Engineering Applications of Artificial Intelligence 52 (2016) 126–134

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Multistage data selection-based unsupervised speaker adaptation for personalized speech emotion recognition

Jae-Bok Kim [a], Jeong-Sik Park [b,*]

[a] Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Drienerlolaan 5, Enschede, The Netherlands
[b] Department of Information and Communication Engineering, Yeungnam University, 280 Daehak-Ro, Gyeongsan, Republic of Korea

ARTICLE INFO

Article history:
Received 18 August 2015
Received in revised form
30 November 2015
Accepted 29 February 2016

Keywords:
Speech emotion recognition
Speaker adaptation
Maximum likelihood linear regression
Universal background model
Acoustic model

ABSTRACT

This paper proposes an efficient speech emotion recognition (SER) approach that utilizes personal voice data accumulated on personal devices. A representative weakness of conventional SER systems is the user-dependent performance induced by the speaker independent (SI) acoustic model framework. But, handheld communications devices such as smartphones provide a collection of individual voice data, thus providing suitable conditions for personalized SER that is more enhanced than the SI model framework. By taking advantage of personal devices, we propose an efficient personalized SER scheme employing maximum likelihood linear regression (MLLR), a representative speaker adaptation technique. To further advance the conventional MLLR technique for SER tasks, the proposed approach selects useful data that convey emotionally discriminative acoustic characteristics and uses only those data for adaptation. For reliable data selection, we conduct multistage selection using a log-likelihood distance-based measure and a universal background model. On SER experiments based on a Linguistic Data Consortium emotional speech corpus, our approach exhibited superior performance when compared to conventional adaptation techniques as well as the SI model framework.

1. Introduction

Nowadays, various personal handheld devices, such as smartphones and tablet PCs, employ more advanced computing capabilities; thus, it is possible to provide users with more intelligent functions regarding human–computer interaction (HCI) (Ballagas et al., 2006). The devices are now extending their functions to identifying emotional states of users by analyzing voice or facial expression (Pittermann et al., 2010; Zhang et al., 2014; Neerincx and Streefkerk, 2003).

Emotion recognition plays a major role in HCI. It enables the devices to deliver more friendly and affectionate interaction with a user by appropriately responding to user demands in accordance with the emotional state of the user. For example, if a smart phone is capable of monitoring human emotions, it could attempt to interact with the user by displaying relevant visual content on the screen or suggesting user-preferred audio content. Emotion is very pertinent to personal feelings that the user might hope to conceal, and therefore, the detection of the user's emotion is more allowable with the user's personal device rather than other public machines.

There are various indicators for identifying human emotions, including tone of voice, facial expressions, and gestures. Among these indicators, a voice interface can be the most effective way of emotion recognition on personal devices, because it delivers direct and natural expression of emotions and does not require expensive equipment. In particular, mobile communications devices steadily provide an amount of personal voice data that can be used for enhancing voice recognition performance.

Although various approaches have been investigated in regard to the speech emotion recognition (SER), they have failed to achieve stable performance for commercial applications. Several studies concluded that the difficulty with SER is derived from domain-oriented characteristics, such as large inter-speaker variations and ambiguity between emotions (Kim et al., 2009; Lopez-Moreno et al., 2009; Grimm et al., 2007). In general, emotional speech data expressed by different speakers demonstrate large variations in acoustic characteristics, even if they intend to express the same emotion. And several pairs of representative emotions tend to have similar acoustic characteristics. For example, voices of sadness and boredom have similar characteristics, thus indicating a large overlap in acoustic feature space. A few studies reported that recognizing the emotion of other persons is not easy, even for humans, demonstrating experimental results where human-classification accuracy for five categories of emotion was just under 70% (Kim et al., 2009; Grimm et al., 2007).

* Corresponding author.
E-mail address: parkjs@yu.ac.kr (J.-S. Park).

http://dx.doi.org/10.1016/j.engappai.2016.02.018
0952-1976/

Approaches to speech emotion recognition can be classified into three categories according to the ways of constructing acoustic emotion models: speaker-independent (SI), speaker-dependent (SD), and speaker-adapted (SA) model frameworks. Among the three frameworks, the standard SI approach reveals apparent weaknesses in the domain-oriented characteristics of emotion recognition. This approach constructs acoustic emotion models by using training data obtained from a specific group of speakers who are not relevant to real users. The SI approach is simple and effective for common applications, but does not always guarantee stable performance because of unmatched acoustic characteristics between speakers in training data and real users. On the other hand, the SD model framework can efficiently handle the inter-speaker variation problem, because the acoustic models are built only using data of the system's user. Nevertheless, this approach has significant limitations in commercial applications owing to the difficulty of collecting a sufficient amount of emotion data from individual users. Finally, the SA model represents a model transformed from the SI model according to speaker adaptation procedures. The adaptation only requires a relatively small amount of data (called adaptation data) obtained from the user (called the target speaker), but produces the user-characterized acoustic model, nearly achieving the performance of the SD model (Matsui and Furui, 1998; Choi et al., 2015).

Speaker adaptation can be performed in either a supervised or an unsupervised manner in accordance with labeling methods. Hereby, the labels refer to transcription of adaptation data. Supervised speaker adaptation requires manual labeling tasks, whereas unsupervised adaptation depends on automatic labeling that is generally performed by recognition of adaptation data. Manual labeling can be characterized as an extremely time-consuming task and, in particular, may produce unreliable labels for emotion data, because it relies on subjective decisions by a human participating in the task. Although manual labels could be regarded as the ground truth, it might not be true in emotion recognition, because a manual annotation task is not a production process but is another perception process (Schuller et al., 2011). For these reasons, this study concentrates on unsupervised speaker adaptation.

The correctness of labels for adaptation data directly affects speaker adaptation performance. Hence, unsupervised adaptation in speech emotion recognition necessarily needs to carefully handle labeling errors, because the SI emotion model may be unreliable, thus generating numerous labeling errors. In this paper, we devise a sophisticated speaker adaptation approach that is not only robust against labeling errors but is also able to reflect the acoustic characteristics of individual speakers.

