Journal of Southeast University (English Edition) Vol. 31, No. 3, pp. 311 − 319 Sept. 2015 ISSN 1003—7985

# Dimensional emotion recognition in whispered speech signal based on cognitive performance evaluation

Wu Chenjian[1] Huang Chengwei[2] Chen Hong[3]

([1] School of Electronic and Information Engineering, Soochow University, Suzhou 215006, China)
([2] College of Physics, Optoelectronics and Energy, Soochow University, Suzhou 215006, China)
([3] School of Mathematical Sciences, Soochow University, Suzhou 215006, China)

**Abstract:** The cognitive performance-based dimensional emotion recognition in whispered speech is studied. First, the whispered speech emotion databases and data collection methods are compared, and the character of emotion expression in whispered speech is studied, especially the basic types of emotions. Secondly, the emotion features for whispered speech is analyzed, and by reviewing the latest references, the related valence features and the arousal features are provided. The effectiveness of valence and arousal features in whispered speech emotion classification is studied. Finally, the Gaussian mixture model is studied and applied to whispered speech emotion recognition. The cognitive performance is also considered in emotion recognition so that the recognition errors of whispered speech emotion can be corrected. Based on the cognitive scores, the emotion recognition results can be improved. The results show that the formant features are not significantly related to arousal dimension, while the short-term energy features are related to the emotion changes in arousal dimension. Using the cognitive scores, the recognition results can be improved.

**Key words:** whispered speech; emotion recognition; emotion dimensional space

**doi:** 10.3969/j.issn.1003 − 7985.2015.03.003

Many disabled people rely on hearing aids to communicate through normal speech[1]. Studies on whispered speech have drawn much attention. The study on speech emotion recognition is a subfield in whispered speech signal processing, and it is closely related to multiple disciplines, including signal processing, pattern recognition, phonetics, and psychology. Therefore, whispered speech emotion recognition has attracted researchers from various backgrounds[2−13]. However, the understanding of how human beings express emotions in whispered speech and how a machine recognizes the emotions is still a great challenge.

Schwartz et al. [14] studied the problem of gender recognition from whispered speech signal. Tartter et al. [15−16] studied the listening perception of vowels in whispered speech. They showed that some of the vowels, such as [a] and [o] may be confused in whispered speech. The early works mainly focused on the whispered speech from the phonetic view and adopted the listening test as an experimental measure. In 2002, Morris[17] studied the enhancement of whispered speech from the signal processing aspect and his work covered many areas, including speech conversion, speech coding, and speech recognition. Gao[18] studied tones in Chinese whispered speech in a perception experiment. Gong et al. [6, 19] studied the detection of endpoint and formant estimation in whispered speech signal. The recognition of tones in whispered speech was also studied in Ref. [7]. Jin et al. [20] studied the whispered speech emotion recognition problem, and established a basic whispered speech emotion database. Gong et al. [5] used formants and speech rate features to classify whispered speech in three emotion categories: happiness, anger, and sadness. Also, they studied the emotional features in whispered speech and found that the time duration and short-term energy may classify anger and sadness.

In this paper, emotion recognition in whispered speech signal is studied. The collection of emotional speech is carried out in an eliciting experiment, in which the cognitive performance is evaluated. Therefore, it is able to fuse two kinds of information, the emotional dimension information and the dynamic change of cognitive ability. The 2D arousal-valence dimensional emotion space is a continuous space for emotions. It can be safely assumed that the whispered speech emotion is also distributed in the same way, as the inner-sate of the subject is the same. Based on the dimensional emotion theory that emotions can be treated as continuous vectors in the 2D space, a system is developed, which can not only recognize the whispered speech emotion, but also model the relationships between the past emotional and the current emotional state. The fundamental belief here is that the emotional state transfer probabilities differ among discrete emotion classes. For instance, rapid shifting between positive and negative emotional states is very likely to be a classification mistake.

**Received** 2014-12-21.

**Biography:** Wu Chenjian(1983—), male, doctor, lecturer, cjwu@suda.edu.cn.

**Foundation item:** The National Natural Science Foundation of China (No. 11401412).

**Citation:** Wu Chenjian, Huang Chengwei, Chen Hong. Dimensional emotion recognition in whispered speech signal based on cognitive performance evaluation[J]. Journal of Southeast University (English Edition), 2015, 31(3): 311 − 319. [doi: 10.3969/j.issn.1003 − 7985.2015.03.003]

## 1 Whispered Speech Database

### 1.1 Overview

A high quality emotion database is the foundation of emotion recognition research. There are many speech emotion databases available, and, however, there is still a lack of whispered speech emotion databases. The establishment of a whispered speech emotion database consists of five major steps: 1) A data collection plan; 2) Whispered speech recording; 3) Data validation and editing; 4) Emotional sample annotation; 5) A listening test. Compared to normal speech emotional data, in view of recognition accuracy, the establishment of a whispered speech emotion database is a great challenge. Various naturalistic data in normal speech is obtained, covering a wide range of emotion types, e. g., frustration, fidgetiness, anxiety. However, in the previous studies in whispered speech, there were not enough types of emotions in the databases. Also, the expression and recognition of emotions in whispered speech is much more difficult than in normal speech. A few examples of whispered speech emotion databases are summarized in Tab. 1.

