

Research Article

Cost-Sensitive Learning for Emotion Robust Speaker Recognition

Dongdong Li,¹ Yingchun Yang,² and Weihui Dai³

¹ School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China

² Department of Computer Science and Technology, Zhejiang University, No. 38, Yuquan Road, Zhejiang 310027, China

³ School of Management, Fudan University, No. 220, Handan Road, Shanghai 200433, China

Correspondence should be addressed to Dongdong Li; ldd@ecust.edu.cn and Weihui Dai; whdai@fudan.edu.cn

Received 19 April 2014; Accepted 22 April 2014; Published 4 June 2014

Academic Editor: Yu-Bo Yuan

Copyright © 2014 Dongdong Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the field of information security, voice is one of the most important parts in biometrics. Especially, with the development of voice communication through the Internet or telephone system, huge voice data resources are accessed. In speaker recognition, voiceprint can be applied as the unique password for the user to prove his/her identity. However, speech with various emotions can cause an unacceptably high error rate and aggravate the performance of speaker recognition system. This paper deals with this problem by introducing a cost-sensitive learning technology to reweight the probability of test affective utterances in the pitch envelop level, which can enhance the robustness in emotion-dependent speaker recognition effectively. Based on that technology, a new architecture of recognition system as well as its components is proposed in this paper. The experiment conducted on the Mandarin Affective Speech Corpus shows that an improvement of 8% identification rate over the traditional speaker recognition is achieved.

1. Introduction

Biometric security systems are based on human exclusive and unique characteristics, such as fingerprints, face, voice, iris, and retina [1, 2]. These systems are used as an extra barrier to prevent unauthorized access to protect data by recognizing the users by their specific physiological or behavioral characteristic. This method is more reliable than the conventional method because it is based on “something one is” rather than “something one knows/has.”

With improved research of vocal signals, people’s interactions through internet and mobile devices, such as phone banking, internet browsing, and secured information retrieval by voice, are becoming popular in a very rapid way [3]. There exists a need for greater security as these human-machine interactions over telephone lines and internet. At the same time, the powerful and ubiquitous handheld devices, such as smart phones and handheld computers, may contain a myriad of sensitive or personal information. All the applications mentioned above put great demand on speaker recognition based on modeling the speaker vocal tract characteristics, providing secure access to financial information

(e.g., credit card information, bank account balance, etc.) or other sensitive customer information (e.g., healthcare records) [4]. Speaker verification provides an extra barrier to prevent unauthorized access to protect data and enhances the security offered by personal identification numbers or user selected passwords. It allows for contactless activation and mitigates the risks of stolen or lost keys, passwords, or keycards.

Automatic speaker recognition can be grouped into the following two classes: speaker verification and speaker identification. Speaker verification is the process to confirm the claim of identity and declare the person to be true or imposter. It is inclined to be used in security system using user specified passcodes for secure user logins. Speaker identification is the process to determine which one best matches the input voice sample from a pool of speakers’ voices. Its main application area is forensics and investigation, where there is a need to determine the identifier of a person. According to the type of spoken utterances, speaker recognition can also be divided into three categories: text-independent, text-dependent, and text-prompted. In text-independent systems,

an arbitrary phrase is uttered to recognize the speaker. In text-dependent systems, a fixed “voice password” is uttered. In “text-prompted” systems, an instruction is given to ask the speaker to utter a certain phrase.

Previous work on security-based speaker recognition systems largely falls within the domain of dealing with interspeaker variables, like channel and background noise. However, intraspeaker variables, like emotion and health state, can also cause an unacceptably high error rate, which limits the commercial viability of speaker recognition systems [5]. Driven by rapid ongoing advances in affective computing, speaker recognition with affective speech (SRAS) is now becoming a popular consideration of modern speaker recognition research. In real life, we cannot expect the speaker to be always in neutral or normal mood. Most of the speaker recognition systems enroll the speaker model with neutral speech. Such systems could distinguish speakers from the others accurately when the speaker provides neutral speech to identify. However, when the recognition step is faced with emotional speech, like angry speech or delighted speech, the systems suffer emotional state mismatch between training and testing stage and the performance deteriorates. We cannot afford to develop the speaker models in all possible emotions for improving the performance, which degrades the user-friendliness of the speaker recognition system. SRAS becomes important because of the difficulty in acquiring large number of affective speeches from the speakers.

In sophisticated human-machine interaction, equipping the computer with the affective computing ability so that it can recognize the identity of the user is urgently needed in many different applications [6]. In telecommunications, the telephone-based speech recognition performance can be enhanced with the SRAS systems. For example, in route emergency call applications which service for high priority emergency calls, the callers experience a panic and scary scene. Their voice is not neutral any more. In the meanwhile, SRAS can also facilitate the applications of call centre. In many cases, the speaker gets disappointed, anxious, or angry when they call to deal with after-sale services problems. SRAS can identify and assess the speaker and help the call centre quickly respond to the disputes as well as achieving the customers' satisfaction.

For such applications, it becomes necessary to take into account the affective impact of speaker recognition so that the speakers could be recognized even when there is only affective speech provided for testing. The focus of this work is to develop a robust intelligent human-machine interface, which is more adaptive and responsive to a speaker's identity in emotional environments. In this paper, we further the research work in [7] and apply cost-sensitive learning to optimize the classification procedure and system performance.

