Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Mismatched feature detection with finer granularity for emotional speaker recognition*

Li CHEN, Ying-chun YANG‡, Zhao-hui WU

(*College of Computer Science & Technology, Zhejiang University, Hangzhou 310027, China*)

E-mail: stchenli@zju.edu.cn; yyc@zju.edu.cn; wzh@zju.edu.cn

Received Jan. 5, 2014; Revision accepted May 20, 2014; Crosschecked Sept. 17, 2014

Abstract: The shapes of speakers' vocal organs change under their different emotional states, which leads to the deviation of the emotional acoustic space of short-time features from the neutral acoustic space and thereby the degradation of the speaker recognition performance. Features deviating greatly from the neutral acoustic space are considered as mismatched features, and they negatively affect speaker recognition systems. Emotion variation produces different feature deformations for different phonemes, so it is reasonable to build a finer model to detect mismatched features under each phoneme. However, given the difficulty of phoneme recognition, three sorts of acoustic class recognition—phoneme classes, Gaussian mixture model (GMM) tokenizer, and probabilistic GMM tokenizer—are proposed to replace phoneme recognition. We propose feature pruning and feature regulation methods to process the mismatched features to improve speaker recognition performance. As for the feature regulation method, a strategy of maximizing the between-class distance and minimizing the within-class distance is adopted to train the transformation matrix to regulate the mismatched features. Experiments conducted on the Mandarin affective speech corpus (MASC) show that our feature pruning and feature regulation methods increase the identification rate (IR) by 3.64% and 6.77%, compared with the baseline GMM-UBM (universal background model) algorithm. Also, corresponding IR increases of 2.09% and 3.32% can be obtained with our methods when applied to the state-of-the-art algorithm i-vector.

Key words: Emotional speaker recognition, Mismatched feature detection, Feature regulation
doi:10.1631/jzus.C1400002 **Document code:** A **CLC number:** TP391.4

1 Introduction

Automatic speaker recognition (ASR) systems cannot achieve satisfactory performance in real environments due to many factors. Most of these factors can be categorized into two groups: inter- and intra-variability (Ghiurcau *et al.*, 2011b). Inter-variability is induced by different external conditions, such as different recording backgrounds (Rose *et al.*, 1994), recording devices (Reynolds, 2003), or recording distances (Jin *et al.*, 2007). Intra-variability indicates changes in the speaker's characteristics from one utterance to another. Long-term intra-variability is caused by aging (Kelly and Harte, 2011), disease (Gadek, 2009), or other permanent physical changes of the vocal organs. Short-term intra-variability is caused by short-term illness, different speaking styles (Shriberg *et al.*, 2008), different emotional states, or imitation. In this paper, we propose to alleviate the negative effects brought about by emotional variability. The application scenario is that only the neutral speech of a speaker is provided in the enrollment stage, whereas the speech of many different emotional states is tested in the evaluation stage. We call this sort of speaker recognition emotional speaker recognition (ESR).

‡ Corresponding author

* Supported by the National Basic Research Program (973) of China (No. 2013CB329504), the National Natural Science Foundation of China (No. 60970080), and the National High-Tech R&D Program (863) of China (No. 2006AA01Z136)

1.1 Previous work

Different emotional states induce different voice pronouncing mechanisms. Physiology research on emotional speech indicates that metabolic oxygen consumption varies under different emotional states, which change indirectly the respiratory frequency and ultimately the speech rate and voice quality (Brady, 2005). As concluded by El Ayadi *et al.* (2011), emotional variability leads to changes in more than 20 acoustic and prosodic features, such as fundamental frequency (F_0), intensity, formant, speech rate, energy, and duration.

Due to the variability induced by emotional effects, the performance of an ASR system deteriorates. This phenomenon was noted as long ago as 1998 (Scherer *et al.*, 1998). Our lab has worked on ESR since 2003 and found the same result (Yang and Chen, 2012) with Mandarin affective speech corpus (MASC). The identification rate (IR) result with MASC is shown in Table 1. The IR was computed by dividing the number of correctly recognized trials by the number of all trials. Table 1 shows clear evidence that the performance of the ASR system deteriorated dramatically with the influence of emotional variability. The IRs under consistent emotional conditions are much higher than those of inconsistent conditions. Ghiurcau *et al.* (2011a) conducted an experiment using the Berlin emotional speech database and showed the important influence of emotional state upon text-independent speaker identification. Jawarkar *et al.* (2012) also implemented experiments on a corpus in Marathi with five emotional states: anger, fear, sadness, happiness, and disgust. The IR dropped by more than 20% under each emotional state, especially under disgust, anger, and happiness.

Table 1 The influence of emotional variability on automatic speaker recognition performance

Emotional state	Identification rate (%)				
	Neutral	Anger	Elation	Panic	Sadness
Neutral	**89.28**	28.27	31.67	26.90	49.74
Anger	21.34	**75.95**	31.99	27.91	17.39
Elation	25.00	18.14	**70.59**	19.84	27.65
Panic	28.30	17.94	27.52	**67.78**	28.82
Sadness	43.01	23.99	21.99	23.86	**89.15**

Each speaker's enrollment data under each emotional state were collected for about 20 s, and the number of test utterances of each speaker was 45 under each emotional state. 13-order mel-frequency cpestral coefficient (MFCC) plus delta and 512-order Gaussian mixture model-universal background model (GMM-UBM) were used

To address the issue of emotional speaker recognition systematically, we propose a deformation compensation (DC) framework, which regards emotional variability as some sort of deformation of feature distribution space, and compensates the deformation by adaptively regulating the corresponding features, models, or scores. There are two strategies to compensate for emotional variability: enrichment and normalization strategies (Yang and Chen, 2012). The enrichment strategy is to synthesize the speaker's emotional model from his/her neutral model using the rules learned from the development corpus. Shan *et al.* (2007) and Shan and Yang (2008) proposed a neutral-emotional transformation algorithm based on the assumption that each emotional Gaussian mixture model (GMM) component is a linear combination of neutral GMM components. The normalization strategy removes the emotional property from the test utterance and can be approached in two ways. The first way is the sub-space normalization method named emotional attribute projection (EAP) (Bao *et al.*, 2007). The main idea is to eliminate the emotional influence by using a new support vector machine (SVM) kernel which subtracts the emotional attribute space from the supervector representing the utterance. The second way is a mismatched feature detection method. Because only some parts of speech express the speaker's emotion, detecting those features severely affected by the emotional states and adopting some strategies to normalize them is feasible. Huang and Yang (2008; 2010) found that the pitch distribution of emotional segments shifts from that of neutral segments, so mismatched segments can be detected through abnormal pitch detection. Pitch-dependent difference detection and modification (PDDM) is proposed to detect the mismatched features and pitch-synchronous overlap and add (PSOLA) is applied to modify the pitches of the mismatched segments. In addition, Shahin (2013) proposed the CSPHMM2s model to integrate both acoustic and suprasegmental information to improve emotional speaker recognition performance.

1.2 Motivations

Emotional variability induces deviation in the distribution of acoustic features. Also, the direction

and scale of the deviations vary among different phonemes. Thus, to enhance the performance of mismatched feature detection, it is necessary to implement the detection method at the phoneme level. However, the accuracy of phoneme recognition is only 70%–80% (Triefenbach *et al.*, 2010). Therefore, we aimed to replace phoneme recognition with three sorts of finer acoustic class detection: phoneme classes, GMM tokenizer, and probabilistic GMM tokenizer. The SVM or fuzzy SVM classifiers are built under each acoustic class to detect these mismatched features. A feature regulation strategy corresponding to each acoustic class is proposed to normalize the emotional features to the distribution of their neutral counterparts.

