PAPER

A Rapid Model Adaptation Technique for Emotional Speech Recognition with Style Estimation Based on Multiple-Regression HMM

Yusuke IJIMA[†*], *Nonmember*, **Takashi NOSE**[†a)], **Makoto TACHIBANA**[†**], *and* **Takao KOBAYASHI**[†b)], *Members*

SUMMARY In this paper, we propose a rapid model adaptation technique for emotional speech recognition which enables us to extract paralinguistic information as well as linguistic information contained in speech signals. This technique is based on style estimation and style adaptation using a multiple-regression HMM (MRHMM). In the MRHMM, the mean parameters of the output probability density function are controlled by a low-dimensional parameter vector, called a style vector, which corresponds to a set of the explanatory variables of the multiple regression. The recognition process consists of two stages. In the first stage, the style vector that represents the emotional expression category and the intensity of its expressiveness for the input speech is estimated on a sentence-by-sentence basis. Next, the acoustic models are adapted using the estimated style vector, and then standard HMM-based speech recognition is performed in the second stage. We assess the performance of the proposed technique in the recognition of simulated emotional speech uttered by both professional narrators and non-professional speakers.

key words: *emotional speech, speaking style, style estimation, multiple-regression HMM, style adaptation, speaker adaptation*

1. Introduction

Speech signals convey not only linguistic information but also paralinguistic and nonlinguistic information, such as the speakers, emotions, and speaking styles. In this context, a wide variety of approaches have been proposed for emotional speech analysis and recognition (e.g., [1]–[3]). The acoustic features of speech are affected by the speaker's emotional states and speaking styles as well as linguistic factors. Such variations can cause mismatches between the acoustic models used in a speech recognition system and the input speech, and could also cause serious deterioration in the recognition performance. A simple approach to this problem is to prepare matched models depending on the respective variations. This might be possible if the variations in the emotion type or style and the degree or intensity of the expressivity are limited and expected. However, in reality the intensity of emotional expressions would change widely. As a result, it is not easy to collect in advance the training data covering all the possible variations of emotional expressions, and thus it would be unrealistic to train a large number of matched models for each variation. In addition, the computational cost of recognition becomes high as the number of emotional intensity variations increases.

One of the more realistic approaches to the problem is to utilize model adaptation. Since variations in emotional expressions appear in every utterance or even in a phrase, it is desirable to perform the model adaptation online. This implies that the model adaptation should be carried out using only a very small amount of data, more specifically, one sentence or one phrase speech. For this purpose, rapid model adaptation techniques based on a small number of control parameters would be more promising than those based on maximum likelihood linear regression (MLLR) [4], because the MLLR generally requires a certain amount of adaptation data to attain considerable performance. Such low-dimensional parameter space-based adaptation techniques include vocal tract length normalization (VTLN) [5], eigenvoice [6], and multiple-regression hidden Markov model (MRHMM) [7].

In this paper, we propose a new rapid model adaptation technique based on a low-dimensional control parameter space for emotional speech recognition. Although the proposed technique utilizes the MRHMM framework, its approach to the modeling of speech is fundamentally different from that of [7]. In the original MRHMM, an additional acoustic feature, that is, fundamental frequency, is used as the explanatory variable of the regression [7]. In contrast, the proposed technique uses the intensity of emotional expressivity that appears in acoustic features of speech as the explanatory variable [8], which is called the *style vector*, rather than the specific acoustic features. The key idea of the technique is based on the style estimation of speech [9] and style control of synthetic speech [10]. In the recognition stage, we first estimate the value of style vector for every sentence of the input speech based on a style estimation technique. Then we conduct the model adaptation by setting the value of the explanatory variable to the estimated style vector and calculating new mean vectors of the probability density functions. After that, we perform standard HMM-based speech recognition.

Manuscript received April 1, 2009.
Manuscript revised July 27, 2009.

[†]The authors are with Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama-shi, 226–8502 Japan.
[*]Presently, with the NTT Cyber Space Laboratories, NTT Corporation, Yokosuka-shi, Japan.
[**]Presently, with Speech Technology Group, Center for Advanced Sound Technologies, YAMAHA Corporation, Japan.
a) E-mail: takashi.nose@ip.titech.ac.jp
b) E-mail: takao.kobayashi@ip.titech.ac.jp
DOI: 10.1587/transinf.E93.D.107

An advantage of the proposed technique is that we can obtain paralinguistic information, that is, the category of the style and its intensity for the input speech as well as linguistic information after the recognition process. In contrast, it is not easy to directory obtain such paralinguistic information when using the eigenvoice technique [6], which is similar to the proposed technique in the sense that the adaptation is based on a low-dimensional parameter vector space. This is because each axis of the eigenspace does not represent a specific emotion or style.

