ON-LINE SPEAKER ADAPTATION BASED EMOTION RECOGNITION USING INCREMENTAL EMOTIONAL INFORMATION

Jae-Bok Kim, Jeong-Sik Park, Yung-Hwan Oh, Member, IEEE

Computer Science Department, Korea Advanced Institute of Science and Technology, Korea

ABSTRACT

This paper proposes a new Speech Emotion Recognition (SER) framework. Compared to the speaker-independent emotion models, speaker-adapted models constructed by using a speaker's emotional speech data can represent the speaker's emotional characteristics more precisely, thus improving SER accuracy. However, it is hard to collect a sufficient amount of personal emotional data at once. For this reason, we propose an MLLR-based on-line speaker adaptation technique using accumulated personal data. Compared to speech models, it is relatively difficult to construct reliable emotion models applicable to MLLR due to the domain-oriented characteristics. Thus, we modify the conventional MLLR procedure by using selective label refinement, which categorizes newly accumulated adaptation data into discriminative and non-discriminative data, and only refines the labels of the discriminative data. On SER experiments based on an LDC emotion corpus, our approach exhibited superior performance when compared to conventional adaptation techniques as well as the speaker-independent model framework.

Index Terms— Speech emotion recognition, on-line speaker adaptation, MLLR

1. INTRODUCTION

Speech Emotion Recognition (SER) is the automatic identification of emotional states of a speaker. This technology is primarily being studied for Human-Computer Interaction and has a variety of applications such as call centers, intelligent automobiles and human robots. Many researchers have introduced various approaches for SER tasks, but unfortunately, they have failed to achieve satisfactory performance due to two critical factors [1][2]. First, different speakers rarely express emotional states in the same way. Since large variations exist in the acoustic characteristics between speakers expressing an emotion, it is undesirable to describe a person's emotional states using conventional Speaker-Independent (SI) models. Second, several pair of emotions such as sadness and boredom have acoustically similar characteristics. This ambiguity causes unreliable recognition results.

To overcome the factors mentioned above, applying speaker adaptation techniques can be a useful approach in SER since the adapted emotion models approximately indicate aspects of Speaker-Dependent (SD) models, in which a speaker's discriminative emotional characteristics are more precisely described than in SI models. The conventional adaptation techniques require a certain amount of speech data spoken by a speaker. However, it is hard to collect personal emotional data unless speakers express each emotional state intentionally. Furthermore, correctly classifying a speaker's speech data according to emotional state is not an easy task owing to the domain-oriented ambiguity. For this reason, it is necessary to devise a more sophisticated adaptation procedure in order to use gradually accumulated personal emotional data.

This paper is organized as follows. Section 2 introduces speaker adaptation techniques for SER. Section 3 describes the proposed on-line speaker adaptation based on MLLR. In Section 4, experimental setups and results are presented and discussed. The paper concludes in Section 5.

2. SPEAKER ADAPTATION FOR SPEECH EMOTION RECOGNITION

Speaker adaptation is a task adjusting the parameters of SI models to a certain amount of adaptation data collected from a specific speaker in order to obtain speaker-adapted models. The Maximum Likelihood Linear Regression (MLLR) and the Maximum A Posteriori (MAP) have been successfully applied to the speech recognition tasks as representative adaptation techniques. In general, MLLR is known as a robust technique against labeling errors [3].

Compared to speech models, it is relatively difficult to construct reliable emotion models due to the domain-oriented ambiguity. Thus, most of the SER experiments have achieved poor accuracy on three or more emotion classes [4][5]. Such unreliable emotion models may generate a great number of labeling errors for adaptation data, degrading adaptation performance. For this reason, MLLR is better suited for SER. Moreover, this technique is quite efficient when the amount of adaptation data is limited.

This work was partially supported by Defense Acquisition Program Administration and Agency for Defense Development under the contract. * Corresponding author: Jeong-sik Park, E-mail: parkjs@kaist.ac.kr

2.1. Conventional MLLR adaptation and its drawback in SER domain

MLLR adaptation updates the parameters of SI models according to transformation matrices [3]. The matrices are estimated to maximize the likelihood of the adapted models observing adaptation data. A principal process of MLLR is to construct a regression class tree, tying acoustically similar Gaussian components into respective classes. The number of classes can be decided by either the amount of adaptation data or reliability of labels. Especially, in unsupervised MLLR that is focused in this paper, careful consideration should be given as to whether the labels are reliable or not.

If there exists a limited amount of adaptation data or the labels are believed to be unreliable, a global transform is recommended. In this case, the components of all SI models are broadly tied into a single class and transformed by a single matrix. On the other hand, if the amount of data is sufficient or the labels are considered to be reliable, a more specific transform is available based on multiple matrices. All components in each class are transformed by the corresponding matrix.

