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

Speaker-independent speech emotion recognition by fusion of functional and accompanying paralanguage features*

Qi-rong MAO†, Xiao-lei ZHAO, Zheng-wei HUANG, Yong-zhao ZHAN

(*Department of Computer Science and Communication Engineering*, *Jiangsu University*, *Zhenjiang 212013, China*)

†E-mail: mao_qr@ujs.edu.cn

Received Dec. 29, 2012; Revision accepted May 12, 2013; Crosschecked June 6, 2013

Abstract: Functional paralanguage includes considerable emotion information, and it is insensitive to speaker changes. To improve the emotion recognition accuracy under the condition of speaker-independence, a fusion method combining the functional paralanguage features with the accompanying paralanguage features is proposed for the speaker-independent speech emotion recognition. Using this method, the functional paralanguages, such as laughter, cry, and sigh, are used to assist speech emotion recognition. The contributions of our work are threefold. First, one emotional speech database including six kinds of functional paralanguage and six typical emotions were recorded by our research group. Second, the functional paralanguage is put forward to recognize the speech emotions combined with the accompanying paralanguage features. Third, a fusion algorithm based on confidences and probabilities is proposed to combine the functional paralanguage features with the accompanying paralanguage features for speech emotion recognition. We evaluate the usefulness of the functional paralanguage features and the fusion algorithm in terms of precision, recall, and F1-measurement on the emotional speech database recorded by our research group. The overall recognition accuracy achieved for six emotions is over 67% in the speaker-independent condition using the functional paralanguage features.

Key words: Speech emotion recognition, Speaker-independent, Functional paralanguage, Fusion algorithm, Recognition accuracy
doi:10.1631/jzus.CIDE1310 **Document code:** A **CLC number:** TP391.4

1 Introduction

Speech emotion recognition plays an important role in the development of human-computer interaction (HCI). In recent years, studies on speaker-dependent speech emotion recognition have achieved considerable success. However, under the condition of natural human-machine interaction, the accuracy of speaker-independent emotion recognition needs to be further improved. Moreover, current studies mostly focus on the speaker's accompanying paralanguage, and overlook other valuable information, such as functional paralinguistic information. The accompanying paralanguages are the non-verbal acoustic features that accompany speech and help convey meaning, such as pitch, loudness of a sound, and speed of speech. The functional paralanguage features are the sudden voice phenomenon in the paralanguages, such as laughter, cry, and sigh. The functional paralanguage and the accompanying paralanguage including emotional information are both called emotional paralanguage.

Ishi *et al.* (2008) indicated that the functional paralinguistic information also conveys important meanings in communication. Li (2004) pointed out that paralanguage is more authentic than language. Many research results indicated that the functional paralanguage research is meaningful and necessary (Li, 2004; Kleckova, 2009).

* Project supported by the National Natural Science Foundation of China (Nos. 61272211 and 61170126), the Natural Science Foundation of Jiangsu Province (No. BK2011521), and the Research Foundation for Talented Scholars of Jiangsu University (No. 10JDG065), China

Nowadays, with the development of artificial intelligence and interactive techniques, there have been many studies on the classification of functional paralinguistic information. The classifications contain expressing intentions, attitudes, and emotions (Ishi *et al.*, 2008). For example, Fujie *et al.* (2003) reported that functional paralinguistic information was used to distinguish positive and negative attitudes. Maekawam (2004) studied the classification of functional paralinguistic items, such as admiration, suspicion, disappointment, and indifference. Hayashi (1999) analyzed functional paralinguistic items like affirmation, asking again, doubt, and hesitation. Additionally, functional paralinguistic information covers some sound phenomena, such as laughter, cry, and sigh. Functional paralinguistic information has been gradually introduced into the emotion research (Devillers and Vidrascu, 2006; Jones and Jonsson, 2008; Kleckova, 2009). Among these studies, laughter is extensively studied, from its feature extraction (Ishi *et al.*, 2006; Szameitat *et al.*, 2007; Bachorowski *et al.*, 2011) to automatic detection (Kennedy and Ellis, 2004; Truong and van Leeuwen, 2005; 2007; Petridis and Pantic, 2008; Li and He, 2011). Some studies focus on human-like laughter (Sundaramb and Narayananc, 2007). In addition to laughter, there are other functional paralinguistic items; for example, Matos *et al.* (2006), Pal *et al.* (2006), and Huq and Moussavi (2012) studied cough, cry, and breath, respectively. Thus far, there is no systematic research concerning the use of these functional paralanguages to improve the emotion recognition accuracy, though some papers postulated that functional paralanguages may contribute to emotion recognition. Therefore, this paper is focused on the contribution of these functional paralanguages to emotion expression. This paper deals with the association of functional paralinguistic information with the accompanying paralanguage, to achieve higher recognition accuracy and more robust speaker-independent recognition.