This paper is organized as follows. Section 2 introduces several previous works related to this study. Section 3 describes the proposed SER approach. In Section 4, experimental setups and results are presented and discussed. The paper concludes in Section 5.

2. Related works

2.1. Acoustic model-based SER

Fig. 1 summarizes the standard SER process that consists of extraction of acoustic feature vectors and identification of an emotional state. Previous studies on SER have concentrated on feature selection and classification approaches (Tato et al., 2002; Ververidis and Kotropoulos, 2006; Park et al., 2015). Feature selection techniques aim to investigate optimal feature sets representing emotional states of the speaker. On the other hand, classification approaches focus on defining distinctive boundaries between emotions. For the classification, various machine learning

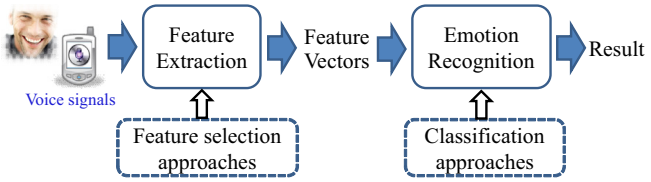

Fig. 1. Standard speech emotion recognition process.

algorithms such as the hidden Markov model (HMM), the Gaussian mixture model (GMM), and the support vector machine (SVM) have been commonly adopted. Among these methods, acoustic model-based classifiers such as GMM are better suited to classify emotions using short-term acoustic features like pitch and energy (Kim et al., 2009; Tato et al., 2002; Huang and Ma, 2006). In GMM-based SER, to identify the emotion type of input utterances, the likelihood of each GMM for an utterance is computed as follows:

$$P(X|\lambda_i) = \sum_{t=1}^{T} P\left(\vec{x_t}|\lambda_i\right) \quad (1)$$

where $X(=\{\vec{x_1},\ldots,\vec{x_T}\})$ means a sequence of feature vectors that are extracted from an input utterance, and a GMM λ_i $(i=1,\ldots,E$ if there are E emotions) indicates an acoustic model corresponding to the ith emotion. Then, a model that has the maximum likelihood of observing the input utterance is chosen as a recognition result.

As introduced in Section 1, acoustic emotion models can be categorized as SI, SD, and SA. SI and SD models have limitations in real applications owing to unreliable recognition accuracy and the difficulty of collecting emotional data, respectively. The SA approach can be an effective model for SER. Several recent studies introduced speaker adaptation-based SER techniques (Ding et al., 2012; Sidorov et al., 2014; Kim et al., 2011). Most of the studies investigated how to derive optimal models for adaptation data from a large speaker pool, taking into account a large speaker variation. However, preparing for a large speaker set is not practical, and error propagation in speaker information may induce unreliable adaptations. For more advanced adaptations, ambiguous properties of adaptation data need to be investigated in SER. Eventually, we propose an efficient adaptation technique that does not require any speaker information or a large speaker set and takes domain characteristics into account.

2.2. MLLR-based speaker adaptation for SER

Several adaptation techniques, such as maximum likelihood linear regression (MLLR) and maximum a posteriori (MAP) have been successfully applied to speech recognition tasks (Leggetter and Woodland, 1995; Woodland et al., 1996; Wang et al., 2009). As addressed in Section 1, SER has limitations in handling supervised adaptation owing to the difficulty of manual labeling of emotional data. Hence, the unsupervised approach is desirable for SER tasks. Among the conventional adaptation techniques, MLLR has been characterized as better suited to unsupervised adaptation because of its robustness against labeling errors (Leggetter and Woodland, 1995; Woodland et al., 1996; Wang et al., 2009). For this reason, this study concentrates on MLLR-based adaptation for SER.

Fig. 2 represents a general procedure for the conventional MLLR adaptation. MLLR adaptation revises the parameters of initial SI models, i.e. Gaussian means and variances, according to transformation matrices. Given adaptation data collected from target speakers and their labels, the transformation matrices are estimated to maximize the likelihood of the adapted models observing the adaptation data, using expectation-maximization (EM)

