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
Available online 22 March 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

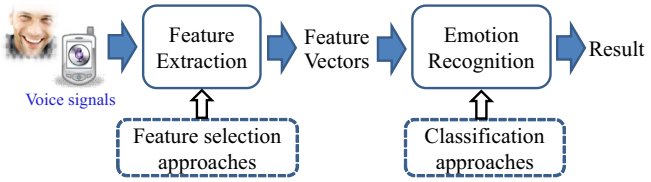


**Fig. 1.** Standard speech emotion recognition process.

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 $\lambda_i$ $(i=1,\ldots,E$ if there are $E$ emotions) indicates an acoustic model corresponding to the $i$th 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)