In general, a considerable amount of speech data of a target speaker is required in advance to train the MRHMM [8]. This leads to difficulty in recognizing arbitrary speakers' speech. Although a possible approach to this problem is to use a speaker-independent MRHMM, this would result in an unsatisfactory performance because the expressiveness of emotions and speaking styles varies sensitively with the individual characteristics. To overcome the problem, we use a speaker-independent neutral style model which can be obtained much more easily than speaker- and style-dependent models for the MRHMM training. The speaker-independent model is adapted to the target speaker's style-dependent models based on simultaneous adaptation of speaker and style with a small amount of speech data uttered by the target speaker [11]. Then, the MRHMM of the target speaker is trained from the obtained style-dependent models. In this paper, we examine the effectiveness of the proposed technique under a condition where the types of emotion are limited, and also the amount of training data of the target speaker is very small.

2. Speech Recognition Based on Multiple-Regression HMM

2.1 Acoustic Modeling of Speech with Multiple Styles Using MRHMM

In the MRHMM-based emotional speech recognition framework [8], the acoustic model is represented by MRHMM, i.e., HMM with Gaussian probability density functions (pdfs) in which the mean vector of each pdf is expressed by a function of a low-dimensional vector called the style vector. Each component of the style vector corresponds to an intensity or quantity that represents how much the acoustic features are affected by a certain emotional expression or speaking style.

Here we consider a Gaussian mixture pdf as the output pdf. Let $\boldsymbol{\mu}_{im}$ be the mean vector of the m-th mixture component at state i. In the MRHMM, the mean vector is assumed to be represented by multiple regression of a style vector $\boldsymbol{v}$ as

$$\boldsymbol{\mu}_{im} = \boldsymbol{h}_0^{(im)} + \boldsymbol{A}_{im}\boldsymbol{v} = \boldsymbol{H}_{im}\boldsymbol{\xi} \tag{1}$$

where

$$\boldsymbol{A}_{im} = \left[\boldsymbol{h}_1^{(im)}, \cdots, \boldsymbol{h}_L^{(im)}\right] \tag{2}$$

$$\boldsymbol{H}_{im} = \left[\boldsymbol{h}_0^{(im)}, \cdots, \boldsymbol{h}_L^{(im)}\right] \tag{3}$$

$$\boldsymbol{v} = [v_1, v_2, \cdots, v_L]^\top \tag{4}$$

$$\boldsymbol{\xi} = \left[1, \boldsymbol{v}^\top\right]^\top. \tag{5}$$

$\boldsymbol{A}_{im}$ and $\boldsymbol{H}_{im}$ are $D \times L$- and $D \times (L+1)$-dimensional regression matrices, and D and L are the dimensionalities of $\boldsymbol{\mu}_{im}$ and $\boldsymbol{v}$, respectively. When training data and corresponding style vectors are given, the regression matrix $\boldsymbol{H}_{im}$ of the MRHMM can be estimated using an EM algorithm. Let $\{\boldsymbol{O}^{(1)}, \cdots, \boldsymbol{O}^{(K)}\}$ and $\{\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(K)}\}$ be sets of observation sequences and style vectors for model training, where K is the total number of observation sequences, $\boldsymbol{O}^{(k)} = (\boldsymbol{o}_1^{(k)}, \cdots, \boldsymbol{o}_{T_k}^{(k)})$ is the k-th observation sequence, T_k is the number of frames of $\boldsymbol{O}^{(k)}$, and $\boldsymbol{v}^{(k)}$ is the style vector that corresponds to $\boldsymbol{O}^{(k)}$. The re-estimation formula of the regression matrix of the MRHMM can be derived in a similar way as that for the single mixture model case [10] based on a maximum likelihood (ML) criterion, and is given as follows.

$$\boldsymbol{H}_{im}^{ML} = \left(\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{m=1}^{M}\gamma_t(i,m)\boldsymbol{o}_t^{(k)}\boldsymbol{\xi}^{(k)\top}\right)\cdot \left(\sum_{k=1}^{K}\sum_{t=1}^{T_k}\sum_{m=1}^{M}\gamma_t(i,m)\boldsymbol{\xi}^{(k)}\boldsymbol{\xi}^{(k)\top}\right)^{-1} \tag{6}$$

where M is the number of mixtures of the MRHMM, $\boldsymbol{o}_t^{(k)}$ is an observation vector at time t in $\boldsymbol{O}^{(k)}$, and $\boldsymbol{\xi}^{(k)} = [1, \boldsymbol{v}^{(k)\top}]^\top$. In addition, $\gamma_t(i,m)$ is the probability of being in the m-th mixture component of state i at time t for given $\boldsymbol{O}^{(k)}$.

2.2 Style Estimation for On-Line Model Adaptation

We consider a problem of estimating the style vector $\boldsymbol{v}$ for an input observation sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ given the trained MRHMM λ whose parameters $\boldsymbol{H}_{im}$ and the covariance matrix $\boldsymbol{\Sigma}_{im}$ are fixed. The optimal style vector $\boldsymbol{v}^*$ for the input observation $\boldsymbol{O}$ is determined based on an ML criterion as

$$\boldsymbol{v}^* = \operatorname*{argmax}_{\boldsymbol{v}} P(\boldsymbol{O}|\lambda, \boldsymbol{v}). \tag{7}$$

The EM algorithm-based re-estimation formula of the style vector for the output pdf is given by

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(i,m)\boldsymbol{A}_{im}^\top\boldsymbol{\Sigma}_{im}^{-1}\boldsymbol{A}_{im}\right)^{-1} \left(\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_t(i,m)\boldsymbol{A}_{im}^\top\boldsymbol{\Sigma}_{im}^{-1}(\boldsymbol{o}_t - \boldsymbol{h}_0^{(im)})\right) \tag{8}$$

where N is the number of states of the MRHMM. The above formula is straightforwardly derived from the single mixture model case [9] where the estimation formula is derived

within a hidden semi-Markov model (HSMM) framework, which is the model having explicit state-duration pdfs.