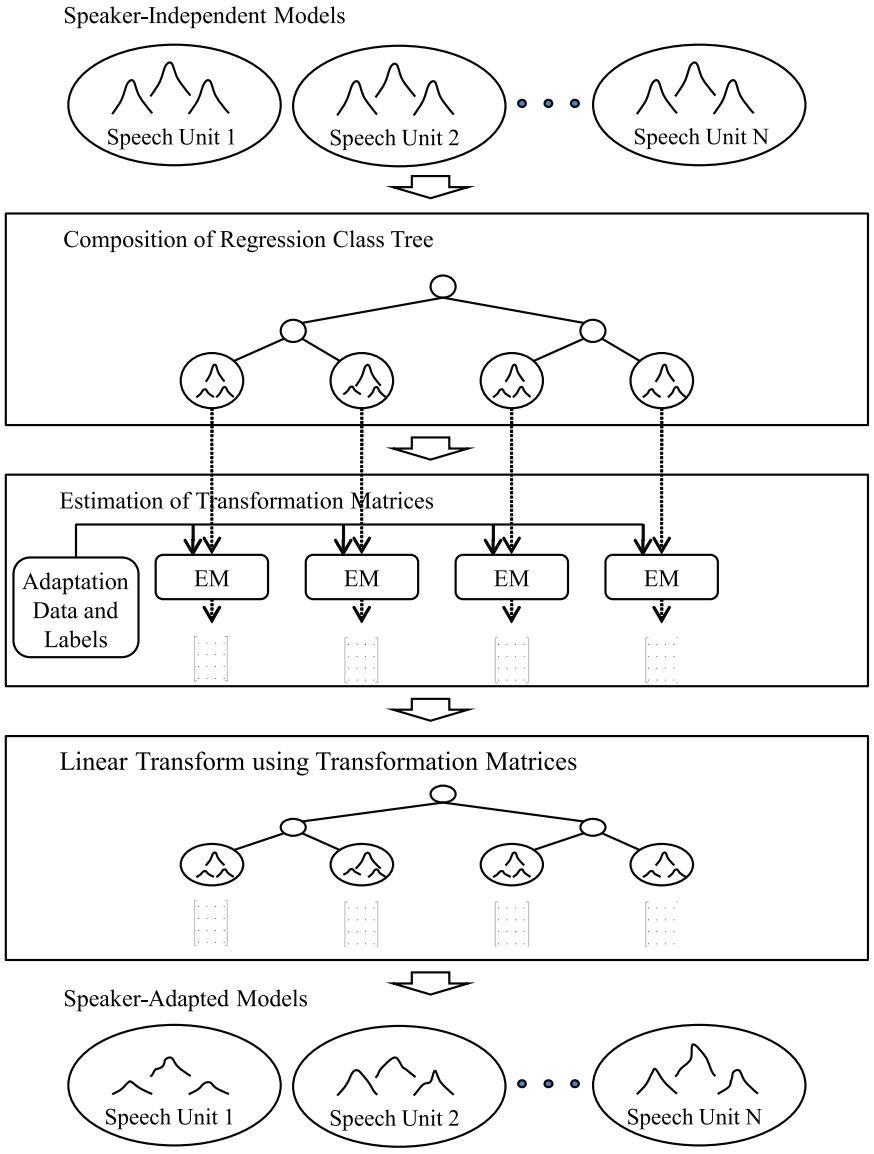

Fig. 2. Procedure for the conventional MLLR adaptation.

algorithm (Leggetter and Woodland, 1995). In unsupervised adaptation, the SI models recognize adaptation data, and the recognition results are used as labels. Hence, the accuracy of the SI model has a strong influence on the performance of MLLR adaptation (Anastasakos and Balakrishnan, 1998). In general, SI models in SER tasks reveal unreliable properties due to the ambiguity of emotional speech (Kim et al., 2009; Lopez-Moreno et al., 2009; Grimm et al., 2007). Therefore, careful consideration should be given as to whether the labels are reliable or not.

A principal process of MLLR is the composition of a regression class tree, tying acoustically similar Gaussian components into a class (Leggetter and Woodland, 1995). For this work, the acoustic similarity between Gaussian components of SI models is estimated using a Euclidean distance measure. The shared transformation matrix linearly transforms the tied components in each class. The total number of classes is decided by either the amount of adaptation data or the reliability of labels. If the labels are assumed to be unreliable, a global transform is recommended (Woodland et al., 1996). In this case, the components of all SI models are broadly tied into a single class. Every component is then linearly transformed by a single transformation matrix, which is estimated using all the adaptation data simultaneously.

On the other hand, if the labels are considered reliable, multiple regressions are available based on multiple transformation matrices. In this case, the components of SI models are specifically tied to multiple classes. For each class, a transformation matrix is estimated using the corresponding adaptation data and their labels. Multiple regressions can handle variations precisely; however, they are vulnerable to labeling errors.

Since global adaptation and multiple regressions (also denoted as multiple adaptations) have their own pros and cons, they are sometimes combined to optimize adaptation performance. Iterative unsupervised adaptation is a representative technique, which refines labels and models in an iterative manner (Woodland et al., 1996). In this paper, we follow this strategy somewhat, but modify the procedure in a more sophisticated way to consider the domain-oriented characteristics of SER.

3. Multistage data selection-based unsupervised speaker adaptation

In this study, we propose an efficient unsupervised speaker adaptation technique based on multistage data selection. Our

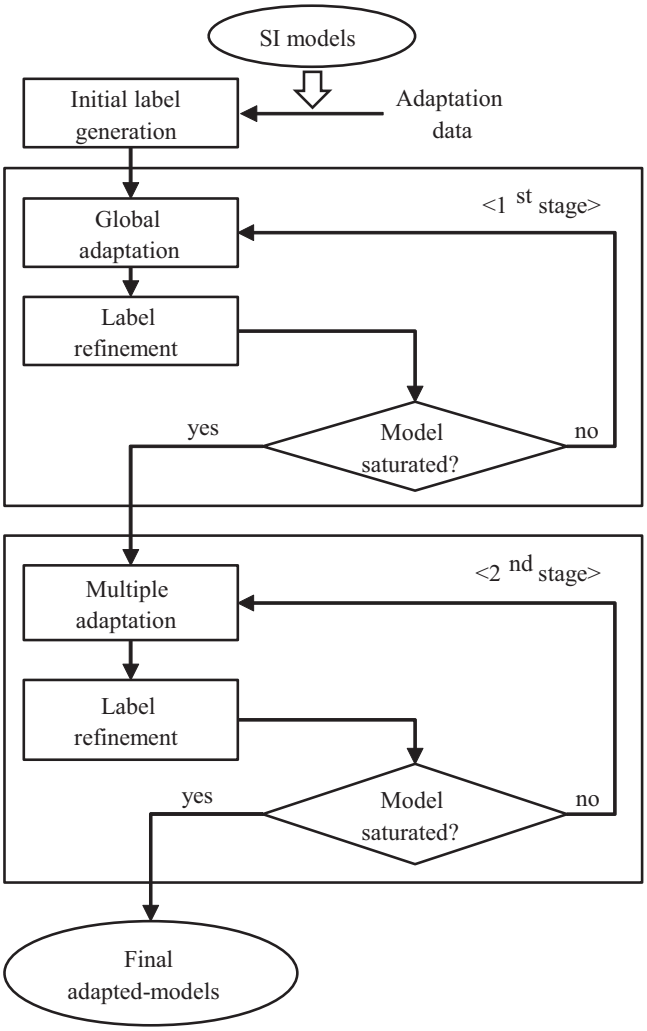

Fig. 3. Procedure for iterative unsupervised speaker adaptation.

approach is characterized as an iterative unsupervised adaptation and carefully considers characteristics of the SER task.

3.1. Iterative unsupervised speaker adaptation

In the proposed approach, we employ an iterative unsupervised adaptation scheme, expecting that the iterative refinements produce further reliable labels in SER. As shown in Fig. 3, this scheme refines emotion models and labels of adaptation data through two iterative stages, which are composed of global adaptation (the first stage) and multiple adaptations (the second stage).

