interspeech 2008 incorporating SST 08

An On-line Adaptation Technique for Emotional Speech Recognition Using Style Estimation 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
{yusuke.ijima, makoto.tachibana, takashi.nose, takao.kobayashi}@ip.titech.ac.jp

Abstract

This paper describes a model adaptation technique for emotional speech recognition based on multiple-regression HMM (MR-HMM). We use a low-dimensional vector called style vector which corresponds the degree of expressivity of emotional speech as the explanatory variable of the regression. In the proposed technique, first, the value of the style vector for input speech is estimated. Then, using the estimated style vector, new mean vectors of the output distributions of HMM are adapted to the input style. The style vector is estimated every input utterance, and an on-line adaptation can be done in each utterance. We perform phoneme recognition experiments for professional narrators' acted speech and evaluate the performance by comparing with style-dependent and style-independent HMMs. Experimental results show the proposed technique reduced the error rates by 11% of the style-independent model.

Index Terms: emotional speech, speaking style, style estimation, multiple-regression HMM

1. Introduction

To bring human-computer or -robot interaction more natural and realistic, we need a speech recognition system that can accept speech with various speaking styles and emotional expressions. Acoustic features of speech are affected by speaking styles and emotions as well as speaker characteristics and linguistic factors. This fact causes serious deterioration of the performance on recognition of emotional or spontaneous speech. One of useful approaches to alleviating such a problem is to adapt the acoustic model. Since the degree of expressivity of speaking styles and emotional expressions would change in every utterance or even in a phrase, it is desirable to perform the model adaptation on line. This implies that we encounter a difficulty of adapting the acoustic model with using only a limited amount of data, more specifically, one sentence or one phrase.

For this purpose, rapid model adaptation techniques based on a small number of control parameters would be promising than that based on maximum likelihood linear regression (MLLR) [1], because 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) [2], Eigenvoice [3], and multiple-regression HMM (MR-HMM) [4]. In this paper, we propose a novel adaptation technique based on a quite small number of control parameters for emotional speech recognition.

The proposed technique utilizes the MR-HMM framework for the model adaptation. However, the approach to the modeling of speech is fundamentally different from that of [4]. In the original MR-HMM, an additional acoustic feature, that is, fundamental frequency (F0), is used as the explanatory variable of the regression. In contrast, the proposed technique uses the degree or intensity of expressivity of emotions and styles appeared in acoustic features of speech, which is called the *style vector*, as the explanatory variable rather than specific acoustic features. The key idea of the technique is based on the style estimation [5] and style control [6] techniques of speech. We first estimate the value of the style vector, then conduct the model adaptation by setting the value of the explanatory variable to the estimated style vector and calculating new mean vectors of the output distribution functions of HMM. As a result, we can obtain para-linguistic information, that is, the degree of expressivity of emotional speech as well as linguistic information after recognition process. In the eigenvoice technique, since each axis of the eigenspace does not represent the degree of expressivity of emotional speech, it is not easy to obtain such para-linguistic information directly. We show the effectiveness of the proposed technique from results of phoneme recognition experiments for acted emotional speech.

2. Multiple-regression HMM

2.1. Style modeling based on MR-HMM

Let $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ be the mean vector and covariance matrix of output distribution of HMM at state i. In this paper, we assume that the mean vector of MR-HMM is modeled using multiple regression 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 a style vector. The component v_k of the style vector represents the degree of expressivity of a certain emotional expressions or speaking style in speech. In addition, $\boldsymbol{H}_i$ is the regression matrix of dimension $M \times (L+1)$ and M is the dimensionality of $\boldsymbol{\mu}_i$.

When the training data and corresponding style vectors are given, the parameters of MR-HMM, i.e., $\boldsymbol{H}_i$ and $\boldsymbol{\Sigma}_i$ can be estimated using EM algorithm. These estimation formulas can be found in [6].