In this study, we assume that the input observation sequence $\boldsymbol{O}$ is a set of acoustic features for one sentence and we estimate the style vector in each sentence.

2.3 Training of MRHMM with a Small Amount of Speech Data Using Model Adaptation

The MRHMM training generally requires a considerable amount of speech data to obtain reliable model parameters. However, it is unrealistic to prepare a sufficient amount of speech data of arbitrary speakers. In the style control and style estimation based on the multiple-regression HSMM (MRHSMM), we have shown that the use of average voice model [12] and simultaneous adaptation of speaker and style is promising for overcoming this problem [11], [13]. Thus we incorporate a similar approach into the MRHMM-based emotional speech recognition.

A block diagram of the model training is illustrated in Fig. 1. First, we train a speaker-independent (SI) neutral style model with a sufficient amount of neutral style speech of many speakers. Next, we adapt the SI neutral style model to a target speaker's respective styles using a model adaptation technique with a small amount of speech data uttered in advance by the target speaker. Then we obtain the target speaker's MRHMM based on least squares estimation from the speaker- and style-adapted HMMs.

Suppose that the adaptation data contains speech uttered in S different styles. Let the mean vector of the m-th mixture pdf at state i of the style-adapted HMM of style s and the corresponding style vector be given by $\boldsymbol{\mu}_{im}^{(s)}$ and $\boldsymbol{v}^{(s)}$, respectively, for $1 \leq s \leq S$. We choose $\boldsymbol{H}_{im}$ that minimizes

$$E = \sum_{s=1}^{S} \left\| \boldsymbol{\mu}_{im}^{(s)} - \boldsymbol{H}_{im} \boldsymbol{\xi}^{(s)} \right\|^2 \tag{9}$$

as the regression matrix of the MRHMM [11], [13]. By differentiating E with respect to $\boldsymbol{H}_{im}$ and equating the result to zero, the optimal regression matrix $\boldsymbol{H}_{im}^{LS}$ is obtained as

$$\boldsymbol{H}_{im}^{LS} = \left(\sum_{s=1}^{S} \boldsymbol{\mu}_{im}^{(s)} \boldsymbol{\xi}^{(s)\top} \right) \left(\sum_{s=1}^{S} \boldsymbol{\xi}^{(s)} \boldsymbol{\xi}^{(s)\top} \right)^{-1}. \tag{10}$$

To improve the performance of the simultaneous adaptation of speaker and style using only a small amount of speech data, we refine the MRHMM parameter $\boldsymbol{H}_{im}$ as follows [11]:

$$\boldsymbol{H}_{im} = \frac{\tau \boldsymbol{H}_{im}^{LS} + \Gamma_{im} \boldsymbol{H}_{im}^{ML}}{\tau + \Gamma_{im}} \tag{11}$$

where $\boldsymbol{H}_{im}^{LS}$ is the regression matrix obtained by Eq. (10) and $\boldsymbol{H}_{im}^{ML}$ is the regression matrix estimated from the adaptation data in ML sense. In addition, τ is a positive parameter for controlling the modification weight and

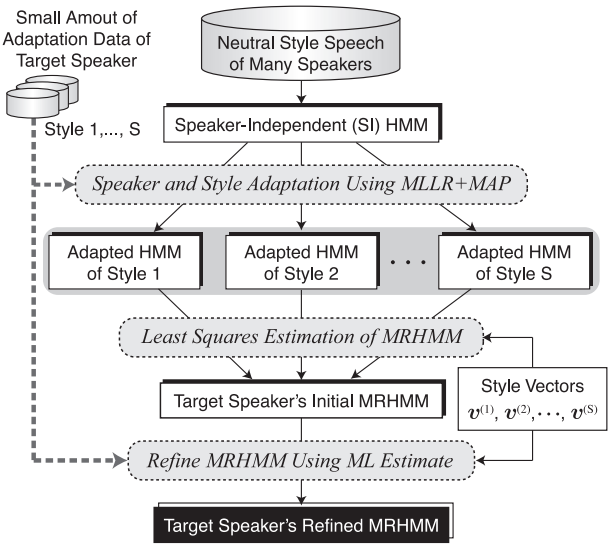


Fig. 1 MRHMM training using SI neutral style model and model adaptation.

$$\Gamma_{im} = \sum_{t=1}^{T} \gamma_t(i, m). \tag{12}$$

It is noted that the regression matrix $\boldsymbol{H}_{im}$ approaches to $\boldsymbol{H}_{im}^{ML}$ when enough adaptation data is available for the m-th mixture component of state i.