This paper is organized as follows. In the next section, we give a review to the related work. The emotional corpus and emotional speech analysis are introduced in Section 3. Section 4 is committed to cost-sensitive learning and its application to speaker recognition. Section 5 discusses the system architecture. The experiments comparison and result discussion are presented in Section 6. We close with a conclusion section.

2. Related Work

In the literature, there are a few studies that focus on speaker recognition with affective speech. Structure training [8, 9] is first proposed and noted as a promising approach to address this problem. The method attempts to elicit different manners of speaking during the enrollment and makes the system become familiar with the variation likely to be encountered in that person's voice. Emotion-added modeling method [10] also attempts to elicit different manners of speaking during the enrollment. The goal of the systems is to learn not only the true distribution of the speaker-dependent speech features but also the influences of various emotions that corrupt this distribution.

Most of such systems model the speakers with a variety of affective speech and achieve great success. Dongdong and Yingchun [11] construct the speaker models with clustered affective speech. This approach aims at the maximum utilization of the limited affective training speech data. The prosodic difference is exploited to cluster affective speech, and the corresponding models are built with the clustered speech for a given speaker.

Along the way, all these methods mentioned above ask users to provide additional reading (emotional) speech in the training stage, which would lead to the unfriendliness of the system.

On the contrary, other researchers aim to promote the SRAS performance with only neutral enrolled for training, by means of adding artificial affective information to neutral training speech or eliminating the affective information in the emotional testing speech. Feature domain compensation aims at adding emotional information to neutral acoustic features prior to model training. One example is the rules based feature modification method based on the statistics of prosodic features [12, 13]. Specifically, the rules of prosodic features modification of duration, pitch, and amplitude parameters are derived from a small number of the content matched source-target pairs. The speaker model is trained with an aggregation of data with all kinds of the converted affective speech and the neutral speech.

Krothapalli et al. [6] believe that performance of the speaker identification system developed using neutral features is better with transformed features compared to emotional features. He proposes neural network based feature transformation framework for mapping the time-aligned syllable level features from any specific emotion to neutral. Shahin investigates emotion identification when the database is biased towards different emotions based on each of HMMs [14] and SPHMMs [15, 16].

Besides, score domain compensation attempts to remove model score scales and shifts caused by varying affective conditions of speakers. An example of score domain compensation techniques is E-Norm [17]. By investigating the pitch distribution variation under different emotional states, Li et al. [7] propose an improved pitch envelope based frame-level score reweighted (PFLSR) algorithm to compensate the affective effect in both speaker verification and identification system. The PFLSR aims to separate the frames that have large

speaker-emotion variability from the ones that are slightly affected by speakers' moods.

Most of the existing speaker recognition systems fail during affective speech due to emotional mismatch in the training and testing environments. Considering both the system friendliness and the algorithm complexity, a probability reweighted score domain compensation approach is proposed. The idea of score normalization has been long acknowledged to speaker verification at both utterance and frame level [18, 19]. It is widely used for its simplification, convenience, and excellent result. This work has furthered the study in [7] and used the supervised learning method to refine the final score.

3. Database and Speech Analysis

3.1. Database. The affective speech database evaluated in this paper is Mandarin Affective Speech Corpus (MASC) [20], which is distributed by the Linguistic Data Consortium. The speech in the database spans five different emotion types, which are neutral (unemotional), panic, anger, sadness, and elation. All the reading material is phonetically balanced, which covers all the phonemes in Chinese.

68 native speakers are elicited to utter 5 phrases and 20 sentences under five emotional states, as well as 2 extra neutral paragraphs speech. Each phrase and sentence is repeated for three times, respectively. Altogether the database contains 5,100 phrases (e.g., 5 phrases * 3 times * 68 subjects * 5 emotional types), 20,400 utterances, and 136 paragraphs. The detailed material is described as follows.

- (i) Five phrases: they are “shi de (yes),” “bu shi (no),” and three nouns as “ping guo (apple),” “huo che (train),” and “wang qiu (tennis ball).” In Chinese, these words contain many different basic vowels and consonants.
- (ii) 20 sentences: these sentences include 12 semantically neutral ones and 2 emotional ones for each type portraying the four emotional states. Different syntactical types are also represented in the sentences, which follow the material design of RUSLANA [21].
- (iii) Two paragraphs: they are two readings selected from a famous Chinese novel, stating a normal fact.

The MASC database is divided into three subsets: development set, training set, and evaluation set. The development set is composed of the speech from the first 18 people. The training set contains 50 speakers, whose 2 paragraphs of neutral speech are used to construct speaker models. The evaluation set is the utterance parts in five types of emotions. There are 50 such speakers, with 15000 authentic tests and 735000 imposter tests.

3.2. Speech Analysis. Referring to the affective speech, the prosody is a medium of emotional expression [22]. Phonetic research demonstrates that prosodic information involves complex articulatory and cognitive controls in speech [23–25]. Promising results of emotion classification have been achieved with the analysis of prosodic feature [26–28].