2 SVM and fuzzy SVM

2.1 Support vector machine

Given a training set of T data points $(\boldsymbol{x}_t, y_t)$ $(t = 1, 2, \cdots, T)$ (where $\boldsymbol{x}_t \in \mathbb{R}^n$ and $y_t \in \mathbb{R}$ are the tth input and output patterns), the support vector machine aims to construct a classifier of the form:

$$y(\boldsymbol{x}) = \sum_{t=1}^{\mathrm{sv}} \alpha_t y_t \Psi(\boldsymbol{x}, \boldsymbol{x}_t) + b,$$

where α_t is a positive real constant and b is a real constant. $\Psi(\cdot, \cdot)$ is the kernel function which maps the input space into a higher dimensional space. sv is the number of support vectors. According to the structural risk minimization principle, the risk bound is minimized by formulating the optimization problem:

$$\min\left\{\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{t=1}^{T}\xi_t\right\}$$
$$\text{s. t.} \begin{cases} y_t(\boldsymbol{w}\cdot\boldsymbol{x}_t + b) \geq 1 - \xi_t, & t = 1, 2, \ldots, T, \\ \xi_t \geq 0, & t = 1, 2, \ldots, T, \end{cases} \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^N$ is a non-zero vector normal to the hyperplane $\boldsymbol{w}\cdot x + b = 0$, C $(C > 0)$ is the penalty coefficient, and ξ_t is used to evaluate the extent of the wrong classification result. The probability output of SVM of each sample is computed as

$$p(\boldsymbol{x}) = \frac{2}{1 + \exp(Ay(\boldsymbol{x}) + B)} - 1, \tag{2}$$

where $y(x) = \boldsymbol{w}\cdot\boldsymbol{x} + b$ is the output of the SVM classifier. $y(\boldsymbol{x})$ is normed to $[-1, 1]$ using Eq. (2) and the corresponding output is $p(\boldsymbol{x})$. The parameters A and B are determined by the fitting sigmoid procedure mentioned by Platt (1999).

In our study, the polynomial kernel is adopted to train the SVM classifier,

$$\Psi(\boldsymbol{x}, \boldsymbol{x}_t) = [1 + (\boldsymbol{x}\cdot\boldsymbol{x}_t)]^q.$$

The penalty coefficient C is set to 0.125, and the degree q of the polynomial kernel is set to 2.

2.2 Fuzzy support vector machine

In traditional SVM classifiers, each sample is labeled strictly as positive or negative. However, in many real-world environments, features may belong to each class only with a certain probability less than 100%. Thus, Lin and Wang (2002) introduced fuzzy membership to represent the concept of probability.

The training set is $F = \{(\boldsymbol{x}_t, y_t, m_t)\}$ $(t = 1, 2, \cdots, T)$, where $\boldsymbol{x}_t$ is the training feature and y_t is the class of the feature. $y_t = 1$ means $\boldsymbol{x}_t$ is a positive sample and $y_t = -1$ means $\boldsymbol{x}_t$ is a negative sample. m_t denotes the fuzzy membership of $\boldsymbol{x}_t$. The optimization problem (1) in traditional SVM is converted into

$$\min\left\{\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{t=1}^{T} m_t\xi_t\right\}$$
$$\text{s. t.} \begin{cases} y_t(\boldsymbol{w}\cdot\boldsymbol{x}_t + b) \geq 1 - \xi_t, & t = 1, 2, \ldots, T, \\ \xi_t \geq 0, & t = 1, 2, \ldots, T, \\ 0 \leq m_t \leq 1, & t = 1, 2, \ldots, T. \end{cases}$$

$C\sum_{t=1}^{T}\xi_t$ in Eq. (1) is converted into $C\sum_{t=1}^{T} m_t\xi_t$. The fuzzy membership m_t of each feature is added as a weight of the extent of the wrong classification result ξ_t.

3 Observations and hypotheses

This section builds the hypotheses for our work. Firstly, we define a mismatched feature using the z-score. Secondly, we observe the distribution of mismatched features and assume that they lie far away from the neutral-emotional classification plane. Thirdly, emotion produces different feature deformations for different phonemes, so we will detect the mismatched features with a finer granularity—acoustic class.

3.1 Observation—emotional effect: deformation of acoustic feature space produces mismatched features

Mismatched features do not preserve speaker identity and negatively affect the performance of an ASR system. According to Cowie and Cornelius (2003), the number of frames influenced by emotional states depends on the emotional intensity. The higher the intensity, the higher the number of corrupted frames. Missing feature theory is effective in partial corruption conditions (Drygajlo and El-Maliki, 1998). It is also the appropriate theory for the emotional speaker recognition task. Thus, finding the distribution of these mismatched or missing features is very important.

The distribution of a speaker's emotional short-time features will deviate from that of his/her neutral features. If a speaker's emotional mel-frequency cpestral coefficient (MFCC) feature has a higher likelihood score on the target speaker than on imposter speakers, the emotional MFCC feature has speaker discriminative ability and makes a positive contribution to our speaker recognition system. Otherwise, it lacks the capability of speaker discrimination and we call it a mismatched feature. The z-score is a measure of percentile rank which takes both the center and dispersion of the distribution into consideration and which is adopted to measure the reliability of the MFCC feature.

Given a feature $\boldsymbol{x}_t$ of the jth speaker under an emotional state, $p(\boldsymbol{x}_t|\lambda_i)$ is the likelihood score of $\boldsymbol{x}_t$ against the ith speaker, denoted as $l_{t,i}$. λ_i is the ith speaker's GMM model. The likelihood scores of $\boldsymbol{x}_t$ on all speakers comprise the set L, $L=\{l_{t,1}, l_{t,2}, \ldots, l_{t,M}\}$. M represents the number of speakers in the corpus. Then, the z-score for the jth speaker's feature $\boldsymbol{x}_t$ is computed on the set L as

$$z_t = \frac{l_{t,j} - \overline{l}}{\sigma_l}, \quad l \in L,$$

where $l_{t,j}$ is the likelihood score against the target speaker. $\overline{l}$ and σ_l are the mean and standard deviation respectively, of the set L. If the z-score is high, it means the gap between the score for the target speaker and the scores for the imposter speakers is large. Moreover, it indicates the feature has an adequate ability to distinguish the target speaker from the imposter speakers. Otherwise, the feature cannot distinguish these two.

3.2 Hypothesis 1: mismatched features lie far away from the classification plane

The z-score of each feature $\boldsymbol{x}_t$ is normalized to the gray scale [0, 255] using the equation

$$\text{Gray}_t = \frac{z_t - \min(z)}{\max(z) - \min(z)} \times 255.$$

The larger the z-score, the larger the gray value and the lighter the color of the point representing the feature. The MFCC feature is reduced to three dimensions using the ISOMAP (Balasubramanian and Schwartz, 2002) manifold algorithm. The distribution of the MFCC feature is shown in Fig. 1.

Fig. 1 illustrates how the z-score varies with the position of the features. The phenomenon indicates that those features far away from the neutral acoustic space have weaker discriminative power. The farther the distance, the weaker the discriminative power. If the discriminative power is less than the threshold, the feature will be regarded as mismatched.

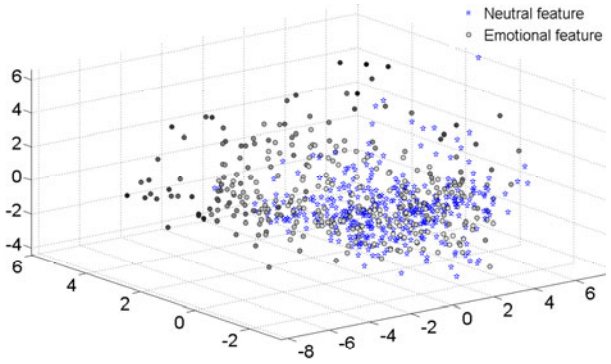

Fig. 1 The z-score value varies with the position of the feature. All the MFCC features of vowel 'a' of a male speaker under neutral and emotional conditions are extracted. These features are reduced to three dimensions using the ISOMAP algorithm (the number of neighbors is set to 10). If the z-score of the emotional feature is larger, the color of the '•' is lighter, indicating that this feature has stronger discriminative power. Otherwise, the color is darker, and the discriminative power of the feature is weaker. The darker '•' spots are located in the area far away from the distribution of neutral features

If an SVM classifier is constructed with emotional features as negative samples and neutral features as positive samples, the mismatched features will lie on the emotional side and be far away from the classification hyperplane (Fig. 2). The classification hyperplane is trained using positive and negative samples. The mismatched features are located

in the triangular area of the emotional region, which is far away from the SVM hyperplane (Chen *et al.*, 2011). Detecting, pruning, and regulating these mismatched features effectively can enhance the performance of an ESR system.