2.2. Style estimation

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 MR-HMM λ whose parameters $\boldsymbol{H}_i$ and $\boldsymbol{\Sigma}_i$ are fixed. The optimal style vector $\overline{\boldsymbol{v}}$ for the input observation $\boldsymbol{O}$ is determined

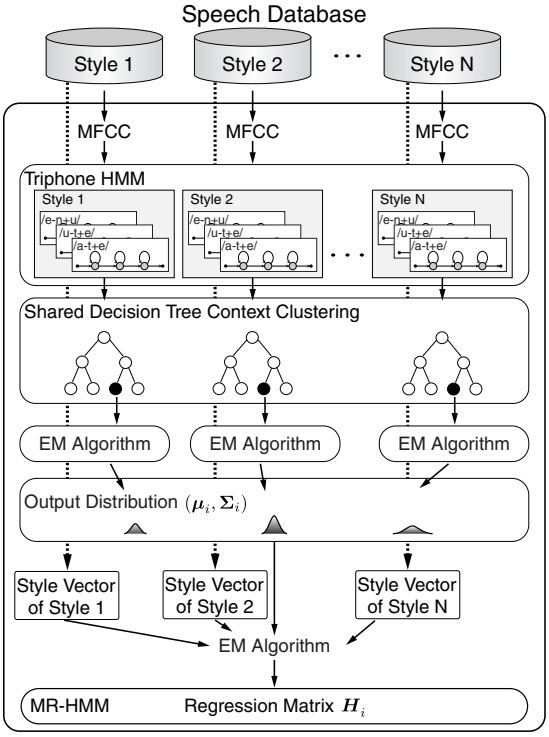


Figure 1: *A block diagram of MR-HMM training.*

in ML sense as

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

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

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

where N is the number of states, and $\gamma_t(i)$ is the probability of being in state i at time t. The derivation of the equation can be found in [5]. Note that the estimation formula in [5] is derived in hidden semi-Markov model (HSMM) framework which has explicit state duration pdfs.

In this study, we assume that the input observation sequence $\boldsymbol{O}$ is a whole sentence, and estimate the style vector in each sentence.

3. Speech recognition using MR-HMM

3.1. MR-HMM training

A block diagram of the training part for the MR-HMM is shown in Fig. 1. We first train triphone HMMs for respective styles, such as neutral, sad, and joyful styles, independently. Then we apply a shared decision tree context clustering (STC) technique [7] to these models to construct a common tree structure for all styles. After that, we further apply re-estimation process based on the EM algorithm to the resultant triphone HMM of each style. Finally, we obtain a single model with the common tree structure for all styles by incorporating the style vector into the re-estimation procedure based on the EM algorithm for MR-HMM.

Table 1: *Experimental conditions.*

Sampling frequency	16 kHz
Frame length	25 ms
Frame shift	10 ms
Analysis window	Hamming window
Feature vector	12 MFCCs $+\Delta$
	Log of power $+\Delta$
Number of monophones	42
Model	left-to-right, 1mix., 3-state triphone HMM

3.2. Overview of the recognition system

When the trained MR-HMM and a specific style vector are given, an HMM having the new mean vectors calculated by Eq.(1) can be obtained. By using the obtained HMM, we can straightforwardly perform ordinary speech recognition based on Viterbi algorithm. In the proposed technique, first, the style vector is estimated using style estimation technique mentioned in Section **2.2**, then, using the estimated style vector, the adapted HMM for recognition is obtained from the MR-HMM. The style vector is estimated every input utterance, and the adapted HMM is modified in each utterance. This recognition process can be viewed as a kind of an on-line adaptation. When we perform the style estimation, we need a label sequence of the input speech. In [5], the triphone label is obtained by a pre-trained style-independent HMM. In this study, for the given MR-HMM, we use a two-pass recognition process summarized as follows:

Step 1 Obtain the *initial* HMM by setting the style vector equal to $\mathbf{0}$ in MR-HMM.

Step 2 Perform phoneme recognition using the initial HMM.

Step 3 Estimate the style vector $\overline{\boldsymbol{v}}$ for input speech using phoneme label of the input speech obtained in **Step 2**.

Step 4 Obtain the *adapted* HMM from MR-HMM by calculating the new mean vectors with the estimated style vector $\overline{\boldsymbol{v}}$.

Step 5 Perform phoneme recognition again using the adapted HMM and obtain the recognition result.

4. Experiments

4.1. Experimental conditions

We used three styles of professional narrators' acted speech – neutral, sad, and joyful styles. Speech database [8] of each style contains 503 phonetically balanced ATR Japanese sentences (about 50 minutes) uttered by two male and one female professional narrators, MMI, MJI, and FTY, respectively. The neutral, sad, and joyful style speech data were not real emotional speech data, but just read speech data with simulated styles. When we recorded the speech data, we directed the speakers not to vary so much the degree of expressivity in each style.

450 sentences were used for the training of MR-HMM and fifty sentences not included in the training data were used as the

Table 2: *Phoneme error rates (%) for neutral style-dependent HMM.*

Input Style	Speaker			
	MMI	FTY	MJI	Average
Neutral	5.47	7.32	4.45	5.75
Sad	5.89	12.36	21.87	13.28
Joyful	14.90	12.32	18.83	15.34

Table 3: *Phoneme error rates (%) for initial and adapted HMMs.*

Input Style	HMM	Speaker			
		MMI	FTY	MJI	Average
Sad	initial	4.79	9.61	10.86	8.42
	adapted	4.23	8.54	8.48	7.07
	correct label	4.20	8.51	8.38	7.03
Joyful	initial	9.44	11.19	8.74	9.79
	adapted	6.69	9.63	6.49	7.60
	correct label	6.56	9.49	6.38	7.48

evaluation data in each style. We performed a 10-fold cross-validation test. A one-dimensional style space was used and the style vectors of training data were set as (-1.0), (0.0), and (1.0) for the sad, neutral, and joyful styles, respectively.

To compare the recognition performance between MR-HMM and HMMs, we also trained style-dependent HMMs (SD) and style-independent HMM (SI). The SD models were trained from 450 sentences of each style and the SI model was a single model trained by 1350 sentences combining 450 sentences of the three styles. In general, the tying topology of HMM affects the recognition performance. To compare the performance of the models under the same condition, we chose the topology of the HMMs and the MR-HMM to be the same tree structure obtained by STC described in Section **3.1**. Other experimental conditions are shown in Table 1.

4.2. Recognition using neutral style HMM

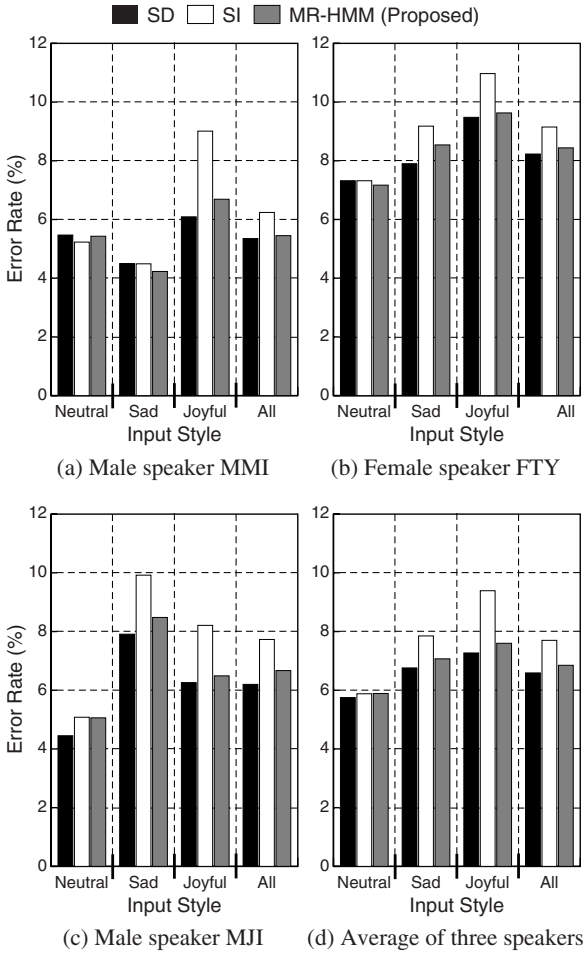
To examine the influence of emotional speech on the recognition rate, we first performed phoneme recognition using the neutral style-dependent HMMs. Table 2 shows the recognition error rate of each speaker and style. In the table, the entry for "Average" means the average result of the three speakers. The error rate was calculated by

$$error(\%) = \left(1 - \frac{H}{H + D + S}\right) \times 100 \qquad (4)$$

where H, S, and D represent the number of correctly recognized phonemes, substitutions, and deletions, respectively. We can see that the error rates increased in sad and joyful styles compared to neutral style. From this result, we confirmed that the acoustic features were different in each style and that degrade the recognition performance.

4.3. Adaptation performance of MR-HMM

Table 3 shows the recognition error rate of the initial HMM obtained in **Step 2**, and the adapted HMM obtained in **Step 4**. In the table, the entry for "correct label" represents the result



(a) Male speaker MMI (b) Female speaker FTY

(c) Male speaker MJI (d) Average of three speakers

Figure 2: *Phoneme recognition error rates (%).*

when using the correct phoneme label of the input speech in the style estimation in **Step 3** and this corresponds to the upper limit of the performance of the proposed technique. It can be seen that the adapted HMM gives higher performance than the initial HMM. Moreover, the result of the adapted HMM is close to that of the "correct label" case.

4.4. Performance comparison between HMM and MR-HMM

Figure 2 shows the recognition error rate of the style-dependent (SD) models, style-independent (SI) model, and MR-HMM. In this figure, "All" represents the average result of the all three styles. As for the SD models, the style of the input speech was assumed to be known and the recognition result was obtained using the SD model of the input style. Thus, it is an ideal case and, in actual use, the style of input speech should be identified by some classification techniques. We can see that the error rates of the proposed technique decreased and were less than or comparable to that of the SD models. The error rates of SI model were reduced compared to that of neutral style-dependent model in Table 2, however, it is still higher than SD models and MR-HMM.

Table 4 shows the error reduction rates of the proposed technique from the SI model. It can be seen that the MR-HMM reduced the error rate by 11.04% on the average. Especially, the

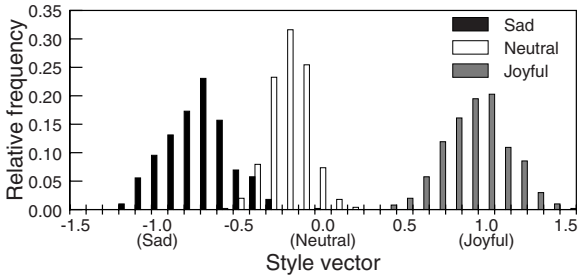
Table 4: *Error reduction rates (%) from the SI model.*

Input Style	Speaker			
	MMI	FTY	MJI	Average
Neutral	-3.82	2.05	0.39	-0.17
Sad	5.79	10.97	14.52	9.97
Joyful	25.75	12.22	20.95	19.06
All	12.66	7.76	13.71	11.04

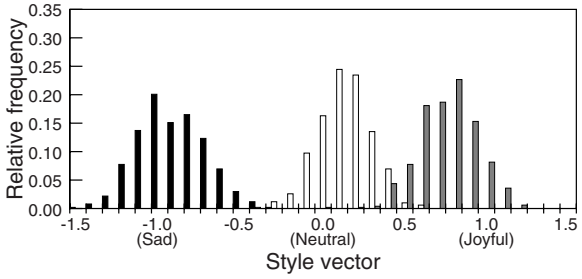
performance is improved in the joyful and sad styles.

4.5. Style estimation using MR-HMM

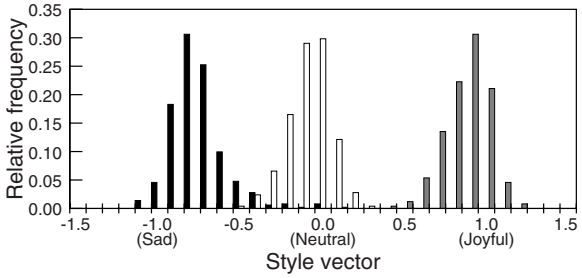
Figure 3 shows the distributions of the estimated value of the style vector of the test speech data. It can be seen that each style gives a different distribution and each distribution is near the value of the style vector that were set in the training. When we chose the classification threshold as from -1.5 to -0.5 for sad, from -0.5 to 0.5 for neutral, and from 0.5 to 1.5 for joyful, about 96% of speech data were classified as the correct style class of the input speech. 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 style included in the database are not completely symmetric and that of neutral style are not absolutely located mid-point between the sad and joyful styles. As a result, the three styles were influenced by each other in the MR-HMM training. Two-dimensional style space in which the sad style and joyful style are located in the independent axis might be one of the approaches to overcoming the problem.



(a) Male speaker MMI



(b) Female speaker FTY



(c) Male speaker MJI

Figure 3: *Histograms of the estimated value of the style vector.*

5. Conclusions

In this paper, we have presented a speech recognition technique considering the degree of expressivity of speaking styles or emotional expressions. This technique utilizes the multiple-regression HMM (MR-HMM) framework for the model adaptation and the style vector which corresponds to the degree or intensity of expressivity of styles as the explanatory variable of the regression. In the recognition stage, we adapt the HMM to the input style using the estimated style vector. We have shown that the proposed technique reduced the error rates by 11% of the style-independent HMM. Furthermore, we can obtain not only linguistic information but also the degree of expressivity of emotional and styled speech from the recognition process. This para-linguistic information would be also useful for human-computer or -robot interaction to detect user's emotional state and respond to the user more natural and appropriately.

In our future work, we will explore effectiveness of the proposed technique using more realistic speech data, such as spontaneous speech. Speaker-independent recognition using speaker adaptation techniques [9] will also done in the future.

6. Acknowledgments

A part of this work was supported by Grant-in-Aid for JSPS Fellows (1910295).

7. References

[1] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," Comput. Speech Lang., 9(2):171–185, 1995.

[2] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," Proc. ICASSP 96, 1:346–348, May 1996.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," IEEE Trans. Speech Audio Process., 8(6):695–707, Nov. 2000.

[4] K. Fujinaga, M. Nakai, H. Shimodaira and S. Sagayama, "Multiple-regression hidden Markov model," Proc. ICASSP 2001, 1:513–516, May 2001.

[5] T. Nose, Y. Kato and T. Kobayashi, "Style estimation of speech based on multiple regression hidden Semi-Markov model," Proc. INTERSPEECH 2007, 1:2285–2288, Oct. 2007.

[6] K. Miyanaga, T. Masuko and T. Kobayashi, "A style control technique for HMM-based speech synthesis," Proc. INTERSPEECH 2004-ICSLP, 1:1437–1440, Oct. 2004.

[7] J. Yamagishi, M. Tamura, T. Masuko, K. Tokuda and T. Kobayashi, "A context clustering technique for average voice models," IEICE Trans. Inf. & Syst., E86-D(3):534–542, Mar. 2003.

[8] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," IEICE Trans. Inf. & Syst., E88-D(3):502–509, Mar. 2005.

[9] 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, 4633–4636, April 2008.