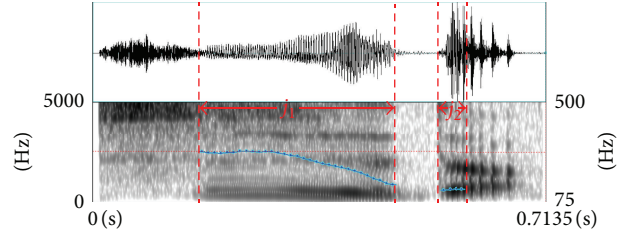


FIGURE 1: Example of segment boundaries estimation for the phrase “shi de.” The vertical bars represent the segment boundaries from the critical points of pitch contours.

Prosody feature refers to the variables (range, contour, and jitter) of pitch, speaking rate, and intensity to convey nonlexical linguistic and high-level cues. Among these features, pitch reveals the frequency at which the vocal folds vibrate and relies greatly on broad classes of sounds [29]. Pitch is investigated in this paper to indicate the characters of different emotion types.

The production of speech is the process of setting the air in rapid vibration. Periodic air pulses passing through vibrating vocal chords make voiced sounds, for example, “a” and “i” while unvoiced sounds such as “s” and “sh” are created by forcing air through a constriction in vocal tract, producing turbulence that is more noise-like. In this case, the pitch of an utterance is discontinuing. We can easily divide speech into voiced and unvoiced regions by detecting the points where the pitch contour appears or disappears. Figure 1 shows the waveform, spectrum, and pitch of phrase “shi de.” The voiced segment alternates with the unvoiced one. The boundaries of voiced and unvoiced speech are represented by vertical dotted bars. The pitch contour of the voiced speech is defined as pitch envelope here. The statistics and analysis of pitch parameters take the pitch envelope as a unit, as it could indicate the average level of the speaker’s voice frequency, which varies greatly under different emotional states.

Definition 1. Let $J = \{j_1, j_2, \dots, j_T\}$ be a pitch sequence of an utterance and let T be the frame numbers. $j_i = 0$ for the pitch of unvoiced segments and $j_i > 0$ for the pitch of voiced segments. The pitch envelope is denoted by $J^* = \{j_i \mid i = n, n + 1, \dots, m\}$ and satisfies the following constraint:

- (1) $j_i \neq 0$,
- (2) $j_{n-1} = 0, j_{m+1} = 0$,
- (3) $0 \leq n \leq i \leq m \leq T$.

The mean value of pitch envelope (PEM) can be calculated as

$$\bar{J}^* = \frac{1}{n - m + 1} \sum_{i=m}^n j_i, \quad (1)$$

where m and n are the numbers of the start and the end frame of a pitch envelope.

The pitch of a man’s voice falls under low pitch frequency (60–250 Hz), whereas woman’s voice is of the high pitch type (120–500 Hz). The distributions of PEM for male and

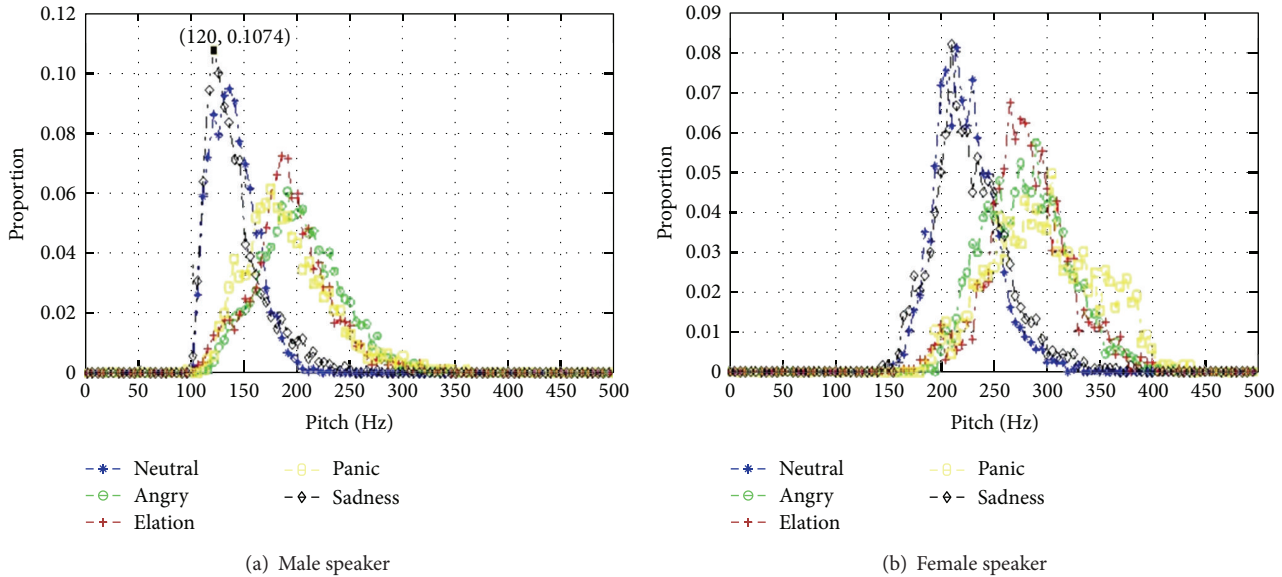


FIGURE 2: The probability distribution of PEM for the male (a) and female speakers (b) under the five emotion states.