**Tab. 1** Comparison of whispered speech emotion database

| Data type | Emotion classes | Name of the database | Recognition rate/% | Data source | Use of database | Applicable algorithm |
|---|---|---|---|---|---|---|
| Whispered speech | Neutrality, anger, surprise, fear, happiness, sadness, tiredness | Soochow University whispered speech emotion database[5-6] | | Acted | Emotion recognition and speaker recognition | Discrimination Analysis |
| | Happiness, anger, sadness, surprise | Southeast University whispered speech emotion database[20-21] | 49 | Acted | Emotion recognition | Gaussian mixture model, support vector machine, *K*-nearest neighbor |
| | Neutrality | Nanjing University Chinese whispered speech database[22] | 59.7 (tone recognition) | Acted | Speech recognition and conversion | Neural network |
| Normal speech | Happiness, neutrality, fidgetiness | Southeast University practical speech emotion database[23] | 70 | Induced | Emotion recognition | Gaussian mixture model, support vector machines |

### 1.2 Whispered emotional speech

The recording of whispered speech emotion is the first step of our research, the quality of the data is essential to the recognition system performance. For normal speech recording, it may only need a quiet laboratory room. However, the recording of whispered speech requires a silent room to avoid the noise.

During the data collection, normal speech in the same text for comparison is also recorded. The normal speech and whispered speech under a neutral state are shown in Fig. 1. The pitch contour of normal speech is demonstrated in Fig. 2(a). The formant frequency of whispered speech is demonstrated in Fig. 2(b). Due to the missing pitch frequency in whispered speech, its formant frequency is especially important. Since the intensity of whispered speech is much lower than the normal speech, the noise influence becomes an important problem in emotion recognition. Schuller et al. [24] first studied noise influence on speech emotion recognition. Tawari et al. [25] applied noise reduction algorithms in in-vehicle emotion recognition applications. Their study showed that the wavelet-based speech enhancement can improve the emotion recognition performance in normal speech.

Under different emotional states, it can be seen that the acoustic parameters of the whispered speech signals have changed. In Fig. 3, the duration of the whispered speech under anger, happiness, and neutrality has changed significantly.

### 1.3 Emotional data analysis in a cognitive task

In this section the relationship between speech emotion and cognitive performance is studied.

Emotions are closely related to cognitive tasks, and detecting the cognitive-related emotion states is particularly interesting in the work environment. Operations in an extreme environment may induce special types of emotions, such as anxiety, tiredness, fidgetiness, etc. Those emotions are closely related to cognitive process and may threat the success of a task. The detection of negative emotions is important for evaluating the operator' s emotional well-being.

The data collection methods can be classified into three categories: naturalistic speech, induced speech and acted speech. For whispered speech, the eliciting methods which are successfully applied to normal speech can be applied. Johnstone et al. [26] used the computer game to induce normal speech emotional data, and established a high quality emotion dataset. For negative emotions with a practical value, such as fidgetiness, sleep deprivation, noise stimulation and repeated cognitive task as the eliciting methods can be used[23].

In this paper, the noise stimulation and repeated math calculations are used to induce negative emotions. The time duration of one particular emotion usually lasts for 1 to 2 min. There is no unhealthy influence on the volunteer subjects.

Many environmental factors can induce negative

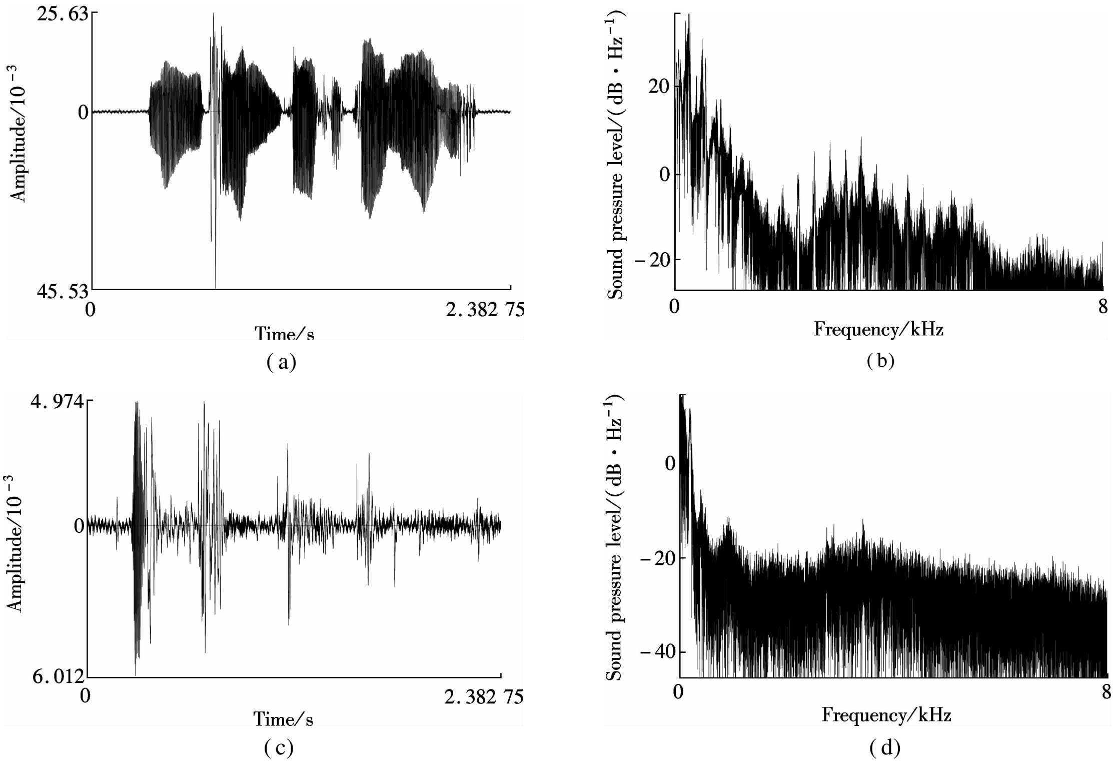

**Fig. 1** Normal and whispered speech under neutral state. (a) Waveform of normal speech signal; (b) Spectrum of normal speech signal; (c) Waveform of whispered speech signal; (d) Spectrum of whispered speech signal

