EMOTIONAL SPEECH RECOGNITION BASED ON STYLE ESTIMATION AND ADAPTATION WITH MULTIPLE-REGRESSION HMM

Yusuke Ijima, Makoto Tachibana, Takashi Nose, Takao Kobayashi

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

ABSTRACT

This paper proposes a technique for emotional speech recognition which enables us to extract paralinguistic information as well as linguistic information contained in speech signal. The technique is based on style estimation and style adaptation using multiple-regression HMM. Recognition process consists of two stages. In the first stage, a style vector that represents the emotional expression category and intensity of its variation of input speech is estimated on a sentence-by-sentence basis. Then the acoustic models are adapted using the estimated style vector and standard HMM-based speech recognition is performed in the second stage. We assess the performance of the proposed technique on the recognition of acted emotional speech uttered by both professional narrators and non-professional speakers and show the effectiveness of the technique.

Index Terms— style estimation, multiple-regression HMM (MRHMM), style adaptation, speaker adaptation

1. INTRODUCTION

Speech signal conveys not only linguistic information but also paralinguistic information such as emotions and speaking styles. Although sate-of-the-art speech recognition systems can achieve a high performance in recognition of neutral style speech, they do not always maintain the same recognition performance for emotional or spontaneous speech. This is because acoustic and prosodic features of speech are affected by emotions and speaking styles as well as speaker characteristics and linguistic factors [1], and such variations cause mismatch between the neutral style model and input speech. A simple approach to alleviating this problem is to prepare matched models depending on respective variations. This might be possible when variations in emotions and speaking styles are limited and expected. However, in reality the degree or intensity of emotional expressions and/or speaking styles would change widely and thus it would be unrealistic to train a large number of matched models that covers all possible variations.

One of realistic approaches to the problem is to use model adaptation. Since the variations in emotional expressions or speaking styles appear in every utterance or even in a phrase, it is desirable to perform the model adaptation on-line. This implies that the model adaptation should be carried out using only a quite small amount of data, more specifically, one sentence or one phrase speech. For this purpose, we have proposed a rapid model adaptation technique based on a low-dimensional control parameter space for emotional speech recognition [2]. The technique utilizes a multiple-regression HMM (MRHMM) framework [3] and takes a similar approach to eigenvoice [4]. We showed that the technique gave the recognition performance comparable with that of using the matched models. However the technique has a problem that a considerable amount of speech data of the target speaker is required in advance to train the MRHMM. This leads to difficulty in recognition of arbitrary speaker's emotional speech. Although a possible approach to this problem is to use a speaker-independent MRHMM, the performance would be unsatisfactory because the emotional or style expressiveness varies sensitively on individual characteristics.

In this paper, we propose a technique that enables us to easily obtain an arbitrary speaker's model and to adapt the model on-line. The on-line adaptation process of the proposed technique is the same as the MRHMM-based rapid model adaptation [2]. However, for the MRHMM training, we use a speaker-independent (SI) neutral style model which can be obtained much easier than speaker-dependent style models. The SI model is adapted to 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. Then, the MRHMM of the target speaker is trained from the obtained style-dependent models. 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 [5]. Then we adapt the model by calculating new mean vectors of the probability density functions and perform standard HMM-based speech recognition. An advantage of the proposed technique is that we can obtain paralinguistic information, that is, the category of emotions and its intensity-related value of the input speech as well as linguistic information after the recognition process.

2. MRHMM-BASED SPEECH RECOGNITION

2.1. Acoustic modeling using MRHMM

In the MRHMM-based emotional speech recognition framework [2], the acoustic model is represented by MRHMM, i.e., HMM having Gaussian probability density functions (pdfs) in which the mean vectors 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 a quantity or intensity that represents how much the acoustic features are affected by a certain emotional expression or speaking style.

