Performance Evaluation of HMM-Based Style Classification with a Small Amount of Training Data

Makoto Tachibana, Keigo Kawashima, Junichi Yamagishi, Takao Kobayashi
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, 226-8502 Japan
{makoto.tachibana, keigo.kawashima, takao.kobayashi}@ip.titech.ac.jp

Abstract
This paper describes a classification technique for emotional expressions and speaking styles of speech using only a small amount of training data of a target speaker. We model spectral and fundamental frequency (F0) features simultaneously using multi-space probability distribution HMM (MSD-HMM), and adapt a speaker-independent neutral style model to a certain target speaker’s style model with a small amount of data using MSD-MLLR which is extended MLLR for MSD-HMM.

We perform classification experiments for professional narrators’ speech and non-professional speakers’ speech and evaluate the performance of proposed technique by comparing with other commonly used classifiers. We show that the proposed technique gives better result than the other classifiers when using a few sentences of target speaker’s style data.

Index Terms: emotional speech, speaking style, speech emotion recognition, classification, MSD-HMM

1. Introduction
Recognition of emotions (e.g. “surprise,” “joy,” “sadness,” and “anger,” etc.) can enhance the capability of human-computer interaction systems, such as communications robots, in-car navigation systems, and operation support systems. For example, in the application to a communications robot, if the robot can detect user’s sad situation from user’s utterances, it will be able to give a more natural and appropriate response to the user by sympathizing with user’s situation or encouraging the user.

It has been proposed a technique for speech emotion recognition based on HMM [1]. In this technique, a high classification rate (more than 80%) was obtained by using speaker-dependent models trained from a large amount of target speaker’s data. However, it is not easy to obtain a large amount of arbitrary target speakers’ speech data including various emotional expressions in the system like a communications robot. A possible approach to solving this problem is to use a speaker-independent model trained from a large amount of multiple speakers’ data. However the performance may become unsatisfactory when multiclass emotion recognition is performed [2]. This is because emotional expressions of speech vary depending on individuals and environments. From this point of view, it would be useful to develop a system that gives higher classification performance using only a small amount of speech data of an arbitrary target speaker.

In this study, we propose a text-independent style classification technique to obtain higher classification performance with a small amount of target speaker’s training data. Since we aim to treat not only emotional expressions appeared in speech but also speaking styles, we refer to the classification process as the style classification. The technique is based on multi-space probability distribution HMM (MSD-HMM) [3] and MSD-MLLR [4]. We simultaneously model spectral and fundamental frequency (F0) information of each phoneme unit using MSD-HMM to obtain a speaker-independent neutral style model.

Then we apply MSD-MLLR, which is an extension of the maximum likelihood linear regression (MLLR) [5] for MSD-HMM, to adapt the speaker-independent neutral style model to a certain target speaker’s style model with a small amount of training data. In order to evaluate the performance of the proposed technique, we compare the proposed technique with several commonly used classifiers including support vector machine, C4.5 decision tree, naive bayes, and logistic regression using professional narrators’ speech and non-professional speakers’ speech.

2. Style Classification System

2.1. Spectrum and F0 Modeling Using MSD-HMM
Speech signal spoken with a certain style is characterized by statistics of feature vectors (e.g. mean value and their range) as well as temporal dynamics. HMM can model such temporal dynamics of the features with the state transition probability. We use phoneme HMMs each of which would be independent of the length of the input sentence. In several studies on classification of emotions [1, 6], spectral information is not always treated as important. On the other hand, we believe that the spectral information is related more or less to the prosodic information and contributes to the classification performance. Moreover, spectral information is essential to the phoneme model. Thus, using MSD-HMM [3], we simultaneously model the spectral feature as well as the F0 feature that is one of the important prosodic features.

The MSD-HMM is an extended HMM for the modeling of sequences of observation vectors with variable dimensionality including zero-dimensional observations like discrete symbols such as the values of F0 of unvoiced region. More specifically, the observation sequence of F0 consists of two spaces by a one-dimensional space corresponding to voiced and a zero-dimensional space corresponding to unvoiced. In addition, for simultaneous use of spectral and F0 information, they are modeled by multi-stream MSD-HMMs in which output distributions for spectral and F0 parts are modeled using continuous probability distribution and multi-space probability distribution, respectively.

2.2. Speaker and Style Adaptation Using MSD-MLLR
In order to obtain the model for a certain style speech of an arbitrarily target speaker, we adapt a speaker-independent neutral
3. Experiments

3.1. Speech Databases

In the following experiments, we used professional narrators’ speech and non-professional speakers’ speech. Professional narrators’ speech database [8] consists of four styles: neutral, sad, joyful, and rough styles in read speech. The rough style is a kind of impolite speaking style and is often perceived as disgusted speech. This database is composed of phonetically balanced 503 sentences of ATR Japanese speech database uttered by two male speakers and a female speaker in each style. All the speakers have some experience in uttering the given sentence with simulated styles. In a human perception test, the correct classification rate ranged between 80 and 100% [8]. In addition, we recorded non-professional speakers’ speech uttered by four male graduate students in six styles: neutral, sad, joyful, anxious, disgusted, and angry styles. The non-professional speakers have little experience of uttering the given sentence with such simulated styles. The speech data in each style contains 100 sentences taken from ATR 503 phonetically balanced sentences. Moreover, we designed 30-sentence sets in which the style can be expressed easily for each style. We recorded these sentences in a quiet room. Each speaker expressed respective styles of speech based on subjectivity, and we did not give specific instructions to the speaker. Each sentence was around three seconds. The correct classification rates of an informal listening test by human were 86%, 74%, 79%, 84%, 56%, and 77% for neutral, sad, joyful, anxious, disgusted, and angry styles, respectively.