2.4 Emotional Speech Recognition Using MRHMM-Based On-Line Model Adaptation

When the trained MRHMM and a specific style vector are given, an HMM having the new mean vectors calculated by Eq. (1) can be obtained. By using this HMM, we can straightforwardly perform ordinary speech recognition based on HMM.

In the proposed technique, first, the style vector is estimated using the style estimation technique mentioned in Sect. 2.2, and then, using the estimated style vector, the adapted HMM for the recognition is obtained from the MRHMM. The style vector is estimated for every input utterance, and the adapted HMM is modified in each utterance. When we perform the style estimation, we need a phoneme label sequence of the input speech to calculate $\gamma_t(i, m)$ by the forward-backward algorithm. For this purpose, we use a two-pass recognition process. The overall recognition process is summarized as follows.

SI model training:

Step 0 Train SI neutral style HMM using neutral style speech data of many speakers.

MRHMM training:

Step 1 Convert the SI neutral style model into the target speaker's respective style models using a model adaptation technique.

Step 2 Construct the target speaker's MRHMM using Eq. (10).

Step 3 Refine the obtained MRHMM using Eq. (11).

MRHMM-based recognition:

Step 4 Obtain *neutral style* HMM by setting the style vector equal to **0**, which is assumed to be the value of the style vector corresponding to the neutral style in the training of the MRHMM.

Step 5 Perform phoneme recognition of input speech using the neutral style HMM.

Step 6 Estimate the style vector $\boldsymbol{\nu}^*$ for the input speech using the trained MRHMM and the phoneme sequence obtained in **Step 5**.

Step 7 Obtain *style-adapted* HMM from the trained MRHMM by calculating the new mean vectors with the estimated style vector $\boldsymbol{\nu}^*$ using Eq. (1).

Step 8 Perform speech recognition using the style-adapted HMM and obtain the final recognition result.

3. Experiments

3.1 Emotional Speech Database

In the following experiments, we used professional narrators' and non-professional speakers' speech. The professional narrators' speech data contains three styles of speech samples with simulated emotions — neutral, sad, and joyful styles, in which 503 phonetically balanced sentences taken from the ATR Japanese speech database were uttered by two males (MMI and MJI) and one female (FTY) narrators in the respective styles. The non-professional speakers' speech data consists of four styles of speech samples — neutral, sad, joyful, and angry styles, uttered with simulated emotions by nine graduate students (eight males and one female). Each style contains 100 sentences chosen from the above 503 sentences. The non-professional speakers had little experience in uttering the given sentence with such simulated styles. All the speech samples were recorded in a quiet room, and the speakers were directed to keep the degree of expressiveness of each style almost constant.

3.2 Experimental Conditions

The SI neutral style model was trained from neutral style speech data of 209 speakers (106 males and 103 females) included in the Japanese Newspaper Article Sentences (JNAS) [14]. These speakers were different from the professional narrators and non-professional speakers mentioned above. The speech data used for the training of the SI neutral style model was about 50 sentences for each speaker, 10498 sentences in total. The parameters of the SI neutral style model were tied using a decision-tree-based context clustering with MDL criterion [15]. The total number of states in the SI neutral style model was 1875.

Table 1 Experimental conditions.

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Analysis window	Hamming window
Feature vector	12 MFCCs (with CMN) +Δ
	Log of power +Δ
Number of monophones	42
Model	left-to-right, 16-mixture, 3-state triphone HMM/MRHMM with diagonal covariance

In the speaker and style adaptation, five sentences (around 20 seconds) of the respective styles were used for each target speaker. To alleviate the dependency of the choice of the adaptation data, the adaptation sentences were randomly chosen and the experiments were conducted twice by changing the adaptation data. As for the model adaptation technique in **Step 1**, we applied a combined approach based on the MLLR and maximum a posteriori (MAP) adaptation (MLLR+MAP) [16]. Since the amount of adaptation data of each target style was small, we used a global transform in the MLLR. In this study, the covariance parameters were not adapted because the amount of adaptation data was very small. We set $\tau = 10$ in Eq. (11) on the basis of preliminary experimental results.

The speech recognition was performed based on the Viterbi algorithm using the decoder of the Hidden Markov Model Toolkit (HTK) [17]. We used phonetic networks based on Japanese phonetic concatenation rules in the recognition. The other experimental conditions are listed in Table 1.

3.3 Performance Evaluation with Professional Narrators

3.3.1 Performance of Speaker and Style Adaptation of the MRHMM

We first evaluated the performance of the speaker and style adaptation by comparing the proposed MRHMM with four types of ordinary HMMs. In this experiment, we used three styles of the professional narrators' speech data. A one-dimensional style space (Fig. 2 (a)) was used for the MRHMM. The style vectors for the adaptation data were set to fixed values, (−1), (0), and (1) for the sad, neutral, and joyful styles, respectively. We performed 10-fold cross-validation tests using 50 test sentences that were not included in the adaptation data.