Let $\boldsymbol{\mu}_i$ be the mean vector of the Gaussian pdf of MRHMM at state i. The mean vector is expressed as

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

where $\boldsymbol{H}_i = [\boldsymbol{h}_0^{(i)}, \cdots, \boldsymbol{h}_L^{(i)}]$, $\boldsymbol{A}_i = [\boldsymbol{h}_1^{(i)}, \cdots, \boldsymbol{h}_L^{(i)}]$, $\boldsymbol{\xi} = [1, \boldsymbol{v}^\top]^\top$, and $\boldsymbol{v} = [v_1, \cdots, v_L]^\top$ is the style vector. For given training data and corresponding style vectors, the parameters of MRHMM, i.e.,

the regression matrix $\boldsymbol{H}_i$ and the covariance matrix $\boldsymbol{\Sigma}_i$ of the output pdf can be estimated in ML sense [6].

2.2. On-line model adaptation based on style estimation

Once an MRHMM has been trained, we can estimate the optimal style vector in ML sense for a given input speech sample using the trained MRHMM. Then, substituting the estimated style vector into (1), we can calculate the new mean vector of the model which is adapted to the input speech style, i.e., a certain emotional expression and/or speaking style [2].

Let λ be the MRHMM and let $\boldsymbol{O} = (\boldsymbol{o}_1, \cdots, \boldsymbol{o}_T)$ be the input observation sequence. Then we estimate the style vector $\boldsymbol{v}$ for $\boldsymbol{O}$ given MRHMM λ. The optimal style vector $\overline{\boldsymbol{v}}$ in ML sense is defined as

$$\overline{\boldsymbol{v}} = \arg\max_{\boldsymbol{v}} P(\boldsymbol{O}|\lambda, \boldsymbol{v}). \tag{2}$$

An EM algorithm-based re-estimation formula of the style vector is given by

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{A}_i \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \gamma_t(i) \boldsymbol{A}_i^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{o}_t - \boldsymbol{h}_0^{(i)}) \right) \tag{3}$$

where N is the number of states, and $\gamma_t(i)$ is the probability of being in state i at time t [5].

2.3. MRHMM training using SI model and model adaptation

The MRHMM training generally requires a considerable amount of speech data, more specifically, several tens minutes speech data of respective styles uttered by the target speaker. However it is unrealistic to prepare sufficient data for arbitrary speakers. In the style control and style estimation based on multiple-regression hidden semi-Markov models (MRHSMMs), we have shown that the use of average voice model and simultaneous adaptation of speaker and style is promising to overcome this problem [7, 8]. Thus we incorporate a similar approach into the MRHMM-based emotional speech recognition.

We first train a speaker-independent (SI) model with a sufficient amount of multiple speakers' neutral speech data. Next we adapt the SI model to target speaker's respective styles using a model adaptation technique with a small amount of speech data uttered by the target speaker in advance. Then we obtain the target speaker's MRHMM based on a least squares estimation from the speaker- and style-adapted HMMs.

Suppose that adaptation data contains speech uttered in S different styles. Let the mean vector of the pdf for style s and corresponding style vector be given by $\boldsymbol{\mu}_i^{(s)}$ and $\boldsymbol{v}_i^{(s)}$, respectively, for $1 \leq s \leq S$. We choose $\boldsymbol{H}_i$ that minimizes

$$E = \sum_{s=1}^{S} \left\| \boldsymbol{\mu}_i^{(s)} - \boldsymbol{H}_i \boldsymbol{\xi}^{(s)} \right\|^2 \tag{4}$$

as the regression matrices of the MRHMM [7, 8]. Differentiating E with respect to $\boldsymbol{H}_i$ and equating the result to zero, we have

$$\overline{\boldsymbol{H}}_i = \left(\sum_{s=1}^{S} \boldsymbol{\mu}_i^{(s)} \boldsymbol{\xi}^{(s)\top} \right) \left(\sum_{s=1}^{S} \boldsymbol{\xi}^{(s)} \boldsymbol{\xi}^{(s)\top} \right)^{-1}. \tag{5}$$

To alleviate a problem due to the fact that the amount of speech data for the simultaneous adaptation of speaker and style uttered by the target speaker in advance is assumed to be small, we refine the MRHMM parameter $\boldsymbol{H}_i$ as follows [8]:

$$\boldsymbol{H}_i = \frac{\tau \overline{\boldsymbol{H}}_i + \Gamma_i \boldsymbol{H}_i^{ML}}{\tau + \Gamma_i} \tag{6}$$

where $\overline{\boldsymbol{H}}_i$ is the regression matrix obtained by (5) and $\boldsymbol{H}_i^{ML}$ 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

$$\Gamma_i = \sum_t \gamma_t(i). \tag{7}$$

It is noted that the regression matrix $\boldsymbol{H}_i$ approaches to $\boldsymbol{H}_i^{ML}$ when enough adaptation data is available at state i,

