Conditions

Speech signals were sampled at a rate of 16 kHz and windowed by a 25-ms Blackman window with a 10-ms shift. The feature vector consisted of 25 mel-cepstral coefficients [25] including the zeroth coefficient, log $F_0$, and their delta coefficients. We used three-state left-to-right triphone MRHSMMs with diagonal covariance matrices. In the style estimation, the initial value of a style vector was set to 0 and updated by the EM algorithm. In the model training, style vectors for training and adaptation data were set to fixed values; i.e., 0 for the neutral style, and 1.0 for the sad and joyful styles, respectively. This means that the style space was modeled as a one-dimensional space, in which the neutral style was positioned at the origin, and the target style — i.e., sad or joyful style — was positioned at 1.0 on a one-dimensional axis. In the MAP-like adaptation of Fig. 1, the value of hyper-parameters for output and state-duration pdfs were set to 1000, as in the previous study [17]. For phoneme recognition, the style-independent triphone HMM of the input speaker is trained in advance using the same data for the MRHSMM training. We used maximum likelihood linear regression (MLLR) [26] as a model adaptation algorithm. In the recognition process, a standard phoneme recognition system [27] was used to obtain a phoneme sequence of an input utterance. The feature vectors consisted of 12 MFCCs with cepstral mean subtraction, normalized log energy, and their deltas. We used three-state left-to-right triphone HMMs with diagonal covariance matrices. In addition, we used phonetic networks based on Japanese phonetic concatenation rules in the recognition. The overall recognition accuracy for test samples of emotional and spontaneous speech was 76.7%.

4.3 Perceptual Labeling of Emotional Intensity for the Speech Database

To evaluate the correspondence between the estimation results and human perception, a perceptual rating on style expressiveness was conducted in advance for all the emotional speech data (except for the neutral style). Nine subjects listened to each utterance from the database in random order and rated the style expressiveness. The rating was done as follows: “1.5” for strong, “1.0” for standard, “0.5” for weak, and “0” for not perceiving as the target style. For the utterances rated as “0”, we asked the subjects how they perceived the speech, and the dominant opinion was that it sounded like the neutral style. Thus, we assumed that “0” corresponded to the neutral style in this study. Table 1 shows the standard deviations of perceptual scores among subjects for respective speakers and styles. The overall average standard deviation was 0.22, and the inter-rater reliability seems to be fair in evaluating emotional intensity. The result for a male professional narrator MMI is shown in Fig. 2, including the average scores and standard deviations of the perceptual rating for the respective subsets. From the figure, we can see that the perceptual rating of style expressiveness was not constant.
4.4 Subjective Evaluation of Estimation Performance for a Professional Narrator’s Emotional Speech

4.4.1 Estimation Result of Style Vectors

For the style estimation of the speaker MMI, we randomly chose 100 samples of each style as test samples from all 500 utterances, and the other 400 utterances of each style were used as training data of the speaker-dependent MRHSMM. We also trained the speaker-adapted MRHSMMs using 5, 10, 25, and 50 utterances of each style, 10, 20, 50, and 100 in total. These utterances for adaptation were chosen randomly from the 400 utterances of each style used as the training data of the speaker-dependent model. When using model adaptation, the estimation performance is sensitive to the choice of adaptation data because the amount of the adaptation data is much smaller than that of the training data of the speaker-dependent model. To alleviate the problem, we prepared four different adaptation data sets for the 5, 10, 25, and 50 utterances, respectively. Then, we repeated the style estimation four times for the same test samples by changing the adaptation data set, and used the average value of these results as the experimental result. The style estimation was conducted for all 500 samples of all styles. Note that the 500 samples of each style to be estimated included the training or adaptation data of the target speaker.