female are studied separately. Figure 2 shows the probability distribution of PEM for all the sentences in MASC under the five emotion states. In particular, Figure 2(a) is the probability distribution of PEM with male’s speech, while Figure 2(a) is the probability distribution of PEM with female’s speech. The whole pitch frequency from 0 Hz to 500 Hz is equally divided into 100 scales with 5 Hz width each. For example, the point “100 Hz” on the abscissa represents the scales from 95 Hz to 100 Hz. The value on the ordinate corresponding to 100 Hz represents the proportion of voiced speech whose PEM falls in (95, 100) for each emotion. The point (120, 0.1074) means that there is 10.74% of PEM that falls in (115, 120) scope. Figure 2 demonstrates the probability distribution of PEM over 5 emotion types that can be divided into two groups. The neutral and sadness speech have similar distribution with smaller mean and variance value. The PEM probability distribution of anger, elation, and panic has larger mean and variance value. In this case, we assume that the voiced speech that has high PEM value is heavily affected by speaker’s emotional mood. We partition the voiced speech into two groups according to a threshold pitch value: (1) the class that is highly different from neutral speech mainly includes the pitch envelop of anger, elation, and panic; (2) the class that is slightly different from neutral speech, mainly includes the pitch envelop of neutral and sadness. Both the male and female’s speech have similar distribution, except that all the PEM of female are much higher than that of males’.

Thus, we can draw two kinds of important information. First, the PEM selection parameters should be set differently for male and female speakers as their dissimilarity distribution. Second, not all frames of the utterance are impacted dramatically by affective speech. In the speaker recognition task, the majority of frames of the test speech give the maximum probability (rank 1 score) to the target speaker (TS). In this case, the utterance could be correctly classified. However, the target speaker could not get the rank 1 score

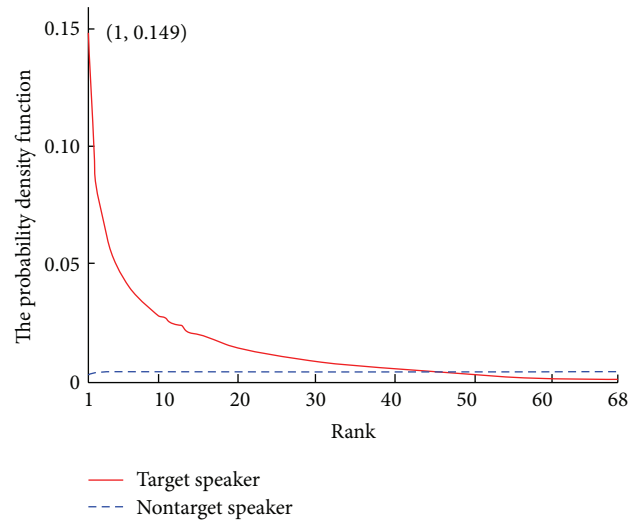


FIGURE 3: The frame-level score rank’s probability density functions for target speakers and nontarget speakers over 68 subjects in MASC.

from all the test frames. In particular when the mood state of speakers shifts from neutral to affective, the vocal and prosody features are also changed. Some test frames give their confidence to a nontarget speaker (NTS) mistakenly because of the mismatch of emotions between the speaker models and the test utterances. With the number of frames that assign the maximum probability to the NTS becoming enormous, the score of the NTS could be comparable or even higher than that of the TS. Figure 3 shows the frame-level score rank’s probability density functions for target speakers and nontarget speakers. The number on the abscissa represents the rank of score for each frame. For instance, the point (1, 0.149) in the red curve of target speaker means 14.9% of frames give the rank 1 score to the target speaker.

The reason why the test utterance is misclassified is not due to a nontarget speaker doing well but rather to a true speaker's model doing poorly. It is assumed that most frames still give the maximum likelihood to the target model, as they are not easily changed with the slight expressive corruption, while part of the frames that have large variations in F_0 may be affected by the emotion state change of speakers, and we define these frames as bad frames in this paper. To overcome the mistaken decisions induced by bad frames, we strengthen the roles of the good frames by giving them weight to exhibit their importance based on cost-sensitive learning.

Combined with the analysis of pitch, we divided the frames into two parts according to the variation of the F_0 value. The voiced part with high PEM that is heavily affected by the expressive speech (HA) is taken as bad frames. On the contrary, the voiced part that is slightly affected (SA) together with the unvoiced is considered as good frames.

4. Cost-Sensitive Speaker Recognition

4.1. Definition of Cost-Sensitive Learning. Cost-sensitive classification is based on a set of weights defining the expected cost when an object is misclassified. First, we give the definition for cost-sensitive classification problem.

