Research Article

INTERNATIONAL JOURNAL OF ADVANCED ROBOTIC SYSTEMS

International Journal of Advanced Robotic Systems
July-August 2017: 1–9
© The Author(s) 2017
DOI: 10.1177/1729881417719836
journals.sagepub.com/home/arx
SAGE

A combined cepstral distance method for emotional speech recognition

Changqin Quan[1], Bin Zhang[2], Xiao Sun[2] and Fuji Ren[3]

Abstract

Affective computing is not only the direction of reform in artificial intelligence but also exemplification of the advanced intelligent machines. Emotion is the biggest difference between human and machine. If the machine behaves with emotion, then the machine will be accepted by more people. Voice is the most natural and can be easily understood and accepted manner in daily communication. The recognition of emotional voice is an important field of artificial intelligence. However, in recognition of emotions, there often exists the phenomenon that two emotions are particularly vulnerable to confusion. This article presents a combined cepstral distance method in two-group multi-class emotion classification for emotional speech recognition. Cepstral distance combined with speech energy is well used as speech signal endpoint detection in speech recognition. In this work, the use of cepstral distance aims to measure the similarity between frames in emotional signals and in neutral signals. These features are input for directed acyclic graph support vector machine classification. Finally, a two-group classification strategy is adopted to solve confusion in multi-emotion recognition. In the experiments, Chinese mandarin emotion database is used and a large training set (1134 + 378 utterances) ensures a powerful modelling capability for predicting emotion. The experimental results show that cepstral distance increases the recognition rate of emotion sad and can balance the recognition results with eliminating the over fitting. And for the German corpus Berlin emotional speech database, the recognition rate between sad and boring, which are very difficult to distinguish, is up to 95.45%.

Keywords

Cepstral distance, emotional speech recognition, two-group classification, principal component analysis

Date received: 10 March 2016; accepted: 16 June 2017

Topic: AI in Robotics
Topic Editors: Chrystopher L Nehaniv and Masahiro Ohka

Introduction

Robot plays an important role in human society: it is not only the right-hand man in human daily life but also the soul companion of us human. Only when intelligence and advanced intelligent robots also have, people will really accept them. Emotion is the particular spiritual activities of human beings. Thus, emotion has become a sign of advanced intelligent in machines. Emotion analysis and decision-making ability are the key to the development of the machine.

Currently, robot with the ability to recognize emotion in speech brings hopes to the lonely elderly, children or people suffering from communication barriers.[1,2] Speech or voice signal are usually the best way to identify human emotion because they are the most common and the most natural way to communicate.[3] The recognition of emotional speech is rather disparate than the recognition of

[1] Graduate School of System Informatics, Kobe University, Kobe, Japan
[2] Department of Computer and Information Science, Hefei University of Technology, Hefei, China
[3] Faculty of Engineering, University of Tokushima, Japan

Corresponding author:
Changqin Quan, Kobe University, 1-1, Rokkodai, Nada, Kobe 6578501, Japan.
Email: quanchqin@gold.kobe-u.ac.jp

speech. Speech recognition is to identify the major components of the language in speech. However, the identification of emotional speech is mainly to distinguish emotional statement based on the differences between these emotions. It is on account of these different understanding that scholars are exploring the method which is appropriated for emotional speech recognition. Whether the improvement of the model or the careful screening for the features, few consideration is taken to the characteristics of emotion.

Early work shows that the voice is able to express emotion because it contains the parameters to reflect characteristics of emotion.[3,4] In order to mirror the characteristics of emotional speech such as speed, stress and tone, feature sets are divided, extended and finally applied.[5] Some paralinguistic properties of the voice including gender, age, voice quality, stress and so forth play a vital role in emotional speech recognition.[6–8] One of the properties which is named as voice quality[6] is widely applied. However, choices for representing voice quality characteristics are varied such as spectral gradients.[9]

To identify emotion in voice, linguistic information[10] or emotional point information[11] is also added to classification models, such as context[12–14] and keywords. Emotion is an important aspect of intelligence. The problem that we want to make the system to distinguish emotion leads scholars to resort to our own emotion cognitive system. Bionics and biology have been used to detect emotion in speech.[15] They utilize the physical structure of the human ear to improve the model to enhance recognition performance or to generate suitable features for human ears perceived characteristics, such as mel-frequency cepstral coefficients (MFCCs), Lyon's cochlear model.[16] These methods which mimic human ears are often easier for people to trust and adopt for they can be better understood. In order to more fully take into account the physical structure of the human ear, adaptability and stability of the system built by this method which mimics the physiological activity are better than the general system which did not use physiological information in speech.