Another problem to be solved is how to fuse the recognition results of the functional paralanguages and to those of the accompanying paralanguages. At present, there are many mature fusion algorithms that are widely used in multiple classifier combinations, such as Dempster-Shafer (DS) evidence theory, fuzzy logic method, weighted fusion algorithm, Bayes fusion method, and the method of voting; however, they all have limitations. If there are conflicts among evidence, DS evidence theory may lead to wrong results (Li *et al.*, 2002). Uncertain problems can be dealt with using the fuzzy logic method to simulate the human reasoning process with inference rules. However, many problems do not have deductive and rigorous reasoning processes or absolute conclusions. In terms of weighted fusion, whether the weight is reasonable directly affects the results, and it is difficult to assess whether the weight is determined reasonably. The Bayes fusion method has the disadvantage of requiring priori knowledge (Berler and Shimony, 1997). The voting method is simple and effective; in our work, however, there are just two channels, and thus the final result cannot be obtained easily using the voting method.

El Ayadi *et al.* (2011) presented detailed surveys on speech emotional databases, but did not mention emotional paralanguage databases. There are some corpuses including only a handful of functional paralanguages. For example, Truong and van Leeuwen (2007) used the database extracted from the International Computer Science Institute (ICSI) Meeting Recorder Corpus data, including mainly laughter, and Matos *et al.* (2006) used the database obtained by recording separate patients and selected events from ambulatory recordings, aiming at cough. Other databases used by researchers also focus on one or two types of functional paralanguages. To date, there is no comprehensive and uniformed emotional paralanguage database.

There are three main contributions in this paper. The first is the speech emotion database containing the emotional paralanguages. The second is that the functional paralanguage is put forward to recognize the emotion of speech. The third is that the paralanguage fusion recognition algorithm (PFRA) combining the functional paralanguages and the accompanying paralanguages is proposed for speech emotion recognition. According to the study requirements, we have recorded one speech emotion database, emotional paralanguages contained speech emotion database (EPSED) containing the functional paralanguages, which covers cry, laughter, doubt, shout, and sigh, widely used in daily life. Specifically, a fusion algorithm is proposed to fuse the functional paralanguages and the accompanying paralanguages. In this algorithm, the functional paralanguage is adopted to assist the speech emotion recognition. Functional paralanguage features and accompanying paralanguage features complement each other and provide more emotion information to classifiers. This increases the speech emotion recognition accuracy. Moreover, as functional paralanguages are insensitive to the change of speakers, the generalization performance of the proposed method is better. Results of speaker-independent speech emotion recognition experiments on EPSED showed that, compared with the single-channel speech emotion recognition method not fusing functional paralanguages, linear weighted method, DS evidence theory, and Bayes fusion method, the proposed fusion algorithm is superior in accuracy and stability.

2 Emotional paralanguage contained speech emotion database

In our daily life, there are multiple functional paralanguages embodying the same emotion; for example, 'laugher' and 'cheer' both embody happiness. In our database, to simplify the relationship, each emotional utterance contains only one kind of functional paralanguage. The relationships between the functional paralanguages and the six typical emotions in our database EPSED are presented in Table 1.

Table 1 Relationships between functional paralanguages and typical emotions

Functional paralanguage	Happiness	Sadness	Surprise	Anger	Fear	Disgust
Laughter	Y	N	N	N	N	N
Sad cry	N	Y	N	N	N	N
Fear cry	N	N	N	N	Y	N
Doubt	N	N	Y	N	N	N
Shout	N	N	N	Y	N	N
Sigh	N	N	N	N	N	Y

Y (N) represents that the functional paralanguage is (not) related with the emotion. One functional paralanguage corresponds to one typical emotion. Sad cry and fear cry have different characteristics, so they correspond to different typical emotions; for example, the energy of sad cry is low while the energy of fear cry is high, and the speed of sad cry is slow while the speed of fear cry is high. Thus, we regard them as different functional paralanguages

The emotional database recorded includes six typical emotions: happiness, sadness, surprise, anger, fear, and disgust, and six kinds of functional paralanguages: laughter, sad cry, fear cry, doubt, shout, and sigh. Each emotion in the database contains 15 scripts. Functional paralanguages were embedded naturally into speech utterances according to the context; for example, happiness is often accompanied by laughter, while sadness is often accompanied by cry. The utterances in this database were recorded by five male and six female actors. Emotional utterances were appropriately embedded with the corresponding functional paralanguages. The recording tools we used included a microphone, a computer, and the Cool Edit Pro software. The EPSED was recorded using the standards proposed by Huang *et al.* (2010). The sampling frequency was 11025 Hz and the samples were stored in wav format.

3 Framework of the fusion method

The method consists mainly of three parts (Fig. 1): the functional paralanguage feature channel, the accompanying paralanguage feature channel, and the fusion algorithm. Both channels contain three components, feature extraction, recognition, and recognition results. These two channels are fused by the fusion algorithm proposed in this study.

4 Functional and accompanying paralanguages fusion algorithm

In the algorithm, both the probability and the confidence are considered.

4.1 Calculation method of confidence

In the fusion algorithm, the calculation of confidence contains two parts of dynamics information: one is based on the decision distance, and the other on the class probability. Since the support vector machine (SVM) is selected as the classifier, the calculation of the distance confidence is based on the one-versus-one SVM classifier.