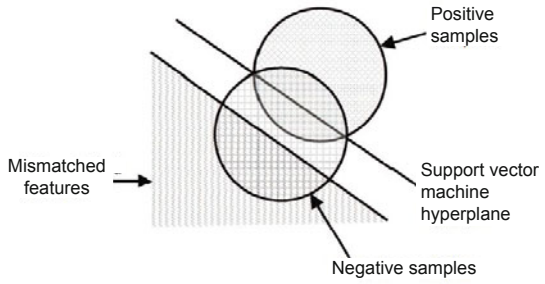

Fig. 2 Schematic plot of the distribution of mismatched features

3.3 Hypothesis 2: emotion variation produces different transformations under different phonemes

Li *et al.* (2010) analyzed the acoustic and articulatory cues of Mandarin vowels by using EMA recordings and found that the articulation changes exhibit differences for different vowels under each emotional state. Lee *et al.* (2004) reached a similar conclusion by observing the distinct constellations: emotions can have different effects under different phonemes. In the light of previous work, we propose the second hypothesis—emotion variation produces different transformations under different phonemes. Thus, it is reasonable to construct a mismatched feature detector for each phoneme.

4 Mismatched feature detection under acoustic class level

Emotional variability causes different deformations of acoustic space for different phonemes. A phoneme is a sound or a group of different sounds perceived to have the same function in a language or dialect (Twaddell, 1935). An acoustic class is a slightly different concept which combines sounds with similar acoustic characteristics into a class. We propose three sorts of acoustic class: phoneme classes, GMM tokenizer, and probabilistic GMM tokenizer. Phoneme classes categorize several similar phonemes into a sort of acoustic class. GMM tokenizer and probabilistic GMM tokenizer categorize similar features under the same Gaussian component as a sort of acoustic class.

4.1 Mismatched feature detection based on phoneme classes (MDPC)

4.1.1 Phoneme classes

Each phoneme class has distinct acoustic properties due to different frequency bands, manners of articulation, and stationarity or non-stationarity of the vocal tract configuration. There are many classification methods. Lee *et al.* (2004) classified phonemes into five sorts: vowel, stop, glide, nasal, and fricative sounds. Bitouk *et al.* (2010) proposed three phoneme classes: stressed vowels, unstressed vowels, and consonants. Considering the specific characteristics of Mandarin, we chose a phoneme classification method with eight phoneme classes: affricate, voiced consonant, fricative, plosive, vowel, semivowel, nasal, and diphthongs. Each phoneme in Mandarin can be categorized into one of the phoneme classes illustrated in Table 2.

Table 2 Mandarin phonemes of each phoneme class

Class name	Phonemes
Affricate	j, q, zh, ch, z, c
Voiced consonant	m, n, l, r
Fricative	f, h, x, sh, s
Plosive	b, p, d, t, g, k
Vowel	a, o, e
Semivowel	i, u, ü
Nasal	an, in, un, en, ün, ang, ing, ong, eng, etc.
Diphthongs	ai, ei, ui, ao, ou, in, ie, üe, etc.

The phoneme class recognition method is implemented following Arslan and Hansen (1994), creating hidden Markov models (HMMs) for each phoneme class using a 3-state left-to-right HMM with five mixtures. The models are trained using the development corpus. The IR of phoneme class recognition can achieve 97.42%.

4.1.2 MDPC

For each phoneme class, a mismatched feature detection system based on SVM is developed. First, the phoneme class of each frame $\boldsymbol{x}_t$ under the development corpus is recognized. Then, under each phoneme class, the neutral features are regarded as the positive samples while the emotional features

are regarded as the negative ones. These features are used to train the SVM classifier to detect the mismatched features. In the evaluation stage, the matching score RS_t for each frame $\boldsymbol{x}_t$ is computed through Eq. (2). If RS_t is less than the threshold, it is regarded as a mismatched feature.

4.2 Mismatched feature detection based on GMM tokenizer (MDGT)

4.2.1 GMM tokenizer

GMM Tokenizer is another method for representing acoustic classes. It was first proposed for language identification based on GMM (Torres-Carrasquillo *et al.*, 1993).

GMM was first proposed by Reynolds and Rose (1995). Then it was applied to the speaker recognition area widely, and is now regarded as the most popular model to describe the distribution of speakers' acoustic features. GMM is represented by the weighted sum of N Gaussian distributions, as shown in Eq. (3). The universal background model (UBM) (Reynolds *et al.*, 2000) is a large-scale GMM representing the distribution of speaker-independent acoustic features.

$$p(\boldsymbol{x}|\lambda)=\sum_{k=1}^{N}\omega_k\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k), \tag{3}$$

where $\mathcal{N}$ is a Gaussian distribution with mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Sigma}_k$, ω_k is the weight of the kth Gaussian distribution, and N is the number of Gaussian components. Each Gaussian distribution represents the distribution of acoustic features of some phonetic event. The mean $\boldsymbol{\mu}_k$ and the variance $\boldsymbol{\Sigma}_k$ represent the mean and the variance of the acoustic distribution of the corresponding phonetic event. Therefore, the component index information can be used to express the phonetic information. The component with the maximum occupation probability is selected to represent each frame's phonetic event in the GMM tokenizer algorithm. The occupation probability $r_k(\boldsymbol{x}_t)$ of the feature $\boldsymbol{x}_t$ on the kth component is computed as

$$r_k(\boldsymbol{x}_t)=\frac{\omega_k\mathcal{N}(\boldsymbol{x}_t|\mu_k,\Sigma_k)}{\sum_{k'=1}^{N}\omega_{k'}\mathcal{N}(\boldsymbol{x}_t|\mu_{k'},\Sigma_{k'})}. \tag{4}$$

The advantage of the GMM tokenizer is that it does not need any phonetically labeled training speech and is less expensive in computation, compared with the phoneme recognition method. In our study, the idea of the GMM tokenizer is adopted to replace the role of a phoneme recognizer to determine the acoustic class of each frame.

4.2.2 MDGT

The procedure of the MDGT method is similar to that of MDPC except that the phoneme class of each frame is replaced by the tokenizer of each frame. Thus, the SVM classifier is trained under each UBM component.

4.3 Mismatched feature detection based on probabilistic GMM tokenizer (MDPGMT)

4.3.1 Probabilistic GMM tokenizer (PGMT)

GMM tokenizer is a 'hard' classification method because it must select a component as the tokenizer. It cannot reflect the probability generation property of a GMM. Thus, probabilistic GMM tokenizer is proposed. It employs the occupation probability vector to represent the acoustic class information of a frame, $\mathrm{PGMT}=[r_1(\boldsymbol{x}_t),r_2(\boldsymbol{x}_t),\ldots,r_N(\boldsymbol{x}_t)]$. PGMT does not categorize the frame into a certain component, but assumes that the feature is generated on all components with a certain probability.

4.3.2 MDPGMT

In the method PGMT, $\boldsymbol{x}_t$ belongs to the kth component with a probability $r_k(\boldsymbol{x}_t)$. Fuzzy SVM should be trained to classify the neutral and emotional features under the kth component. In our context, the fuzzy membership m_t means the probability that $\boldsymbol{x}_t$ belongs to the kth component, which is also the meaning of $r_k(\boldsymbol{x}_t)$. The range of the possible fuzzy membership is $[0,1]$, the same as the range of $r_k(\boldsymbol{x}_t)$. Thus, we use $r_k(\boldsymbol{x}_t)$ as the fuzzy membership of $\boldsymbol{x}_t$ under the kth component.

For MDPGMT, a fuzzy SVM-based mismatched feature detection model under each component is constructed. All the component-level fuzzy SVM classifiers comprise the detection model.

Each feature $\boldsymbol{x}_t$ under the kth component is represented by a triple $\{\boldsymbol{x}_t,y_t,r_k(\boldsymbol{x}_t)\}$. All these features in the development corpus with $r_k(\boldsymbol{x}_t)\geq 0.01$ are used to train the fuzzy SVM under the kth component.

In the evaluation process, the matching score RS_t for each feature $\boldsymbol{x}_t$ is calculated as follows:

1. Compute the occupation probability vector for each feature $\boldsymbol{x}_t$ under each component: $[r_1(\boldsymbol{x}_t), r_2(\boldsymbol{x}_t), \ldots, r_N(\boldsymbol{x}_t)]$.