Although the global transform guarantees robustness to labeling errors in MLLR, the single matrix may fail to precisely represent variations in the Gaussian components of SI models. Unfortunately, there exist large variations in Gaussian components of emotion models [6], and therefore, the global transform may operate ineffectively in SER tasks. On the other hand, the specific transform using multiple matrices can accurately reflect variations of the components.

3. ON-LINE MLLR ADAPTATION FOR EMOTION RECOGNITION

In SER domain, maintaining multiple matrices is necessarily required in order to reflect large variations existing in emotion models. However, multiple matrices are not quite available for tasks with a frequent occurrence of labeling errors. Moreover, a sufficient amount of adaptation data is necessary. To overcome these drawbacks, we propose a label refinement technique to reduce labeling errors in adaptation procedure. Based on this technique, we perform on-line speaker adaptation iteratively, using accumulated personal emotional data.

3.1. Selective label refinement

A fundamental objective of the label refinement is to improve the accuracy of the transformation matrices by correctly refining the labels of adaptation data. Due to the domain-oriented ambiguity, emotional adaptation data collected from a target speaker may retain ambiguous emotional information, depending on the speaker’s capacity to express emotions. Such ambiguous data may induce labeling errors since it preserves non-discriminative emotional characteristics. We believe that incorrect labels of such data are not likely to be revised accurately, whereas a correct label may be unintentionally revised into incorrect one by the refinement task. For this reason, this study does not refine the labels of all adaptation data. Instead, we classify adaptation data into discriminative and non-discriminative data to refine only the labels of discriminative data.

For the classification, this study uses a confidence measure based on N-best likelihood results. In general, the non-discriminative emotional data demonstrate a similar likelihood on each emotion model since they rarely have discriminative emotional characteristics. On the other hand, the discriminative data tend to indicate further distances between the likelihood results corresponding to each rank in the N-best list. Based on this property, we classify the adaptation data as follows. Note that there are E emotion models and \( x_i \) (where \( i = 1, 2, ..., T \)) is one piece of \( T \) adaptation data. Let us denote \( R_j(x_i) \) as the emotion model index (ranging from 1 to E) at the \( r \)-th rank in the N-best list obtained from all E emotion models with a given \( x_i \).

We calculate the confidence measure for each piece of data as follows:

\[
CM_{select}(x_i) = \frac{1}{E-1} \sum_{r=1}^{E-1} \left\{ P(x_i \mid \hat{\lambda}_{R_j(x_i)}^{\text{r}}} - P(x_i \mid \hat{\lambda}_{R_j(x_i)}^{\text{r}}} \right\}
\]

(1)

where \( \hat{\lambda}_{R_j(x_i)}^{\text{r}} \) and \( P(x_i \mid \hat{\lambda}_{R_j(x_i)}^{\text{r}}} \) indicate the emotion model corresponding to the index and the likelihood result at the \( r \)-th rank in the N-best list, respectively. This confidence measure considers the likelihood results at overall ranks in the N-best list, calculating the average of distances between the likelihood at the \( r \)-th rank and that at the \( (r+1) \)-th rank.

Most of the confidence measures use an empirically determined threshold as a classification criterion [7][8]. Such a threshold has to be estimated prior to the adaptation procedures and may induce incorrect decisions depending on the consistency of adaptation data. For this reason, the fixed threshold is impractical in on-line speaker adaptation, where various types of adaptation data could be given. Instead, this study uses the average of \( CM_{select}(x_i) \) calculated for each adaptation data, that is, \( \sum_{i=1}^{T} CM_{select}(x_i) / T \). If \( CM_{select}(x_i) \) is lower than the average, \( x_i \) is categorized as non-discriminative data. Otherwise, the data is regarded as discriminative one.

After finishing the classification, a task refining only the labels of discriminative data is carried out. In consideration of the reliability of the discriminative data, we measure the confidence of each label using all discriminative data and use it to refine the labels. If a piece of discriminative data preserves the acoustic characteristics of an emotion model
that corresponds to a label, it will indicate a higher likelihood result for that model than it would for other models. According to this property, the confidence of each label corresponding to a model can be measured using the average likelihood of all discriminative data for the model and be formulated as follows:

\[
CM_{\text{refine}}(x_{D,k}, \lambda_k) = P(x_{D,k} \mid \lambda_k) = \frac{1}{T_{Dk}} \sum_{i=1}^{E} P(x_{D_k,i} \mid \lambda_k)
\]  