And for multi-class emotional speech signals, no classification tools are especially suitable for multi-class classification. Thus, optimized classification tools and perfected classification strategies[9] will improve accuracy of emotion recognition. To solve these problems, fusions of multiple classifiers[11,17,18] and multi-stage classification[9,19] are adopted, and the effect is satisfactory.

Typically, cepstral distance is applied to endpoint detection in speech signal. Because of its exceptional ability to characterize the similarity between speech frames, it is also used for isolated word recognition which is a part of speech recognition.[20] But it is the first time to apply it to measure the similarity of emotion in speech signal. In this study, cepstral distance which measures similarity between two signals is applied to characterize differences in the emotion space. The voice quality is added in feature set in this study. Voice quality features include formant feature and harmonic–noise ratio (HNR). Formant is an import feature that reflect resonance characteristic of channel. Formant is resonance section in speech spectral and convey a direct information about the sound source. In addition to expresses the strength of vowel phonemes, formant characterizes speaker's pronunciation. It also express the quality of vowel phonemes. Thus, describing channel information formant is added to the feature set. HNR is the ratio of the energy of harmonic part and energy of a noise portion in speech signals. This ratio objectively and quantitatively descripts the crack sound in voice. For example, when men express grief (sad) mood, their voices usually become soft, choked, and at this time, they cannot even control his breathing. Then it pronounce incomplete vowel which should have been made into a complete waveform vowel. The vowel even becomes a consonant in sad mood. In this case, listeners often hear a sound that just like vocal cords are tearing. This harsh voice is called crack voice. And this is the sign of the coming of strong emotion. In the classification process, we decompose the problem of multi-class classification into the problem of tri-classification and four emotion classification. Using the strategy and the model, directed acyclic graph support vector machine (DAG-SVM),[21] the recognition rate can be improved significantly.

The rest of this article is organized as follows. 'Related works' section talks about the related works. 'The method' section describes the materials and methods proposed and applied in this article. The experiments and results are shown and analysed in 'Experiment and analysis' section. The comparisons between related works with our method are given in 'Conclusion and future work' section, and we come to a conclusion and discussion of some future work.

Related works

Affective interaction can significantly improve the efficiency of a reinforcement learning robot.[22] From the behaviour of the robot, we find evidence to prove that the ability to perceive emotions strongly affects the ability of robot learning and decision-making.[23] Therefore, more and more robotics involve affective computing. Yilmazyildiz et al.[24] apply multi-modal emotional information containing video and audio to identify the emotion in affect human–robot interaction. This article discusses the emotion recognition problem in speech, and this speech model will eventually be integrated as one of multi-modal emotion recognition system in humanoid robot.

Unbalanced train data result in a low recognition rate of classification model. This problem exists in two-class classification and it is particularly evident in the multi-class classification. In other words, in multi-category emotion classification, the accuracy is generally low. The reason is probably that the variation of features is not one-to-one mapping with emotion. A given feature variation tends to be associated with a cluster of affective attributes.[6] This article is focused on a balanced outcome when classifying multiple emotion. For this purpose, we propose cepstral distance parameter which is as a benchmark of neutral emotion. It can identify the differences between the average neutral voice and other emotional voice in each frame and be regarded as a measure of emotion.

Researchers have tried on a large number of methods. These methods can be divided into two categories: one is to improve the classification model,[10,19,25–28] including a comparison in the performance of different classifiers, a fusion of multiple classifier, and another is selection of the integration strategies.[28] SVM, k-nearest neighbour (KNN) algorithm, C4.5, multiple layers perception, artificial neural network and hidden Markov model are the models approbated by scholars whose performances in other areas are quite good. However, there is no sole model that is particularly suitable for a multi-class emotions recognition in speech. In most of the literatures, the majority still agreed with this order of classifier performance: KNN<C4.5<ANN<SVM.[26,29] For SVM, the effect of different kernel functions varies greatly. And deep neural networks model has recently been used in speech recognition by a large number of research institutions[30] and its effect is quite good. Yu,[31] who conducts tests to compare the effects of different SVM kernel functions, finds that the effects of radial basis function (RBF) kernel no matter on the training set or on the test set are the best. In the end, it is suggested to apply the model fusion and multi-level classifier. If the fusion strategy is proper, model integration would be better than separate model.

Another is the discovery and selection[27,32–39] for the feature set. Gharavian et al.[5] investigate the effect on using a rich set of features to improve the performance. He concludes that the recognition rate is different when testing on different feature set combinations in experiments. So in his method, a feature selection strategy is applied to select the appropriate feature sets, making feature has complementary advantages and minimizes mutual interference.