Figure 3 shows the estimation result for all 500 samples. The average scores of the estimated values are shown for the respective subsets with standard deviations. In the figure, ADAPT-5, -10, -25, and -50 represent the results of the style estimation using the adaptation data of 5, 10, 25, and 50 utterances, respectively, and SD-400 represents the result using the speaker-dependent model. From the estimation results by the proposed technique, similar tendencies of the average scores can be seen over the subsets between perceptual and estimated scores. This implies that the estimation results have a close relation to human perception. However, we also found that there was some bias between the estimated values and the perceptual scores. We will discuss this problem in Sect. 5.

4.4.2 Evaluation of Estimation Accuracy of Style Expressiveness by Correlation and Classification

To evaluate the estimation performance of the style expressiveness for emotional speech by the proposed technique, we show quantitatively the relation between estimated values and perceptual scores by calculating the correlation coefficients between the estimated and perceptual values obtained in Sect. 4.3. We also compared the performance of the proposed technique with a conventional approach based on a linear regression analysis (LRA). In the LRA, we used eight explanatory variables for each utterance by computing the following statistics: maximum, mean, and standard deviation of log power, minimum, maximum, mean, and standard deviation of log $F_0$, and mean of mora duration. The regression coefficients were calculated using the same

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Standard deviations of perceptual rating among subjects.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Professional</td>
</tr>
<tr>
<td></td>
<td>Male</td>
</tr>
<tr>
<td>Sad</td>
<td>0.26</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.19</td>
</tr>
</tbody>
</table>

![Fig. 2](image.png)

**Fig. 2** Average scores of perceptual rating of (a) sad style and (b) joyful style for each subset of a professional speaker MMI.

![Fig. 3](image.png)

**Fig. 3** Average scores of estimated style vectors for all 500 samples of (a) sad style and (b) joyful style for each subset of a professional speaker MMI.
Table 2 Correlation coefficients between perceptual intensity of styles and estimated style vectors for a professional narrator.

<table>
<thead>
<tr>
<th>Style</th>
<th>Method</th>
<th>Number of sentences</th>
<th>SD-400</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Sad</td>
<td>LRA</td>
<td>0.169</td>
<td>0.473</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>0.662</td>
<td>0.685</td>
</tr>
<tr>
<td>Joyful</td>
<td>LRA</td>
<td>0.146</td>
<td>0.537</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>0.614</td>
<td>0.616</td>
</tr>
</tbody>
</table>

Table 3 Average classification rates (%) for emotional expressions of a professional narrator.

<table>
<thead>
<tr>
<th>Style</th>
<th>Method</th>
<th>Number of sentences</th>
<th>SD-400</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>Sad</td>
<td>SVM</td>
<td>93.1</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>94.4</td>
<td>95.3</td>
</tr>
<tr>
<td>Joyful</td>
<td>SVM</td>
<td>98.9</td>
<td>98.8</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>100</td>
<td>99.9</td>
</tr>
</tbody>
</table>

data as used for the adaptation of the MRHSMM. Table 2 shows the correlation scores for the 100 test samples of each style. From the result, we can see that the proposed technique gives better correlation with human perception than the conventional one especially when the amount of adaptation data is very limited. Although the correlation drops as the adaptation data decreases in the proposed technique, the correlation coefficients are still over 0.6 for all styles.

Next, we evaluated the proposed technique by a style classification test. We classified the test samples using the following classification criterion. If the value of the estimated style vector was less than a pre-determined threshold, the input speech was classified into the neutral style. If the value was greater than or equal to the threshold, the input speech was classified into the sad or joyful style. To set the classification threshold, we estimated the style vector for the samples of each training data in advance. Then, the mean value of the estimated style vectors was calculated and used as the threshold in the style classification. We also compared the classification performance with a conventional powerful classifier based on the support vector machine (SVM) supported by Weka toolkit [28]. We used the same features and training data as used in the LRA. The classification result of styles is shown in Table 3. It is found that the proposed technique gives comparable performance to the SVM.

