An Estimation Technique of Style Expressiveness for Emotional Speech Using Model Adaptation Based on Multiple-Regression HSMM

Takashi Nose, Yoichi Kato, Makoto Tachibana, Takao Kobayashi

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

Abstract

This paper describes a technique of estimating style expressiveness for an arbitrary speaker’s emotional speech. In the proposed technique, the style expressiveness, representing how much the emotions and/or speaking styles affect the acoustic features, is estimated based on multiple-regression hidden semi-Markov model (MRHSMM). In the model training, we first train average voice model using multiple speakers’ neutral style speech. Then, the speaker- and style-adapted HSMMs are obtained based on linear transformation from the average voice model with a small amount of the target speaker’s data. Finally, MRHSMM of the target speaker is obtained using the adapted models. For given input emotional speech, the style expressiveness is estimated based on maximum likelihood criterion. From the experimental results, we show that the estimated value gives good correspondence to the perceptual rating.

Index Terms: emotional expression, estimation of expressiveness, multiple-regression HSMM, model adaptation

1. Introduction

Recently, various approaches to emotion recognition of expressive speech have been proposed [1, 2]. Most of the techniques focus only on the classification of emotional states. However, it is important to detect the variability within a certain emotion, e.g., “a little sad” or “very sad,” as well as the emotion categories. This is because we often milden or emphasize such emotional expression flexibly depending on the situation in actual human speech communication.

Several attempts have been made for utilizing the degree of emotion in analysis and synthesis of emotional speech [3, 4]. In [3], the prosodic features, fundamental frequency (F0) and speech rate, were analyzed for expressive speech with different degree of emotion. It was also reported that the emotional speech can be synthesized by formulating the rules between the prosodic features and the degree of emotion [4]. However, it is not easy to estimate quantity representing how strong the emotion appears in speech when using these rule-based approaches. To address this problem, we have proposed statistical techniques for style modeling [5] and style estimation [6]. In these techniques, emotional expressions and/or speaking styles are called the styles, and the quantity how much the style affects the acoustic features of speech, which will be referred to as style expressiveness in this paper, is used as the model parameter of multiple-regression hidden semi-Markov model (MRHSMM).

In MRHSMM, the mean vector of output and state duration distributions in each state is given by multiple regression of a low dimensional vector, called the style vector, which represents a point in a space, called the style space. Moreover, the spectral, F0 and duration are simultaneously modeled based on multi-space probability distribution [7]. However, in the conventional MRHSMM-based technique [6], one of the problems is that a sufficient amount of speech data for each style of the target speaker is required to train the MRHSMM. And thus it is not easy to conduct estimation of arbitrary speaker’s speech. Although a possible approach to this problem is the speaker-independent one, the performance would be unsatisfactory because the style expressiveness varies sensitively on individuals.

In this paper, we propose an alternative approach based on average voice model and model adaptation. The proposed technique uses a small amount of target speaker’s data in the training of MRHSMM to estimate the style expressiveness for arbitrary speakers. We have already shown the effectiveness of this approach in speech synthesis application [8]. Here we apply it to the estimation of the style expressiveness based on MRHSMM. In the proposed technique, first the target speaker’s style-dependent models are obtained using constrained structural maximum a posteriori linear regression (CSMAPLR) adaptation [9] from average voice model which is trained with multiple speakers’ neutral style speech. Then, the MRHSMM of the target speaker is trained from the style-dependent model. When the input speech is given, the style vector is estimated based on maximum likelihood (ML) criterion. We conduct experiments to confirm that the estimation result matches the human perception.

2. Estimating style expressiveness based on MRHSMM

2.1. Style modeling of speech based on MRHSMM

In the MRHSMM-based style modeling of speech [5], we use HSMM framework [10] in which the i-th state output and duration distributions are given by Gaussian density functions, and their mean vectors \( \mu_i \) and \( m_i \) are assumed to be modeled using multiple regression as

\[
\mu_i = H_{b_i} \xi \tag{1}
\]

\[
m_i = H_{p_i} \xi \tag{2}
\]

where

\[
\xi = [1, v_1, v_2, \ldots, v_L]^T = [1, v^T]^T \tag{3}
\]

and \( v \) is the style vector, and its components \( \{v_k\} \) are explanatory variables, called style components, each of which represents the degree or intensity of a certain style, e.g., emotional expression or speaking style, appeared in the acoustic features of speech. \( H_{b_i} \) and \( H_{p_i} \) are regression matrices of dimension
Using the least square method, we obtain the following:

\[ \mu_i = H_i \xi + b_i \]

(4)

where \( \xi \) denotes the style vector to be estimated, and \( H_i = [h_i^{(b_1)}, \ldots, h_i^{(b_L)}] \)

(5)

and \( A_i = [h_i^{(b_1)}, \ldots, h_i^{(b_L)}] \)

(6)

and \( \nu = [\nu_1, \ldots, \nu_L]^\top \).