However, when trying on both features and models, there must be a major part and the other for the secondary. In brief, one serves for another such as that selecting the appropriate feature set optimizes the performance in some separate classifiers or multiple-layer classifiers.[13,33,40]

Inspiring by the previous researches, we try to find the feature in theory to ensure a balanced recognition result. From the relationship between different emotions in speech frames, we explore the similarity between frames and find a certain distance measure to characterize the similarity at last. Cepstrum is a kind of feature that reflects the sound perception of human ear. The distance certainly reflects the differences between different emotions. Based on these, we propose a method combined cepstral distance feature to measure the similarity between frames in emotional speech. It can solve the imbalance of accuracy for classification to some extent.

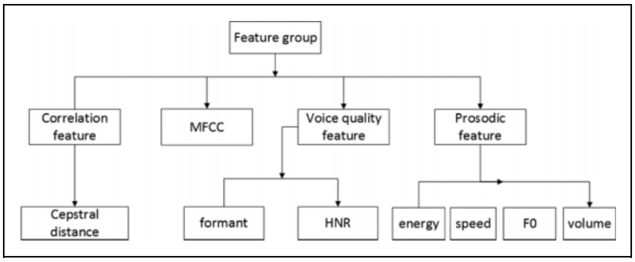

Figure 1. The organization of feature group in this experiment.

The method

Main feature set applied in the experiment

The goal of the proposed approach is to recognize sentiment in emotional speech. We apply four kinds of feature sets in the experiment: (1) the correlation feature including cepstral distance, (2) MFCC, (3) quality features and (4) prosodic features.

All feature groups are shown in Figure 1. MFCC and prosodic features are widely applied in emotional speech recognition and called traditional feature set in this article. The expressiveness is strong on prosodic feature by reason that it conveys information of the voice such as speed, tone and mood, so we apply it in experiment. More details about the traditional feature sets can be found in the study by Koolagudi and Rao[35] and Chen et al.[40] And the new features are introduced in the following.

The application of cepstral distance in the field of speech is endpoint detection as well as medical judgement for morbidity. The cepstrum is gained by the inverse discrete Fourier transform of discrete Fourier transform of the input speech signal after taking logarithm of the inverse discrete Fourier transform. Cepstrum $c(n)$ is defined as

$$c(n) = \frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \log|X(e^{jw})| e^{jwn} d_w \tag{1}$$

The cepstrum forms a natural basis for comparing patterns in speech recognition because of its stable mathematical characterization for speech signals. A typical 'cepstral distance measure' is of the form

$$D = \sum_{n=1}^{n_w} \left(c[n] - \overline{c[n]} \right) \tag{2}$$

where $c[n]$ and $\overline{c[n]}$ are cepstral sequences corresponding to signal frames, and D is the cepstral distance between the pair of sequences.

We calculate the cepstral distance between different emotions and emotion neutral. For each sentence is recorded in six emotions, we can calculate the cepstral distance between emotional utterances and their corresponding

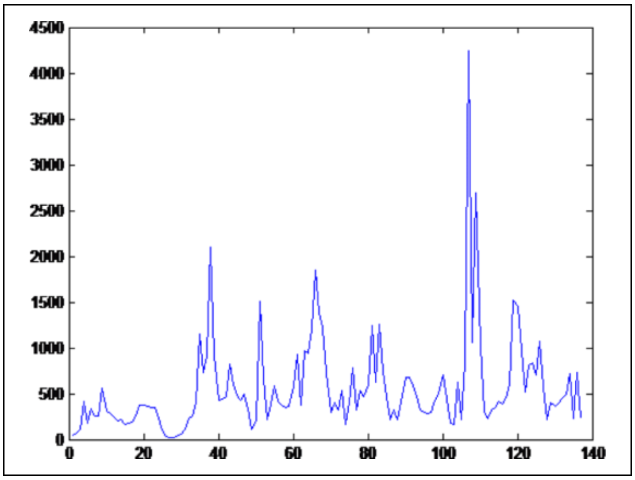

Figure 2. The cepstral distance from emotion neutral of one angry utterance in time domain. And the ordinate represents the amplitude and the abscissa stands for samples.

neutral utterance. First, getting statistical average of each sentences of emotion neutral and then computing the distance between a certain emotion and emotion neural. Finally, the cepstral distance is used as an important supplement for the other feature set. In general, cepstral distance is applied to measuring the similarity between two frames of signals. In this article, it represents the similarity between emotions. Figure 2 shows cepstral distance between one angry utterance and its corresponding neutral utterance.

Classification model

The SVM classification model is applied for emotional speech recognition in the experiments. We perform DAG-SVM[21] for multi-classification. The basic idea of DAG-SVM is that for an N-class problem, it constructs N(N-1)/2 classifiers, one for each pair of classes. In addition, principal component analysis (PCA) algorithm is combined with grid search method to optimize parameters for obtaining the classification model.