4.4.3 Effectiveness of Prosodic Features in Style Estimation

We next evaluated the effectiveness of prosodic features in the style estimation. For the same 100 test samples as in Sect.4.4.2, we estimated the style vector by using the MRHSMM trained with (a) only mel-cepstrum, (b) mel-cepstrum and $F_0$, and (c) all features (proposed). The results are shown in Table 4. It is obvious that the modeling of both $F_0$ and duration contributes to the improvement of the estimation performance in terms of the correlation with perceptual emotional intensity.

4.5 Subjective Evaluation of Estimation Performance for Non-professional Speakers’ Emotional Speech

We also evaluated the estimation performance of the proposed technique under a little more realistic condition. We used the speech data of non-professional speakers described in Sect. 4.1. We randomly chose 50 samples of each style as test samples from all 100 utterances, and the speaker-adapted MRHSMMs of target speakers were trained using 5, 10, 25, and 50 utterances of each style chosen randomly from the rest of the test samples. We conducted the estimation four times by changing the utterances used in the adaptation, and the average scores of the estimated values were used as the estimation results. The style spaces were the same as in Sect. 4.4. Table 5 shows the correlation coefficients between the estimated and perceptual values: (a) is the result for a male speaker MKA and (b) is that for a female speaker FHS. For comparison, the result obtained by the LRA is also shown in the table. The explanatory variables and the other conditions are the same as in Sect. 4.4.2. From the result, although the correlation coefficients of these non-professional speakers were lower than those of a professional narrator shown in Table 3, the pro-
Table 6  Average classification rates (%) for emotional expressions of non-professional speakers.

(a) Male speaker MKA

<table>
<thead>
<tr>
<th>Style</th>
<th>Method</th>
<th>Number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>SVM</td>
<td>98.0</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>97.8</td>
</tr>
<tr>
<td>Joyful</td>
<td>SVM</td>
<td>98.5</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>100</td>
</tr>
</tbody>
</table>

(b) Female speaker FHS

<table>
<thead>
<tr>
<th>Style</th>
<th>Method</th>
<th>Number of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>SVM</td>
<td>97.0</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>95.0</td>
</tr>
<tr>
<td>Joyful</td>
<td>SVM</td>
<td>99.5</td>
</tr>
<tr>
<td></td>
<td>MRHSMM</td>
<td>98.2</td>
</tr>
</tbody>
</table>

The proposed technique still gives better correlation than the LRA in all conditions except for the joyful style of the speaker MKA.

We also evaluated the proposed technique by a style classification test. The classification criterion and thresholds were given in the same way as in Sect. 4.4.2. We compared the classification performance with the SVM under the same condition in Sect. 4.4.2. The classification result of styles is shown in Table 6. Although the correlations become worse compared to those of the professional narrator, the classification performance of the proposed technique is still high and comparable to that of the SVM.

4.6 Evaluation of Proposed Technique in Speaking Style Classification of Spontaneous Speech

In the above experiments, we evaluated the simulated speech with emotions. For a more realistic application, we applied the proposed technique to the classification of speaking styles of the spontaneous speech described in Sect. 4.1. The target speakers’ 100 samples of each style were divided into five subsets of 20 samples. Then, we chose one subset as adaptation data from these subsets, and used the other 80 samples as test data. The same procedure was also followed for the other four subsets. This means that we evaluated all 100 samples as adaptation data by changing the pairs of the adaptation and test data. As the style space, we used a one-dimensional space where the reading style and academic presentation style were positioned at 0.0 and 1.0, respectively. Figures 4 and 5 show histograms of the estimated style vectors for the male and female speakers. From the figures, it can be seen that the style vectors of the two styles had different distributions.

Figures 6 and 7 show the classification error rates for the test samples of each style by changing the threshold from 0.0 to 1.0 with an increment of 0.01. The equal error rates are 2.6% and 0.3% for the male and female speakers, respectively. In the classification, the thresholds were given in the same way as in Sect. 4.4.2. The average values of the determined thresholds and correct classification rates for each speaker are shown in Table 7. For comparison, the classification result obtained by the SVM is also shown in the table. The feature vector and the other conditions are the same in
Table 7  Average values of classification thresholds and correct rates (%) for speaking styles of spontaneous speech.