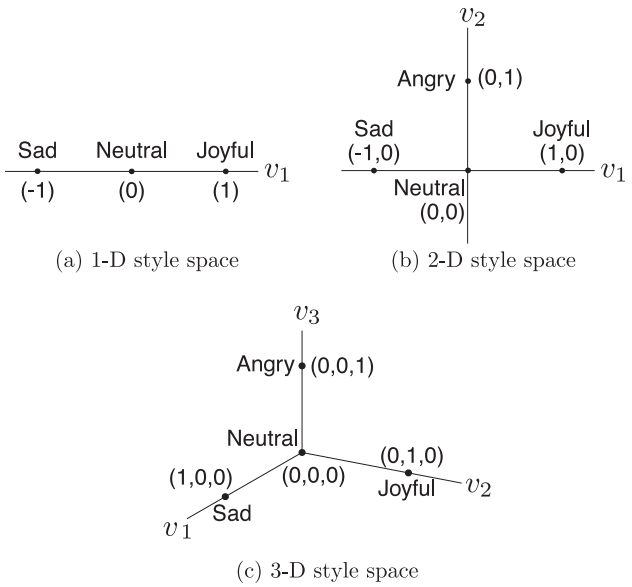
Table 2 shows the average scores of the three speakers' phoneme recognition error rates, and the entry for "Overall" represents the average score of all the styles. The error rate was calculated by

$$error(\%) = \left(1 - \frac{H}{H + D + S}\right) \times 100 \tag{13}$$

where H, S, and D represent the numbers of correctly recognized phonemes, substitutions, and deletions, respectively.

Table 2 Comparison of phoneme error rates (%) between ordinary HMMs and MRHMM.

Model	HMM				1-D MRHMM
Speaker	Independent	Adapted	Adapted	Adapted	Adapted
Style	Neutral	Neutral	Independent	Adapted	Adapted
Neutral	12.3	8.6	8.8	8.6	8.6
Sad	16.8	13.4	11.8	11.5	11.1
Joyful	19.4	16.7	14.9	13.7	13.5
Overall	16.2	12.9	11.8	11.2	11.0



(a) 1-D style space (b) 2-D style space

(c) 3-D style space

Fig. 2 Style spaces for MRHMM.

In the table, the speaker-independent HMM is the SI neutral style model obtained in **Step 0**. The speaker-adapted neutral style HMM is the one adapted from the SI neutral style model using MLLR+MAP with the target speaker's five sentences of neutral style speech. Similarly, the speaker-adapted style-independent HMM is the one adapted from the SI neutral style HMM using MLLR+MAP with the target speaker's five sentences for each style, 15 sentences in total. The style-adapted HMMs are the ones obtained in **Step 1** using the target speaker's five sentences of the respective styles. It is noted that we assumed that the style of the input speech was known when using the style-adapted HMMs, and unknown for the other models. From the result, we can see that the error rates of the MRHMM significantly decreased compared with the speaker-independent HMM. Moreover, we confirmed the improvement of recognition performance to be statistically significant at the 1% level between the MRHMM and the ordinary HMMs except for the style-adapted HMMs. It should be again noted that the results for the style-adapted HMMs were obtained under the condition where the input speech's style was known. It has been found that the recognition performance of the style-adapted HMMs becomes worse when the style of input speech is unknown.

Table 3 Classification rates (%) for professional narrators' emotional speech.

Input Style	Classified Style		
	Neutral	Sad	Joyful
Neutral	98.2	0.4	1.4
Sad	14.8	85.2	0.0
Joyful	15.6	0.0	84.4

3.3.2 Results of Style Estimation and Classification

We also evaluated the performance of the proposed technique in terms of the style estimation. The style classification test was conducted for the test speech samples using the following classification criterion: if the value of the style vector is less than -0.5, then the input speech is classified into sad style; if it is greater than 0.5, then joyful style; otherwise, neutral style. Table 3 shows the average classification rates of the respective styles for the test speech samples of three speakers. In total, about 89% of the speech data were classified as the correct style class of the input speech. This would be promising results in the sense that we estimated the degree of expressivity of the input speech without using prosodic features.