The emotional speech classification model is trained based on the combination of MFCC feature, prosodic features, voice quality features and cepstral distance features. In terms of no single features which are complement with each other by overlapping and influencing, they also plays a key role in the recognition. Considering the unbalanced feature data, we employ C-support vector classification (C-SVC) which is from the Library for SVMs (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). C-SVC is a cost-sensitive SVM, and different penalty functions are used for positive and negative class samples. The parameter c is a penalty factor parameter, which helps implement a penalty on the misclassifications that are performed while separating the classes. The model parameter 's' is set as the value of '0' which means set type of SVM. Another model parameter 't' is set as the value of '2' which means set the RBF kernel.

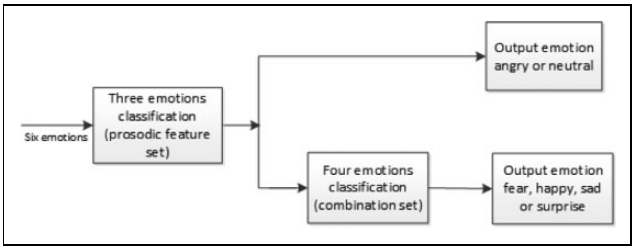
Six emotions
Three emotions classification (prosodic feature set)
Output emotion angry or neutral
Four emotions classification (combination set)
Output emotion fear, happy, sad or surprise

Figure 3. The process of classification in this experiment. First recognizes three emotions. When the emotion is angry or neutral output the result. If not, conduct classification in four emotions.

Classification strategy

Among a lot of experiments on emotional speech recognition, we truly feel the interaction and overlap of these feature sets and try to balance the results by applying characteristics of emotions.

We categorize the emotions into several categories and, then in different emotional categories, we apply the feature set whose performance is the best to identify those emotions in that category. All emotions are sorted into several groups. Grouping principle is grouping emotions in the groups which is easy to recognize, and the number of groups is within a reasonable range. The process in this article is showed in Figure 3.

Experiment and analysis

The whole experiment is implemented in platform MATLAB on a standard desktop with Intel Core i5 2.4 GHz processor with 8.0 GB RAM.

Materials

It involves two databases in this research: one is continuous Mandarin Chinese emotional speech corpus and another is German Berlin[41] speech database. The former is in Chinese for train and test, while the latter is used to compare with the existing methods.

The experiments are conducted on the emotional speech database recorded by Chinese Academy of Social Sciences. Among the six speakers, the female speaker called LiuChang is chosen. Six emotions are included, such as angry, fear, happy, sad, surprise and neutral. Each emotion corresponds to 252 utterances, a total of 1512 utterances. The signals are sampled at 16 kHz and transcribed in mono. Each sampling point is represented with 16 bit. The utterances are between 1 and 2 s to reserve prominent parts in emotions.

As used herein, the German Berlin[41] speech database is composed by about 490 utterances in six emotions: happiness, anger, sadness, anxiety, boredom and neutral. These

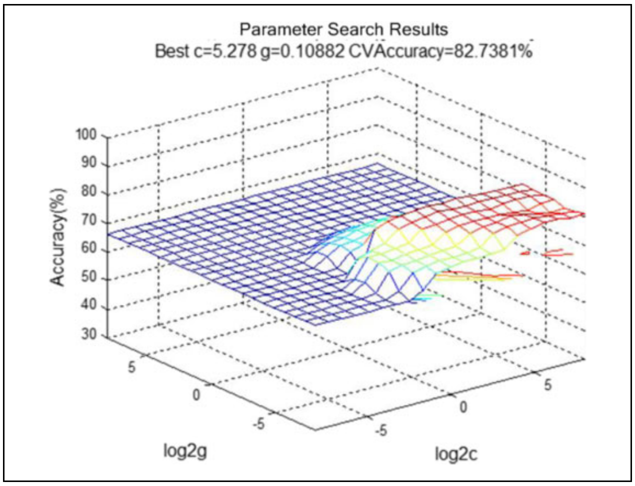

Figure 4. 3D view parameter search results for feature set.

signals are recorded by 10 different people including men and women and sampled at 16 kHz. And the utterances are short enough to ensure that one sentence contains only one emotion as much as possible.

To obtain data for facilitating subsequent processing, we conduct a series of operations. First of all, detecting signal endpoint with short-time energy of the speech and then pre-emphasizing the speech signal. Framing utterances with the frame length of 400 samples (25 ms) and a frame shift of 100 samples according to short characteristics of the speech. Finally, windowing the signal with hamming windows.

Procedure

We perform emotion recognition by applying DAG-SVM for multi-classification. After preprocessing, extract features from the original data. The feature set is composed of MFCC feature, prosodic features, voice quality features and cepstral distance features. The combinations of these features are also applied. A total of 1134 utterances selected randomly are used as training data, and 378 utterances are used as testing data. Training and testing data are normalized between [−1, 1]. After the process of the 0.95 PCA, optimize parameters for classification model SVM with grid search method and obtain the parameters c and g when attaining the best classification accuracy in cross-validation. The parameters are utilized to train a DAG-SVM classifier.