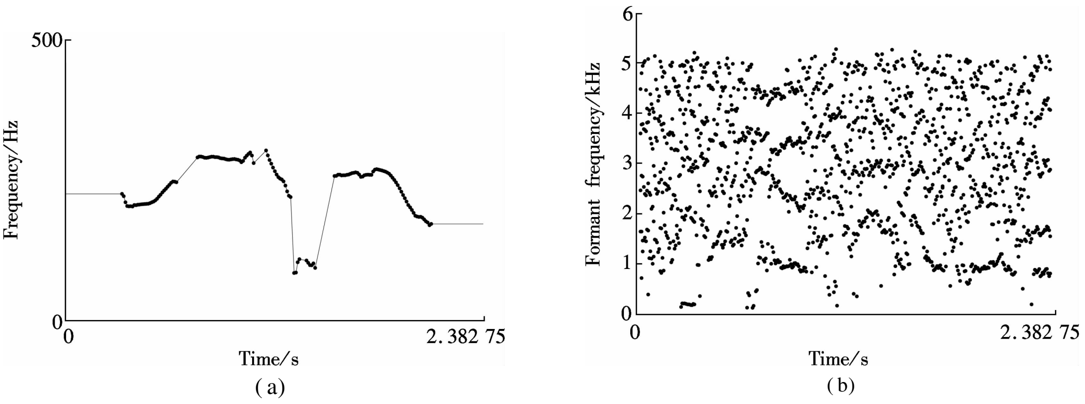

**Fig. 2** Speech parameters of normal speech and whispered speech. (a) Pitch contour of normal speech; (b) Formant frequency of whispered speech

emotions. Noise is a common cause that induces negative emotions in extreme environments. For example, in the Russian Mir space station, the noise level is between 59 to 72 dB, which can cause a series of stimulated emotions and hearing problems.

The repeated boring task is a commonly used technique to induce negative emotions in psychology experiments. The subject is required to do math calculations repeatedly and report orally. At the same time, the answers are recorded and evaluated. Correct answers add scores to the cognitive performance. In Fig. 4, the relationship between negative emotions and the false answers is analyzed which reflects the subject's cognitive working ability changing over time.

## 2 Annotation and Validation of Whispered Emotional Speech

After the recording of the original whispered speech data, a listening test is needed. The validation of the speech data relies on the listening perception. For each utterance, the emotions at five different intensity levels may be labeled with the scales of 1, 3, 5, 7, 9 corresponding to very weak, weak, ordinary, strong, and very strong.

Then an evaluation result $\boldsymbol{E}_i$ on each utterance from each listener is obtained.