4.1.1 Confidence calculation method based on classification distance

Assuming there are N' training samples and N emotions, the optimal distances of sample x_i to be

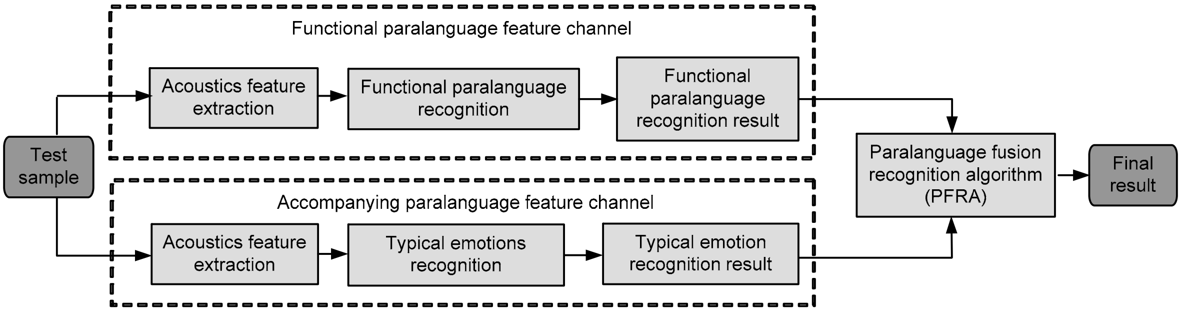
Functional paralanguage feature channel
Acoustics feature extraction
Functional paralanguage recognition
Functional paralanguage recognition result
Test sample
Paralanguage fusion recognition algorithm (PFRA)
Final result
Accompanying paralanguage feature channel
Acoustics feature extraction
Typical emotions recognition
Typical emotion recognition result

Fig. 1 Recognition process fusing the functional paralanguages and the accompanying paralanguages

identified to the SVM classification surface is given as follows:

$$d(x_i)=\frac{\operatorname{sgn}(wx+b)}{|w|}=\frac{\operatorname{sgn}\left(\sum_{j=1}^{N'} y_j\alpha_j K(x_j,x_i)+b\right)}{|w|}, \tag{1}$$

where $w=\sum_{j=1}^{N'}\alpha_j y_j x_j$, $y_j\in\{\pm1\}$, α_j denotes the Lagrange coefficient, b denotes the intercept, and $K(x_j, x_i)$ is the radial basis function (RBF) kernel function. If sample x_i is the support vector, then the distance is the maximum distance of all samples to the SVM classification surface. If x_i is a sample to be identified, the distance is the actual distance of the sample to the SVM classification surface.

It is known that multi-SVMs are based on the vote mechanism. The more the valid votes one class obtains, the more credible it is, and the higher the confidence of this class. The result of multi-SVMs is determined by the number of valid votes of $N(N-1)/2$ one-versus-one SVM, and the class that obtains the most votes is judged as the final result. A one-versus-one SVM chooses one class each time to vote. In our experiments, if the determination distance of this class is greater than a certain threshold t, the vote is credible, and it is remarked as the valid vote. Namely, when $d(x_i)>t$, the vote is valid, where t is the average of all distances. Then the distance confidence of sample x_i belonging to emotion E_j is defined as

$$\mathrm{conf}_{i,j}=v_{i,j}/V_{i,j}, \tag{2}$$

where $v_{i,j}$ and $V_{i,j}$ denote the numbers of valid votes and all votes of sample x_i belonging to emotion E_j, respectively.

4.1.2 Confidence calculation based on output probability

If there are N emotions and n samples, and the samples are identified by the classifiers, the probability matrix M_p will be obtained as follows:

$$M_\mathrm{p}=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,N} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,N} \\ \vdots & \vdots & & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,N} \end{bmatrix}, \tag{3}$$

where $p_{i,j}$ denotes the probability of sample x_i belonging to emotion E_j. The probability confidence of sample x_i belonging to emotion E_j can then be calculated by

$$\mathrm{conf}'_{i,j}=p_{i,j}-\frac{1}{N-1}\sum_{\substack{k=1\\k\neq j}}^{N}p_{i,k}. \tag{4}$$

It is clear that the higher the maximum probability, the higher the confidence of the sample belonging to the class.

The final confidence of sample x_i belonging to emotion E_j is

$$\mathrm{Tconf}_{i,j}=\frac{\mathrm{nconf}_{i,j}+\mathrm{nconf}'_{i,j}}{2}, \tag{5}$$

where $\mathrm{nconf}_{i,j}$ and $\mathrm{nconf}'_{i,j}$ denote the normalized $\mathrm{conf}_{i,j}$ and $\mathrm{conf}'_{i,j}$, respectively.