No single feature set plays a key role in the recognition. Different feature sets complement with each other by overlapping and influencing. And note that the model chosen is C-SVC, and kernel function type is RBF, corresponding to the model parameters '$-s$ 0' and '$-t$ 2'.

By means of grid searching, we record the best c and best g and obtained the classification model with the best performance. The result of parameter optimization is demonstrated in Figure 4. It means that when the parameter c is assigned to 5.278 and g is assigned to 0.10882, the classification model is optimal and the best accuracy is 82.7381%. The accuracy is calculated by the following formula

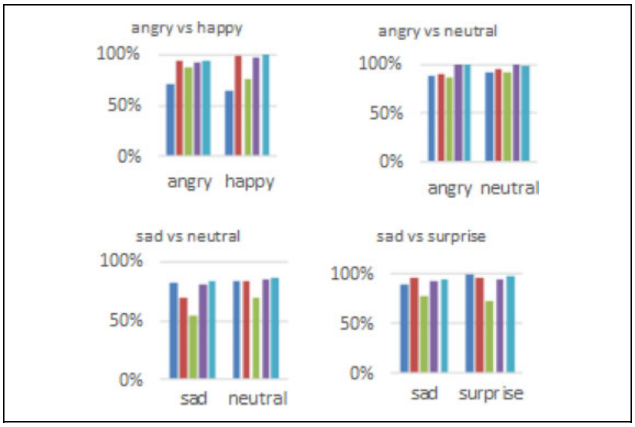

Figure 5. The two-class recognition accuracy of the voice on different feature set combinations. Each series from left to right in the four figure represents: MFCC, prosodic feature, voice quality feature, MFCC + prosodic feature + voice quality and MFCC + prosodic feature + voice quality + cepstral distance. And the ordinate represents the recognition accuracy and the abscissa stands for emotions. Different colours show different feature set combinations. MFCC: mel-frequency cepstral coefficient.

$$\text{Accuracy} = \sum_{i=1}^{m} a_i \bigg/ \sum_{j=1}^{n} t_j \tag{3}$$

where the numerator is the total number of certain emotional sentence predicted by classification and the denominator is that total number in the test set we acquired in advance.

Two-class recognition results

Figure 5 shows the two-class recognition results on feature sets.

Accuracy is high enough in two-class emotions recognition. In two-class recognition, different feature sets such as MFCC, prosodic feature, voice quality feature, MFCC + prosodic feature + voice quality and MFCC + prosodic feature + voice quality + cepstral distance are applied. Among these feature sets, the performance of the last feature set (MFCC + prosodic feature + voice quality + cepstral distance) is the best. In the comparison of feature sets MFCC + prosodic feature + voice quality and MFCC + prosodic feature + voice quality + cepstral distance, the latter is higher than the former by 3% in each class pair. The average recognition rate of the latter is higher than 91%. Except emotion pair sad and neutral, the other three groups are recognized with more than 95% and some even up to 100%. For emotion angry and happy which are hard to distinguish, the result is still good on the latter feature set which combines cepstral distance.

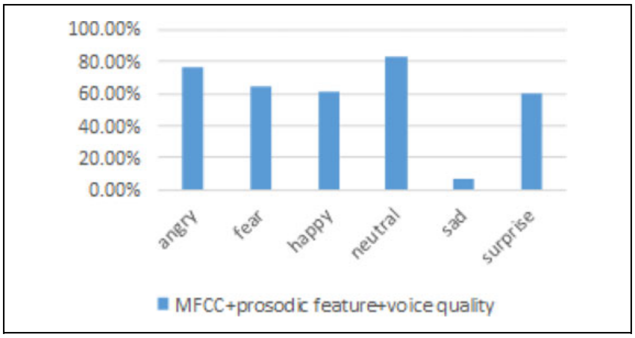

Figure 6. Six-class recognition accuracy of the voice on different feature set combinations.

Table 1. Three emotions classification results.

Emotion	Accuracy (%)	Best *c* and *g*
Anger	82.54	$c = 48.5029$
Neutral	63.67	$g = 0.1088$
Emotion mixture	86.18	

Table 2. Four emotions classification results.

Emotion	Accuracy (%)	Best *c* and *g*
Fear	81.54	$c = 3.0314$
Happy	83.61	$g = 0.0068$
Sad	87.14	
Surprise	62.50	

However, it is necessary to discuss the situation in multi-class recognition on emotional speech.