Let $x \in \mathfrak{R}^n$ be a feature vector, let $\{1, 2, \dots, N\}$ be the label set of N classes, and let C be a $N * N$ cost matrix with entries $c_{i,j}$. $c_{i,j}$ are the cost of misclassifying an example of class i to class j ; both i and j belong to $\{1, 2, \dots, N\}$. $c_{i,j} > 0$ if $i \neq j$ and $c_{i,j} = 0$ if $i = j$:

$$C = \begin{bmatrix} 0 & c_{1,2} & \dots & c_{1,N} \\ c_{2,1} & 0 & \dots & c_{2,N} \\ \dots & c_{i,j} & 0 & \dots \\ c_{N,1} & \dots & c_{N,N-1} & 0 \end{bmatrix}. \quad (2)$$

Here, $c_{i,i} = 0$ is the correct classification. The expectation cost of class i can be computed by

$$c_j = \sum_{i=1}^N c_{i,j}. \quad (3)$$

The cost-sensitive learning can be defined as follows.

Definition 2. Let $P(X | Y)$ be the unknown joint distribution of X and Y . Let F be a set of mappings from X to Y . The cost-sensitive learning procedure is to select a mapping $f (f \in F)$, to minimize the risk functional $R(f)$, defined as

$$R(f) = E_{P(X|Y)} c_{y,f(x)} = \int \left[\sum_{y=1}^N c_{y,f(x)} P(y | X) \right] p(x) dx. \quad (4)$$

It is easy to recognize that when given $c_{i,j} = 1$ if $i \neq j$, (4) reduces to the standard classification.

4.2. Speaker Recognition Task. In the speaker recognition task, given a group of N known speakers model $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_N\}$ and a sample of test speech with T frames,

the likelihood of X that belongs to the i th speaker can be written as $P(X | \lambda_i)$ according to Bayes' rule. In the case that frames in the utterance are independent, $P(X | \lambda_i)$ could be expressed as

$$P(X | \lambda_i) = \prod_{t=1}^T p(x_t | \lambda_i). \quad (5)$$

Obviously, there are relations between frames and frames, and the $P(X | \lambda_i)$ can be rewritten as

$$P(X | \lambda_i) = \prod_{t=1}^T f(p(x_t | \lambda_i)). \quad (6)$$

Equation (5) is a special case of (6) when $f(x) = x$.

In the process of computing the test utterance on the speaker model, the score is always mapped to the log domain for calculation facilitation as follows:

$$\begin{aligned} \text{score}(X | \lambda_i) &= \log p(X | \lambda_i) \\ &= \log \prod_{t=1}^T p(x_t | \lambda_i) \\ &= \sum_{t=1}^T \log p(x_t | \lambda_i). \end{aligned} \quad (7)$$

According to the frame selection conducted in Section 3, the frames in the SA part and unvoiced part are reweighted to strengthen their confidence to the maximum likelihood model. Thus, for a special speaker i , the utterance level score of a T frame speech sequence is defined as

$$\text{score}(X | \lambda_i) = \sum_{t=1}^{T_1} f(\log p(x_t | \lambda_i)) + \sum_{t=1}^{T_2} \log p(x_t | \lambda_i), \quad (8)$$

where T_1 is the number of frames in SA and unvoiced part, T_2 is the number of frames in HA part, respectively, and $T = T_1 + T_2$.

Given the frame vector x_t in the SA part and the speaker i , the cost-sensitive function can be assumed as

$$f(\log p(x_t | \lambda_i)) = \sum_{y=1}^N c_{y,i} * \log p(x_t | \lambda_i). \quad (9)$$

Note that cost matrix C only needs to be computed once when function L is defined.

It is obvious that the reweight function should meet the rule that the frame score and the cost matrix L are in direct ratio.

4.3. Cost-Sensitive Parameters Learning. To deal with the class-dependent cost-sensitive learning problem, the data space expansion technique is adapted [30, 31]. Each example is expanded to N examples of different classes. The weights of N examples are decided based on the loss of the corresponding misclassifications. When a test utterance is compared with

a certain speaker model, the sum loss of its expanded N examples is in proportion to the loss of classifying it to that speaker model. The details of the expansion technique are given as follows.

Assume that in a speaker classification task, sample X is assigned to speaker model Y by classifier $f(X)$. C_y is positive as well as being not less than the largest misclassification loss in order to keep weights of expanded examples positive. The loss of $f(X)$ on the example (X, Y) is defined as

$$\begin{aligned} c_{y,f(x)} &= \sum_{i=1}^N c_{y,i} - \sum_{i=1}^N c_{y,i} I(f(x) \neq i) \\ &= \sum_{i=1}^N c_{y,i} - \sum_{i=1}^N c_{y,i} I(f(x) \neq i) + (K-1)c_y - (K-1)c_y \\ &= \sum_{i=1}^N (c_y - c_{y,i}) I(f(x) \neq i) - \sum_{i \in \{1, \dots, N\} \setminus y} (c_y - c_{y,i}), \end{aligned} \quad (10)$$

where $I(x)$ is a step function with value 1 if the condition in the parenthesis is true and 0 otherwise.

The expanded examples (X^n, Y^n) with weights $w_{y,n}$ are defined as

$$X^n = X, \quad Y^n = n, \quad w_{y,n} = (c_y - c_{y,n}). \quad (11)$$

Substituting (11) into (10), we can get

$$c_{y,f(x)} = \sum_{i=1}^N w_{y,i} I(f(x) \neq i) - \varphi(y), \quad (12)$$

where $\varphi(y) = \sum_{i \in \{1, \dots, N\} \setminus y} w_{y,i}$.