4.2 Fusion recognition algorithm

To make the fusion algorithm more general, the condition of multiple functional paralanguages corresponding to one typical emotion is taken into consideration. Some symbols used in this fusion algorithm are defined as follows: ParaChannel denotes the functional paralanguage feature channel, and SpeechChannel represents the accompanying paralanguage feature channel. $X=\{x_1, x_2, \ldots, x_n\}$ represents the test sample set to be identified, and the results are saved into array R that marks which emotional class of each test sample belongs to after samples are recognized. PP_i, SP_i, Pconf_i, and Sconf_i denote the output probability sets and confidence sets of sample x_i in ParaChannel and SpeechChannel, respectively. η_p and η_s denote probability thresholds for ParaChannel and SpeechChannel, respectively. E_i denotes the ith emotion state, and P_i denotes the ith functional paralanguage. Then, detailed procedures of the fusion algorithm are described as follows. PC and SC are used to store the emotion labels of test samples recognized in the functional paralanguage channel and the accompanying paralanguage channel, respectively. The fusion algorithm is as shown in Algorithm 1.

In Algorithm 1, Vote is a set of times of each emotion class appearing in the two channels. For each emotion in Vote, a vote higher than 1 means that the recognition results for this sample of the two channels are overlapping. If the vote of the emotion class in set Vote is equal to or greater than 2, the new probability and the new confidence of this emotion class are determined as the average values of the probabilities and the confidences of this class in the two channels, respectively. For the emotion class in set Vote, whose vote is equal to 1, the new probability and the new confidence of this emotion class are unchanged, keeping the original probability and confidence of the class in the channels they come from. Then the new probability, confidence, vote of candidate emotions are saved into set PCv_i. Firstly, the elements in PCv_i are sorted by a descending order of their votes. Then, for the elements with the same vote, they are sorted by a descending order of their probabilities. Denote $\mathrm{PCv}_i=\{(\mathrm{np}_1, \mathrm{nc}_1, v_1, E_1), (\mathrm{np}_2, \mathrm{nc}_2, v_2, E_2), \ldots, (\mathrm{np}_m, \mathrm{nc}_m, v_m, E_m)\}$, where $(\mathrm{np}_k, \mathrm{nc}_k, v_k, E_k)$ $(k=1, 2, \ldots, m)$ denotes that the new probability of sample x_i belonging to emotion class E_k is equal to np_k, the confidence is nc_k, and the vote is v_k. It is assumed that v_max denotes the maximum number of votes in set PCv_i. The emotion classes in set PCv_i are checked in order, starting from the emotion with the maximum number of votes and the highest probability.

Algorithm 1 Functional and accompanying paralanguages fusion recognition algorithm (PFRA)

Input: test sample set $X=\{x_i|i=1, 2, \ldots, n\}$.

Output: recognition results of the samples $R(N)$.

Definition:

```
i=0;
while i<n do
  i=i+1;
// Step 1: Obtain the recognition probabilities of two
           channels for test sample $x_i$
  Put $x_i$ into PM and SM, and obtain $\mathrm{PP}_i=\{\mathrm{pp}_1, \mathrm{pp}_2, \ldots, \mathrm{pp}_i, \ldots, \mathrm{pp}_P\}$ and $\mathrm{SP}_i=\{\mathrm{sp}_1, \mathrm{sp}_2, \ldots, \mathrm{sp}_i, \ldots, \mathrm{sp}_N\}$;
  Remark classes in $\mathrm{PP}_i$ using corresponding traditional
      emotional classes $\{E_1, E_2, \ldots, E_P\}$;
// Step 2: Obtain the final recognition result of test sample $x_i$;
  $\mathrm{pp}_k=\max(\mathrm{PP}_i)$, $\mathrm{sp}_h=\max(\mathrm{SP}_i)$;
  if $k==h$ then
    $R(i)=E_k$;
    continue;
  else
    Calculate the confidences of sample $x_i$ in ParaChannel
      and SpeechChannel using Eq. (5), denoted as
      $\mathrm{Pconf}_i=\{\mathrm{pconf}_{i,1}, \mathrm{pconf}_{i,2}, \ldots, \mathrm{pconf}_{i,P}\}$ and
      $\mathrm{Sconf}_i=\{\mathrm{sconf}_{i,1}, \mathrm{sconf}_{i,2}, \ldots, \mathrm{sconf}_{i,N}\}$, respectively;
    Set thresholds $\eta_\mathrm{p}$, $\eta_\mathrm{s}$ for ParaChannel, SpeechChannel
      by the average of channel output probabilities,
      respectively;
    if $\mathrm{pconf}_{i,j}>\eta_\mathrm{p}$ then
      Save $E_j$ in set PC;
    end
    if $\mathrm{sconf}_{i,j}>\eta_\mathrm{s}$ then
      Save $E_j$ in set SC;
    end   // retain the classes having higher probability
    Receive emotion candidates $\mathrm{PC}=\{E_1, E_2, \ldots, E_s\}$, $s<P$,
      and $\mathrm{SC}=\{E_1, E_2, \ldots, E_{s'}\}$, $s'<N$;
    Check the times $v_i$ of emotion classes in PC and SC;
    Sort results by in descending order, denoted as
      $\mathrm{Vote}=\{(v_1, E_1), (v_2, E_2), \ldots, (v_m, E_m)\}$, $m\le N$;
    Calculate $\mathrm{PCv}_i$;
    $j=v_\mathrm{max}$;
    while $j>1$ do
      Calculate the average values $\mathrm{TP}_j$, $\mathrm{TC}_j$ of the
        probability and the confidence of emotion
        classes with vote $j$;
      Check $\mathrm{PCv}_i$ in descending order;
      if $v_k==j$ and $\mathrm{np}_k\ge\mathrm{TP}_j$ and $\mathrm{nc}_k\ge\mathrm{TC}_j$ then
       // There are emotion classes with vote j in $\mathrm{PCv}_i$.
        $R(i)=E_k$; break;
      end
      $j=j-1$;
    end
    if $j==1$ then
      if $\mathrm{sp}_l=\max(\mathrm{SP}_i)$ then
        $R(i)=E_l$;
      end
    end
  end
end
```