Multiple class recognition results

The recognition results on feature set MFCC + prosodic + quality are shown in Figure 6. The imbalance is obviously in the results. So the cepstral distance is added. Meanwhile, we apply the two-group classification strategy. The six emotions are sorted into two categories: one contains angry and neutral and another includes the rest emotions (fear, happy, sad, surprise) which are, respectively, corresponding to prosodic feature set and MFCC + prosodic + quality + distance feature set. In the implementation process, firstly, extracting prosodic features of the six emotions, meanwhile generating labels. Notice that the labels in the first stage are divided into three categories: angry utterances are marked as one, neutral utterances are marked as two and the rest of the utterances are labelled as three. In the second stage, the utterances, which are labelled as three in the first stage, are classified. The process of this method is showed in Figure 3 and its result is expounded in Tables 1 to 3. And the best *c* and *g* parameters are shown in these tables.

Table 3. Final classification result (six emotions).

Emotion	Accuracy (%)	Best *c* and *g*
Anger	85.07	$c = 9.1896$
Fear	60.00	$g = 0.0625$
Happy	49.12	
Neutral	64.41	
Sad	78.67	
Surprise	46.67	

Table 4. Compare with the study by Wu and Liang.[17]

Emotion	Accuracy (%) (Reference 17)	Accuracy (%) (results in this article)
Neutral	72.98	77.62
Happy	78.81	88.89
Sad	81.01	71.67
Angry	79.83	90.14

The result is acceptable in the first recognition (Table 1). The accuracy of the mixture of emotions (i.e. four emotions) is 86.18% that lay a solid foundation for the second part of the recognition. In regard to the second recognition, the correct rate is also considerable. Hence, the final results for six emotions classification are balanced and they are better than these of above-mentioned feature combinations. This is mainly due to the addition of cepstral distance.

When Tables 1 to 3 are analysed, that is, the accuracy of ‘emotions mixture’ is 86.18% in Table 1 and, in Table 2, the results of the emotions belonging to ‘emotions mixture’ are shown, there may arise a question: for example, why the final result of emotion fear (60%) is not equal to the result that the accuracy of ‘emotions mixture’ (86.18%) multiplied by the accuracy of ‘fear’ in Table 2 (81.54%)? This is because the train sets in Tables 1 to 3 are selected randomly. So the recognition rate in the three tables represents the ability of the model rather than the results for a single test.

From the above, we can conclude that cepstral distance feature is efficient to distinguish emotion fear, happy, neutral, sadness and surprise. Especially for emotion sadness, its ability is outstanding among the present feature set. Besides, it is conducive for the improvement and balance of recognition rate to incorporating several specific feature sets according to the pertinence of different feature set. And to be fair discussion, the results of this method on Berlin on German database are listed in Tables 4 to 6. And they will be analysed in the discussion part.

Comparison and discussion

Due to different experimental conditions, completely fair is not to exist,[29] but we can still conclude from these comparisons. Table 4 shows the results of this method to identify four emotions on Chinese Mandarin emotional speech database which is compared with the study by Wu and Liang.[17]

Table 5. Compare with the study by Paeschke et al.[41] (no silence).

Emotion	Accuracy (%) (in the study by Paeschke et al.[41])	Accuracy (%) (results in this article)
Neutral	None	93.55
Happy	71.80	90.63
Sad	68.40	77.20
Angry	90.90	73.91

Table 6. Compare with the study by Xiao et al.[19] for multi-group classification.

Emotion pair	Accuracy (%) (in the study by Xiao et al.[19])	Accuracy (%) (results in this article)
Neutral versus Fear	83.84	96.67
Happy versus Angry	70.55	73.47
Sad versus Boring	82.12	95.45

For multiple classifiers comparison as shown in Table 4, the results after adding cepstral distance are significantly better than the results in the study by Wu and Liang.[17] The second column of Table 5 displays the results in the same conditions except deleting all the silence in the speech which is superior to the results in Table 4. In our experiment, the average accuracy with no silence is 59.0% for six emotions and is higher than the accuracy with silence, which is consistent with the study by Zheng et al.[42] in the trend.

Furthermore, for the comparison of multi-group emotion classification, the results of this method are given in Table 6. Unlike most works in the literature which mainly rely on classical frequency and energy-based features along with a single global classifier for emotion recognition, Xiao et al.[19] propose some new harmonic and Zipf-based features for better speech emotion characterization in the valence dimension. The Zipf law is an empirical law proposed by Zipf.[43] Zipf features is a kind of linguistic features which characterize rhythmic and prosodic aspects of vocal expressions. And a multi-stage classification scheme driven by a dimensional emotion model for better emotional class discrimination is applied in the study by Xiao et al.[19] Xiao et al.[19] explain and cite multi-stage classification from the perspective of the valence dimension of feature. In this article, emotions, which are easily confused, are acquired by adaptive learning. Another difference compared with the study by Xiao et al.[19] is that feature in this article is designed to measure the similarity between emotion, and the study by Xiao et al.[19] is raised harmonic and Zipf-based features. Each pair in Table 6 locates in the same arousal-valence space[4]; therefore, they are easily confused and this confusion can be seen everywhere in the experiment.[4,32,34,41] The validity of that hierarchical classification is proved. And the results in this article are competitive.