The loss of $f(x)$ on the example (X, Y) could be computed by a weighted loss of $f(x)$ on expanded examples minus a variable irrelevant to $f(x)$. The weights can modify the distribution on (X, Y) and produce a new one as well. In other words, cost-sensitive learning can be reduced to the standard classification [30].

5. System Architecture

Our previous work presented a pitch envelope based frame-level score reweighted speaker recognition framework [7]. The main contribution of this work is to introduce the cost-sensitive learning to reweigh the score. The testing process of the proposed speaker recognition system relies on 3 modules: gender identification, PEM based pitch envelop selection, and frame-level probability reweighted.

The purpose of the gender identification is to set different PEM threshold for male speakers and female speakers. This process is taken before frame selection. Given an utterance, the Mel frequency cepstral coefficients (MFCC) feature is extracted and tested with both male and female GMM models to produce the likelihood scores. The utterance is classified to the gender that has higher likelihood score. Corresponding

frame selection thresholds are set and adopted based on the result of gender identification.

In the process of the PEM based pitch envelop selection, the variation of pitch distribution under different emotional states is analyzed and compared with PEM threshold. The voiced envelop frames whose mean pitch value is smaller than threshold and the unvoiced part are chosen for reweighting.

The score reweight step aims to strengthen the confidence of the selected speech segments and optimize the final accumulated frame scores over the whole test utterance.

6. Experiment and Discussion

6.1. Experiment Settings. The evaluation task conducted in the experiments is text-independent and closed-set speaker recognition. The front end processing of speech single is as follows. A 13-dimensional Mel frequency cepstral coefficients (MFCC) vector is extracted from the preemphasized speech signal every 16 ms using a 32 ms Hamming window. A simulated triangular filter bank on the DFT spectrum is used to compute the Mel cepstral vector. Delta-cepstral coefficients are then computed and appended to the 13-dimensional cepstral vector, producing a 26-dimensional feature vector. The speaker classification, the GMM, consists of 32 mixture components. In the gender identification, two 1024 mixture GMMs, male and female model, are built with MAP method using the speech from the development subset. The statistical F_0 thresholds of PEM for the female and male speakers are set as 289 Hz and 189 Hz, respectively.

6.2. The Baseline System with Neutral Models. The aim of the first experiment is to capture the fluctuation in the system performance with various affective speeches. The speaker models are built with paragraph part on neutral speech, and the test utterances are in anger, elation, panic, sadness, and neutral, respectively. Figure 4 gives the verification results with the five types of affective speech tested independently with neutral speaker models. The verification performance will decline greatly when the system is involved with affective utterances for testing. It is clear that the consistence affective state of the training and testing utterances is important. The verification performance for speech in anger, elation, and panic drops more sharply than that in sadness. It is reported that the speakers would have a much higher arousal level mood when they are in the emotion of anger, elation, and panic than that of sadness [32]. This is one of the possible reasons that the EER of test speech in sadness state goes down to 26.1%, while the EER of test speech in other three affective states have a sharper drop.

6.3. Experiment Results and Discussions. The identification rates (IR) of the standard accumulated approach and the CSSR on emotional speech of anger, elation, neutral, panic, and sadness are shown in Table 1. The enhancement of IR for speech in anger, elation, and panic achieves 11.94%, 13.53%, and 9.84%, respectively, which is significantly greater than that achieved for speech in sadness and neutral. A possible reason is that when speakers are in the emotion of anger,

TABLE 1: Comparison of system performance under different types of affective speech (%).

Method	Baseline	CSSR
Anger	21.80	33.74
Elation	22.70	36.23
Neutral	94.40	95.63
Panic	26.30	36.14
Sadness	51.13	54.67
Total	43.27	51.28

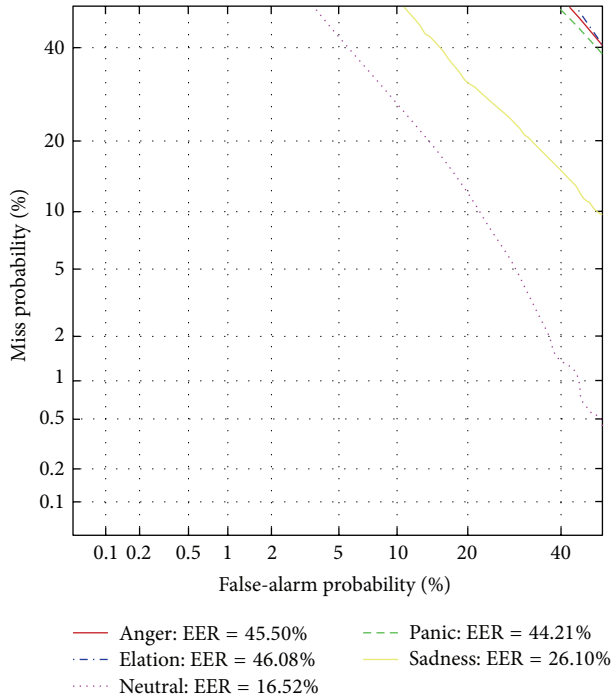


FIGURE 4: DET curves for the traditional speaker models trained with neutral speech only.

elation, and panic, they would have a much higher arousal level mood which makes more speech envelopes with high PEM. In other words, the speech in anger, elation, and panic has a much greater number of the bad frames. Once the confidence of good frames are strengthened for the speech, the identification rates of speaker recognition are easily promoted.