$$\boldsymbol{E}_{ij} = \{e_1^{ij},\ e_2^{ij},\ e_3^{ij},\ e_4^{ij}\} \tag{1}$$

where $j$ is the index of emotional utterances; $i$ denotes the

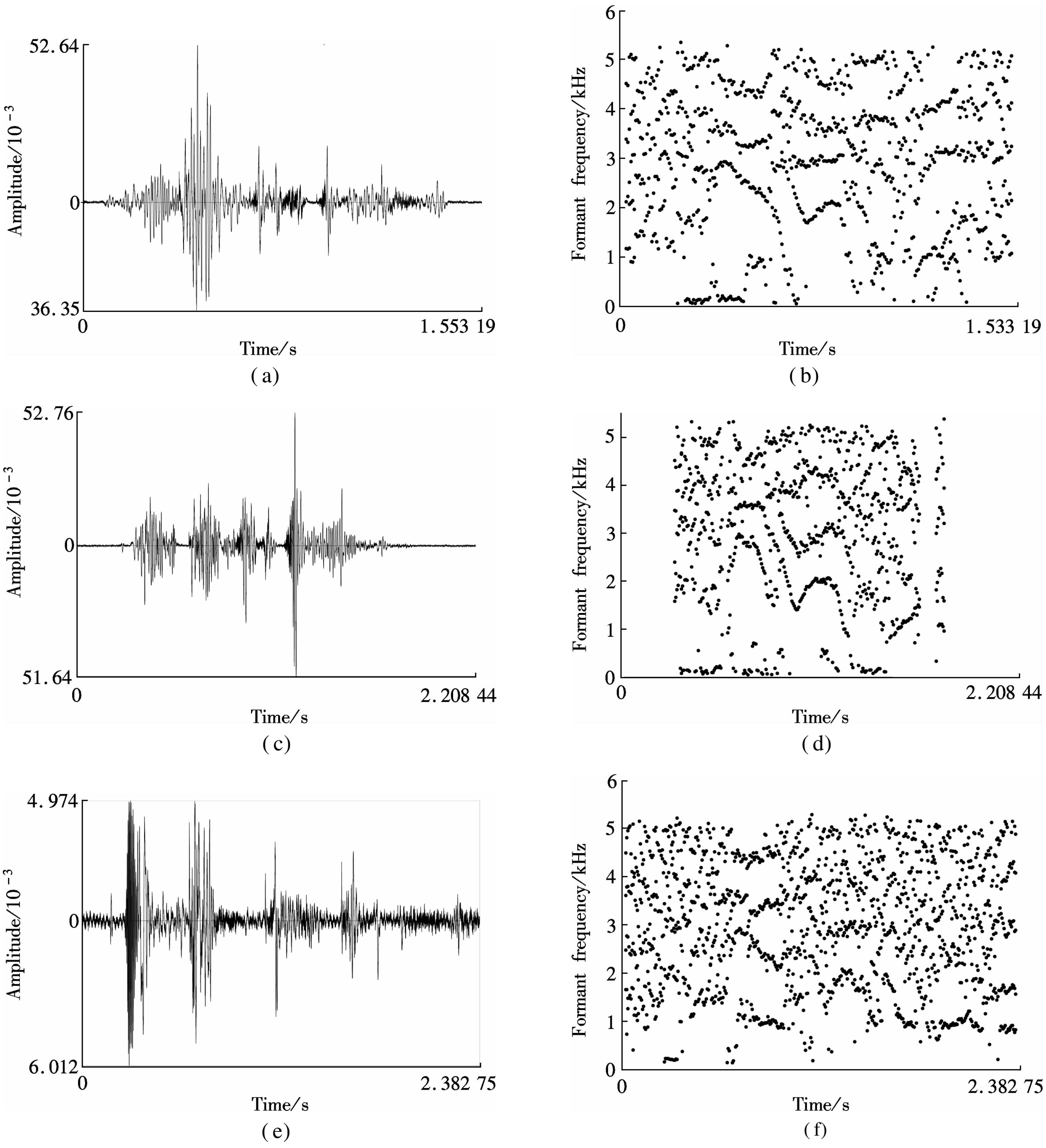

**Fig. 3** Whispered speech under emotional states. (a) Waveform of speech of happiness; (b) Formant of speech of happiness; (c) Waveform of speech of anger; (d) Formant of speech of anger; (e) Waveform of speech of neutrality; (f) Formant of speech of neutrality

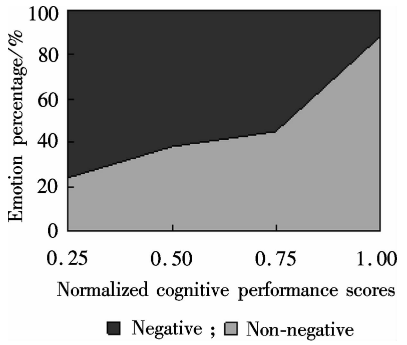

**Fig. 4** Correlation between negative emotion and cognitive score in a cognitive experiment

listener; $e$ represents the evaluation score.

In the multiple listener case, to achieve the evaluation result on the $j$-th sample, the listening results need to be combined,

$$\boldsymbol{E}_j = \sum_{i=1}^{M} a_i \boldsymbol{E}_{ij} \tag{2}$$

where $a_i$ is the fusion weight; $M$ is the number of listeners. The weight represents the confidence in each listener, and it is noticed that

$$\sum_{i=1}^{M} a_i = 1 \tag{3}$$

For annotation results from different listeners, a fusion method[15] can be adopted to achieve the final result. For the $j$-th sample, the similarity between the two listeners $p$ and $q$ can be represented as[23]

$$\rho_j^{pq} = \prod_{i=1}^{K} \frac{\min\{e_i^{pj}, e_i^{qj}\}}{\max\{e_i^{pj}, e_i^{qj}\}} \tag{4}$$

where $K$ is the number of emotion classes. For each evaluation, the agreement matrix is

$$\rho^{pq}=\frac{1}{N}\sum_{j=1}^{N}\rho_j^{pq}=\frac{1}{N}\sum_{j=1}^{N}\prod_{i=1}^{K}\frac{\min\{e_i^{pj},\ e_i^{qj}\}}{\max\{e_i^{pj},\ e_i^{qj}\}} \tag{5}$$

where $N$ is the total number of the data samples. Based on the similarity between two listeners, the agreement matrix $\boldsymbol{\rho}$ is achieved, in which each element represents the degree of agreement between two listeners, and $M$ is the total number of the listeners.

$$\boldsymbol{\rho}=\begin{bmatrix}1 & \rho^{12} & \cdots & \rho^{1M}\\ \rho^{21} & 1 & \cdots & \rho^{2M}\\ \vdots & \vdots & & \vdots\\ \rho^{M1} & \rho^{M2} & \cdots & 1\end{bmatrix} \tag{6}$$

The averaged value represents the degree of agreement between the $i$-th listener and the others:

$$\bar{\rho}^{i}=\frac{1}{M-1}\sum_{j=1,\ j\neq i}^{M}\rho^{ij} \tag{7}$$

The normalized value is adopted as the fusion weight $a_i$,

$$a_i=\tilde{\rho}^{i}=\frac{\bar{\rho}^{i}}{\sum_{i=1}^{M}\rho^{i}} \tag{8}$$

Substituting Eq. (8) into Eq. (2), the final evaluation result $\boldsymbol{E}_j$ of each utterance is achieved.

## 3 Emotional Feature Analysis for Whispered Speech

### 3.1 Acoustic features of normal speech signal

The speech feature that can reflect emotional change in speech has been an essential question in emotion research for a long time. In past decades, researchers studied the emotional features from phonetics and psychology. Emotional speech features can be classified into two groups, prosodic features and voice quality features. Prosodic features include intensity, duration, pitch, accent, tone, intonation, etc. In the early research, in normal speech emotion recognition, prosodic features are the most commonly used emotional features. Among them the pitch parameter is the most important. However, in whispered speech, pitch is missing. Voice quality features include formant frequency, harmonic-to-noise ratio, linear prediction coefficients, etc. Also, voice quality features can be essential for classifying valence dimensional information in normal speech.