2. Compute the matching score $\mathrm{RS}_k(\boldsymbol{x}_t)$ of the fuzzy SVM model of the kth component.

3. Combine all matching scores $[\mathrm{RS}_1(\boldsymbol{x}_t), \mathrm{RS}_2(\boldsymbol{x}_t), \cdots, \mathrm{RS}_N(\boldsymbol{x}_t)]$ into the total matching score RS_t, as shown in Eq. (5). $r_k(\boldsymbol{x}_t)$ is also used as the weight of the matching score under the kth component.

$$\mathrm{RS}_t = \sum_{k=1}^{N} r_k(\boldsymbol{x}_t)\mathrm{RS}_k(\boldsymbol{x}_t). \tag{5}$$

If the matching score RS_t of the feature $\boldsymbol{x}_t$ is less than the threshold, it is regarded as a mismatched feature.

4.4 Neutral-emotional discriminative power under acoustic class level

The discriminative power of MFCC for distinguishing neutral from emotional features is calculated and we compare the neutral-emotional discriminative power under two conditions. One is the total feature, i.e., the ordinary MFCC feature, which mixes the features under all acoustic classes together to train only one classifier. The other is the acoustic class feature, which constructs the classifier under each acoustic class using the features under the corresponding acoustic class.

We use the F-ratio to compare the discriminative power of the total feature with that of the feature under each acoustic class represented by the PGMT method. The F-ratio is computed as the ratio of the between-class distance and the within-class distance. There are two 'classes': neutral and emotional features. The larger the F-ratio, the stronger the discriminative power of the feature. The F-ratio of the feature under the kth acoustic class (i.e., the kth UBM component) is calculated as

$$d_k = \frac{(\overline{\boldsymbol{x}}_{\mathrm{n}}^k - \overline{\boldsymbol{x}}_{\mathrm{e}}^k)^{\mathrm{T}}(\overline{\boldsymbol{x}}_{\mathrm{n}}^k - \overline{\boldsymbol{x}}_{\mathrm{e}}^k)}{(\sigma_{\mathrm{n}}^k)^2 + (\sigma_{\mathrm{e}}^k)^2}, \tag{6}$$

where $\overline{\boldsymbol{x}}_{\mathrm{n}}^k = \sum_{t=1}^{T_{\mathrm{n}}} r_k(\boldsymbol{x}_t)\boldsymbol{x}_t / \sum_{t=1}^{T_{\mathrm{n}}} r_k(\boldsymbol{x}_t)$, $\boldsymbol{x}_t \in x_{\mathrm{n}}$ represents the mean of neutral features under the kth acoustic class, x_{n} is the neutral feature set, and T_{n} is the total number of neutral features. Correspondingly, $\overline{\boldsymbol{x}}_{\mathrm{e}}^k$ represents the mean of the emotional MFCC feature under the kth acoustic class. $(\sigma_{\mathrm{n}}^k)^2$ and $(\sigma_{\mathrm{e}}^k)^2$ are the variances of the neutral and emotional features respectively, under the kth acoustic class.

$$(\sigma_{\mathrm{n}}^k)^2 = \frac{\sum_{t=1}^{T_{\mathrm{n}}} r_k(\boldsymbol{x}_t)(\boldsymbol{x}_t - \overline{\boldsymbol{x}}_{\mathrm{n}}^k)^{\mathrm{T}}(\boldsymbol{x}_t - \overline{\boldsymbol{x}}_{\mathrm{n}}^k)}{\sum_{t=1}^{T_{\mathrm{n}}} r_k(\boldsymbol{x}_t)}, \ \boldsymbol{x}_t \in x_{\mathrm{n}}.$$

For the total feature, $\overline{\boldsymbol{x}}_{\mathrm{n}} = \frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}} \boldsymbol{x}_t$, $\sigma_{\mathrm{n}}^2 = \frac{1}{T_{\mathrm{n}}}\sum_{t=1}^{T_{\mathrm{n}}}(\boldsymbol{x}_t - \overline{\boldsymbol{x}}_{\mathrm{n}})^{\mathrm{T}}(\boldsymbol{x}_t - \overline{\boldsymbol{x}}_{\mathrm{n}})$, and $d = (\overline{\boldsymbol{x}}_{\mathrm{n}} - \overline{\boldsymbol{x}}_{\mathrm{e}})^{\mathrm{T}}(\overline{\boldsymbol{x}}_{\mathrm{n}} - \overline{\boldsymbol{x}}_{\mathrm{e}})/(\sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{e}}^2)$.

F-ratio values of the total feature and features under each acoustic class were calculated and plotted (Fig. 3). The features under each acoustic class have more discriminative power than the total features. Thus, it is reasonable to construct the mismatched feature detector under each acoustic class.

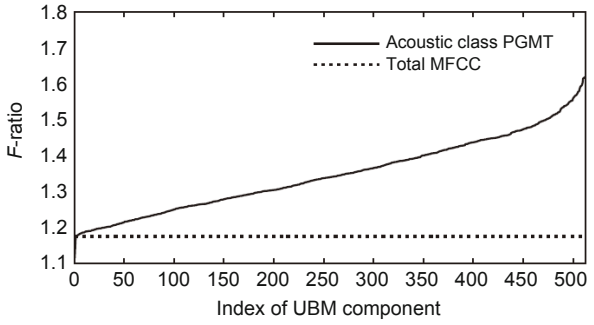

Fig. 3 Comparison of *F*-ratios between the total features and the acoustic class level features using PGMT. The UBM was trained by all the neutral and panic utterances of the development corpus and the Gaussian order was set to 512. The *F*-ratio was calculated between neutral and panic features

5 Mismatched feature regulation

The purpose of feature regulation is to retain the speaker's characteristics, and eliminate the negative effects caused by emotional variability. In other words, the mismatched feature needs to be converted to obey the distribution of the target speaker's neutral features, and be far away from the distribution of imposters' neutral features. As mismatched feature detection, feature regulation is processed under each acoustic class to improve the performance of feature regulation. We will introduce our feature regulation

method on the basis of acoustic class represented by PGMT.

Suppose there are $T_{s,\mathrm{n}}$ neutral features of the sth speaker and $T_{s,\mathrm{e}}$ emotional features of the sth speaker. The tth feature of the sth speaker is denoted as $\boldsymbol{x}_{s,t}$ and the mean of the neutral features under the kth acoustic class is represented as

$$\overline{\boldsymbol{x}}_s^k = \frac{\sum_{t=1}^{T_{s,\mathrm{n}}} r_k(\boldsymbol{x}_t)\boldsymbol{x}_{s,t}}{\sum_{t=1}^{T_{s,\mathrm{n}}} r_k(\boldsymbol{x}_t)}, \quad \boldsymbol{x}_{s,t} \in x_{\mathrm{n}}.$$

The transformation matrix under the kth acoustic class $\boldsymbol{A}_k$ is to be built and the mismatched feature $\boldsymbol{x}_{s,t}$ is transformed into the matched feature $\boldsymbol{A}_k\boldsymbol{x}_{s,t}$. The distance between the transformed features and the target speaker's neutral features is the within-class distance

$$D_{\mathrm{w},k} = \sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}} r_k(\boldsymbol{x}_{s,t})(\boldsymbol{A}_k\boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}_s^k)^{\mathrm{T}} \cdot (\boldsymbol{A}_k\boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}_s^k), \quad \boldsymbol{x}_{s,t} \in x_{\mathrm{e}},$$

where S denotes the total number of speakers in the development corpus and x_{e} is the emotional feature set. The distance between $\boldsymbol{A}_k\boldsymbol{x}_{s,t}$ and the imposter speakers of the development corpus is labeled as the between-class distance

$$D_{\mathrm{b},k} = \frac{1}{S-1}\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}}\sum_{\hat{s}=1,\hat{s}\neq s}^{S} r_k(\boldsymbol{x}_{s,t})(\boldsymbol{A}_k\boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}_{\hat{s}}^k)^{\mathrm{T}} \cdot (\boldsymbol{A}_k\boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}_{\hat{s}}^k), \quad \boldsymbol{x}_{s,t} \in x_{\mathrm{e}}.$$