We also apply the proposed method to speaker verification task and compare it with other score normalization methods, as shown in Figure 5. The performance measured by the detection error tradeoff function (DET) as well as equal error rate (EER). The EER is calculated as the operating point on the DET curve where the false-alarm and missed-detection rates are equal. Figure 5 shows the promise of the new approach. Evaluation results clearly show that CSSR technique outperforms the standard accumulated approach, T-norm, and ENORM methods for speaker verification on affective speech.

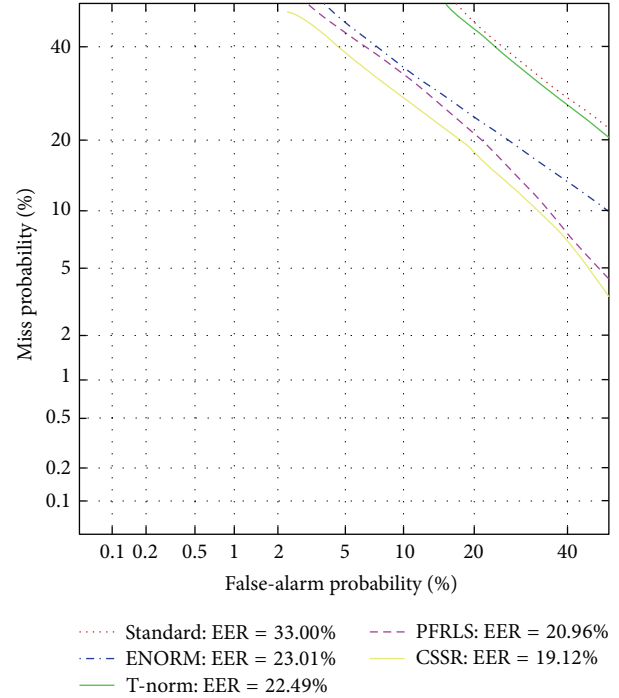


FIGURE 5: DET curves for the baseline, T-norm, ENORM, PFLSR, and CSSR based speaker verification system.

7. Conclusion and Discussion

In this paper, we introduce cost-sensitive learning to speaker recognition system with emotional speech. Based on an emotion robustness framework, cost-sensitive parameters are used to refine the probability of the slightly affected envelopes and to strengthen the confidence of the target speakers. Promising results are achieved in both speaker identification and speaker verification system. In future work, more effective algorithms of the frame selection and clustering recognition method [33] may be suggested to be employed in this system. On the other hand, the emotional parameters associated with specific speakers can also be considered as the characteristics in the recognition of their speeches [34].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research was supported by Nature Science Foundation of Shanghai Municipality, China (no. 11ZR1409600) and partly supported by the Natural Science Foundation of China (no. 91324010), the 2nd Regular Meeting Project of Science and Technology Cooperation between China and Serbia (no. 2-9), and China Scholarship Council (no. 201308310364).