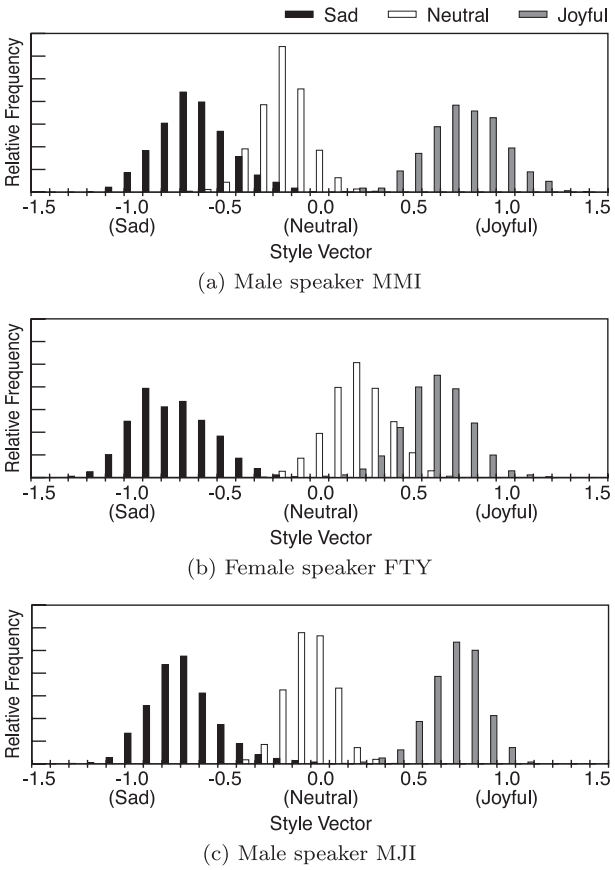
Figure 3 shows the histograms of the estimated values of the style vectors for the test speech samples. It can be seen that different styles give different distributions and the estimated values of the style vector are distributed around the values that were set in the training. However, there is a slight displacement between the mode of each distribution and the value of the style vector assumed in the training. This is because the acoustic features of the sad and joyful styles included in the speech database are not completely symmetric and those of the neutral style are not absolutely located on the mid-point between the sad and joyful styles. As a result, the three styles were influenced by each other in the MRHMM training.

3.4 Performance Evaluation with Non-professional Speakers

We next assessed the performance of the proposed technique using non-professional speakers' speech which is a little more realistic situation than focusing on the professional narrators' speech. We used four styles of the nine

Table 4 Phoneme error rates (%) for non-professional speakers' emotional speech with different style spaces.

Model	HMM				2-D MRHMM	3-D MRHMM
Speaker	Independent	Adapted	Adapted	Adapted	Adapted	
Style	Neutral	Neutral	Independent	Adapted	Adapted	
Neutral	15.1	11.2	11.4	11.2	11.1	10.9
Sad	18.6	15.7	14.8	14.0	13.5	13.4
Joyful	19.4	16.4	15.3	15.3	14.7	14.7
Angry	23.4	20.6	19.0	18.8	17.7	17.7
Overall	19.1	16.0	15.1	14.8	14.3	14.2



(a) Male speaker MMI

(b) Female speaker FTY

(c) Male speaker MJI

Fig. 3 Histograms of the estimated values of the style vectors.

non-professional speakers' speech with simulated emotion. Two different style spaces, namely a two-dimensional space (Fig. 2 (b)) and a three-dimensional one (Fig. 2 (c)) were used for modeling MRHMMs. In the two-dimensional space, the style vectors for adaptation data were set to $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(-1, 0)$ for the neutral, joyful, angry, and sad styles, respectively. In the three-dimensional space, the style vectors for adaptation data were set to $(0, 0, 0)$, $(1, 0, 0)$, $(0, 1, 0)$, and $(0, 0, 1)$ for the neutral, sad, joyful, and angry styles, respectively. We performed two-fold cross-validation tests using 50 test sentences that were not included in the adaptation data.

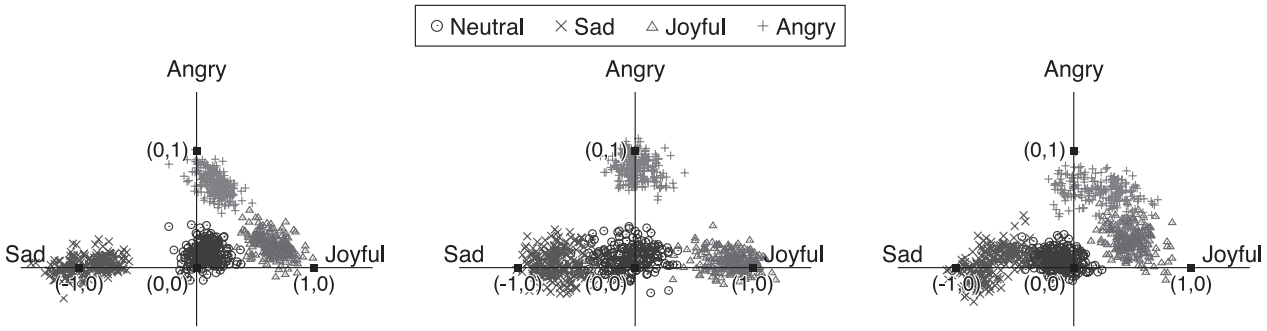
3.4.1 Effect of the Choice of Style Spaces for Speaker and Style Adaptation Performance

We first examined whether the choice of style spaces affect the recognition performance. Table 4 shows the average scores of the nine speakers' phoneme recognition error rates of respective styles. In the table, the entries for "2-D" and "3-D" represent the results for the MRHMM with the two-dimensional and three-dimensional style spaces, respectively. For comparison, we also evaluated the four types of ordinary HMMs described in Sect. 3.3.1. The speaker-adapted style-independent HMM was obtained using adaptation data of the target speaker's five sentences for each style, 20 sentences in total. We again assumed that the style of the input speech was known when using the style-adapted HMMs, and unknown for the other models. It can be seen that both of the 2-D and 3-D MRHMMs gave lower error rates than the ordinary HMMs. It was found that there are significant differences at the 1% level between the ordinary HMMs and the MRHMMs. As for the style spaces, the error rates are comparable in scores between the 2-D and 3-D style spaces, and it seems that the recognition performance is not sensitive to the choice of style spaces.