To satisfy the requirement of the transformation matrix (maximizing the between-class distance and minimizing the within-class distance) mentioned above, an optimization equation is proposed.

$$\min \Phi_k, \quad \Phi_k = D_{\mathrm{w},k} - \alpha D_{\mathrm{b},k},$$

where α is the relax coefficient to adjust the weight between maximizing the between-class distance and minimizing the within-class distance. According to $\partial\Phi_k/\partial\boldsymbol{A}_k = \boldsymbol{0}$, we can obtain

$$\boldsymbol{A}_k\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}} r_k(\boldsymbol{x}_{s,t})(1-\alpha)\boldsymbol{x}_{s,t}\boldsymbol{x}_{s,t}^{\mathrm{T}} = \sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}} r_k(\boldsymbol{x}_{s,t})(\overline{\boldsymbol{x}}_s^k\boldsymbol{x}_{s,t}^{\mathrm{T}} - \frac{\alpha}{S-1}\sum_{\hat{s}=1,\hat{s}\neq s}^{S}\overline{\boldsymbol{x}}_{\hat{s}}^k\boldsymbol{x}_{s,t}^{\mathrm{T}}). \tag{7}$$

The deduction is presented in the Appendix. Set

$$\boldsymbol{B}_k = \sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}} r_k(\boldsymbol{x}_t)(1-\alpha)\boldsymbol{x}_{s,t}\boldsymbol{x}_{s,t}^{\mathrm{T}}.$$

$$\boldsymbol{C}_k = \sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}} r_k(\boldsymbol{x}_{s,t})(\overline{\boldsymbol{x}}_s^k\boldsymbol{x}_{s,t}^{\mathrm{T}} - \frac{\alpha}{S-1}\sum_{\hat{s}=1,\hat{s}\neq s}^{S}\overline{\boldsymbol{x}}_{\hat{s}}^k\boldsymbol{x}_{s,t}^{\mathrm{T}}).$$

Then, the transformation matrix of the kth acoustic class is

$$\boldsymbol{A}_k = \boldsymbol{C}_k\boldsymbol{B}_k^{-1}.$$

All the transformation matrices $\boldsymbol{A}_k$ comprise the entire feature regulation model. Only the mismatched features should be regulated. The feature regulation procedure in the evaluation corpus is similar to the mismatched feature detection procedure:

1. Compute the occupation probability vector for each feature $\boldsymbol{x}_t$ under each acoustic class: $[r_1(\boldsymbol{x}_t), r_2(\boldsymbol{x}_t), \ldots, r_N(\boldsymbol{x}_t)]$.

2. Regulate the mismatched features under each acoustic class:

$$\boldsymbol{x}'_{t,k} = \boldsymbol{A}_k\boldsymbol{x}_t. \tag{8}$$

3. Add all the regulated features in each acoustic class into the regulated feature $\boldsymbol{x}'_t$. $r_k(\boldsymbol{x}_t)$ is used as the weight.

$$\boldsymbol{x}'_t = \sum_{k=1}^{N} r_k(\boldsymbol{x}_t)\boldsymbol{x}'_{t,k}.$$

Similarly, feature regulation methods based on phoneme classes and GMM tokenizer are evaluated and $r_k(\boldsymbol{x}_t)$ is set to 1 for each feature under each phoneme class or UBM component.

6 Algorithm framework

The framework of our algorithm is illustrated in Fig. 4. The mismatched feature detection module is trained according to the procedure mentioned in Section 4. The feature regulation matrices are trained using the procedure introduced in Section 5.

In the enrollment stage, a GMM for each speaker is adapted from the UBM through his/her neutral speech. In the test stage, MFCC features $[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T]$ are extracted and their acoustic classes are determined. Then, the matching score RS_t of each feature is computed. If RS_t is greater than the threshold δ, the feature is regarded as a

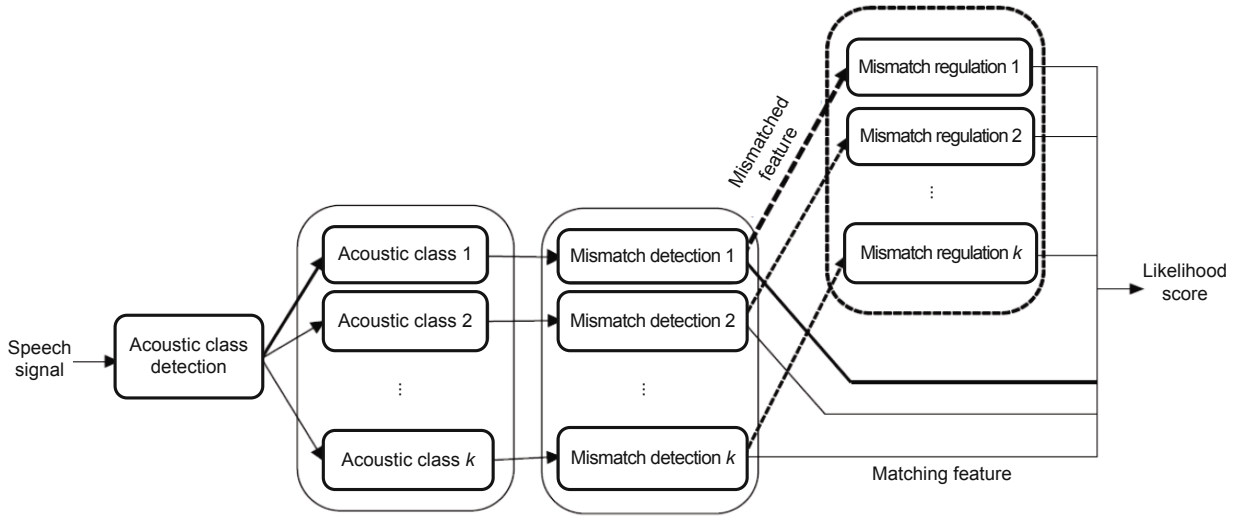

Fig. 4 Framework of our algorithms

matched feature. Otherwise, it is regarded as a mismatched feature.

A matched feature can be directly used to compute the likelihood score for the target speaker and imposter speakers. Techniques are needed to deal with mismatched features. In our study, we use feature pruning and feature regulation methods. A feature pruning method simply removes the mismatched features. The score computing formula for our feature pruning method against the sth speaker is

$$\text{Score}(s) = \sum_{t=1}^{T} w_t p(\boldsymbol{x}_t|\lambda_s),$$

where

$$w_t = \begin{cases} 1, & \text{RS}_t \geq \delta, \\ 0, & \text{RS}_t < \delta. \end{cases}$$

As for our feature regulation method, the mismatched features are regulated so that they obey the distribution of matched features. Then the likelihood score for the regulated feature is computed:

$$\text{Score}(s) = \sum_{t=1}^{T} w_{t,1} p(\boldsymbol{x}_t|\lambda_s) + w_{t,2} p(\boldsymbol{x}'_t|\lambda_s),$$
$$\begin{cases} w_{t,1} = 1, w_{t,2} = 0, & \text{RS}_t \geq \delta, \\ w_{t,1} = 0, w_{t,2} = 1, & \text{RS}_t < \delta. \end{cases} \quad (9)$$

Here, $\boldsymbol{x}_t$ is the original feature and $\boldsymbol{x}'_t$ is the regulated feature.

7 Experimental results

7.1 Corpus

The development and evaluation corpuses of our experiments were subsets of MASC. This corpus can be obtained from the Linguistic Data Consortium (LDC). MASC contains 68 native Mandarin speakers, 45 of whom are male and 23 female. Each speaker utters 2 paragraphs (each lasting about 30 s) in neutral, 5 phrases (each lasting about 1 s), and 20 utterances (each lasting about 3–10 s) under five emotional states (neutral, anger, elation, panic, and sadness) for three times. The speech is recorded with the same device in the same peaceful environment. The speech content covers the whole set of vowels and consonants in Mandarin.

7.2 Experimental protocol

In our experiments, the first 18 speakers were selected as the development corpus, and the remaining speakers were used as the evaluation corpus. Among the evaluation data, the two neutral paragraphs were used to adapt the speaker's GMM and the utterances under all emotional states were used to evaluate the performance of our algorithms.