References

- [1] O. A. Esan, S. M. Ngwira, and I. O. Osunmakinde, "Bimodal biometrics for financial infrastructure security," in *Proceedings of the Information Security for South Africa*, pp. 1–8, IEEE, August 2013.
- [2] S. Rane, W. Ye, S. C. Draper, and P. Ishwar, "Secure biometrics: concepts, authentication architectures, and challenges," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 51–64, 2013.
- [3] A. Alarifi, I. Alkurtass, and A. S. Als Salman, "SVM based Arabic speaker verification system for mobile devices," in *Proceedings of the International Conference on Information Technology and e-Services (ICITeS '12)*, pp. 1–6, March 2012.
- [4] K. S. Rao, A. K. Vuppala, S. Chakrabarti, and L. Dutta, "Robust speaker recognition on mobile devices," in *Proceedings of the International Conference on Signal Processing and Communications (SPCOM '10)*, pp. 1–5, July 2010.
- [5] I. R. Murray and J. L. Arnott, "Synthesizing emotions in speech: is it time to get excited?" in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '96)*, Invited Paper, pp. 1816–1819, Philadelphia, Pa, USA, October 1996.
- [6] S. R. Krothapalli, J. Yadav, S. Sarkar, S. G. Koolagudi, and A. K. Vuppala, "Neural network based feature transformation for emotion independent speaker identification," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 335–349, 2012.
- [7] D. Li, Y. Yang, and T. Huang, "Pitch envelope based frame level score reweighted algorithm for emotion robust speaker recognition," in *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII '09)*, pp. 1–4, September 2009.
- [8] K. R. Scherer, "Vocal communication of emotion: a review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, 2003.
- [9] G. Klasmeyer, T. Johnstone, T. Bänziger, C. Sappok, and K. R. Scherer, "Emotional voice variability in speaker verification," in *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 212–217, Belfast, Ireland, 2000.
- [10] T. Wu, Y. Yang, and Z. Wu, "Improving speaker recognition by training on emotion-added models," in *Affective Computing and Intelligent Interaction: Proceedings of the 1st International Conference (ACII '05)*, Beijing, China, October 22–24, 2005, vol. 3784 of *Lecture Notes in Computer Science*, pp. 382–389, 2005.
- [11] L. Dongdong and Y. Yingchun, "Emotional speech clustering based robust speaker recognition system," in *Proceedings of the 2nd International Congress on Image and Signal Processing (CISP '09)*, pp. 1–5, Tianjin, China, October 2009.
- [12] L. Dongdong, Y. Yingchun, W. Zhaohi, and W. Tian, "Emotion-state conversion for speaker recognition," in *Affective Computing and Intelligent Interaction: Proceedings First International Conference (ACII '05)*, Beijing, China, October 22–24, 2005, vol. 3784 of *Lecture Notes in Computer Science*, pp. 403–410, 2005.
- [13] W. Zhaohui, L. Dongdong, and Y. Yingchun, "Rules based feature modification for affective speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, pp. I661–I664, Toulouse, France, May 2006.
- [14] I. Shahin, "Using emotions to identify speakers," in *Proceedings of the 5th International Workshop on Signal Processing and Its Applications (WoSPA '08)*, Sharjah, United Arab Emirates, 2008.
- [15] I. Shahin, "Speaker identification in the shouted environment using Suprasegmental Hidden Markov Models," *Signal Processing*, vol. 88, no. 11, pp. 2700–2708, 2008.
- [16] I. Shahin, "Speaker identification in emotional environments," *Iranian Journal of Electrical and Computer Engineering*, vol. 8, no. 1, pp. 41–46, 2009.
- [17] W. Wu, T. F. Zheng, M.-X. Xu, and H.-J. Bao, "Study on speaker verification on emotional speech," in *Proceedings of the INTERSPEECH and 9th International Conference on Spoken Language Processing (INTERSPEECH '06—ICSLP)*, pp. 2102–2105, September 2006.
- [18] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, 1995.
- [19] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using non-linear frame likelihood transformation," *Speech Communication*, vol. 24, no. 3, pp. 193–209, 1998.
- [20] T. Wu, Y. Yang, Z. Wu, and D. Li, "MASC: a speech corpus in mandarin for emotion analysis and affective speaker recognition," in *Proceedings of the IEEE Workshop on Speaker and Language Recognition (Odyssey '06)*, pp. 1–5, San Juan, Puerto Rico, June 2006.
- [21] V. Makarova and V. Petrushin, "RUSLANA: Russian language affective speech database," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '02)*, pp. 2041–2044, 2002.
- [22] S. Kemal, S. Elizabeth, H. Larry, and W. Mitchel, "Modeling dynamic prosodic variation for speaker verification," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '98)*, pp. 3189–3192, Sydney, Australia, 1998.
- [23] R. W. Frick, "Communicating emotion. The role of prosodic features," *Psychological Bulletin*, vol. 97, no. 3, pp. 412–429, 1985.
- [24] Santiago-Omar and Caballero-Morales, "Recognition of emotions in Mexican Spanish speech: an approach based on acoustic modelling of emotion-specific vowels," *The Scientific World Journal*, vol. 2013, Article ID 162093, 13 pages, 2013.
- [25] J. Hirschberg, "Communication and prosody: functional aspects of prosody," in *Proceedings of ESCA Workshop Dialogue and Prosody*, pp. 7–15, 1999.
- [26] N. Minematsu and S. Nakagawa, "Modeling of variations in cepstral coefficients caused by Fo changes and its application to speech processing," in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP '98)*, pp. 1063–1066, Sydney, Australia, 1998.
- [27] M. Schröder, "Emotional speech synthesis: a review," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 561–564, Aalborg Congress and Culture Centre, Aalborg, Denmark, September 2001.
- [28] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, 2004.
- [29] P. Yang, Y. Yang, and Z. Wu, "Exploiting glottal information in speaker recognition using parallel GMMs," in *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '05)*, pp. 804–812, Hilton Rye Town, NY, USA, July 2005.
- [30] F. Xia, Y.-W. Yang, L. Zhou, F. Li, M. Cai, and D. D. Zeng, "A closed-form reduction of multi-class cost-sensitive learning to weighted multi-class learning," *Pattern Recognition*, vol. 42, no. 7, pp. 1572–1581, 2009.
- [31] N. Abe, B. Zadrozny, and J. Langford, "An iterative method for multi-class cost-sensitive learning," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '04)*, pp. 3–11, August 2004.

- [32] R. Cowie, R. Corive, E. Douglas-Cowie et al., "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.
- [33] W. Dai, S. Liu, and S. Liang, "An improved ant colony optimization cluster algorithm based on swarm intelligence," *Journal of Software*, vol. 4, no. 4, pp. 299–306, 2009.
- [34] Y. Wang, X. Hu, W. Dai, J. Zhou, and T. Guo, "Vocal emotion of humanoid robots: a study from brain mechanism," *The Scientific World Journal*, vol. 2014, Article ID 216341, 7 pages, 2014.