(7)

The optimal style vector \( \xi \) for the input observation \( O \) is defined in ML sense as

\[ \xi = \arg \max_\nu P(O|\lambda, \nu). \]

(8)

The obtained values of style components give quantities how much each style affects the acoustic features including spectral and prosodic information compared to those of the training data in ML sense. As a result, the estimated values of the style components can be used to detect the style expressiveness on speech [6].

3. Training of MRHSMM using average voice model and model adaptation

An outline of the proposed technique is shown in Fig. 1. The proposed technique utilizes the average voice model trained by multiple speakers’ neutral style speech.

3.1. Simultaneous adaptation of speaker and style

We adapt the average voice model to target speaker’s style-dependent HSMMs using a technique for simultaneous adaptation of speaker and style [8]. In the adaptation, the mean vector of average voice model and covariance matrix of output pdf, \( \mu_s, \Sigma_s, m_s, \sigma_s^2 \), are linearly transformed as follows:

\[ \hat{\mu}_i = \xi \mu_i - \epsilon, \]

(9)

\[ \hat{\Sigma}_i = \xi \Sigma_i \xi^\top \]

(9)

\[ \hat{m}_i = \chi m_i - \nu, \]

(10)

\[ \hat{\sigma}_i^2 = \chi \sigma_i^2 \chi^\top \]

(10)

where \( \xi, \epsilon \) are transformation matrix and bias vector for output pdf, and \( \chi, \nu \) are transformation coefficient and bias term, respectively. As the linear transformation-based adaptation, we use HSMN-based constrained maximum a posteriori linear regression (CSMAPLR) algorithm [9].

3.2. Training MRHSMM based on least square method

Using speaker- and style-adapted HSMMs obtained in Sect. 3.1, the target speaker’s MRHSMM is estimated based on the least square method. Suppose that speech database contains \( S \) styles and mean vector of each style and corresponding style vector is given by \( \mu_s^{(s)} (1 \leq s \leq S), \xi_s^{(s)} \), respectively. We choose \( H_t \) that minimizes

\[ E = \sum_{s=1}^{S} \left\| \mu_s^{(s)} - H_t \xi_s^{(s)} \right\|^2 \]

(11)

as the regression matrices of the MRHSMM. By differentiating \( E \) with \( H_t \) and setting the result zero, we have

\[ \tilde{H}_{b_t} = \left( \sum_{s=1}^{S} \mu_s^{(s)} \xi_s^{(s)} \xi_s^{(s)} \right)^{-1}. \]

(12)

The regression matrices for the state duration pdf \( \tilde{H}_{b_t} \) can be estimated using the same criterion. In [8], the obtained MRHSMM is refined using MLLR-based adaptation and ML estimation. However, in the estimation of the style expressiveness, from a preliminary experiment, we have found that the results did not change so great before and after the refinement of MRHSMM adaptation, then we used the least square method for estimating the MRHSMM parameters in this study.

4. Experiment

4.1. Experimental conditions

We used phonetically balanced 503 ATR Japanese sentences composed of ten subsets — subsets A, B, ..., and J. The number of sentences included in each subset except for J was 50, and subset J had 53 sentences. For the training of the average voice model, we used 5 male and 4 female speakers in ATR Japanese speech database (Set B). The training data were 450 sentences for each speaker, 4050 sentences in total. In the training stage of the average voice model, the shared-decision-tree-based context clustering algorithm and the speaker adaptive training [11] were applied.
In the training of MRHSMM, we used three styles of read speech — neutral, sad, joyful styles. As the target speakers, we used three speakers, MMI, MKA, and FHS. The male speaker MMI was a professional narrator who had some experience in speaking the given sentence with simulated styles. MMI uttered whole 503 ATR sentences in each style [12]. The male and female speakers, MKA and FHS, were non-professional speakers who had little experience of speaking with simulated styles. We used speech data of 100 sentences, subset of ATR sentences, for MKA and FHS. For each target speaker, we used 50 test sentences not included in neither training nor adaptation data.

Speech signals were sampled at a rate of 16kHz and windowed by a 25-ms Blackman window with a 5-ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. The feature vector consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency, and their delta and delta-delta coefficients. The MRHSMM was 5-state left-to-right model with diagonal covariance. The style space was modeled as one-dimensional space, and we used two style spaces, “neutral – sad” and “neutral – joyful”. Style vectors of training and adaptation data were set to fixed values, namely, 0 for the neutral style, 1.0 for the sad and joyful styles, respectively. In the MRHSMM training and the estimation of the style expressiveness, we used the context-dependent label including context information described in [12].

4.2. Perceptual rating of style expressiveness for speech database

To evaluate the correspondence between the estimation results and the human perception, a perceptual rating experiment of the style expressiveness was conducted for whole speech data of the target speakers and styles except for the neutral style. Nine subjects listened each sentence in random order from the database and rated their style expressiveness. The rating was done as follows: “1.5” for strong, “1.0” for standard, “0.5” for weak, and “0” for not perceiving as the target style. About the utterances rated as “0”, we asked the subjects how they perceived the speech. The dominant opinion was that the speech sounded like the neutral style. Thus we assumed that “0” meant the neutral style in this study. Here we show the result for speaker MMI in Fig. 2. In the figure, the average scores of the perceptual rating for the respective subsets are shown. From the figure, we can see that the perceptual rating of style expressiveness is not constant.