The speech was pre-emphasized and framed with a 16 ms Hamming window. A 13-order MFCC feature plus delta was extracted to train the UBM, adapt the GMM, and compute the likelihood score. The neutral paragraphs and utterances of the first

18 speakers were used to train the UBM. The total length of the development corpus was 3081.1 s.

The UBM was firstly trained with the expectation maximization (EM) algorithm, and then the speaker model GMM was adapted from the UBM via the maximum a posterior (MAP) algorithm. The number of Gaussian components of the UBM was set to 512. IR was used to evaluate the performance of our algorithms.

In our corpus, the neutral features and the sadness features were similar (Huang and Yang, 2010). So, these two emotions were labeled as 'neutral' and the other three types as 'emotional'.

7.3 Experimental results

7.3.1 Feature pruning method

Fig. 5 indicates the performance of our feature pruning method when the threshold δ was set within the range $[-0.4, 0.4]$ (the effect on IRs was small when $\delta < -0.4$, and the IRs dropped dramatically when $\delta > 0.4$). We can draw three conclusions from Fig. 5. First, the IRs of our feature pruning method under all emotional states except neutral increased when a fraction of the features were pruned. However, when the number of pruned features increased, the performance deteriorated. Second, the trend in the variation of average IR with different numbers of pruned features was similar to that of the IR under each emotional state. Third, the average IR achieved a peak value of 55.19% when the threshold δ was set to -0.19.

When the threshold δ was set to -0.19, the distribution of z-scores for the pruned features was as shown in Fig. 6. We computed all the z-scores for the matched features and detected mismatched features of the first speaker's vowel 'a'. The mismatched features detected usually had smaller z-score values. This phenomenon illustrates the effectiveness of our pruning method.

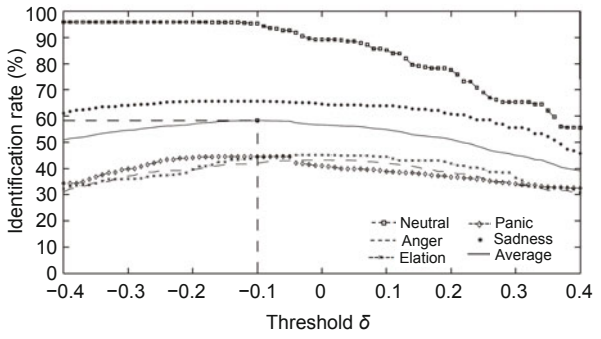

Fig. 5 Relationship between IR and the threshold δ of the feature pruning method

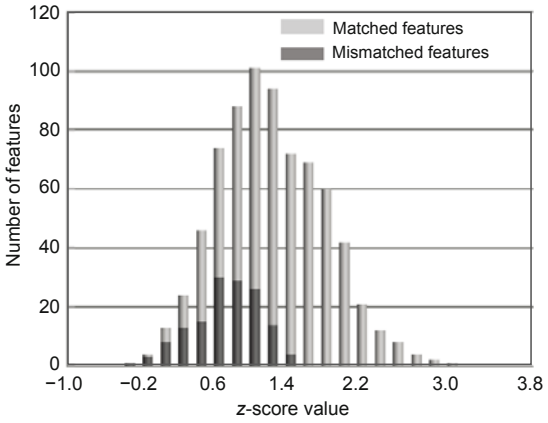

Fig. 6 Distribution of the z-score values of the detected mismatched features with threshold $\delta = -0.19$

Table 3 compares the IR performance of our method with the PDDM (Huang and Yang, 2010) and the baseline GMM-UBM algorithms. The IR increases achieved using MDPC, MDGT, and MDPGMT were 0.54%, 2.42%, and 3.64%, respectively, compared with GMM-UBM. These three methods performed better than PDDM because MFCC contains more information than pitch. The dimension of MFCC is 26 whereas that of pitch is only one. So, using MFCC to distinguish mismatched features is more effective. MDGT and MDPGMT are better than MDPC because their acoustic class numbers are higher. MDGT and MDPGMT have 512 acoustic classes whereas MDPC has only eight. Therefore, MDGT and MDPGMT are better. MDPGMT is much better than MDGT because a 'hard' categorization method will cause information loss.

Table 3 Comparison between the baseline algorithm, the PDDM (Huang and Yang, 2010) method, and our pruning methods

Method	Identification rate (%)					
	Neutral	Anger	Elation	Panic	Sadness	Average
GMM-UBM	96.23	31.50	33.57	35.00	61.43	51.55
PDDM	95.20	34.33	34.47	35.10	59.83	51.79
MDPC	94.73	35.03	35.20	35.90	59.60	52.09
MDGT	95.07	36.27	40.93	37.13	60.47	53.97
MDPGMT	95.50	37.60	39.47	39.77	63.63	**55.19**

7.3.2 Feature regulation method

In the feature regulation method, two parameters affect the performance of our methods. One is the threshold δ, which decides whether a feature is matched or mismatched. The other is the relax coefficient α, which determines the weight for maximizing the between-class distance and minimizing the within-class distance. The threshold δ was fixed to -0.19 as before. The relax coefficient α was determined through the following experiment. Table 4 shows the IR performance of the feature regulation method under different α values.

Table 4 Relationship between IR and the relax coefficient α

α	IR (%)	α	IR (%)	α	IR (%)
1/10	55.23	1/2	56.41	3	**57.28**
1/5	55.52	1	56.74	5	57.09
1/3	56.03	2	57.06	10	56.44

According to the data in Table 4 and other data not listed, a quadratic curve is suitable to simulate the relationship between IR and α (Fig. 7).

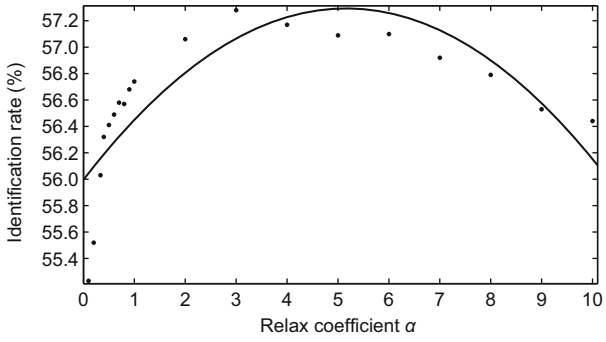

Fig. 7 Quadratic curve fitting the relationship between IR and the relax coefficient α

The IR increases first and then decreases with the increase in α. When α is small, it means the weight for maximizing the between-class distance is less than the weight for minimizing the within-class distance. The transformed features can obey the distribution of neutral features, but cannot retain the speaker's characteristics well. When the value of α is large, the transformed features can retain the speaker's characteristics but cannot fit the distribution of neutral features well. In other words, in extreme conditions (when the value of α is too large or too small), our methods will not perform well. When α is set around 3, the method achieves its peak because maximizing the between-class distance and minimizing the within-class distance achieve a balance.

After the relax coefficient was fixed, the threshold was varied from -0.4 to 0.4 to select the best threshold (Fig. 8).

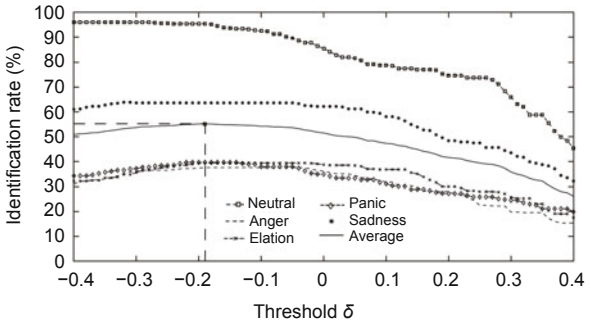

Fig. 8 The relationship between the IR and the threshold δ under the feature regulation method

The performance of the feature regulation method peaked when the threshold was set to -0.1, and the IR variation trend was similar to that shown in Fig. 5.

Then, we compared our feature regulation method with the baseline system and the pitch modification method (Huang and Yang, 2008) based on PSOLA. As shown in Table 5, our feature regulation method can enhance the ASR performance. The average IR increases achieved by our methods under phoneme classes, GMM tokenizer, and PGMT acoustic class levels, were 3.30%, 4.22%, and 6.77%, respectively, compared with that of the baseline GMM-UBM algorithm. Our methods also performed better than the pitch modification method (Huang and Yang, 2010).