Algorithm 1 shows the pseudo code of PFRA, which contains two steps:

1. Obtain the recognition probabilities of two channels for test sample x_i. In this step the output vectors of two channels are obtained for the test sample, including feature extraction, classification of the functional paralanguages, and classification of the typical emotions.

2. Obtain the final recognition result of test sample x_i. In this step the final recognition result of the test sample is obtained, including the calculation of the distance confidence and the probability confidence, and then the confidence of the output of the two channels is calculated. Finally, the emotion with the highest confidence is selected as the emotion to which the test sample belongs.

It can be seen from the algorithm that PFRA is suitable for such cases as several functional paralanguages corresponding to one typical emotion and one functional paralanguage corresponding to one typical emotion.

5 Experimental

Since Matlab facilitates the processing of the digital signal, and Visual Studio (VS) facilitates the processing cycle and other operations, we adopt the mixed programming of Matlab 7.0 and VS 2005, which can improve the system performance. The SVM classifier is used to train and recognize samples. The SVM model is achieved by VS, and feature extraction and integration of operations are implemented in Matlab. VS uses the Matlab engine to call Matlab programs.

5.1 Experimental setup

5.1.1 Experiment datasets

Our experiments were set up on the speaker-independent datasets selected from the database EPSED recorded by our research group, as described in Section 2. Based on EPSED, two training sets and one testing set were sorted out. One training dataset was collected for the functional paralanguage feature channel to train the functional paralanguage recognition classifiers, recorded as Tp; similarly, the other was collected for the accompanying paralanguage feature channel to train the typical emotion classifiers, recorded as Ts. The testing dataset was used for testing the model performance, recorded as Te. Tp and Ts both contained 540 samples, which were recorded by three male and two female actors. Dataset Tp contained six emotional paralanguages, and each sample contained one kind of functional paralanguage. Ts included six typical emotions with some samples containing functional paralanguages, while Te had 240 test samples recorded by another three male and two female actors, with some samples containing functional paralanguages. This phenomenon is called the ‘paralanguage coverage’. There were certain degrees of paralanguage coverages in Ts and Te. In our experiments, 50% samples were text-dependent between training and testing sets.

5.1.2 Model training

The SVM classifiers in the functional paralanguage channel and the accompanying paralanguage channel were trained using Tp and Ts, respectively. For the training sets, 101-dimensional features (Mao *et al.*, 2010) were extracted. After selecting features and adjusting the values of γ and C of SVM, the best models of the functional paralanguage channel and the accompanying channel were obtained. The values of γ and C were determined by the method in Huang and Wang (2006). For feature selection, we adopted the sequential floating forward search (SFFS) method (Pudil *et al.*, 1994).

5.2 Results and analyses

5.2.1 Analysis of confidence calculation methods

The paralanguage coverage was set to 40% for Ts and Te. When γ=0.01 and C=8.5, and the dimension of features was reduced to 60, the average recognition accuracy of the functional paralanguage model reached the highest value. For the accompanying paralanguage model, when γ=0.06 and C=5, and the dimension of features was reduced to 60, its recognition accuracy reached the highest value. These parameters were used in the following experiments.

To analyze the confidence calculation method of PFRA, we conducted the experiments adopting the confidence calculation methods described in Section 4.1 (Fig. 2). In Fig. 2, Eqs. (2), (4), and (5) refer to the confidence calculation methods shown in Eqs. (2), (4), and (5) of Section 4.1, respectively, and SERBF denotes the single-channel based speech emotion recognition method, which used only the accompanying paralanguages.

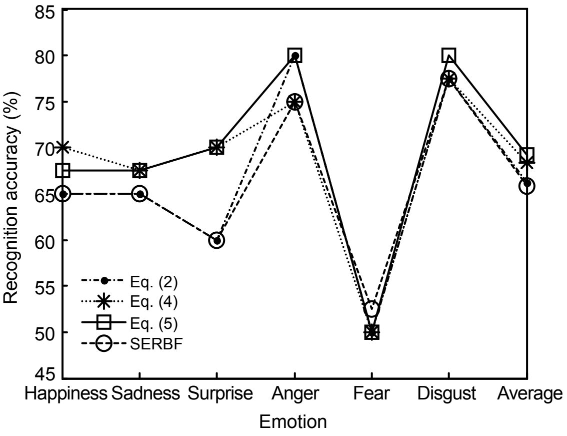

Fig. 2 Recognition accuracies of the methods using different confidence calculations

SERBF: single-channel based speech emotion recognition method

Fig. 2 shows that the recognition accuracies using different confidence calculation methods are all higher than the result of the single-channel based method, SERBF. Moreover, the average recognition accuracy is the highest when the confidence calculation method shown in Eq. (5) is adopted, because the confidence calculation method given in Eq. (5) combines the distance confidence and the probability confidence. Therefore, the confidence calculation method given in Eq. (5) is used to calculate the confidence in PFRA.