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Method</th>
<th>Threshold</th>
<th>Reading</th>
<th>Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>SVM</td>
<td>–</td>
<td>82.8</td>
<td>71.5</td>
</tr>
<tr>
<td>Male</td>
<td>MRHSMM</td>
<td>0.485</td>
<td>98.0</td>
<td>95.5</td>
</tr>
<tr>
<td>Female</td>
<td>SVM</td>
<td>–</td>
<td>95.8</td>
<td>75.5</td>
</tr>
<tr>
<td>Female</td>
<td>MRHSMM</td>
<td>0.478</td>
<td>98.3</td>
<td>96.8</td>
</tr>
</tbody>
</table>

Table 8  Correlation with perceptual scores using different style vectors of training data for a professional narrator MMI.

<table>
<thead>
<tr>
<th>Style</th>
<th>Fixed</th>
<th>Perceptual</th>
<th>Estimated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad</td>
<td>0.774</td>
<td>0.826</td>
<td>0.802</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.493</td>
<td>0.712</td>
<td>0.699</td>
</tr>
</tbody>
</table>

Sect. 4.4.2. Although the classification criterion of the proposed technique is quite simple, the results are promising and outperform the conventional one.

5. Discussion

Through the experimental results of style estimation for emotional speech in Sect. 4.4, we showed that the estimated style vectors had correlation with the perceptual intensity of emotional expressions. However, the estimated values showed some bias against the perceptual scores in Fig. 3. There are several factors which cause such bias. A possible reason is that the perceptual scores for the training data of sad and joyful styles had some variation and the mean of the distribution did not coincide with 1.0, though the style vectors for the training data of these styles were assumed to be fixed at 1.0.

To investigate the mismatch between the perceptual scores and fixed style vectors for training data, we estimated style vectors for the test samples using the speaker-dependent MRHSMM of a speaker MMI trained with style vectors labeled by the perceptual scores instead of the fixed ones used in Sect. 4.4. Correlation coefficients between the estimated style vectors and the perceptual scores are shown in Table 8 for each style. The results show that the estimation performance was improved by using the perceptual labels.

In Sect. 4.4, we confirmed that the estimated style vectors had correlation with the perceptual labels. We expect that updating style vectors in the model training will also improve the estimation performance if the estimated style vectors are used for the training data. More specifically, first we train the MRHSMM using fixed style vectors for each style. Then, we update the style vectors for training data by style estimation, and the parameters of the MRHSMM are re-estimated using the updated style vectors. We conducted the style estimation for the same test samples as above using the re-estimated MRHSMM. The correlation coefficients between the estimated style vectors and perceptual scores are shown in Table 8. From these results, it can be seen that the estimation performance was improved by using the style vectors correlating with human perception.

6. Conclusion

In this paper, we have presented a technique for estimating the degree or intensity of emotional expressions and speaking styles appearing in the acoustic features of speech. The technique is based on the simultaneous modeling of both spectral and prosodic information using the MRHSMM, and the explanatory variables of the MRHSMM are estimated for input speech utterances based on an ML criterion.

We have shown through experimental results that the proposed technique can quantitatively detect the expressivity of emotional expressions for simulated emotional speech. For a professional narrator’s speech, we obtained the correlation coefficients over 0.6 for sad and joyful styles when using only five adaptation sentences of each style. It was also shown that the proposed technique performed well in speaking style classification of spontaneous speech.

Future work will include evaluating the recognition performance of the proposed technique by applying it to spontaneous emotional speech. We will also evaluate the proposed technique for various emotions and speaking styles such as angry and dialogue styles.

Acknowledgment

The authors thank Dr. Makoto Tachibana, Yoichi Kato, Shinsuke Izawa, Takeshi Matsubara, and Yusuke Ijima of the Tokyo Institute of Technology for their valuable helps and discussions on this work. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (B) 21300063.

References


[27] The Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk/