INTERSPEECH 2009 BRIGHTON

Speaking Style Adaptation for Spontaneous Speech Recognition Using Multiple-Regression HMM

Yusuke Ijima, Takeshi Matsubara, Takashi Nose, Takao Kobayashi

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

10.21437/Interspeech.2009-198

Abstract

This paper describes a rapid model adaptation technique for spontaneous speech recognition. The proposed technique utilizes a multiple-regression hidden Markov model (MRHMM) and is based on a style estimation technique of speech. In the MRHMM, the mean vector of probability density function (pdf) is given by a function of a low-dimensional vector, called style vector, which corresponds to the intensity of expressivity of speaking style variation. The value of the style vector is estimated for every utterance of the input speech and the model adaptation is conducted by calculating new mean vectors of the pdf using the estimated style vector. The performance evaluation results using "Corpus of spontaneous Japanese (CSJ)" are shown under a condition in which the amount of model training and adaptation data is very small.

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

1. Introduction

Achieving high recognition accuracy for spontaneous speech still remains a difficult problem to be solved in speech recognition research area [1]. The acoustic features of speech are affected by speaking styles and speaker's emotional state as well as linguistic factors. This fact leads to serious deterioration in recognition performance due to the mismatch between the acoustic models and the input speech. A simple approach to this problem is to adapt the acoustic model to the input speaking style and/or emotional expression. Since variations in speaking styles and emotional expressions appear in every utterance or even in a phrase, it is desirable to perform the model adaptation on-line and, therefore, the adaptation must be done using only a small amount of data. Although it has been reported that there exist distinct differences in acoustic features between read speech and spontaneous speech [2, 3], it has not been shown clearly how these results can be utilized for adapting the acoustic models and improving the recognition performance.

As for emotional speech recognition, we have proposed a rapid model adaptation technique based on a quite small number of control parameters [4] and also shown that the proposed approach would be promising under a condition in which available adaptation data is limited [5]. This technique utilizes a multiple-regression hidden Markov model (MRHMM) [6] framework for the model adaptation, but takes a different approach to the modeling from the original MRHMM, that is, a low-dimensional vector, which represents the intensity of emotional expressivity of speech, is used as the explanatory variable. The key idea of the technique is based on the style estimation of speech [7] and style control of synthetic speech [8]. In the recognition process, the value of the style vector is estimated for every sentence of the input speech using 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 (pdfs). An advantage of the proposed technique is that we can obtain paralinguistic information, that is, the category of the emotional expression and its intensity for the input speech as well as linguistic information after the recognition process.

In this paper, we reformulate the MRHMM-based adaptation framework into that for multi-mixture models. Then we apply it to spontaneous speech recognition and examine the basic recognition performance of the proposed technique using "Corpus of spontaneous Japanese (CSJ)" [9]. Although it has been shown that a large amount of training and adaptation data is generally required for obtaining reliable acoustic models for spontaneous speech recognition [1], we cannot always obtain a sufficient amount of data of every target speaker in a realistic situation. Hence here we focus on the performance evaluation under the condition where the amount of available data of each target speaker for both model training and adaptation is very small, more specifically, five utterances of each speaking style for the model training, and only one input utterance for the speaking style adaptation. We also evaluate the ability of the proposed technique to discriminate the speaking style of the input speech.

2. Speech recognition based on MRHMM

2.1. Acoustic modeling of speech with multiple styles

In the MRHMM-based speech recognition framework [4, 5], the acoustic model is represented by MRHMM, i.e., HMM with Gaussian pdf in which the mean vector 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 $\boldsymbol{v} = [v_1, v_2, \cdots, v_L]^\top$ as

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

where $\boldsymbol{\xi} = \left[1, \boldsymbol{v}^\top\right]^\top$, $\boldsymbol{H}_{im} = \left[\boldsymbol{h}_0^{(im)}, \cdots, \boldsymbol{h}_L^{(im)}\right]$, and $\boldsymbol{A}_{im} = \left[\boldsymbol{h}_1^{(im)}, \cdots, \boldsymbol{h}_L^{(im)}\right]$. In addition, $\boldsymbol{A}_{im}$ and $\boldsymbol{H}_{im}$ are $D \times L$- and $D \times (L+1)$-dimensional regression matrices, and D is

the dimensionality of $\boldsymbol{\mu}_{im}$. 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 [8] based on a maximum likelihood (ML) criterion, and is given as follows.

$$\overline{\boldsymbol{H}}_{im} = \left(\sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{o}_t^{(k)} \boldsymbol{\xi}^{(k)\top} \right) \cdot \left(\sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{\xi}^{(k)} \boldsymbol{\xi}^{(k)\top} \right)^{-1} \quad (2)$$

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, $\gamma_t(i,k)$ 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

$$\boldsymbol{v}^* = \underset{\boldsymbol{v}}{\operatorname{argmax}} \, P(\boldsymbol{O}|\lambda, \boldsymbol{v}). \quad (3)$$

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

$$\overline{\boldsymbol{v}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{A}_{im}^\top \boldsymbol{\Sigma}_{im}^{-1} \boldsymbol{A}_{im} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_t(i,m) \boldsymbol{A}_{im}^\top \boldsymbol{\Sigma}_{im}^{-1} (\boldsymbol{o}_t - \boldsymbol{h}_0^{(im)}) \right) \quad (4)$$

where N is the number of states of the MRHMM. The above formula is straightforwardly derived from the single mixture model case described in [7] where the estimation formula is derived within a hidden semi-Markov model (HSMM) framework.

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 every sentence.

2.3. Training of MRHMM with a small amount of speech data using model adaptation

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, we have shown that the use of speaker-independent (SI) model and simultaneous adaptation of speaker and style is promising for overcoming this problem [11, 12]. Thus we incorporate a similar approach into the MRHMM-based speech recognition [5].

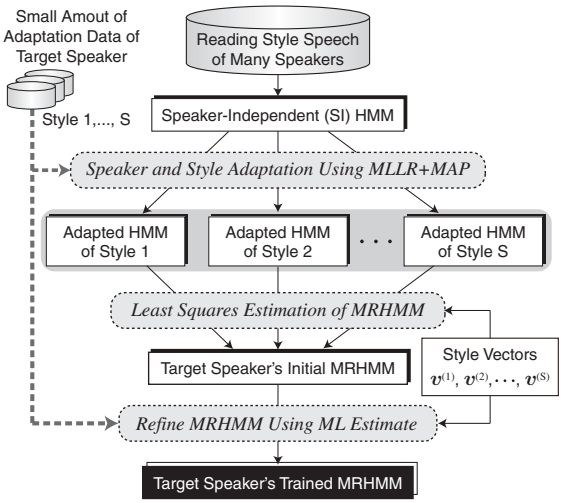

Figure 1: MRHMM training using SI model and adaptation.

A block diagram of the model training is illustrated in Fig. 1. First, we train an SI reading style model with a sufficient amount of read speech data of many speakers. Next, we adapt the obtained SI model to 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 a least squares estimation from the 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 $1 \leq s \leq S$. Then the optimal regression matrix $\boldsymbol{H}_{im}^{LS}$ is given as [11, 12]

$$\boldsymbol{H}_{im}^{LS} = \left(\sum_{s=1}^{S} \boldsymbol{\mu}_{im}^{(s)} \boldsymbol{\xi}^{(s)\top} \right) \left(\sum_{s=1}^{S} \boldsymbol{\xi}^{(s)} \boldsymbol{\xi}^{(s)\top} \right)^{-1}. \quad (5)$$

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]:

$$\boldsymbol{H}_{im} = \frac{\tau \boldsymbol{H}_{im}^{LS} + \Gamma_{im} \boldsymbol{H}_{im}^{ML}}{\tau + \Gamma_{im}} \quad (6)$$

where $\boldsymbol{H}_{im}^{LS}$ is the regression matrix obtained by Eq. (5) 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

$$\Gamma_{im} = \sum_{t=1}^{T} \gamma_t(i,m). \quad (7)$$

2.4. 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) is obtained. Using this HMM, we can straightforwardly perform ordinary speech recognition based on HMM. More specifically, first, the style vector for the input utterance is estimated using the style estimation technique described in **2.2**, and

then, using the estimated style vector, the on-line adapted HMM for the recognition is obtained from the MRHMM by calculating the new mean vectors of pdfs. The style vector is estimated for every input utterance. To perform the style estimation, we need a phoneme label sequence of the input utterance [4]. For this purpose, we use the following two-pass recognition process.

In the first pass, we obtain an HMM by setting the style vector of the MRHMM equal to $\mathbf{0}$ and perform phoneme recognition of input speech using the obtained HMM. Then we estimate the style vector $\boldsymbol{v}^*$ for the input utterance using the MRHMM and the resultant phoneme sequence. In the second pass, we obtain on-line adapted HMM from the MRHMM by calculating the new mean vectors with the estimated style vector $\boldsymbol{v}^*$ using Eq. (1). After that, we perform speech recognition using the style-adapted HMM and obtain the final recognition result.

3. Experiments

3.1. Speech database

For the performance evaluation of the proposed technique, we used the corpus of spontaneous Japanese (CSJ) [9]. We chose speech data uttered by six non-professional speakers, specifically, three male speakers (speaker ID: 423, 471, and 685) and three female speakers (speaker ID: 19, 463, and 514) with two types of speaking styles — reading and academic presentation (AP) styles. From the database of each speaker and style, we extracted 100 utterances segmented by silences with 200 ms or longer and having over ten moras and no fillers for test samples. We also extracted 15 utterances of each style for the MRHMM training.

For the SI model training, we used reading 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 those who were used for the test samples mentioned above.

3.2. Experimental conditions

The SI model was trained using about 50 phonetically balanced sentences for each speaker, 10498 sentences in total. We used three-state left-to-right triphone HMMs with 16-mixture and diagonal covariance pdfs. The parameters of the SI model were tied using a decision tree based context clustering with MDL criterion. The total number of states in the SI model was 1875.

Five utterances of the respective styles, i.e., reading and AP styles, taken from the CSJ and not included in the test samples, were used for each target speaker's MRHMM training. We applied a combined approach based on the MLLR and maximum a posteriori (MAP) adaptation [13] for the purpose of the speaker and style adaptation. Since the amount of the adaptation data of each target style was very small, we used a global transform in the MLLR. Moreover, to alleviate the dependency of the choice of the adaptation data, we divided the 15 training utterances into three subsets and repeated the same experiments three times by changing the adaptation data subset.

We used a one dimensional style space to represent two speaking styles as shown in Fig. 2. The style vectors for the adaptation data in the MRHMM training were set to fixed values, 0 and 1 for the reading and AP styles, respectively. For the MRHMM refinement, we set $\tau = 15$ in Eq. (6) on the basis of preliminary experimental results. In addition, we did not adapt the covariance parameters.

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

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) $+\Delta$
	Log of power $+\Delta$
Number of monophones	42

Figure 2: 1-D style space representing reading and academic presentation (AP) styles.

3.3. Performance evaluation in speaking style classification

We first examined the style classification performance of the proposed technique. We classified the test samples into two categories, reading and AP styles, 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 reading style. If the value was greater than or equal to the threshold, the input speech was classified into the AP style. Figures 3 shows the classification error rates for the test samples of each speaking style for respective target speakers by changing the threshold from 0.0 to 1.0 with an increment of 0.01.

Since the optimal threshold that gives the equal error rate depends on various factors such as speakers, model training and adaptation data, and input utterances, it is not known in advance. Hence, we divided the target speaker's 100 test samples of each style into 10 subsets of 10 samples and chose one subset to determine the threshold. We calculated the mean value of the estimated style vectors for the chosen subsets of respective styles and used it as the threshold. The other 90 samples of each style were used as test data for speaking style classification. The same procedure was followed for the other nine subsets. The average values of correct classification rates for each speaker are shown in Table 2. Although the classification criterion was quite simple, it can be seen that the results are promising.

3.4. Performance evaluation in phoneme recognition

We assessed the performance of the proposed technique in terms of the phoneme recognition error rate. For comparison, we also evaluated the performance of two types of ordinary HMMs which were adapted from the SI model using the same data and the same adaptation technique used for the MRHMM training, but not adapted on-line. The one is a set of style-dependent models that was adapted using target speaker's five utterances of the respective styles and the other is a style-independent model that was adapted using the target speaker's five utterances for each style, 10 utterances in total. It is noted that we assumed that the speaking style of the input speech was known when using the style-dependent models, and unknown for the other

Table 2: Average values of correct classification rates (%) for speaking styles of CSJ.

Input Style	Speaker ID					
	423	471	685	19	463	514
Reading	98.3	99.2	97.3	96.0	97.2	99.9
AP	94.9	99.4	97.9	88.3	96.2	99.4

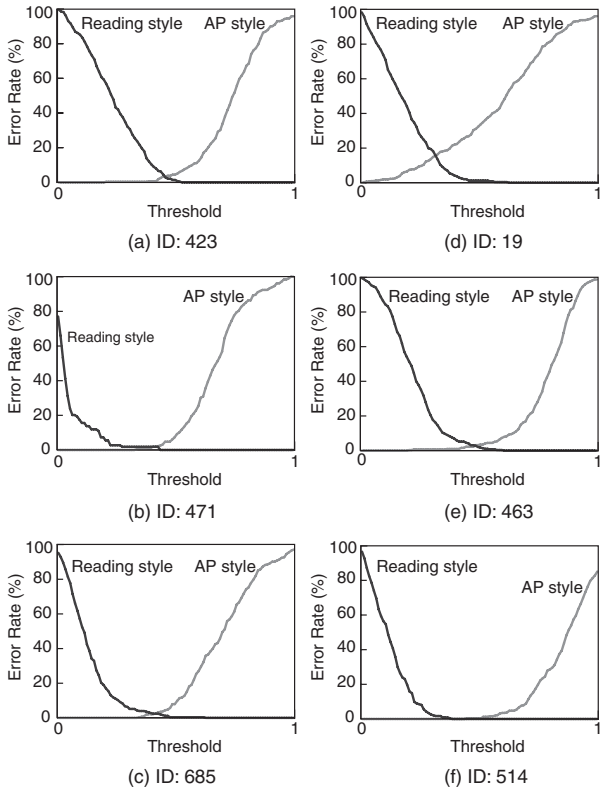

Figure 3: Classification error rates as functions of the threshold.

models. Figure 4 shows the average phoneme recognition error rates. The error rate was calculated based on the number of correctly recognized phonemes, substitutions, and deletions. We can see that MRHMM gives the highest performance in all the styles. Moreover, it is emphasized that we can obtain the paralinguistic information in addition to the linguistic information when using MRHMM.

4. Conclusions

We have proposed a technique for spontaneous speech recognition using rapid model adaptation, in which paralinguistic information can be obtained as well as linguistic one. The technique utilizes a multiple-regression HMM (MRHMM) framework, and is based on style estimation and adaptation. Using a speaker-independent reading 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 speaking style of input speech from the trained MRHMM using the estimated style vector. From the experimental results using the CSJ, we found that the performance of the proposed technique in both speech recognition and style estimation is promising. In our future work, we will explore the effectiveness of the proposed technique using other speaking styles.

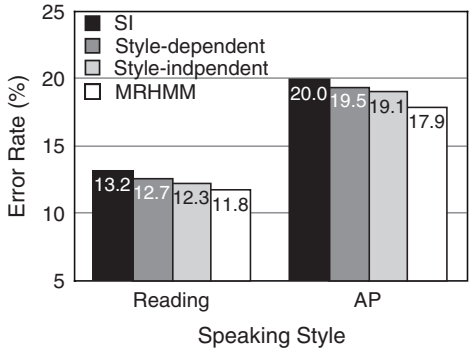

Figure 4: Comparison of phoneme error rates (%).

5. Acknowledgments

The authors would like to thank Dr. Makoto Tachibana of the Tokyo Institute of Technology for his valuable discussions with us. A part of this work was supported by JSPS Grant-in-Aid for Scientific Research 21300063.

6. References

[1] S. Furui, M. Nakamura, T. Ichiba, and K. Iwano, "Analysis and recognition of spontaneous speech using *Corpus of Spontaneous Japanese*," Speech Communication, 47(1-2):208–219, 2005.

[2] R.J.J.H. van Son and L.C.W. Pols, "An acoustic description of consonant reduction," Speech Communication, 28(2):125–140, 1999.

[3] M. Nakamura, K. Iwano, and S. Furui, "The effect of spectral space reduction in spontaneous speech on recognition performances," Proc. ICASSP 2007, 4:473–476, 2007.

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

[5] Y. Ijima, M. Tachibana, T. Nose, and T. Kobayashi, "Emotional speech recognition based on style estimation and adaptation with multiple-regression HMM," Proc. ICASSP 2009, 4157–4160, 2009.

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

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

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

[9] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," Proc. IEEE Workshop on SSPR, 7–12, 2003.

[10] T. Nose, Y. Kato and T. Kobayashi, "A speaker adaptation technique for MRHSMM-based style control of synthetic speech," Proc. ICASSP 2007, 4:833–836, 2007.

[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, 4633–4636, 2008.

[12] 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 HSMM," Proc. INTERSPEECH 2008, 2759–2762, 2008.

[13] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," IEEE Trans. Speech and Audio Process., 4(4):294–300, 1996.

[14] JNAS: Japanese Newspaper Article Sentences, http://www.mibel.cs.tsukuba.ac.jp/jnas/instruct.html