3.4.2 Results of Style Estimation and Classification Using Different Style Spaces

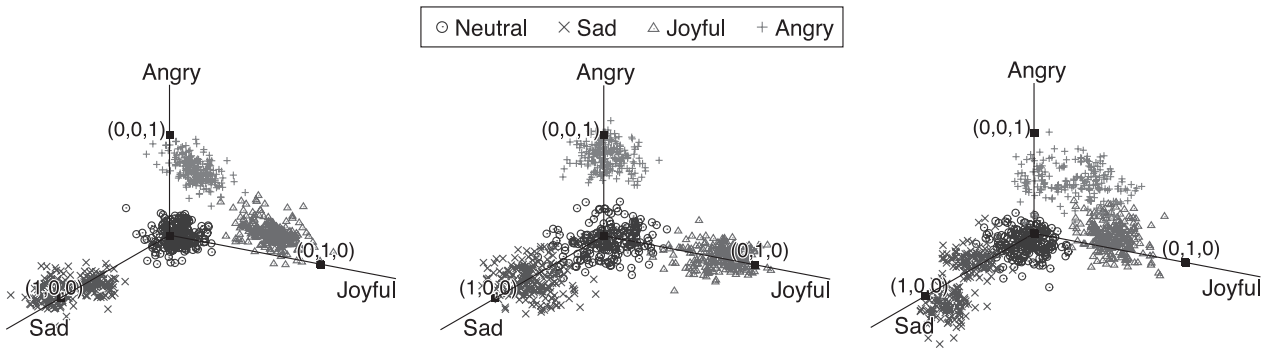
Next, we compared the estimation performance of the degree of emotional expressivity between the 2-D and 3-D style spaces. Figure 4 shows the distributions of the estimated values of the style vectors using the 2-D style space for all the test samples of one female and two male speakers who were arbitrarily chosen from the nine speakers, and Fig. 5 shows those using the 3-D style space. We can see that the distribution of the estimated style vectors belonging to the same style differs from those of other styles.

Table 5 shows the average classification rates of the styles for the test speech samples of nine speakers. The input speech samples were classified based on the Euclidean distance between the predetermined style vector used in the training (see Figs. 2 (b) and (c)) and the estimated style vector. The overall correct classification rates of the MRHMM were 86.8% and 91.3% for the 2-D and the 3-D style spaces, respectively.



(a) Male speaker MKY (b) Male speaker MKA (c) Female speaker FHS

Fig. 4 Examples of the distributions of the estimated style vector with 2-D style space.



(a) Male speaker MKY (b) Male speaker MKA (c) Female speaker FHS

Fig. 5 Examples of the distributions of the estimated style vector with 3-D style space.

Table 5 Classification rates (%) for non-professional speakers' emotional speech with different style spaces.

(a) 2-D style space

Input Style	Classified Style			
	Neutral	Sad	Joyful	Angry
Neutral	99.6	0.3	0.1	0.0
Sad	6.2	93.8	0.0	0.0
Joyful	37.3	0.0	61.2	1.5
Angry	4.9	0.0	2.5	92.6

(b) 3-D style space

Input Style	Classified Style			
	Neutral	Sad	Joyful	Angry
Neutral	99.4	0.5	0.1	0.0
Sad	1.1	99.9	0.0	0.0
Joyful	22.7	0.0	76.3	1.1
Angry	7.6	0.0	2.0	90.4

3.4.3 Performance Evaluation in Continuous Speech Recognition

We examined the performance of the proposed technique for non-professional speakers in terms of the word error rate in continuous speech recognition. Adaptation data and other experimental conditions are described in Sect. 3.1 and Sect. 3.2. We performed two-fold cross-validation tests using 50 test sentences that were not included in the adaptation data. The style space for the MRHMM was the 2-D one. For comparison, we again evaluated the speaker-independent HMM and the speaker-adapted style-independent HMM in Sect. 3.4.1. We used Julius (ver. 4.1) [18] as a decoder. We used one of the sets of lexicons and language models contained in [19]. The vocabulary set of the lexicon contains 60k words, and consists of the most frequent words in Mainichi newspaper articles from the year 1991 to 1994 (45 months). The language models were bigram and backward trigram for the first and second pass, respectively, and

Table 6 Comparison of word error rates (%) between ordinary HMMs and MRHMM.

Model	HMM		2-D MRHMM
Speaker	Independent	Adapted	Adapted
Style	Neutral	Independent	Adapted
Neutral	29.0	23.4	23.2
Sad	36.4	30.4	29.0
Joyful	37.2	31.3	30.1
Angry	44.2	38.2	35.9
Overall	36.7	30.8	29.5

obtained from above newspaper corpus. Although there are some out-of-vocabulary words in the lexicon for test speech sentences, we did not add any words to the lexicon.

Table 6 shows the word error rates for the respective models. We can see that the MRHMM gave the highest performance in all styles. The difference between the speaker-adapted style-independent HMM and the MRHMM is statistically significant at the 1% level. These results show that the proposed technique would also be effective in LVCSR.