5.2.2 Effectiveness analysis of PFRA

To analyze the effectiveness of the proposed PFRA, we compared it with the commonly used linear-weighted method, DS evidence theory, and Bayes fusion method. The weight coefficient of the linear-weighted method was set to 0.5. The DS evidence theory adopted the method introduced by Li *et al.* (2002). In the DS method, the values of distribution function *M* were assigned by the probability of each model. For the Bayes fusion method, the priori knowledge was the paralanguage coverage. We regarded the functional paralanguage feature channel and the accompanying paralanguage feature channel as mutual independent. In the following experiments, LW stands for the linear-weighted fusion method, DS stands for the DS theory, Bayes stands for the Bayes fusion method, and PFRA stands for the method proposed in this paper.

1. Comparisons with the methods adopting only confidence or probability

To verify the effectiveness of PFRA, experiments were conducted for comparison with other three methods. In these experiments, the confidence was calculated using Eq. (5), and the paralanguage coverage was set to 40% for Ts and Te. The comparison results are shown in Fig. 3. In these fusion methods, the recognized functional paralanguages must be remarked using their corresponding typical emotion classes.

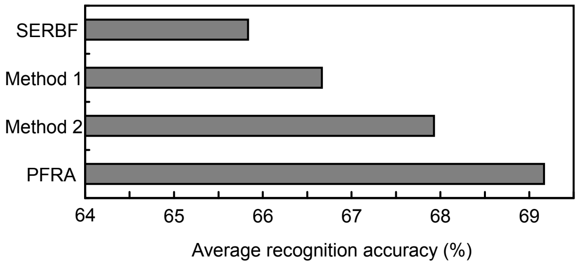

Fig. 3 Average recognition accuracies of different methods

SERBF does not use functional paralanguages. The result of SERBF is the recognition accuracy of using only the accompanying paralanguage channel.

In Method 1, the new discrimination criterion of each emotion class was defined as the ratio of its confidence to its probability. The emotion category with the highest ratio was chosen as the category to which the test sample belongs.

In Method 2, the voting method was used to fuse the first three emotion classes with the higher probability of two channels. The emotion class that appeared most often in the two channels was regarded as the final recognition result. If there were several emotions receiving the most votes, the emotion having the highest probability was taken as the final recognition result.

As can be seen from Fig. 3, the average recognition accuracies of fusion methods are higher than that of the method using a single channel. This indicats that fusing the functional paralanguage feature channel and the accompanying paralanguage feature channel is effective. Fig. 3 also shows that the average recognition accuracy of Method 2 is higher than that of Method 1, and the average recognition accuracy of PFRA is the highest. Thus, PFRA is superior to Methods 1 and 2. It is shown that the probability confidence and the distance confidence both contribute to the decision.

2. Comparisons under different paralanguage coverages

In practice, the functional paralanguage coverage is variable. Therefore, the recognition accuracies of methods in different functional paralanguage coverages are discussed.

In Table 2, the recognition accuracies of DS and LW methods are very similar, and the recognition accuracy of the Bayes method is higher than those of the former two methods, but lower than that of PFRA. It can be concluded that the proposed PFRA is effective in any paralanguage coverage testing set. Specifically, when the training set contains 30% or 40% functional paralanguage, the recognition accuracies under different functional paralanguage coverage testing sets are improved greatly. The maximum and average accuracies are improved by 4.58% and 3.82%, respectively. This is because the weight coefficient of LW is difficult to choose (in our experiment the weight coefficient was set to 0.5), and the Bayes and DS methods are both sensitive to false identification of functional paralanguages. Moreover, the Bayes method requires the functional paralanguage proportion in a testing set as priori knowledge. But in practice, functional paralanguage proportion cannot be exactly known in advance. In PFRA, both the output probability and confidence are considered.

Table 2 Recognition accuracy differences between the methods and the accompanying paralanguage feature channel under different functional paralanguage coverages