4.3. Subjective evaluation of proposed technique for professional narrator

We estimated the style vectors for speaker MMI using the proposed technique. To alleviate the dependency of the choice of the adaptation data, the adaptation utterances were chosen randomly from whole subsets. The numbers of adaptation sentences were 5, 10, 20, and 50. In addition, to obtain a more reliable result, we conducted the estimation three times by changing the sentences used in adaptation. Then we used the average of the estimated values for these sets as the estimation result. For comparison, we also estimated the style vectors for whole 503 sentences of the target styles using conventional speaker-dependent MRHSMM trained with 450 sentences of each style. Note that the 503 sentences to be estimated included the adaptation or training data of the target speaker.

Figure 3 shows the average scores of the estimated values for the respective subsets. In the figure, ADAPT-5, 10, 20, and 50 represent the results of the proposed technique using the adaptation data of 5, 10, 20, and 50 sentences, respectively, and SD-450 represents the results of the conventional speaker-dependent MRHSMM trained with 450 sentences of each style. Note that the 503 sentences to be estimated included the adaptation or training data of the target speaker.

From the estimation results by conventional and proposed techniques, similar tendency of the average scores can be seen over the subsets between perceptual and estimated scores. It implies that the estimation results have close relation to the human perception. However, we can also find that the estimated value has the bias to the perceptual score. There are several factors which cause such bias. A possible reason is that the perceptual scores for training data have some variation and the mean of the distribution does not coincide with 1.0, though the style vector for the training data were assumed to be 1.0.
Table 1: Correlation coefficients between perceptual scores and style component values for a professional narrator MMI.

<table>
<thead>
<tr>
<th>Style</th>
<th>Conventional</th>
<th>Number of adaptation sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>0.722</td>
<td>0.516</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.661</td>
<td>0.511</td>
</tr>
</tbody>
</table>

(b) Fifty test sentences

<table>
<thead>
<tr>
<th>Style</th>
<th>Conventional</th>
<th>Number of adaptation sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>0.727</td>
<td>0.619</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.332</td>
<td>0.621</td>
</tr>
</tbody>
</table>

Table 2: Correlation coefficients between perceptual scores and style component values for non-professional speakers.

<table>
<thead>
<tr>
<th>Style</th>
<th>Number of adaptation sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>0.562</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.488</td>
</tr>
</tbody>
</table>

(b) Female speaker FHS

<table>
<thead>
<tr>
<th>Style</th>
<th>Number of adaptation sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sad</td>
<td>0.495</td>
</tr>
<tr>
<td>Joyful</td>
<td>0.312</td>
</tr>
</tbody>
</table>

fixed to 1.0.

4.4. Correlation between perceptual and estimated scores for professional narrator

To show the correlation quantitatively, we calculated the correlation coefficients between the estimated and perceptual values. Table 1 shows the results; (a) is the result for the whole 503 sentences of the database of the target speaker MMI, and (b) is that for the 50 test sentences described in 4.1. From Table 1-(a), we can see that the correlation coefficients for all styles are over 0.6 when using the conventional training technique. Although the results of the proposed technique do not outperform those of the conventional one, the correlation coefficients are still over 0.5 for all styles. Moreover, the proposed technique also gives the estimation results with medium correlation. From Table 1-(b), we can see the effectiveness of the proposed technique for the unseen sentences.

4.5. Correlation between perceptual and estimated scores for non-professional speakers

We next evaluated the estimation performance of the proposed technique under a more realistic condition where speech was uttered by non-professional speakers MKA and FHS. Table 2 shows the correlation coefficients between the estimated and perceptual values; (a) is the result for the male speaker MKA and (b) is that for the female speaker FHS. It can be seen that the correlation coefficients are almost over 0.5 for male speaker MKA. However, the results of female speaker FHS are slightly worse than MKA. The utterances of FHS seem to be relatively difficult to perceive the variability of style expressiveness compared to MKA, and this can lead to the degradation of the estimation performance.

5. Conclusions

In this paper, we have proposed a technique for estimating the style expressiveness on speech with a small amount of target speaker’s data. The proposed technique is based on multiple-regression hidden semi-Markov model (MRHSMM) which is trained using average voice model and model adaptation. From the experimental results, we have shown that the estimation results of the style expressiveness give good correspondence to the perceptual rating. For a professional narrator’s simulated styled speech, the correlation coefficients are over 0.6 using only 5 adaptation sentences. Future work will focus on the performance evaluation of the proposed technique using spontaneous speech database in various emotions and speaking styles.

6. Acknowledgments

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

7. References