4. Conclusion

This paper proposed a technique for emotional speech recognition using rapid model adaptation, in which paralinguistic as well as linguistic information can be obtained. The technique utilizes a multiple-regression HMM (MRHMM) framework, and is based on style estimation and adaptation. Using a speaker-independent neutral style model, the MRHMM is trained with a small amount of target speaker's data. Furthermore, the acoustic models for speech recognition are adapted to the style of input speech from the trained MRHMM using the estimated style vector. From the experimental results of phoneme and continuous speech recognition, we found that the performance of the proposed technique in both speech recognition and style estimation is promising for simulated emotional speech. In our future work, we will explore the effectiveness of the proposed technique using more realistic speech data, such as spontaneous speech, and also develop a technique that would be effective for unknown emotions.

Acknowledgment

A part of this work was supported by JSPS Grant-in-Aid for Scientific Research (B) 21300063.

References

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

[2] L. Bosch, "Emotions, speech and the ASR framework," Speech Commun., vol.40, no.1-2, pp.213–225, April 2003.

[3] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: Clarifying the issues and enhancing performance," Neural Netw., vol.18, no.4, pp.437–444, May 2005.

[4] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput. Speech Lang., vol.9, no.2, pp.171–185, 1995.

[5] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," Proc. ICASSP 96, pp.346–348, May 1996.

[6] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process., vol.8, no.6, pp.695–707, Sept. 2000.

[7] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," Proc. ICASSP 2001, pp.513–516, May 2001.

[8] Y. Ijima, M. Tachibana, T. Nose, and T. Kobayashi, "An on-line adaptation technique for emotional speech recognition using style estimation with multiple-regression HMM," Proc. INTERSPEECH 2008, pp.1297–1300, Sept. 2008.

[9] T. Nose, Y. Kato, and T. Kobayashi, "Style estimation of speech based on multiple regression hidden semi-Markov model," Proc. INTERSPEECH 2007, pp.2285–2288, Aug. 2007.

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

[11] M. Tachibana, S. Izawa, T. Nose, and T. Kobayashi, "Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis," Proc. ICASSP 2008, pp.4633–4636, April 2008.

[12] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," IEICE Trans. Inf. & Syst., vol.E90-D, no.2, pp.533–543, Feb. 2007.

[13] T. Nose, Y. Kato, M. Tachibana, and T. Kobayashi, "An estimation technique of style expressiveness for emotional speech using model adaptation based on multiple-regression HMM," Proc. INTERSPEECH 2008, pp.2759–2762, Sept. 2008.

[14] JNAS: Japanese Newspaper Article Sentences, http://www.milab.is.tsukuba.ac.jp/instruct.html

[15] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn. (E), vol.21, no.2, pp.79–86, March 2000.

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

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

[18] Open-Source Large Vocabulary CSR Engine Julius, http://julius.sourceforge.jp/

[19] K. Shikano, K. Ito, T. Kawahara, K. Takeda, and M. Yamamoto, IT text: Speech recognition system (accompanying CD-ROM), Ohmsha, 2001, ISBN 4-274-13228-5.

Yusuke Ijima received the B.E. degree in electric and electronics from National Institution for Academic Degrees and University Evaluation by graduation from Yatsushiro National College of Technology, Japan, in 2007, and the M.E. degree in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 2009, respectively. He is currently with the NTT Cyber Space Laboratories, NTT Corporation, Yokosuka, Japan. His research interests include speech synthesis, speech recognition and speech analysis. He is a member of ASJ.

Takashi Nose received the B.E. degree in electronic information processing, from Kyoto Institute of Technology, Kyoto, Japan, in 2001, and the Dr.Eng. degree in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 2009, respectively. He was an intern researcher at ATR spoken language communication Research Laboratories (ATR-SLC) in 2008. He is currently a Assistant Professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. His research interests include speech synthesis, speech analysis, and speech recognition. He is a member of IEEE, ISCA, and ASJ.

Makoto Tachibana received the B.E. degree in computer science, M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 2003, 2005, and 2008, respectively. He held a research fellowship from the Japan Society for the Promotion of Science (JSPS) during 2007 - 2008. He is currently with Speech Technology Group, Center for Advanced Sound Technologies, YAMAHA Corporation, Japan. His research interests include speech synthesis, speech analysis, and speech recognition. He is a member of IEEE, ISCA and ASJ.

Takao Kobayashi received the B.E. degree in electrical engineering, the M.E. and Dr.Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 1977, 1979, and 1982, respectively. In 1982, he joined the Research Laboratory of Precision Machinery and Electronics, Tokyo Institute of Technology as a Research Associate. He became an Associate Professor at the same Laboratory in 1989. He is currently a Professor of the Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan. He is a co-recipient of both the Best Paper Award and the Inose Award from the IEICE in 2001, and the TELECOM System Technology Prize from the Telecommunications Advancement Foundation Award, Japan, in 2001. His research interests include speech analysis and synthesis, speech coding, speech recognition, and multimodal interface.