Our approach originates from a general tendency of the SER task: the standard SI emotion models are considered unreliable. This tendency leads to incorrect labels of adaptation data, and therefore, only global adaptation is available, which generates inadequate adapted models compared to multiple adaptations. The main idea of our approach is to construct initial adapted emotion models by performing global adaptation in the first stage and enhance the adapted models by performing more precise adaptations, which are the multiple adaptations in the second stage.

Prior to the first stage, SI models and a regression class tree are constructed using a set of training data collected from a sufficient number of speakers. Initial labels for adaptation data are then generated from the SI models. Based upon the SI models, adaptation data, and the initial labels, a global adaptation process is performed to construct adapted models. Then, likelihoods for all adaptation data are estimated using the adapted models. We believe that the sum of the likelihood results can be used to determine whether the models are more reliable than the SI models, because more reliable models certainly produce higher likelihoods when compared using the same data set. So, if the likelihood sum obtained from the adapted models is higher than the sum from the SI models, we draw the conclusion that global adaptation generates more reliable models compared to the SI models. Based on this view, we proceed to perform global adaptation processes consecutively until the likelihood sum obtained from adapted models in a certain process does not increase any more. At that time, the models can be considered to be saturated and are characterized as the most reliable adapted models for the given adaptation data.

Adapted models that are finally obtained in the first stage are passed to the second stage, in which multiple adaptations are performed with an expectation that the emotion models preserve sufficiently discriminative emotional characteristics to construct a multiple regression class. To achieve more precisely adapted models, the adaptation process is repeated in an iterative manner similar to the first stage. Compared to the first stage, the multiple adaptations are capable of improving the adapted models by refining the labels of the adaptation data. As shown in Fig. 3, if the models are considered less saturated, the labels are refined using the adapted models and then used in the next adaptation process.

3.2. Necessity of data selection in SER tasks

In iterative unsupervised adaptation, global adaptation is expected to significantly advance the standard SI models. Nevertheless, we still doubt whether all the adaptation data of an unspecified target speaker always operate the multiple adaptation in a desirable way or not. It is hard to expect general speakers to express emotional speech that preserves discriminative acoustic characteristics while recording emotional speech or even communicating with someone in natural situations. Such ambiguous emotional speech may negatively affect the multiple adaptations.

To handle this problem, we carefully extend our iterative adaptation, employing a data selection technique. Data selection classifies all of the adaptation data into two categories: discriminative data and indiscriminative data. The discriminative data are considered to preserve discriminative emotional characteristics, whereas adaptation data indicating ambiguous characteristics are categorized as indiscriminative data. The main objective of data selection is to correctly select the discriminative data and submit only those data to the second stage, expecting that those discriminative data estimate multiple transformation matrices in a more sophisticated way, compared to the general case where all data are used. For reliable data selection, we propose a multistage data selection approach.

3.3. Multistage data selection

The proposed multistage approach conducts two consecutive procedures for data selection. The log-likelihood distance-based procedure is followed by the model adaptation-based approach.

3.3.1. Log-likelihood distance-based data selection

In acoustic model-based SER, a model that has the maximum log-likelihood with a given input utterance is determined to be a recognition result. In speech recognition, the larger distance between the maximum log-likelihood and that at the following rank is regarded as the result conveying higher confidence (Jiang, 2005). We attempt to apply this general tendency of speech recognition to the data selection in our task. But the distance between such a pair of log-likelihood results may be relatively

small in SER because the domain-oriented characteristics of several pairs of emotions, such as sadness and boredom, have similar acoustic characteristics. For this reason, we carefully extend the range of log-likelihood results to *N*-best results, which means a list of *N* candidate models ranked in order of the log-likelihood computed during the recognition process.

Emotionally discriminative data are regarded as the data that preserve explicit acoustic characteristics of the relevant emotion. Thus, compared to indiscriminative data, discriminative data are expected to demonstrate larger distances between the *N*-best log-likelihood results. Based on this property, we use the *N*-best results as a confidence measure to classify the adaptation data into discriminative and indiscriminative data. Given *E* emotion models and *T* adaptation data, let us $R_r(X_i)$ denote 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 the *i*th adaptation data, X_i. We conduct the classification for each piece of adaptation data according to the following log-likelihood distance-based confidence measure (LDM):

$$LDM(X_i) = \frac{1}{E-1}\sum_{r=1}^{E-1}\{\log P(X_i|\lambda_{R_r(X_i)}) - \log P(X_i|\lambda_{R_{r+1}(X_i)})\}^2 \quad (2)$$

where $\lambda_{R_r(X_i)}$ and $\log P(X_i|\lambda_{R_r(X_i)})$ indicate the emotion model corresponding to the model index and the log-likelihood result at the *r*th rank in the *N*-best list, respectively. This measure calculates the average distance between the likelihood at the *r*th rank and that at the $(r+1)$th rank, while considering the likelihood results at overall ranks in the *N*-best list.

Most studies on confidence measures have relied on a threshold that is empirically pre-determined to determine whether to accept or reject the given results (Anastasakos and Balakrishnan, 1998; Gollan and Bacchiani, 2008; Wallace et al., 2009). However, relying on this static value has limitations in covering acoustic characteristics of adaptation data collected from unspecified target speakers. In particular, the amount of adaptation data for the respective emotions can be different among target speakers, and thus, the static threshold may lead to misclassification of adaptation data. For this reason, we replace the static threshold with a dynamic threshold. The proposed threshold is dynamically updated according to adaptation data collected from a target speaker, and it is estimated differently for respective emotion models. So, we designate it as a model-based dynamic threshold (MDT). The MDT is defined as follows:

$$MDT(\lambda_{R_1(X_i)}) = \frac{1}{T_e}\sum_{j=1}^{T_e} LDM(X_j) \quad (3)$$

When $\lambda_{R_1(X_i)}$ refers to *e*, T_e means the total number of adaptation data recognized as *e*. In this equation, the threshold for an emotion model is the average value of LDM for all adaptation data belonging to the corresponding emotion. Based on this criterion, if $LDM(X_i) > MDT(\lambda_{R_1(X_i)})$, X_i is considered to sufficiently preserve the characteristics of the relevant emotion type, it is categorized as discriminative data. If not, it is categorized as indiscriminative.