(2)

where \( \lambda_k \) is an emotion model corresponding to the index \( k \) among \( E \) emotions, and \( T_{Dk} \) indicates the total number of discriminative data \( (x_{D_k}) \) corresponding to the \( k \)-th emotion. This equation calculates the relative likelihood of the \( i \)-th discriminative data, which means the difference between the likelihood of a given data \( x_{D,k} \) for the \( k \)-th emotion model and the average likelihood of all discriminative data whose label at the first rank represents the \( k \)-th emotion. A piece of discriminative data whose label is more confident gives higher likelihood compared with the average likelihood of all data classified as the label, indicating larger difference between them. Based on this measure, the model that demonstrates the maximum relative likelihood is regarded as the best model, and the index of the model is determined as a newly refined label of \( x_{D,k} \), as follows:

\[
k^* = \arg \max_{\lambda_k \in \mathcal{F}} (CM_{\text{refine}}(x_{D,k}, \lambda_k))
\]  

(3)

Selective label refinement uses only recognition results of adaptation data, focusing on acoustically discriminative characteristics of the data from a target speaker. Since this technique does not require any prior knowledge, our approach is completely applicable to on-line adaptation.

3.2. On-line speaker adaptation

Fig. 1 illustrates the proposed MLLR procedure based on selective label refinement. First, each piece of adaptation data collected from the target speaker is recognized using SI emotion models. The recognition results, the N-best likelihood, are then obtained. Second, each piece of data is categorized as discriminative or non-discriminative using \( CM_{\text{select}} \). In the third step, the labels of discriminative data are refined by \( CM_{\text{refine}} \). Next, both discriminative and non-discriminative data are used to estimate the transformation matrices according to the respective labels and each Gaussian component is linearly transformed into a component of the adapted models according to its corresponding matrix.

If the amount of emotional data gradually increases, it is available to build more correct adapted models. However, if there is quite a considerable difference in the amount of data collected for respective emotions, it becomes ambiguous to define the discriminative emotional characteristics. Thus, it takes quite a long time to collect a sufficient amount of data for respective emotions. For this reason, this study proposes an iterative on-line speaker adaptation technique, which performs the adaptation procedures iteratively, using gradually accumulated emotional data. Once a certain amount of adaptation data is collected from a user, they are added to previously accumulated data. Using the newly organized adaptation data, MLLR adaptation constructs speaker-adapted models in accordance with the proposed adaptation procedure described in Fig. 1. The adapted models are used as a substitute for SI models at the next iteration. Such an iterative manner is expected to improve the SER performance, enhancing the adapted models.

4. EXPERIMENTAL RESULTS

4.1. Experimental setups

We performed SER experiments on the Emotional Prosody Speech of LDC [9]. We used a phrase as the basic unit of the adaptation data and performed experiments according to 7-fold leave-one-speaker-out cross validation. In each experiment, half of the utterances spoken by a single speaker were used for adaptation while the rest were used for verification. The proportion of the total adaptation data is less than 7.5% of total 738 utterances and the amount of adaptation data of a single speaker is about 1 minute in terms of duration. To implement the on-line adaptation, we divided adaptation data of each speaker into three sets and
accumulated each set in iteration, carrying out three times of iterations in total. To investigate the SER performance according to the number of emotions, we composed four types of emotion categories with five representative emotions as table 1 shows [4]. 'Neutral' was chosen as the most common emotion and used in every set. We used log energy, 12 dimensional MFCCs, pitch, and their first and second derivatives as a feature vector.

### 4.2. Experimental results and discussion

We investigated the SER performance according to several types of speaker-adapted emotion models constructed by respective adaptation techniques: the proposed MLLR using Selective Label Refinement ('MLLR_SLR') and representative techniques ('MLLR' and 'MAP'). In addition, the SI emotion models ('Baseline') were also evaluated for the purpose of performance comparison. We investigated the SER results on respective emotion classes according to table 1. As presented in Fig. 2, MLLR_SLR demonstrated superior performance compared to that of other approaches over all emotion categories. MLLR and MAP demonstrated abnormal performance in the four-class and five-class emotion recognition, where their SER accuracy deteriorated although the amount of adaptation data increased. Especially, in the five-class SER, they yielded even lower accuracy than Baseline. Such results explain that the conventional techniques were adversely affected by the unreliable SI emotion models, which may generate incorrect labels of adaptation data during unsupervised adaptation.

An exceptional result is investigated on the two-class recognition, where MLLR_SLR showed the same performance as MLLR. In this case, because of sufficiently reliable SI models, each piece of adaptation data can be correctly labeled; therefore, most of the adaptation data preserve their labels without refinement.

### 5. CONCLUSION

We propose an MLLR-based online speaker adaptation for emotional speech recognition. The proposed approach has been shown to provide superior performance compared to conventional methods. Future work will involve applying the proposed method to real-world applications and further improving the accuracy of emotion recognition.