### 3.2 Acoustic features of whispered speech signals

Emotional feature analysis is an essential part of emotion recognition. In this section, the characters of whispered speech signal and the extracted basic emotional features are analyzed. In whispered speech, the vocal cords do not vibrate normally since it is an unvoiced mode of phonation. In normal speech, air from the lungs causes the vocal folds of the larynx to vibrate, exciting the resonances of the vocal tract. In whispered speech, the glottis is opened and turbulent flow created by exhaled air passing through this glottal constriction provides a source of sound[17].

The acoustic features of normal speech and whispered speech is studied. Among the commonly used speech features, pitch is the most important feature to classify emotions in normal speech signals. However, there is no pitch feature in the whispered speech signal. Therefore, this parameter cannot be applied to the whispered speech emotion recognition. The formant frequency, Mel frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) are the important speech features and they can be applied to emotion classification in whispered speech signal. These features are generally related to the valence dimension. Short-term energy, speech rate and time duration are related to the arousal level. Experimental results show that short-term energy and speech rate are effective for classifying emotions in whispered speech signals[5]. The Teager energy operator (TEO) has also been applied to whispered speech emotion analysis[21].

Based on the basic acoustic parameters of whispered speech, proper features for modeling can be constructed. Generally speaking, the speech emotional features can be grouped into two categories, the static features and the temporal features. Since the temporal features largely rely on the phonetic changes in the text, the global static features are chosen to construct utterance level features. Difference, mean, maximum, minimum and variance are used to construct higher level text-independent features.

## 4 Recognition Methodology

### 4.1 Overview of classification methods

In this section, the general speech emotion recognition methods are discussed, as shown in Tab. 2. Several algorithms that are successfully applied to whispered speech emotion recognition are also studied. Many of the pattern recognition algorithms, such as the hidden Markov model (HMM), the Gaussian mixture model (GMM), and the support vector machines (SVM), have been studied for emotion recognition.

### 4.2 Emotion recognition for whispered speech

Hultsch et al. [2-3] discovered that the expressions of happiness and fear were more difficult in whispered speech. Shortly after, Cirillo et al. [4] studied this problem again in a listening perception experiment, and came to a similar conclusion. In their research, happiness in whispered speech was easily confused with fear or neutrality. Further spectrum analysis showed that the confusion of these emotions might be caused by the quality decrease of tones in whispered speech. In a low voice quality speech listening experiment, the whispered speech

**Tab. 2** Speech emotion recognition algorithms

| Algorithms | Modeling ability on emotional data | Reported recognition rate | Characters |
|---|---|---|---|
| GMM[27–28] | Very strong | High | Strong ability in data distribution modeling and highly dependent on the training data |
| SVM[29–30] | Strong | High | Suitable for small sample size and unsuitable for large number of training utterances |
| KNN[31] | Strong | Median | Simple to implement |
| HMM[22, 29] | Average | High | Suitable for time sequence and affected by phonetic information |
| Decision Trees[32] | Average | Median | Simple to implement and suitable for multiple emotion classes |
| ANN[33–34] | Very strong | Median | Suitable for modeling non-linear relations and may be over-fitted to training data. |
| Fuzzy methods[35] | Average | High | Suitable for practical requirements with new emotion types |
| Shuffled frog leaping algorithm[34, 36] | Strong | High | Strong optimization ability |

signal was sent by telephone line and happiness was confused with sadness or neutrality. Cirillo and his colleagues[4] found that sadness and fear were easy to classify in whispered speech and the acoustic analysis also support this conclusion.

Using a set of discriminant features on a simple dataset, many popular pattern recognition algorithms can succeed in speech emotion recognition. However, up to now the studies on whispered speech emotion recognition have been very rare, and which algorithm is suitable to whispered speech emotion recognition is still an open question. Whispered speech emotions are successfully classified[6, 15]. Quantum neural networks are discussed in whispered speech emotion recognition in Ref. [21]. By applying quantum genetic algorithms to back-propagation neutral networks, the connection weights are optimized and the robustness of the neutral network is improved.

The GMM is the state-of-the-art speaker and language identification algorithm[28], and theoretically, the GMM can be used to model any probability distribution. In practical terms, the GMM parameters need to be empirically set to achieve good performance. In this paper, the GMM-based classifiers to whispered speech emotion recognition is applied.

The GMM is the weighted sum of $M$ members,

$$p(\boldsymbol{X}_t \mid \lambda) = \sum_{i=1}^{M} a_i b_i(\boldsymbol{X}_t) \tag{9}$$

where $\boldsymbol{X}$ is a $D$-dimensional vector, $b_i(\boldsymbol{X})$ ($i = 1, 2, \ldots, M$) is the Gaussian distribution of each member; $a_i$ ($i = 1, 2, \ldots, M$) is the mixture weight.

$$b_i(\boldsymbol{X}_t) = \frac{1}{(2\pi)^{D/2} \left| \sum_i \right|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{X}_t - U_i)' \sum_i^{-1} (\boldsymbol{X}_t - U_i) \right\} \tag{10}$$

where the mixture weight satisfy

$$\sum_{i=1}^{M} a_i = 1 \tag{11}$$

The complete GMM parameters can be represented as

$$\lambda_i = \left\{ a_i, U_i, \sum_i \right\} \qquad i = 1, 2, \ldots, M \tag{12}$$

According to the Bayes theory, the classification can be made by maximizing the posterior probability:

$$E = \underset{k}{\operatorname{argmax}} \{ p(\boldsymbol{X}_t \mid \lambda_k) \} \tag{13}$$

## 4.3 Error correction based on context information

The current detection decisions are made by the fusion of the current outputs of the emotion classifiers, the previous outputs of the emotion classifiers and the current cognitive performance score. The system block diagram is shown in Fig. 5. The detection outputs are the likelihoods of the classifier, representing negative and non-negative emotions. The previous emotion states are used for inferring the current emotion states since emotion state is treated as a continuously changing variable in the time domain. Cognitive performance is modeled by the correctness of the answers the subjects made during the math calculation experiment. Cognitive performance information is presented in the system as a total test score dropping or rising, which is based on the current answer being correct or incorrect.