The LDM-MDT-based data selection approach reflects the acoustic characteristics of adaptation data dynamically. However, the MDT depends on the average value of the LDM for classification of adaptation data. Thus, approximately half of all adaptation data are determined to be indiscriminative and are disregarded. This property becomes an obvious disadvantage when the amount of adaptation data is insufficient. In addition, adaptation data preserving discriminative emotional characteristics may be discarded after being categorized as indiscriminative data. For this reason, we continue to conduct an additional data selection procedure for the data categorized as indiscriminative during LDM-MDT-based data selection.

3.3.2. Model adaptation-based data selection

To select extra discriminative data from among indiscriminative data, we investigate the similarity between discriminative data and indiscriminative data categorized during the first data selection procedure. If a piece of indiscriminative data represents acoustic characteristics similar to discriminative data, the data can be added to the discriminative data set. For estimating similarity, we construct two types of acoustic model (a discriminative data model and an indiscriminative data model) using corresponding discriminative and indiscriminative data sets. The main idea of the second data selection procedure is to select indiscriminative data that demonstrate more specific characteristics on a discriminative data model rather than an indiscriminative data model, and then add them to a discriminative data set.

It should be remembered that an insufficient amount of data induces an inaccurate acoustic model. To overcome drawbacks caused by insufficiency in discriminative data relevant to respective emotions, we adapt the discriminative data to a generalized GMM designated as the universal background model (UBM). The UBM has been efficiently used to build a GMM-based speaker model with a small amount of data in speaker verification or segmentation tasks (Reynolds et al., 2000; Park et al., 2010, 2012). Instead of using the initial SI models as a UBM, we construct a new model using all adaptation data collected from a target speaker, expecting that using the model as a UBM derives more definite characteristics of the target speaker. After constructing the UBM, we adapt the discriminative data of each emotion type to the UBM. The adapted models preserve more definite acoustic characteristics of relevant emotion, as well as of the target speaker, in comparison with discriminative data models. The adapted discriminative data model is designated herein as ADM. In the same way, the indiscriminative data of each emotion is adapted to the UBM, and the adapted indiscriminative data model (AIM) is obtained. Fig. 4 illustrates the procedure for constructing the ADM and the AIM.

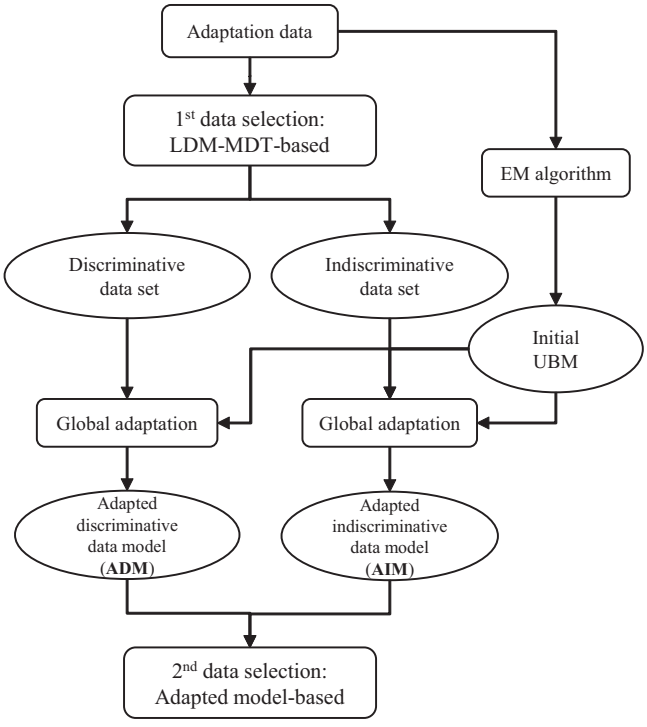

Fig. 4. Procedure for constructing the adapted discriminative and indiscriminative data models.

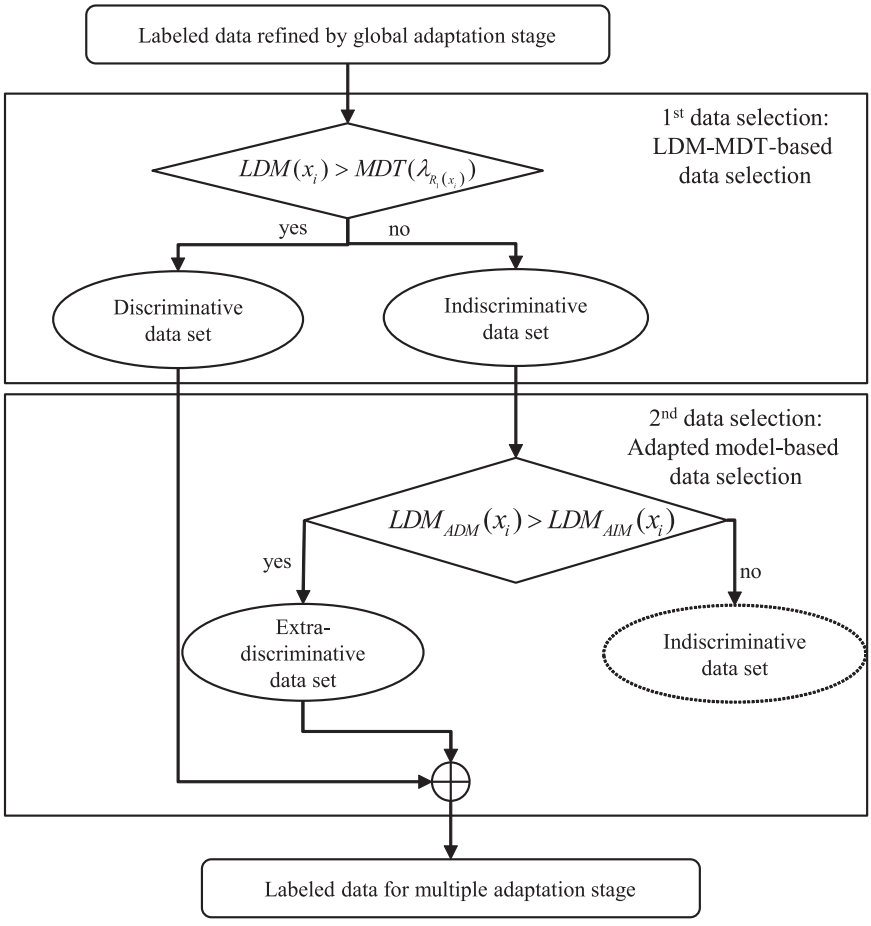

Fig. 5. Procedure for multistage data selection.