Conclusion and future work

As the primary carrier in robot to communicate with people, emotional speech is critical for robot socialization. Emotion recognition affects decision-making and learning process of the robot. In this article, a combining cepstral distance features two-group multi-class classification for emotional speech is proposed. The classification is based on the Chinese Mandarin speech database. By entering a different feature set, reducing the dimensionality of features and optimizing parameters *c* and *g*, the model is built by DAG-SVM. On the circumstance of well-behaved two-class classification, we expound the multi-class classification on six emotions. Adding cepstral distance features improves the extreme low recognition rate for emotion sad and balances the combination of the feature set. In addition to reduce bad influence between feature sets, we decompose the problem of multi-class classification into the problem of tri-classification and four emotions classification. By fusing two models, the recognition rate for six emotion identification is enhanced.

Our future work mainly includes two parts. The first part is to improve and evaluate our combined cepstral distance model-based emotional speech recognition by combining linguistic-based emotion recognition[44] and affective dialogue management.[45] The second part is to develop a useful emotion-sensitive human–computer interaction system with intelligent functions for some real applications such as nursing home robots and intelligent tutoring system.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was partially supported by National Natural Science Foundation of China under grant 61472117.

References

1. Shahamiri SR and Binti Salim SS. Artificial neural networks as speech recognisers for dysarthric speech: identifying the best-performing set of MFCC parameters and studying a speak-er-independent approach. *Adv Eng Inf* 2014; 28(1): 102–110.
2. Planet S and Iriondo I. Children's emotion recognition from spontaneous speech using a reduced set of acoustic and linguistic features. *Cogn Comput* 2013; 5(4): 526–532.