Table 5 Comparison between the baseline algorithm, pitch modifying method (Huang and Yang, 2008), and our feature regulation methods

Method	Identification rate (%)					
	Neutral	Anger	Elation	Panic	Sadness	Average
GMM-UBM	96.23	31.50	33.57	35.00	61.43	51.55
Pitch-mod	95.90	37.90	39.53	38.56	60.83	54.54
MDPC	94.80	38.73	41.20	39.20	60.33	54.85
MDGT	95.37	38.40	43.20	40.07	61.80	55.77
MDPGMT	95.20	41.93	44.33	44.50	65.63	**58.32**

Pitch-mod: pitch modification

7.3.3 Cross validation

Sets of cross validation experiments were also conducted to verify the robustness of our feature regulation method, i.e., MDPGMT. We conducted four cross validation experiments:

(I) The first 18 speakers were selected as the development corpus and the remaining speakers as the evaluation corpus (the result of this experiment is shown above).

(II) Speakers 18–35 were selected as the development corpus and the remaining speakers as the evaluation corpus.

(III) Speakers 35–52 were selected as the development corpus and the remaining speakers as the evaluation corpus.

(IV) Speakers 51–68 were selected as the development corpus and the remaining speakers as the evaluation corpus.

The experimental results are shown in Table 6. The IR performances of the baseline GMM-UBM algorithm on different subsets were different. However, it is clear that our method can enhance the performance under different conditions: the average increase in the IR varied between 6% and 8%. This phenomenon indicates the robustness of our method.

7.3.4 Model training with different speech contents

In the MASC corpus, the utterances of each speaker have the same text under different emotional states. Actually, our methods can train the mismatched feature detection and regulation model without this restriction. We select the first 10 sentences under neutral and sadness for neutral features and the last 10 sentences under anger, elation, and panic for emotional features. Thus, the speeches of development data under neutral and emotional states are different. These features are used to train the fuzzy SVM model and feature transformation matrix $\boldsymbol{A}_k$.

IRs with different speech contents drop slightly, compared with those with the same speech content. But IRs drop by only 0.13% for feature pruning (from 55.19% at Table 3 to 55.06% at Table 7) and 0.22% for feature regulation (from 58.32% at Table 5 to 58.10% at Table 7). So, our methods also work for development data with different speech contents.

7.3.5 Performance on i-vector

The current state-of-the-art speaker recognition algorithm is the i-vector (Dehak *et al.*, 2011). We measured the performance of our feature pruning and regulation methods on i-vector. In our experiments, the total variability matrix was trained by the first 18 speakers. The dimension of the i-vector was set to 300. The feature pruning and regulation methods were based on acoustic classes represented by PGMT and we trained the i-vector with the pruned or regulated features for test utterance. The conventional i-vector with linear discriminant analysis (LDA) + within class covariance normalization (WCCN) compensation technique was applied for comparison with our methods. The IRs of these algorithms are listed in Table 8.

As Table 8 shows, training the i-vector with pruned or regulated features can enhance the performance of the conventional i-vector. The IR increases were 2.09% for the feature pruning method and 3.32% for the feature regulation method, compared with that of the baseline i-vector algorithm.

8 Conclusions

A speaker's varying emotional states will cause the degradation of an ASR system's performance due to mismatched features. Therefore, we propose a mismatched feature detection method based on three sorts of acoustic class: phoneme classes, GMM tokenizer, or PGMT. Feature pruning and feature reg-

Table 6 Results of cross-validation experiments

Experiment	Identification rate (%)					
	Neutral	Anger	Elation	Panic	Sadness	Average
II	96.53 (−0.87)	23.13 (11.33)	25.00 (11.80)	24.93 (7.47)	56.80 (3.67)	**45.28 (6.68)**
III	96.47 (−0.53)	35.00 (10.83)	31.80 (12.60)	29.60 (8.73)	60.40 (4.80)	**50.65 (7.29)**
IV	97.13 (−1.13)	32.53 (7.87)	34.13 (11.33)	29.27 (8.93)	54.47 (5.07)	**49.91 (6.41)**

The data outside the brackets are the IRs of the baseline system, and the data inside the brackets are the IR increase achieved by our feature regulation method under the PGMT acoustic class level

ulation methods are presented to process the mismatched features. The experiments conducted on MASC showed that these two methods can effectively reduce the negative effects caused by emotional variability. Compared with the baseline GMM-UBM algorithm, the IR increased by 3.64% using the feature pruning method, and by 6.77% using the feature regulation method. So, our feature regulation method can reduce emotional effects more effectively. In this study, the emotional state was simulated and usually exaggerated. The feature mismatch was significant, and our algorithm achieved a high IR increase. In the future, we will focus on applying our algorithms to the problem of gentle emotional variation in real-life situations.

Table 7 Results for development data with different speech contents

Method	Identification rate (%)					
	Neutral	Anger	Elation	Panic	Sadness	Average
Feature-P	95.47	37.33	39.27	39.63	63.60	**55.06**
Feature-R	95.03	41.67	44.10	44.37	65.33	**58.10**

Feature-P: feature pruning; Feature-R: feature regulation

Table 8 Algorithm performance on i-vector

Method	Identification rate (%)					
	Neutral	Anger	Elation	Panic	Sadness	Average
i-vector	95.53	43.37	44.17	50.53	62.60	59.24
Feature-P	95.07	46.47	46.73	53.13	65.27	61.33
Feature-R	95.40	47.87	47.60	55.80	66.13	**62.56**

Feature-P: feature pruning; Feature-R: feature regulation

References

Arslan, L.M., Hansen, J.H.L., 1994. Minimum cost based phoneme class detection for improved iterative speech enhancement. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.45-48. [doi:10.1109/ICASSP.1994.389722]

Balasubramanian, M., Schwartz, E.L., 2002. The isomap algorithm and topological stability. *Science*, **295**(5552):7. [doi:10.1126/science.295.5552.7a]

Bao, H.J., Xu, M.X., Zheng, T.F., 2007. Emotion attribute projection for speaker recognition on emotional speech. Proc. 8th Annual Conf. of the Int. Speech Communication Association, p.601-604.

Bitouk, D., Verma, R., Nenkova, A., 2010. Class-level spectral features for emotion recognition. *Speech Commun.*, **52**(7-8):613-625. [doi:10.1016/j.specom.2010.02.010]

Brady, M.C., 2005. Synthesizing affect with an analog vocal tract: glottal source. Toward Social Mechanisms of Android Science: a CogSci Workshop, p.45-49.

Chen, L., Yang, Y.C., Yao, M., 2011. Reliability detection by fuzzy SVM with UBM component feature for emotional speaker recognition. Proc. 8th Int. Conf. on Fuzzy Systems and Knowledge Discovery, p.458-461. [doi:10.1109/FSKD.2011.6019484]

Cowie, R., Cornelius, R.R., 2003. Describing the emotional states that are expressed in speech. *Speech Commun.*, **40**(1-2):5-32. [doi:10.1016/S0167-6393(02)00071-7]

Dehak, N., Kenny, P., Dehak, R., *et al.*, 2011. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.*, **19**(4):788-798. [doi:10.1109/tasl.2010.2064307]

Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.121-124. [doi:10.1109/ICASSP.1998.674382]

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: features, classification schemes, and databases. *Patt. Recog.*, **44**(3):572-587. [doi:10.1016/j.patcog.2010.09.020]

Gadek, J., 2009. Influence of upper respiratory system disease on the performance of automatic voice recognition systems. *Comput. Med. Act.*, **65**:211-221. [doi:10.1007/978-3-642-04462-5_21]

Ghiurcau, M.V., Rusu, C., Astola, J., 2011a. A study of the effect of emotional state upon text-independent speaker identification. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.4944-4947. [doi:10.1109/ICASSP.2011.5947465]

Ghiurcau, M.V., Rusu, C., Astola, J., 2011b. Speaker recognition in an emotional environment. Proc. Signal Processing and Applied Mathematics for Electronics and Communications, p.81-84.