Next, given the ADM and the AIM, the second data selection is conducted for indiscriminative data. In this procedure, we employ the LDM described in (2) as a measure for determining whether an indiscriminative data preserves acoustic characteristics of the ADM. If the indiscriminative data indicates a higher LDM on ADM than on AIM, the data is regarded as preserving acoustically discriminative emotional characteristics and is added to discriminative data set. Otherwise, the data is still considered indiscriminative. Fig. 5 summarizes the procedures for the proposed multistage data selection. Only the data categorized as discriminative from this procedure are submitted to the multiple adaptation stage, whereas indiscriminative data are disregarded.

3.4. *Personalized SER using multistage data selection*

As shown in Fig. 6, we propose a new SER framework in terms of a personalized SER. In an offline process, SI emotion models and regression class trees are constructed. The following iterative unsupervised speaker adaptation can be conducted in an online manner, once a sufficient amount of adaptation data are obtained from a target speaker. After completing global adaptation, multistage data selection is performed to select discriminative data from among adaptation data and submit them to the next process. The multiple adaptation process is then followed, and speaker-adapted emotion models are finally obtained. The models preserve acoustic characteristics relevant to the target speaker, thus enabling personalized SER for the speaker.

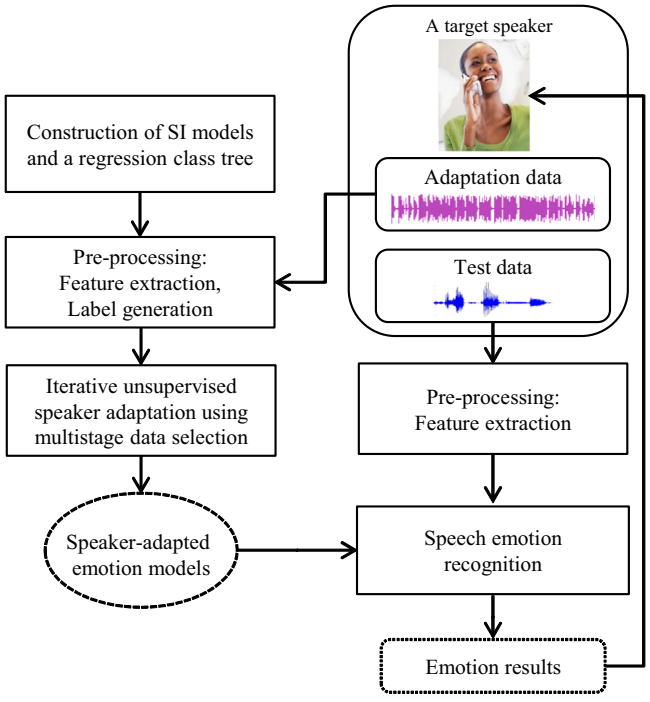

Fig. 6. Proposed framework for personalized SER.

4. Experimental results and discussion

4.1. Experimental setups

To evaluate the proposed adaptation approach, we performed emotion recognition experiments on emotional speech data obtained from the Emotional Prosody Speech of the LDC (Liberman et al.). This corpus consists of speech recorded by seven professional actors and actresses expressing emotions while reading short phrases of dates and numbers. We used a phrase as the basic unit of the adaptation data and performed seven folds of experiments using Leave-One-Speaker-Out-Cross-Validation (LOSOCV). In each fold, half of the utterances spoken by a test speaker are used for adaptation while the other half are used for testing. All utterances of other remaining speakers are used for training SI models. As a result, the proportion of the total adaptation data is less than 7.5% of the total 738 utterances, and the amount of adaptation data from a single speaker is about 1 min long.

To investigate SER performance according to the number of emotions, we composed four types of emotion categories with five different emotions: neutral, happiness, hot-anger, boredom, and sadness as the representative human emotions, and generally used categories in 5-class SER tasks (Kim et al., 2009; Tato et al., 2002; Ververidis and Kotropoulos, 2006). For a fair evaluation, we estimated the SER performance on a variety of emotion categories. Table 1 represents the emotion categories composed for our evaluation. “Neutral” is chosen as the most common emotion and is used in every emotion set.

We used log energy, 12-dimensional mel-frequency cepstral coefficients (MFCC), pitch, and their first and second derivatives, yielding a 42-dimensional feature vector. All feature vectors were extracted within frames of 40 ms with a Hamming window shifted by 10 ms, and 40 ms considered a minimum duration for reliable estimation of emotion characteristics (Kim et al., 2009). Acoustic models were trained as GMMs, with each GMM having 16 Gaussian mixtures.

We investigated SER performance according to several types of speaker-adapted emotion models constructed by conventional adaptation techniques, such as MLLR and MAP. The proposed adaptation scheme using multistage data selection based on LDM-MDT and UBM is denoted as LDM-BM-MLLR. A conventional MLLR-based iterative unsupervised adaptation scheme is denoted as MLLR. Conventional MAP-based unsupervised adaptation is denoted as MAP. In addition, the SI model-based SER was also tested for the purpose of performance comparison and is denoted as Baseline.

4.2. Performance comparison on overall emotion sets

Fig. 7 summarizes the overall performance of each approach. In all emotion categories, LDM-BM-MLLR demonstrated significant performance improvements with respect to other approaches. In 2-class SER, LDM-BM-MLLR achieved an error rate of 4.5%, which is commercially applicable performance.