2.4. MRHMM-based emotional speech recognition

For the given MRHMM, speech recognition process can be done straightforwardly. First, the style vector for the input speech is estimated on a sentence-by-sentence basis. Then, using the estimated style vector, an adapted HMM for recognition is obtained from the MRHMM. After that, the process is the same as the standard HMM-based speech recognition. When performing the style estimation, we need phone transcription of input speech [2, 5]. For this purpose, we use a two-pass recognition process. Overall recognition process is summarized in the following.

SI model training:

Step 0 Train SI model using multiple speakers' neutral style speech data.

MRHMM training:

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

Step 2 Construct the target speaker's MRHMM using (5).

Step 3 Refine the obtained MRHMM using (6).

MRHMM-based recognition:

Step 4 Obtain neutral style HMM by setting the style vector equal to $\mathbf{0}$ in the trainded MRHMM.

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

Step 6 Estimate the style vector $\overline{\boldsymbol{v}}$ for the input speech using the phoneme sequence obtained in **Step 5**.

Step 7 Obtain adapted HMM from the trained MRHMM by calculating the new mean vectors with the estimated style vector $\overline{\boldsymbol{v}}$.

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

3. EXPERIMENTS

3.1. Experimental conditions

In the following experiments, we used professional narrators' and non-professional speakers' speech. Professional narrators' speech database is the identical one that used in the previous study [2]. It contains three styles of speech samples with simulated emotions — neutral, sad, and joyful styles, in which phonetically balanced 503 sentences taken from the ATR Japanese speech database were uttered by two males and one female, MMI, MJI, and FTY, respectively, in each style. Non-professional speakers' speech data consists of four styles of speech samples — neutral, sad, joyful, and

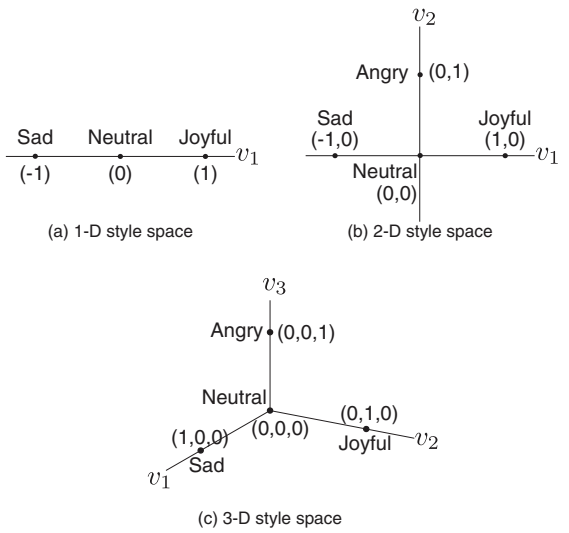


Fig. 1. *Style spaces for MRHMM.*

angry styles, uttered with simulated emotions by eight male and one female graduate students. Each style contains a subset of 100 sentences chosen from the ATR 503 phonetically balanced sentence set. The non-professional speakers have little experience of uttering the given sentence with such simulated styles. All speech samples were recorded in a quiet room, and speakers were directed to keep the degree of expressiveness of each style almost constant.

Speech signals were sampled at a rate of 16kHz and windowed by a 25 ms Hamming window with a 10 ms frame shift. The feature vectors consisted of 12 MFCCs, log energy, and their first-order deltas. We used 42 phonemes including silence and pause. Phonemes were modeled by 3-state single-mixture left-to-right triphone HMMs. Parameter tying of the triphone HMMs were done using decision-tree-based clustering. The same decision tree structure was used in all the models except for SD-MRHMM which will be described in **3.2**.

The speaker-independent (SI) model was trained from six male and four female speakers' neutral style speech data included in the ATR Japanese speech database (Set B). These ten speakers were different from the professional narrators and non-professional speakers mentioned above. Speech data used for the SI model training were 450 sentences for each speaker, 4500 sentences in total.