Functional paralanguage coverage rate of testing set	Fusion method	Improved recognition accuracy compared with SERBF (%)						
		30%*	40%*	50%*	60%*	70%*	80%*	Average
30%	LW	0.42	0.83	0.42	1.25	1.67	2.08	1.11
	DS	−0.42	0.83	0.42	1.25	1.67	1.67	0.90
	Bayes	0.83	0.83	0.83	1.67	1.67	2.50	1.39
	PFRA	2.50	3.33	3.75	4.17	4.58	4.58	3.82
40%	LW	1.25	2.08	2.50	1.25	2.08	2.50	1.94
	DS	0.83	2.08	2.50	1.67	1.25	2.50	1.81
	Bayes	1.67	2.08	2.50	1.67	2.50	2.92	2.22
	PFRA	2.50	3.33	3.33	3.75	3.75	4.17	3.47
50%	LW	1.25	1.25	1.25	1.25	1.67	1.67	1.39
	DS	0.42	1.25	0.83	1.25	2.50	1.67	1.32
	Bayes	1.25	1.25	1.67	1.67	2.08	2.92	1.81
	PFRA	1.67	2.08	2.08	3.33	3.75	3.33	2.71
60%	LW	0.83	1.25	1.67	2.08	1.25	2.50	1.60
	DS	0.83	1.25	1.67	2.50	1.25	2.50	1.67
	Bayes	0.83	2.08	1.67	1.67	1.67	2.92	1.81
	PFRA	2.08	2.92	2.08	3.33	3.33	3.75	2.92
70%	LW	0.42	1.67	1.25	1.25	0.83	0.42	0.97
	DS	−0.42	1.67	0.42	1.67	0.83	1.25	0.90
	Bayes	1.25	2.92	2.50	1.25	1.67	0.83	1.74
	PFRA	2.08	3.33	3.33	3.33	2.50	2.92	2.92
80%	LW	0.83	1.25	1.67	1.25	0.83	1.25	1.18
	DS	0.42	1.25	2.08	1.67	0.42	0.83	1.11
	Bayes	1.25	1.67	1.67	1.67	1.25	1.67	1.53
	PFRA	1.67	2.92	2.92	2.08	2.50	2.50	2.43

* The 30%–80% represent the paralanguage coverages of trainning sets. LW: linear-weighted fusion method; DS: Dempster-Shafer (DS) evidence theory; Bayes: Bayes fusion method; PFRA: paralanguage fusion recognition algorithm (PFRA) proposed in this paper; SERBF: single-channel based speech emotion recognition method

3. Stability comparisons with other fusion methods

We took the 40% functional paralanguage coverage in training set Ts to further test the classification performances of different fusion methods. Different functional paralanguage coverages (30%–80%) in testing set Te were adopted. The fusion methods were compared in terms of precision and recall (i.e., accuracy and their harmonic mean F1-measurement (F1)) (Yang and Liu, 1999). Meanwhile, the amplitude of each measure for each method was computed (Table 3).

Table 3 Comparison results of different fusion methods

Fusion method	Precision (%)	Recall (%)	
		Accuracy	F1-measurement
LW	67.66±2.09	65.31±2.15	66.49±2.17
DS	66.93±2.99	64.67±2.65	65.80±2.87
Bayes	67.57±2.01	65.92±1.73	66.75±1.89
SERBF	65.29±1.95	64.17±1.85	64.73±1.90
PFRA	68.94±1.99	67.32±1.65	68.13±1.87

The fluctuations of the LW fusion algorithm and DS method were evident (Table 3). However, PFRA fluctuates slightly and its amplitude is lower than those of the other fusion methods. Thus, it can be concluded that the stability of the proposed PFRA algorithm is superior to those of the other algorithms.

The proposed PFRA achieved the highest precision, recall, F1-measurement, and relatively high stability (Table 3). Compared with SERBF, using the proposed PFRA algorithm, the precision rate was improved by 3.65%, the accuracy by 3.15%, and the F1-measurement by 3.40%.

According to the results, it can be concluded that the effectiveness of the proposed PFRA is superior to those of other fusion methods, in terms of recognition accuracy and algorithm stability. The proposed PFRA has three advantages: first, functional paralanguages are fused by the fusion method based on confidence and probability in PFRA, and functional paralanguages include much emotional information, improving the recognition accuracy. Second, as the distance confidence and probability distance are both considered in the confidence calculation of the PFRA, the number of wrongly identified samples decreased. The recognition accuracy is further improved. Third, as functional paralanguages are not sensitive to the change of speakers, the stability and the generalization of the algorithm are guaranteed. Therefore, PFRA is effective in fusing the information of functional paralanguages with the accompanying paralanguage, greatly improving the speaker-independent speech emotion recognition accuracy.

6 Conclusions and future work

To improve the speaker-independent speech emotion recognition accuracy, functional paralanguages features are introduced, and PFRA is proposed to combine the functional paralanguages and the accompanying paralanguage features. This method makes up for the lack of accompanying paralanguage features. It provides a reference for the future study of speech emotion recognition. Experimental results show that the proposed algorithm is superior to other fusion methods in terms of stability and the recognition accuracy. Compared with the single-channel based recognition method, the speaker-independent speech emotion recognition accuracy of the proposed PFRA is improved by 3.15%. In the future, we will separate the functional paralanguages and speech automatically and increase the number of functional paralanguage samples and categories of functional paralanguages, to further improve the speech emotion recognition accuracy.

References

Bachorowski, J.A., Smoski, M.J., Owren, M.J., 2011. The acoustic features of human laughter. *J. Acoust. Soc. Am.*, **110**(3):1581-1597. [doi:10.1121/1.1391244]

Berler, A., Shimony, S.E., 1997. Bayes Networks for Sonar Sensor Fusion. Proc. 13th Conf. on Uncertainty in Artificial Intelligence, p.14-21.