Table 1
Emotion categories.

Number of emotions	Classification sets
5	{neutral, hot-anger, happiness, sadness, boredom}
4	neutral, 3-combination from {hot-anger, happiness, sadness, boredom}/total 4 sets
3	neutral, 2-combination from {hot-anger, happiness, sadness, boredom}/total 6 sets
2	neutral, 1-combination from {hot-anger, happiness, sadness, boredom}/total 4 sets

MAP and MLLR also successfully improved Baseline performance in most classes. It greatly supports a main motivation in this study that speaker adaptation techniques efficiently reduce inter-speaker variation among domain-oriented characteristics in SER. In the 2-class SER experiment, MAP showed performance similar to LDM-BM-MLLR because the 2-class SI models are sufficiently reliable for generation of correct labels.

Careful consideration is necessary for other classes, in which performance from MAP and MLLR deteriorated significantly in comparison to LDM-BM-MLLR. We believe that as the number of emotion sets increases, the amount of overlapped feature vectors increases in the acoustic feature space, thus inducing indiscriminative data. This tendency is confirmed in Table 2, which shows the proportion of discriminative feature vectors classified by LDM-BM-MLLR. As shown in this table, the amount of discriminative vectors notably decreased over increasing emotion sets. Loss of discriminative vectors may negatively affect the correctness of label refinement results.

To analyze the effect on adaptation performance of a decreasing number of discriminative vectors, we investigated relative performance improvement of the proposed adaptation approach to the conventional approaches. As shown in Table 3, LDM-BM-MLLR achieved superior performance compared to Baseline and MLLR-based and MAP-based approaches, even though the improvement ratio decreased over increasing emotion sets. This experimental result explains why the conventional adaptation techniques fail to maintain the adaptation performance, whereas the proposed approach successfully copes with the domain-oriented tendency by efficiently maintaining emotionally discriminative vectors via multistage data selection.

4.3. Performance comparison on negative emotion sets

In general, correctly detecting negative emotions such as anger, sadness, and boredom is of great importance to commercial application of an emotion recognition system, because it is more

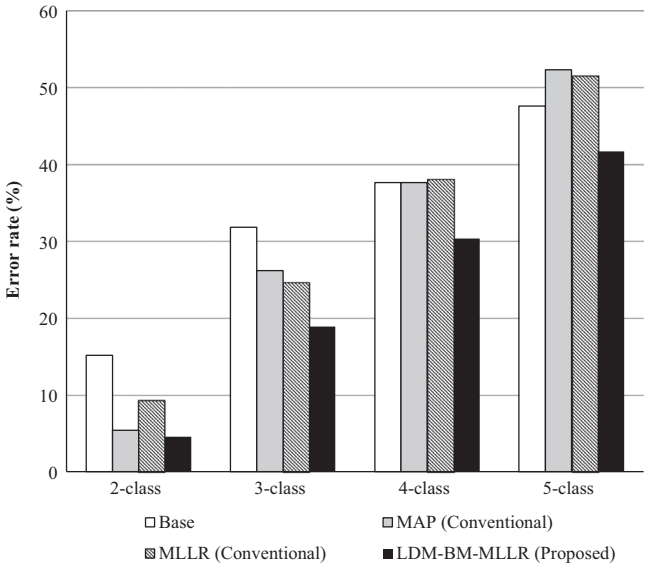

Fig. 7. Error rate (%) of each approach according to the number of emotions.

Table 2
Proportion (%) of discriminative feature vectors.

	2-class	3-class	4-class	5-class	Average
Proportion	91.9	84.4	82.6	76.8	83.9

Table 3
Relative Improvement (RI; %) of LDM-BM-MLLR to each approach.

Approach	2-class	3-class	4-class	5-class	Average
RI to Baseline	70.0	40.6	19.3	12.4	35.6
RI to MLLR	51.1	23.3	20.3	19.0	28.4
RI to MAP	15.4	27.8	19.3	20.3	20.7

Table 4
Error rate (%) of each approach according to negative emotion sets and proportion (%) of discriminative feature vectors.

Approach	NAB	NAS	NBS	Average
Baseline	15.3	22.2	49.0	28.9
MLLR	6.1	23.3	47.6	25.7
LDM-BM-MLLR	2.0	17.5	41.8	20.4
Proportion	94.0	84.0	76.0	84.7

necessary to provide appropriate feedback to users feeling negative emotions rather than users who have positive emotions. For this reason, we attempted to compare performance focusing on negative emotion sets. In this experiment, four emotion types were involved: anger (A), boredom (B) and sadness (S) as representative negative emotions, and neutral (N) as a non-negative emotion. Thus, recognition results for three emotion categories (NAB, NAS, NBS), were investigated respectively. In this experiment, we disregarded the MAP-based approach that provided the worst performance in the previous experiments.

Table 4 demonstrates the recognition results of each approach according to negative emotion sets and also provides the proportion of discriminative feature vectors. As expected, the proposed approach showed significant performance improvement over other approaches for each emotion set. Relative improvement of LDM-BM-MLLR was about 28.9%, 25.7%, and 20.4% with respect to Baseline and MLLR. This result confirms that the LDM-BM-MLLR approach effectively selects acoustically discriminative vectors for negative emotion sets, thus more correctly estimating multiple transformation matrices.

We need to concentrate on the NBS set among others, because sadness and boredom are reported to have acoustically similar characteristics and to induce recognition errors (Grimm et al., 2007). This tendency was investigated in our experimental results, in which the NBS set showed an even higher error rate than other sets. A main reason for this result can be explained with regard to the amount of emotionally discriminative feature vectors. As shown in this table, the NBS set provided the lowest proportion of discriminative vectors among three sets. It means that sadness and boredom have acoustically similar characteristics, thus reducing the amount of discriminative vectors. Nevertheless, the proposed LDM-BM-MLLR approach significantly improved performance compared to Baseline and MLLR for the NBS set.