Huang, T., Yang, Y.C., 2008. Applying pitch-dependent difference detection and modification to emotional speaker recognition. Proc. 9th Annual Conf. of the Int. Speech Communication Association, p.2751-2754.

Huang, T., Yang, Y.C., 2010. Learning virtual HD model for bi-model emotional speaker recognition. Proc. 20th Int. Conf. on Pattern Recognition, p.1614-1617. [doi:10.1109/ICPR.2010.399]

Jawarkar, N.P., Holambe, R.S., Basu, T.K., 2012. Text-independent speaker identification in emotional environments: a classifier fusion approach. *Front. Comput. Educ.*, **133**:569-576. [doi:10.1007/978-3-642-27552-4_77]

Jin, Q., Schultz, T., Waibel, A., 2007. Far-field speaker recognition. *IEEE Trans. Audio Speech Lang. Process.*, **15**(7):2023-2032. [doi:10.1109/tasl.2007.902876]

Kelly, F., Harte, N., 2011. Effects of long-term ageing on speaker verification. Proc. European Workshop on Biometrics and ID Management, p.113-124. [doi:10.1007/978-3-642-19530-3_11]

Lee, C.M., Yildirim, S., Bulut, M., *et al.*, 2004. Effects of emotion on different phoneme classes. *J. Acoust. Soc. Am.*, **116**:2481. [doi:10.1121/1.4784911]

Li, A., Fang, Q., Hu, F., *et al.*, 2010. Acoustic and articulatory analysis on Mandarin Chinese vowels in emotional speech. Proc. 7th Int. Symp. on Chinese Spoken Language Processing, p.38-43. [doi:10.1109/ISCSLP.2010.5684866]

Lin, C.F., Wang, S.D., 2002. Fuzzy support vector machines. *IEEE Trans. Neur. Netw.*, **13**(2):464-471. [doi:10.1109/72.991432]

Platt, J.C., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classifiers*, **10**(3):61-74.

Reynolds, D.A., 2003. Channel robust speaker verification via feature mapping. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.53-56. [doi:10.1109/ICASSP.2003.1202292]

Reynolds, D.A., Rose, R.C., 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, **3**(1):72-83. [doi:10.1109/89.365379]

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted Gaussian mixture models. *Digit. Signal Process.*, **10**(1-3):19-41. [doi:10.1006/dspr.1999.0361]

Rose, R.C., Hofstetter, E.M., Reynolds, D.A., 1994. Integrated models of signal and background with application to speaker identification in noise. *IEEE Trans. Speech Audio Process.*, **2**(2):245-257. [doi:10.1109/89.279273]

Scherer, K., Johnstone, T., Bänziger, T., 1998. Automatic verification of emotionally stressed speakers: the problem of individual differences. Proc. Int. Conf. on Speech and Computer, p.233-238.

Shahin, I., 2013. Speaker identification in emotional talking environments based on CSPHMM2s. *Eng. Appl. Artif. Intell.*, **26**(7):1652-1659. [doi:10.1016/j.engappai.2013.03.013]

Shan, Z.Y., Yang, Y.C., 2008. Learning polynomial function based neutral-emotion GMM transformation for emotional speaker recognition. Proc. 19th Int. Conf. on Pattern Recognition, p.1-4. [doi:10.1109/icpr.2008.4761647]

Shan, Z.Y., Yang, Y.C., Ye, R.Z., 2007. Natural-emotion GMM transformation algorithm for emotional speaker recognition. Proc. 8th Annual Conf. of the Int. Speech Communication Association, p.782-785.

Shriberg, E., Graciarena, M., Bratt, H., *et al.*, 2008. Effects of vocal effort and speaking style on text-independent speaker verification. Proc. 9th Annual Conf. of the Int. Speech Communication Association, p.609-612.

Torres-Carrasquillo, P.A., Reynolds, D.A., Deller, J.R.Jr., 1993. Language identification using Gaussian mixture model tokenization. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, p.757-760. [doi:10.1109/ICASSP.2002.5743828]

Triefenbach, F., Jalalvand, A., Schrauwen, B., *et al.*, 2010. Phoneme recognition with large hierarchical reservoirs. Proc. 24th Annual Conf. on Neural Information Processing Systems, p.2307-2315.

Twaddell, W.F., 1935. On defining the phoneme. *Language*, **11**(1):5-62. [doi:10.2307/522070]

Yang, Y.C., Chen, L., 2012. Toward emotional speaker recognition: framework and preliminary results. Proc. 7th Chinese Conf. on Biometric Recognition, p.235-242. [doi:10.1007/978-3-642-35136-5_29]

Appendix: Deduction of Eq. (7)

$$\begin{aligned}\Phi_k =&D_{\mathrm{w},k}-\alpha D_{\mathrm{b},k}\\ =&\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}} r_k(\boldsymbol{x}_{s,t})\Big[(\boldsymbol{A}_k\boldsymbol{x}_{s,t}-\overline{\boldsymbol{x}}_s^k)^{\mathrm{T}}(\boldsymbol{A}_k\boldsymbol{x}_{s,t}-\overline{\boldsymbol{x}}_s^k)\\ &-\frac{\alpha}{S-1}\sum_{\widehat{s}=1,\widehat{s}\neq s}^{S}(\boldsymbol{A}_k\boldsymbol{x}_{s,t}-\overline{\boldsymbol{x}}_{\widehat{s}}^k)^{\mathrm{T}}(\boldsymbol{A}_k\boldsymbol{x}_{s,t}-\overline{\boldsymbol{x}}_{\widehat{s}}^k)\Big].\end{aligned}$$

$$\begin{aligned}&\frac{\partial\Phi_k}{\partial\boldsymbol{A}_k}=\\ &\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}}2r_k(\boldsymbol{x}_{s,t})\Big[(\boldsymbol{A}_k\boldsymbol{x}_{s,t}-\overline{\boldsymbol{x}}_s^k)\boldsymbol{x}_{s,t}^{\mathrm{T}}\\ &-\frac{\alpha}{S-1}\sum_{\widehat{s}=1,\widehat{s}\neq s}^{S}(\boldsymbol{A}_k\boldsymbol{x}_{s,t}-\overline{\boldsymbol{x}}_{\widehat{s}}^k)\boldsymbol{x}_{s,t}^{\mathrm{T}}\Big]\\ =&\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}}2r_k(\boldsymbol{x}_{s,t})\Big[(\boldsymbol{A}_k\boldsymbol{x}_{s,t}\boldsymbol{x}_{s,t}^{\mathrm{T}}-\overline{\boldsymbol{x}}_s^k\boldsymbol{x}_{s,t}^{\mathrm{T}})\\ &-\frac{\alpha}{S-1}\sum_{\widehat{s}=1,\widehat{s}\neq s}^{S}(\boldsymbol{A}_k\boldsymbol{x}_{s,t}\boldsymbol{x}_{s,t}^{\mathrm{T}}-\overline{\boldsymbol{x}}_{\widehat{s}}^k\boldsymbol{x}_{s,t}^{\mathrm{T}})\Big]\\ =&\boldsymbol{A}_k\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}}2r_k(\boldsymbol{x}_{s,t})\big[(1-\alpha)(\boldsymbol{x}_{s,t}\boldsymbol{x}_{s,t}^{\mathrm{T}})\big]\\ &-\sum_{s=1}^{S}\sum_{t=1}^{T_{s,\mathrm{e}}}2r_k(\boldsymbol{x}_{s,t})\left(\overline{\boldsymbol{x}}_s^k\boldsymbol{x}_{s,t}^{\mathrm{T}}-\frac{\alpha}{S-1}\sum_{\widehat{s}=1,\widehat{s}\neq s}^{S}\overline{\boldsymbol{x}}_{\widehat{s}}^k\boldsymbol{x}_{s,t}^{\mathrm{T}}\right)\\ &\frac{\alpha}{S-1}\sum_{\widehat{s}=1,\widehat{s}\neq s}^{S}(A_kx_{s,t}x_{s,t}^{\mathrm{T}})=\alpha(A_kx_{s,t}x_{s,t}^{\mathrm{T}})\end{aligned}$$

Eq. (7) can be obtained by setting $\partial\Phi_k/\partial\boldsymbol{A}_k=\boldsymbol{0}$ and transposing the terms of the deduced equation.