Speech signals were sampled at a rate of 16kHz and windowed by a 25ms Blackman window with a 5ms shift. Then mel-cepstral coefficients were obtained by mel-cepstral analysis. F0 values were obtained using the instantaneous frequency amplitude spectrum [9]. The feature vectors consisted of 13 mel-cepstral coefficients including the zeroth coefficient and logF0. We used 42 phonemes including silence and pause [8]. The phoneme HMMs were 3-state left-to-right monophone HMMs.

3.2. Other Classification Techniques

To evaluate the performance of the proposed technique, we compared it with other commonly used classifiers: support vector machine using polynomial kernel (SVM), C4.5 decision tree (C4.5), naive bayes (Naive Bayes), and penalized logistic regression (Logistic Regression). We utilized the Weka toolkit.

The style registration stage, the SI phoneme models are adapted to speaker- and style-dependent phoneme HMMs using a small amount of the target style speech data uttered by the target speaker. Then a set of adapted phoneme HMMs is obtained as the reference style model for classification. We utilize the standard MLLR [5] and the MSD-MLLR for spectrum and F0 adaptation, respectively.

In the classification stage, phoneme recognition of input speech is performed using each reference style model. We use a simple word level network as a language model for phoneme recognition. Then the accumulated likelihood of the phoneme-concatenation model of each style is calculated. Finally, the style of the input speech is determined by selecting the style whose model gives the maximum accumulated likelihood among the reference models.
Figure 2: Classification error rates for large amount of training data.

In these techniques, it is not easy to model dynamic features directly. Therefore, we simply used a fixed-length feature vector in each utterance by computing the following statistics: maximum, minimum, median, mean, standard deviation, mode, and quartiles of each mel-cepstral coefficient and logF0. The values of logF0 in unvoiced regions were excluded. The dimensionality of the feature vector was 98 for each utterance. From preliminary experimental results, we obtained slightly better performance by combining logF0 with mel-cepstral coefficients than using only mel-cepstral coefficients. Since we aimed to evaluate the relative baseline performance of the classifiers, we did not further tune up the features and thus the choice of these statistics was not necessarily the best for the style classification.

3.3. Performance Evaluation with a Large Amount of Data

We first compared the performance of the MSD-HMM-based classifier with that of other classifiers when a large amount of training data of each style was available. We performed 4-class style classification using the speech data uttered by the professional narrators. The training and evaluation data were 500 out of the 503 sentences. We chose one out of three professional narrators as a target speaker to be evaluated and repeated the same test three times changing the target speaker. Each reference style model for classification was the speaker- and style-dependent model directly trained from 400 sentences of target speakers’ speech by ML estimation. In this experiment, we did not use SI model or MSD-MLLR. The other 100 sentences of the target speaker not included in the training data were used as the evaluation data. We performed a 5-fold cross-validation test.

Figure 2 shows classification error rates. The error rates for respective styles are also shown. It can be seen that the classification error rate of SVM is the lowest and followed by Logistic Regression and MSD-HMM. The correct classification rate of MSD-HMM was about 98%. We obtained high classification performance comparable to human subjective evaluation for all classifiers. This is because the degree of expressivity of style in the professional speakers’ speech is relatively stable and an appropriate classification model might be made.

3.4. Performance Evaluation with a Small Amount of Data

We evaluated the performance of MSD-MLLR when only a small amount of training data of the target speaker was available. The evaluation data and the target speakers were the same as the previous experiment. For each target speaker, five sentences of each style were used as the adaptation data for MSD-MLLR. The SI model was trained using 800 sentences uttered by two professional narrators not including the target speaker. In the other classifiers, we used only the adaptation data for training. Fifty sentences not included in the training nor adaptation data were used as the evaluation data. We performed a 5-fold cross-validation test.

Figure 3 shows classification error rates. We can see that the error rate of MSD-HMM is the lowest and the performance difference between the adapted model using five sentences and the speaker- and style-dependent model trained using 400 sentences (Fig. 2) is smaller than the other classifiers. A possible reason for the relatively high performance of MSD-HMM is to use a larger amount of training data for the SI model. As a result, the proposed classifier becomes more effective when only a small amount of target speaker’s style data is available.

3.5. Performance Evaluation Using Non-professional Speakers’ Speech

Finally, we evaluated the performance of the proposed technique using non-professional speakers’ speech for slightly realistic situation. We used each of the four non-professional narrators as a target speaker and performed 6-class style classification. The SI model was trained from 1200 sentences uttered by three professional narrators. From one to 20 sentences taken from ATR phonetically balanced sentences of target speaker’s speech data were used as the adaptation data for MSD-MLLR in each style. In the other classifiers, we used only the adaptation data for training. The 30-sentence sets that we designed were used as the evaluation data.

Figure 4 shows correct classification rates of each style as functions of the number of adaptation sentences for MSD-HMM and the number of training sentences for the other classifiers. It can be seen that the improvement of the correct classification rate of MSD-HMM is higher than the other classifiers in a few adaptation sentence. Moreover, the classification performance is stable without depending on the number of sentences. In the other classifiers, SVM and NaiveBayes gave good results to some extent. Using adaptation techniques or choosing the more optimal features including velocity and acceleration would improve the performance of these techniques.

4. Conclusions

In this paper, we have proposed a text-independent style classification technique based on MSD-HMM and MSD-MLLR. In comparison with other classifiers, we have shown that the proposed classification technique is promising when only a small
amount of training data of a certain style of a target speaker is available. A future task is to explore effectiveness of the proposed technique in more realistic situations. The consideration of linguistic information through the result of the phoneme recognition is also a future task.

5. References