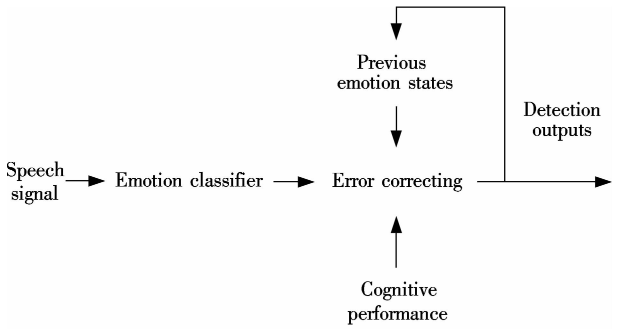

**Fig. 5** System block diagram

The likelihoods of the GMM emotion classifiers can form an emotion vector $\boldsymbol{E}_i = \{p_1, p_2, \ldots, p_m\}$. Here, $i$ is the sampling of time; $m$ is the number of emotion classes and $p_i$ is the likelihood of the classifier. Considering the previous emotion states and the cognitive performance $P$, the emotion vector $\{E_i, E_{i-1}, E_{i-2}, \ldots, E_{i-n}, P_i\}$is extended. Error correcting is then achieved by using the naive Bayes classifier trained on the instances of the extended emotion vector.

By using the context information, the emotional state transfer between neighboring utterances can be modeled. The affective state generally lasts for a certain period of time, and, therefore, it is safe to assume that the neighboring emotion recognition results are dependent on each other. Error correction is, therefore, believed to be effective.

## 5 Experimental Results

### 5.1 Experiments on arousal features and valence features recognition

In the emotion dimensional model, it is generally believed that for normal speech the prosodic features are related to the arousal level and the voice quality features are related to the valence level. It is noted that for whispered speech this correlation has not been proved yet. Therefore, the GMM-based recognition experiment is carried out to demonstrate the possible relationship between arousal-valence dimension space and speech feature space in whispered speech.

In the whispered speech emotion database[20], sadness and anger are located on the negative side of the valence dimension, and happiness is located on the positive side of the valence dimension. In the arousal dimension, anger, happiness, and surprise are located on the positive side, while sadness is on the negative side. Based on the GMM classifiers, 200 utterances are chosen for each emotion category. The training and testing ratio is 3∶1. Voice quality feature (formant frequency) and prosodic features (short-term energy, speech rate, duration) for the recognition test are used, and the cross-validation results are shown in Tab. 3.

**Tab. 3** Recognition rates using arousal features and valence features in whispered speech signal

| Emotion recognition rate | Voice quality features | | Prosodic features | | Related dimension |
|---|---|---|---|---|---|
| | Positive | Negative | Positive | Negative | |
| Surprise and sadness | 62 | 66 | 58 | 48 | Arousal |
| Anger and sadness | 70 | 66 | 72 | 76 | Arousal |
| Happiness and anger | 60 | 64 | 70 | 72 | Valence |

Surprise and sadness are not well classified, while anger can be easily recognized. When formant features are used alone, the recognition result is not satisfactory. On the other hand, short-term energy and other prosodic features are proved effective. In this experiment, it can be seen that voice quality features are not effective for classifying arousal level, while prosodic features are obviously related to both dimensions.

### 5.2 Recognition experiments comparison

In the recognition test of our whispered emotional speech database, three popular machine learning algorithms, the GMM, SVM and $K$-nearest neighbor, are adopted. The GMM is a widely used modeling tool for the estimation of underlining the probability density function. For the emotional data, the GMM-based classifiers are developed based on the expectation maximum (EM) algorithm. A GMM emotional model consists of several Gaussian members, and the mixture number is set experimentally. In our case, it is set to be 6, due to the limited instances. The SVM is a powerful learning algorithm in a small sample set. Since the basic SVM is designed for two-class classification, a decision tree structure to classify multiple emotion classes in the one-against-all fashion is used. However, in the decision tree, how to stop the error from accumulating is an important issue. The SVM classifiers are configured to form the tree structure according to the rank of error rates. The highest error rate appears at the bottom of the tree. In this way, the error can be prevented from spreading, the resulting emotion tree structure is shown in Fig. 6. The KNN is a simple but powerful classification algorithm, and a basic KNN classifier for comparison is also adopted. The number of dimensions is set to be ten, and the radial basis function is chosen for the SVM classifier. The GMM is initialized with $K$-means clustering, and $K$ is set as the number of emotion classes. In the expectation-maximization algorithm for parameter estimation, the maximum iteration is set to be 50, which is enough for convergence in our application.