The next analysis focuses on the other sets involving anger. In general, anger has acoustically different characteristics from sadness and boredom. Hence, in the NAB and NAS sets, the adaptation process is expected to enhance performance over Baseline. An unexpected result was shown in the NAS set, in which the conventional MLLR approach failed to improve performance. A possible reason is incorrect label refinement that leads to construction of unreliable adapted models. In comparison to the MLLR approach, the proposed approach successfully achieved better performance over Baseline.

Our experimental results demonstrate that the proposed adaptation approach solves two problems that should be addressed when the conventional unsupervised adaptation approach is applied to SER: incorrect label refinement in a large number of emotions, and sparseness of acoustically discriminative feature vectors. By selecting discriminative data based on a multistage data selection method, it is possible to construct more robust speech emotion models, thus reducing recognition errors.

5. Conclusion

This paper proposed an efficient speaker adaptation approach for personalized SER using individual user voice data. To solve drawbacks of the conventional adaptation approaches such as MAP and MLLR, we proposed a multistage data selection method that maintains the amount of discriminative feature vectors, and thus enables us to construct reliable adapted models. In emotion recognition experiments, the proposed approach exhibited superior performance to that of the conventional approaches, achieving commercially applicable performance for 2-class tasks, and even some 3-class tasks, involving negative emotions.

In future work, we will apply the proposed approach to other emotional speech databases for further verification. In addition, performance comparison with other adaptation techniques, such as the eigenvoice approach, will be considered.

Acknowledgments

This research was supported by EC's 7th FP grant agreement 611153 (TERESA – Telepresence Reinforcement-learning Social Agent), the 2014 Yeungnam University Research Grant, and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2014R1A1A2057751).

References

Anastasakos, T., Balakrishnan, S.V., 1998. The use of confidence measures in unsupervised adaptation of speech recognizers. In: ICSLP, 1998.
Ballagas, R., Borchers, J., Rohs, M., Sheridan, J.G., 2006. The smart phone: a ubiquitous input device. IEEE Pervasive Comput. 5 (1), 70–77.
Choi, D., Park, J., Oh, Y., 2015. Unsupervised rapid speaker adaptation based on selective eigenvoice merging for user-specific voice interaction. Eng. Appl. Artif. Intell. 40, 95–102.
Ding, N., Sethu, V., Epps, J., Ambikairajah, E., 2012. Speaker variability in emotion recognition-an adaptation based approach. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5101–5104.
Gollan, C., Bacchiani, M., 2008. Confidence scores for acoustic model adaptation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, ICASSP 2008. IEEE, pp. 4289–4292.
Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Commun. 49 (10), 787–800.
Huang, R., Ma, C., 2006. Toward a speaker-independent real-time affect detection system. In: Proceedings of the International Conference on Pattern Recognition (ICPR), vol. 1, pp. 1204–1207.
Jiang, H., 2005. Confidence measures for speech recognition: a survey. Speech Commun. 45 (4), 455–470.
Kim, J.-B., Park, J.-S., Oh, Y.-H., 2011. On-line speaker adaptation based emotion recognition using incremental emotional information. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4948–4951.
Kim, J., Park, J., Oh, Y., 2009. Feature vector classification based speech emotion recognition for service robots. IEEE Trans. Consum. Electron. 55 (3), 1590–1596.
Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Comput. Speech Language 9 (2), 171–185.
Liberman, M., Davis, K., Grossman, M., Martey, N., Bell, J. Emotional prosody speech and transcripts. In: Proceedings of the Linguistic Data Consortium, Philadelphia.
Lopez-Moreno, I., Ortego-Resa, C., Gonzalez-Rodriguez, J., Ramos, D., 2009. Speaker dependent emotion recognition using prosodic supervectors. In: INTERSPEECH, pp. 1971–1974.
Matsui, T., Furui, S., 1998. N-best-based unsupervised speaker adaptation for speech recognition. Comput. Speech Language 12 (1), 41–50.

Neerincx, M., Streefkerk, J.W., 2003. Interacting in desktop and mobile context: emotion, trust, and task performance. In: Ambient Intelligence. Springer, pp. 119–132.

Park, K., Park, J.-s., Oh, Y.-H., 2010. Gmm adaptation based online speaker segmentation for spoken document retrieval. IEEE Trans. Consum. Electron. 56 (2), 1123–1129.

Park, K.-M., Park, J.-S., Bae, J.-H., Oh, Y.-H., 2012. Online speaker diarization for multimedia data retrieval on mobile devices. Int. J. Pattern Recognit. Artif. Intell. 26 (08), 1260011.

Park, J., Kim, J., Oh, Y., 2015. Emotional information processing based on feature vector enhancement and selection for human–computer interaction via speech. Telecommun. Syst.

Pittermann, J., Pittermann, A., Minker, W., 2010. Handling Emotions in Human–computer Dialogues. Springer.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. Digit. Signal Process. 10 (1), 19–41.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Commun. 53 (9), 1062–1087.

Sidorov, M., Ultes, S., Schmitt, A., 2014. Emotions are a personal thing: Towards speaker-adaptive emotion recognition, In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4803–4807.

Tato, R., Santos, R., Kompe, R., Pardo, J.M., 2002. Emotional space improves emotion recognition. In: INTERSPEECH.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features, and methods. Speech Commun. 48 (9), 1162–1181.

Wallace, R., Thambiratnam, K., Seide, F., 2009. Unsupervised speaker adaptation for telephone call transcription. In: IEEE International Conference on Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009). IEEE, pp. 4393–4396.

Wang, H., Zhang, X., Xiao, X., Zhang, J., Yan, Y., 2009. Combining MAP and MLLR approaches for SVM based speaker recognition with a multi-class MLLR technique. In: International Symposium on Information Science and Engineering, pp. 447–450.

Woodland, P.C., Pye, D., Gales, M., 1996. Iterative unsupervised adaptation using maximum likelihood linear regression. In: Proceedings of the International Conference on Spoken Language (ICSLP), vol. 2, pp. 1133–1136.

Zhang, S., Wang, X., Zhang, G., Zhao, X., 2014. Multimodal emotion recognition integrating affective speech with facial expression. WSEAS Trans. Signal Process. 10, 526–537.