In the speaker and style adaptation, five sentences (around 20 seconds) for each style were used for each target speaker. To alleviate the dependency of the choice of the adaptation data, the adaptation sentences were chosen randomly and the experiments were conducted twice by changing the adaptation data. As the model adaptation technique, we applied a combined approach based on the maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) adaptation [9]. Since the amount of adaptation data of a target style was small, we used a global transform in the MLLR. We set $\tau = 60$ on the basis of preliminary experimental result.

3.2. Performance with speaker and style adaptation

We first evaluated the performance of speaker and style adaptation by comparing the proposed speaker- and style-adapted MRHMM (SA-MRHMM) with the SI neutral style model (SI-HMM) and speaker-dependent MRHMM (SD-MRHMM). In this experiment, we used three styles of the professional narrators' speech data. A one-dimensional style space (Fig. 1(a)) was used. SD-MRHMM was trained using 450 sentences for each style, 1350 sentences in total, for each target speaker. In the SA-MRHMM training, the initial model was the SI-HMM and adapted to the target speaker's respective style models using five sentences for each style. We performed a 10-fold cross-validation test. It is noted that these experimental conditions are the same as the SD-MRHMM case described in [2].

Table 1. *Comparison of phoneme error rates (%) for SI-HMM, SD-MRHMM, and SA-MRHMM.*

Input Style	Model		
	SI-HMM	SD-MRHMM	SA-MRHMM
Neutral	20.24	5.91	10.62
Sad	27.95	7.07	14.13
Joyful	23.92	7.60	14.99
Overall	24.03	6.87	13.25

Table 2. *Phoneme error rates (%) for non-professional speakers' emotional speech.*

	SI-HMM	2-D	3-D
Neutral	22.69	15.78	15.55
Sad	29.17	19.14	19.19
Joyful	25.14	18.51	18.63
Angry	30.32	21.92	21.96
Overall	26.83	18.84	18.83

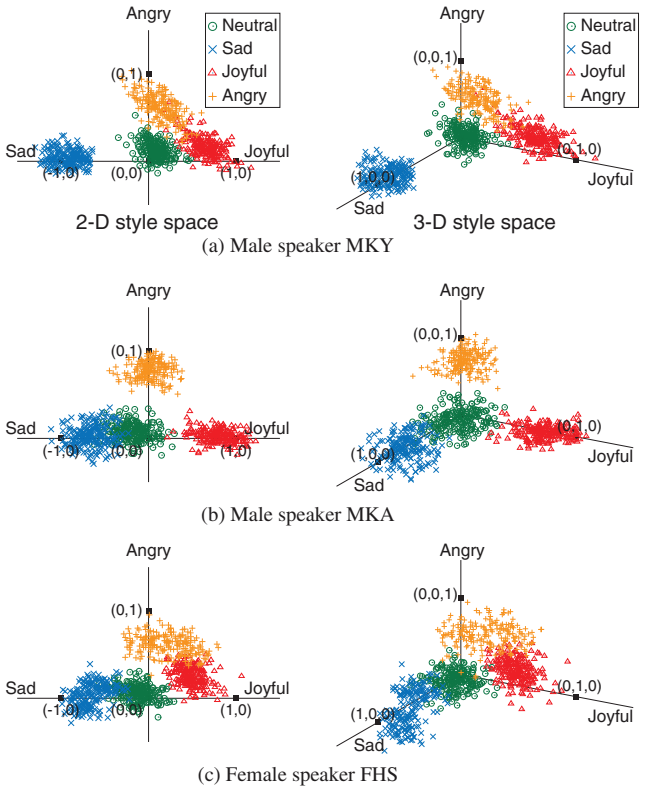
Table 1 shows the average scores of the three speakers' phoneme recognition error rates. The error rate was calculated based on the number of correctly recognized phonemes, substitutions, and deletions. "Overall" represents the average score of all the styles. Although the error rates for SA-MRHMM are larger than SD-MRHMM, it is obvious that they are significantly smaller than SI-HMM. Moreover it should be emphasized that the number of target speaker's utterances used for the model training was only five sentences of each style for SA-MRHMM, whereas 450 sentences for SD-MRHMM. In addition, as is the case with SD-MRHMM [2], we have found that the phoneme recognition errors in the first pass (**Step 5**) do not affect the final recognition result significantly.