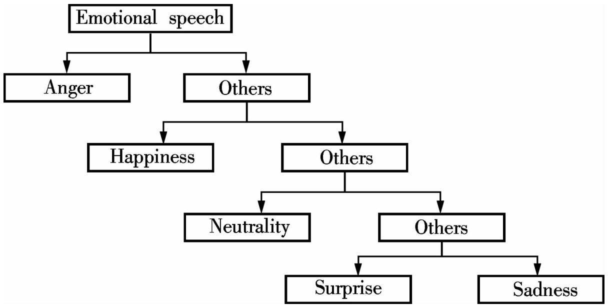

**Fig. 6** A depiction of the decision tree for the SVM multi-class classifier

As shown in Tab. 4, it can be seen that the GMM outperforms the other classifiers in average. The highest recognition occurs in the classification of anger using the GMM-based classifier. The GMM is a powerful modeling algorithm, given a proper setting of mixture numbers and sufficient training data, it can achieve an accurate performance compared to other state-of-the-art machine learning algorithms. Among different speakers, the expressions of anger in whispered speech are perhaps closer to each other. The lowest rate occurs in the classification of surprise. With the absence of pitch frequency feature, the modeling of surprise is much difficult. Theoretically, the GMM classifier can present any probability distribution and it has a strong ability to fit the training data.

**Tab. 4** Comparison of recognition results %

| Classifier | Happiness | Sadness | Surprise | Anger |
|---|---|---|---|---|
| GMM | 65 | 68 | 59 | 71 |
| SVM | 66 | 60 | 51 | 67 |
| KNN | 62 | 62 | 56 | 55 |

The SVM classifier performs better than the KNN classifier in the acted data under a small training sample. The SVM classifier has a strong learning ability when the training data is limited.

The error correcting method using the context information proposed in this paper brings an improvement in recognition, as shown in Tab. 5. Based on the previous detected emotional states, the current sample can be classified effectively. Using the cognitive performance scores as a correlated factor, the useful context information for the recognition of subject' s inner emotional state is provided. The close relationship between the cognitive process and emotional states has been accepted for a long time, and in this experiment it is able to correct the emotion recognition results using the cognitive scores.

**Tab. 5** Emotion detection rates before and after error correction %

| Emotion | GMM-based classifier | After error correction | Improvement |
|---|---|---|---|
| Negative | 72 | 76 | 4 |
| Non-negative | 65 | 70 | 5 |

## 6 Conclusion

Automatic detection of human emotion is important in many human-computer interaction applications. The detection of emotions is the first step for evaluating the human factor in a man-machine system or in a special operation. Based on the emotion monitor, psychological intervention for people to cope with negative emotions can be adopted.

Studies on whispered speech can lead to future applications in intelligent human-computer interaction, especially for natural interaction and person-dependent interaction. In surveillance technology and security, studying whispered speech can help with detecting potential dangerous situations and gathering valuable information. For future emotion recognition systems based on whispered speech, multi-modal information including acoustic features, linguistic features and context features will be adopted.

## References

[1] Liang R, Xi J, Zhao L, et al. Experimental study and improvement of frequency lowering algorithm in Chinese digital hearing aids[J]. *Acta Physica Sinica*, 2012, **61**(13): 134305-1 − 134305-11.

[2] Hultsch H, Todt D, Ziiblke K. *Einsatz und soziale interpretation gefliisterter signale, umwelt und verhalten*[M]. Bern, Switzerland: Huber Verlag, 1992: 391 − 406.

[3] Tartter V C, Braun D. Hearing smiles and frowns in normal and whisper registers[J]. *Journal of Acoustic Society of America*, 1994, **96**(4): 2101 − 2107.

[4] Cirillo J, Todt D. Decoding whispered vocalizations: Relationships between social and emotional variables[C]// *Proceedings of the* 9*th International Conference on Neural Information Processing*. Singapore, 2002: 1559 − 1563.

[5] Gong C, Zhao H, Tao Z, et al. Feature analysis on emotional Chinese whispered speech[C]//2010 *International Conference on Information Networking and Automation*. Kunming, China, 2010: 137 − 141.

[6] Gong C, Zhao H, Wang Y, et al. Development of Chinese whispered database for speaker verification[C]// 2009 *Asia Pacific Conference on Postgraduate Research, Microelectronics & Electronics*. Shanghai, China, 2009: 197 − 200.

[7] Gong C, Zhao H. Tone recognition of Chinese whispered speech[C]//2008 *Pacific-Asia Workshop on Computational Intelligence and Industrial Application*. Wuhan, China, 2008: 418 − 422.

[8] Tartter V C. Identifiability of vowels and speakers from whispered syllables[J]. *Perception and Psychophysics*, 1991, **49**(4): 365 − 372.

[9] Takeda T K, Itakura F. Acoustic analysis and recognition of whispered speech[C]//*Proceedings of IEEE International Conference on Acoustics, Speech and Signal.* Orlando, FL, USA, 2002: 389 − 392.

[10] Yang L, Li Y, Xu B. The establishment of a Chinese whisper database and perceptual experiment [J]. *Journal of Nanjing University*: *Natural Science*, 2005, **41**(3): 311 − 317.

[11] Huang C, Jin Y, Zhao L, et al. Speech emotion recognition based on decomposition of feature space and information fusion [J]. *Signal Processing*, 2010, **26**(6): 835 − 842.

[12] Huang C, Jin Y, Zhao Y, et al. Recognition of practical emotion from elicited speech [C]//*Proceedings of ICISE.* Nanjing, China, 2009: 639 − 642.

[13] Huang C, Jin Y, Zhao Y, et al. Speech emotion recognition based on re-composition of two-class classifiers [C]//*Proceedings of ACII.* Amsterdam, Netherland, 2009: 1 − 3.

[14] Schwartz M F, Rine M F. Identification of speaker sex from isolated, whispered vowels[J]. *Journal of Acoustical Society of America*, 1968, **44**(6): 1736 − 1737.