3. Hasrul MN and Hariharan M. Human affective (Emotion) behavior analysis using speech signals: a review. In: *Proceedings of 2012 International conference on biomedical engineering*, 27–28 February 2012, pp. 217–222. IEEE.
4. Ramakrishnan S, Ibrahiem MM and Emary El. Speech emotion recognition approaches in human computer interaction. *Telecommun Syst* 2013; 52(3): 1467–1478.
5. Gharavian D, Sheikhan M, Nazerieh A, et al. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. *Neural Comput Appl* 2011; 21(8): 1–12.
6. Gobl C and Ní Chasaide A. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun* 2003; 40(1-2): 189–212.
7. Yang B and Lugger M. Emotion recognition from speech signals using New Harmony features. *Signal Process* 2010; 90(5): 1415–1423.
8. Li M, Han KJ and Narayanan S. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Comput Speech Language* 2013; 27(1): 151–167.
9. Lugger M and Yang B. Psychological motivated multi-stage emotion classification exploiting voice quality features. In: Mihelic F and Zibert J (eds) *Speech Recognition*. InTech, 2008. DOI: 10.5772/6383.
10. Wei W, Wu C, Wu CH, et al. Exploiting psychological factors for interaction style recognition in spoken conversation. *IEEE/ACM Trans Audio Speech Language Process* 2014; 22(3): 659–670.
11. Chen L, Mao X, Wei P, et al. Mandarin emotion recognition combining acoustic and emotional point information. *Appl Intell* 2012; 37(4): 602–612.
12. Tawari A and Trivedi MM. Speech emotion analysis: exploring the role of context. *IEEE Trans Multimedia* 2010; 12(6): 502–509.
13. Wöllmer M, Schuller B, Eyben F, et al. Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. *IEEE J Sel Topics Signal Process* 2010; 4(5): 867–879.
14. Johar S. Paralinguistic profiling using speech recognition. *Int J Speech Technol* 2014; 17:1–5. DOI: 10.1007/s10772-013-9222-4.
15. Drolet M, Schubotz RI and Fisch-er J. Explicit authenticity and stimulus features interact to modulate BOLD response induced by emotional speech. *Cogn Affect Behav Neurosci* 2013; 13: 318–329.
16. Caponetti L, Buscicchio CA and Castellano G. Biologically inspired emotion recognition from speech. *EURASIP J Adv Signal Process* 2011; 1: 1–10.
17. Wu CH and Liang WB. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Trans Affect Comput* 2011; 2(1): 10–21.
18. Milton A and Tamil Selvi S. Class-specific multiple classifiers scheme to recognize emotions from speech signals. *Computer Speech Language* 2014; 28: 727–742.
19. Xiao Z, Dellandrea E, Dou W, et al. Multi-stage classification of emotional speech motivated by a dimensional emotion model. *Multimedia Tools Appl* 2010; 46: 119–145.
20. Tohkura Y. A weighted cepstral distance measure for speech recognition. *IEEE Trans Acoust Speech Signal Processing* 1986; 35(10): 761–764.
21. Chen P and Liu S. An improved DAG–SVM for multi–class classification. In: *Proceedings of Fifth international conference on natural computation*, Piscataway, NJ, USA, 14–16 August 2009, pp. 460–462. IEEE.
22. Broekens J. Emotion and reinforcement: affective facial expressions facilitate robot learning. *Lect Notes Comput Sci* 2007; 4451: 113–132.
23. Berridge KC. Pleasures of the brain. *Brain Cogn* 2003; 52(1): 106–128.
24. Yilmazyildiz S, Henderickx D, Soetens E, et al. Multi–modal emotion expression for affective human–robot interaction. In: *Proceedings of 2013 workshop on affective social speech signals*, 22–23 August 2013, pp.1–5.
25. Albornoz EM, Milone DH and Ru-finer HL. Spoken emotion recognition using hierarchical classifiers. *Computer Speech Language* 2011; 25: 556–570.
26. Iliev AI, Scordilis MS, Papa JP, et al. Spoken emotion recognition through optimum-path forest classification using glottal features. *Computer Speech Language* 2010; 24: 445–460.
27. Busso C, Mariooryad S, Metal-linou A, et al. Iterative feature normalization scheme for auto-matic emotion detection from speech. *IEEE Trans Affect Comput* 2013; 4(4): 386–397.
28. Polzehl T, Schmitt A, Metze F, et al. Anger recognition in speech using acoustic and linguistic cues. *Speech Commun* 2011; 53: 1198–1209.
29. Sheikhan M, Bejan M and Gharavian D. Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. *Neural Comput Appl* 2013; 23(1): 215–227.
30. Xu Y, Du J, Dai L, et al. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Process Lett* 2014; 21(1): 65–68.
31. Yu B. Speech emotion recognition based on optimized support vector machine. *J Softw* 2012; 4(4): 2726–2733.
32. Ntalampiras S and Fako-takis N. Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *IEEE Trans Comput* 2012; 3(1): 116–125.
33. Ooi CS, Seng KP, Chew LW, et al. A new approach of audio emotion recognition. *Expert Syst Appl* 2014; 41: 5858–5869.
34. Wu S, Falk TH and Chan W. Automatic speech emotion recognition using modulation spectral features. *Speech Commun* 2011; 53: 768–785.
35. Koolagudi SG and Rao KS. Emotion recognition from speech using source, system, and prosodic features. *Int J Speech Technol* 2012; 15: 265–289.
36. Xiao Z, Dellandrea E, Dou W, et al. Features extraction and selection for emotional speech classification. In: *Proceedings of 2005 IEEE conference on advanced video and signal based surveillance*, 15–16 September 2005, pp. 411–416. IEEE.

37. Yeh JH, Pao TL, Lin CY, et al. Segment-based emotion recognition from continuous Mandarin Chinese speech. *Comput Hum Behav* 2011; 27: 1545–1552.
38. Sheikhan M, Gharavian D and Ashoftedel F. Using DTW neural–based MFCC warping to improve emotional speech recognition. *Neural Comput Appl* 2012; 21: 1765–1773.
39. Bozkurt E, Erzin E, Erdem CE, et al. Formant position based weighted spectral features for emotion recognition. *Speech Commun* 2011; 53: 1186–1197.
40. Chen L, Mao X, Xue Y, et al. Speech emotion recognition: features and classification models. *Digit Signal Process* 2012; 22: 1154–1160.
41. Paeschke BF, Rolfes A, Sendlmeier M, et al. A database of German emotional speech. In: *Proceedings of interspeech 2005, international speech communication association*, Lisboa, Portugal, 4–8 September 2005, pp. 1517–1520.
42. Zheng W, Xin M, Wang X, et al. A novel speech emotion recognition method via incomplete sparse least square regression. *IEEE Signal Process Lett* 2014; 21(5): 569–572.
43. Zipf GK. *Human behavior and the principle of least effort: an introduction to human ecology*. Cambridge, Massachusetts: Addison-Wesley Press, 1949.
44. Quan C and Ren F. Weighted high-order hidden Markov models for compound emotions recognition in text. *Inf Sci* 2016; 329: 581–596.
45. Ren F, Wang Y and Quan C. A novel factored POMDP model for affective dialogue management. *J Intell Fuzzy Syst* 2016; 31: 127–136.