3.3. Performance evaluation with non-professional speakers

We next assessed the performance of the proposed technique using non-professional speaker's speech which leads to a more realistic situation than the professional narrators' speech. We used four styles of non-professional speakers' speech with simulated emotion. Two different style spaces, namely two-dimensional space (Fig. 1(b)) and three-dimensional ones (Fig. 1(c)) were used for modeling MRHMMs. In this experiment, we performed a two-fold cross-validation test using 50 test sentences that were not included in the adaptation data.

Table 2 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 SA-MRHMM with the two-dimensional and three-dimensional style spaces, respectively. It is seen that the error rates decreased significantly in both cases of SA-MRHMM compared with the SI neutral style model. The results for 2-D and 3-D style spaces are comparable in scores.

Figure 2 shows the distributions of the estimated values of the style vector for the whole test speech samples of one female and two male speakers who are arbitrarily chosen from nine speakers. We can see that the distribution of the estimated style vectors belonging



(a) Male speaker MKY

(b) Male speaker MKA

(c) Female speaker FHS

Fig. 2. *Distributions of estimated values of the style vector for non-professional speakers' test samples.*

to the same style differs from those to other styles. This result seems to be promising despite of using no prosodic features and giving no guarantee that the style space is orthogonal. In other words, we would obtain certain paralinguistic information as well as the speech recognition result. In fact, the average correct style classification rates of nine speakers based on the Euclidean distance in the 3-D style space were 99.1, 97.6, 65.3, and 86.2% for neutral, sad, joyful, and angry styles, respectively.

3.4. Performance comparison with multi-mixture HMMs

We further compared the performance of MRHMM with ordinary HMMs. The evaluation test was the same as described in **3.3**, and the style space for the MRHMM (SA-MRHMM) was the 3-D one. We used speaker- and style-adapted HMMs (SA-HMMs) obtained in **Step 1** using target speaker's five sentences of each style as the baseline. We also trained the following style-independent HMMs. Style-independent single-mixture HMM (1-M HMM) is a model adapted from the SI-HMM using the target speaker's five sentences for each style, 20 sentences in total. In contrast, style-independent four-mixture HMM (4-M HMM) is a combined model obtained by collecting Gaussian components of four SA-HMMs. The mixture weights were given uniformly. Both of the style-independent HMMs were refined by the MAP adaptation using the target speaker's adaptation data. It is noted that we assumed that the style of the input speech was known when using SA-HMMs, and unknown for other models.

The result is shown in Table 3. The performance of the standard HMMs was improved and comparable with that of the SA-MRHMM as the number of mixtures of the HMM was increased. However, the number of parameters of SA-MRHMM is smaller than 4-M HMM. Moreover, it is again emphasized that we can obtain the paralinguistic information in addition to the linguistic information.

Table 3. *Comparison of phoneme error rates (%) of respective models for non-professional speakers' emotional speech,*

	SA-HMM	1-M HMM	4-M HMM	SA-MRHMM
Neutral	16.11	16.09	15.98	15.55
Sad	20.14	20.58	19.60	19.19
Joyful	19.34	19.75	18.48	18.63
Angry	23.30	23.41	22.41	21.96
Overall	19.72	19.96	19.12	18.83

4. CONCLUSIONS

In this paper, we have presented a technique for emotional speech recognition which can obtain paralinguistic information as well as linguistic information. The technique utilizes the multiple-regression HMM (MRHMM) framework and is based on style estimation and adaptation. Using a speaker-independent neutral style model, MRHMM is trained with a small amount of target speaker's data. Furthermore, the acoustic model for speech recognition is adapted to input speech style from the trained MRHMM. We have shown that the performance of proposed technique for both the style estimation and speech recognition is promising. In our future work, we will explore effectiveness of the proposed technique using more realistic speech data, such as spontaneous speech.

5. REFERENCES

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

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

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

[4] 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, Nov. 2000.

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

[6] 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.

[7] 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," in *Proc. INTERSPEECH 2008*, Sept. 2008, pp. 2759–2762.

[8] 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," in *Proc. ICASSP 2008*, Apr. 2008, pp. 4633–4636.

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