[15] Tartter V C. Identifiability of vowels and speakers from whispered syllables[J]. *Perception and Psychophysics*, 1991, **49**(4): 365 − 372.

[16] Higashikawa M, Minifie F D. Acoustical-perceptual correlates of “whisper pitch” in synthetically generated vow-

els[J]. *Speech Lung Hear Res*, 1999, **42**(3):583 - 591.

[17] Morris R W. Enhancement and recognition of whispered speech[D]. Atlanta, USA: School of Electrical and Computer Engineering, Georgia Institute of Technology, 2002.

[18] Gao M. Tones in whispered Chinese: articulatory and perceptual Cues[D]. Victoria, Canada: Department of Linguistics, University of Victoria, 2002.

[19] Huang C, Jin Y, Bao Y, et al. Whispered speech emotion recognition embedded with Markov networks and multi-scale decision fusion[J]. *Signal Processing*, 2013, **29**(1): 98 - 106.

[20] Jin Y, Zhao Y, Huang C, et al. The design and establishment of a Chinese whispered speech emotion database [J]. *Technical Acoustics*, 2010, **29**(1): 63 - 68.

[21] Zhao Y. Research on several key technologies in speech emotion recognition and feature analysis [D]. Nanjing: School of Information Science and Engineering, Southeast University, 2010.

[22] New T L, Foo S W, Silva L C D. Speech emotion recognition using hidden Markov models[J]. *Speech Communication*, 2003, **41**(4): 603 - 623.

[23] Huang C, Jin Y, Zhao Y, et al. Design and establishment of practical speech emotion database[J]. *Technical Acoustics*, 2010, **29**(4): 396 - 399.

[24] Schuller B, Arsic D, Wallhoff F, et al. Emotion recognition in the noise applying large acoustic feature sets[C]// *The 3rd International Conference on Speech Prosody.* Dresden, Germany, 2006: 276 - 289.

[25] Tawari A, Trivedi M M. Speech emotion analysis in noisy real-world environment[C]//*Proceedings of the 20th International Conference on Pattern Recognition.* Washington DC, USA, 2010:4605 - 4608.

[26] Johnstone T, van Reekum C M, Hird K, et al. Affective speech elicited with a computer game [J]. *Emotion*, 2005, **5**(4): 513 - 518.

[27] Zou C, Huang C, Han D, et al. Detecting practical speech emotion in a cognitive task[C]//*20th International Conference on Computer Communications and Networks.* Hawaii, USA, 2011:1 - 5.

[28] Kockmann M, Burget L, Cernocky J H. Application of speaker-and language identification state-of-the-art techniques for emotion recognition[J]. *Speech Communication*, 2011, **53**(9/10):1172 - 1185.

[29] Lin Y, Wei G. Speech emotion recognition based on HMM and SVM[C]//*Proceedings of 2005 International Conference on Machine Learning and Cybernetics*. Bonn, Germany, 2005:4898 - 4901.

[30] Jin Y, Huang C, Zhao L. A semi-supervised learning algorithm based on modified self-training SVM[J]. *Journal of Computers*, 2011, **6**(7): 1438 - 1443.

[31] Dellaert F, Polzin T, Waibel A. Recognizing emotion in speech[C]//*The Fourth International Conference on Spoken Language*. Pittsburgh, PA, USA, 1996:1970 - 1973.

[32] Lee C, Mower E, Busso C, et al. Emotion recognition using a hierarchical binary decision tree approach [J]. *Speech Communication,* 2011,**53**(9/10):1162 - 1171.

[33] Nicholson J, Takahashi K, Nakatsu R. Emotion recognition in speech using neural networks[J]. *Neural Computing &Applications*, 2000, **9**(4):290 - 296.

[34] Yu H, Huang C, Zhang X, et al. Shuffled frog-leaping algorithm based neural network and its application in speech emotion recognition[J]. *Journal of Nanjing University of Science and Technology*, 2011, **35**(5):659 - 663.

[35] Wang Z. Feature analysis and emotino recognition in emotional speech [D]. Nanjing: School of Information Science and Engineering, Southeast University, 2004.

[36] Yu H, Huang C, Jin Y, et al. Speech emotion recognition based on modified shuffled frog leaping algorithm neural network [J]. *Signal Processing*, 2010, **26** (9): 1294 - 1299.

# 基于认知评估的多维耳语音情感识别

吴晨健[1] 黄程韦[2] 陈 虹[3]

([1] 苏州大学电子信息学院,苏州 215006)
([2] 苏州大学物理与光电·能源学部,苏州 215006)
([3] 苏州大学数学科学学院,苏州 215006)

**摘要:**研究了基于认知评估原理的多维耳语音情感识别. 首先,比较了耳语音情感数据库和数据采集方法,研究了耳语音情感表达的特点,特别是基本情感的表达特点. 其次,分析了耳语音的情感特征,并通过近年来的文献总结相关阶特征在效价维和唤醒维上的特征. 研究了效价维和唤醒维在区分耳语音情感中的作用. 最后,研究情感识别算法和应用耳语音情感识别的高斯混合模型. 认知能力的评估也融入到情感识别过程中,从而对耳语音情感识别的结果进行纠错. 基于认知分数,可以提高情感识别的结果. 实验结果表明,耳语音信号中共振峰特征与唤醒维度不显著相关,而短期能量特征与情感变化在唤醒维度相关. 结合认知分数可以提高语音情感识别的结果.

**关键词:**耳语音;情感认知;情感维空间

**中图分类号:**TP391.4