Devillers, L., Vidrascu, L., 2006. Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs. Proc. Interspeech, p.801-804.

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

Fujie, S., Ejiri, Y., Matsusaka, Y., Kikuchi, H., 2003. Recognition of Paralinguistic Information and Its Application to Spoken Dialogue System. Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, p.231-236. [doi:10.1109/ASRU.2003.1318446]

Hayashi, Y., 1999. Recognition of Vocal Expression of Emotions in Japanese: Using the Interjection eh ‘Korean’. Proc. Int. Conf. on Phonetic Sciences, p.2355-2359.

Huang, C.L., Wang, C.J., 2006. A GA-based feature selection and parameters optimization for support vector machines. *Expert Syst. Appl.*, **31**(2):231-240. [doi:10.1016/j.eswa.2005.09.024]

Huang, C.W., Jin, Y., Zhao, Y., Yu, Y.H., Zhao, L., 2010. Design and establishment of practical speech emotion database. *Techn. Acoust.*, **29**(4):396-399 (in Chinese).

Huq, S., Moussavi, Z., 2012. Acoustic breath-phase detection using tracheal breath sounds. *Med. Biol. Eng. Comput.*, **50**(3):297-308. [doi:10.1007/s11517-012-0869-9]

Ishi, C.T., Ishiguro, H., Hagita, N., 2006. Evaluation of Prosodic and Voice Quality Features on Automatic Extraction of Paralinguistic Information. Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, p.9-15. [doi:10.1109/IROS.2006.281786]

Ishi, C.T., Ishiguro, H., Hagita, N., 2008. Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality. *Speech Commun.*, **50**(6):531-543. [doi:10.1016/j.specom.2008.03.009]

Jones, C., Jonsson, I.M., 2008. Using paralinguistic cues in speech to recognize emotions in older car drivers. *LNCS*, **4868**:229-240. [doi:10.1007/ 978-3-540-85099-1_20]

Kennedy, L.S., Ellis, D.P.W., 2004. Laughter Detection in Meetings. Proc. Int. Conf. on Acoustics, Speech, and Signal Processing Meeting Recognition Workshop, p.118-121.

Kleckova, J., 2009. Important Nonverbal Attributes for Spontaneous Speech Recognition. 4th Int. Conf. on Systems, p.13-16. [doi:10.1109/ICONS.2009.41]

Li, C.G., 2004. Paralinguistic Studying. MS Thesis, Heilongjiang University, Harbin, China (in Chinese).

Li, Y.C., Wang, B., Wei, J., Qian, C., Huang, Y., 2002. An efficient combination rule of evidence theory. *J. Data Acquis. Process.*, **17**(1):33-36 (in Chinese).

Li, Y.X., He, Q.H., 2011. Detecting laughter in spontaneous speech by constructing laughter bouts. *Int. J. Speech Technol.*, **14**(3):211-225. [doi:10.1007/s10772-011-9097-1]

Maekawam, K., 2004. Production and Perception of ‘Paralinguistic’ Information. Int. Conf. on Speech Prosody, p.367-374.

Mao, Q.R., Wang, X.J., Zhan, Y.Z., 2010. Speech emotion recognition method based on improved decision tree and layered feature selection. *Int. J. Human. Rob.*, **7**(2):245-261. [doi:10.1142/S0219843610002088]

Matos, S., Birring, S.S., Pavord, I.D., Evans, D.H., 2006. Detection of cough signals in continuous audio recordings using hidden Markov models. *IEEE Trans. Biomed. Eng.*, **53**(6):1078-1083. [doi:10.1109/TBME.2006.873548]

Pal, P., Iyer, A.N., Yantorno, R.E., 2006. Emotion Detection from Infant Facial Expressions and Cries. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.721-724. [doi:10.1109/ICASSP.2006.1660444]

Petridis, S., Pantic, M., 2008. Audiovisual Discrimination Between Laughter and Speech. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, p.5117-5120. [doi:10.1109/ICASSP.2008.4518810]

Pudil, P., Novovicova, J., Kittler, J., 1994. Flating search methods in feature selection. *Pattern Recogn. Lett.*, **15**(11):1119-1125. [doi:10.1016/0167-8655(94)90127-9]

Sundaramb, S., Narayananc, S., 2007. Automatic acoustic synthesis of human-like laughter. *J. Acoust. Soc. Am.*, **121**(1):527-535. [doi:10.1121/1.2390679]

Szameitat, D.P., Darwin, C.J., Szameitat, A.J., 2007. Formant Characteristics of Human Laughter. Interdisciplinary Workshop on the Phonetics of Laughter, p.4-5.

Truong, K.P., van Leeuwen, D.A., 2005. Automatic Detection of Laughter. Proc. 9th European Conf. on Speech Communication and Technology, p.485-488.

Truong, K.P., van Leeuwen, D.A., 2007. Automatic discrimination between laughter and speech. *Speech Commun.*, **49**(2):144-158. [doi:10.1016/j.specom.2007.01.001]

Yang, Y.M., Liu, X., 1999. A Re-examination of Text Categorization Methods. Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval, p.42-49.