Neurocomputing 398 (2020) 257–264

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Wavelet packet analysis for speaker-independent emotion recognition

Kunxia Wang[a,b], Guoxin Su[c], Li Liu[d,*], Shu Wang[e,*]

[a] School of Electronic and Information Engineering, Anhui Jianzhu University, Hefei, Anhui, China
[b] Cognition Lab, Texas A& M University, Texas, USA
[c] School of Computing and Information Technology, University of Wollongong, Australia
[d] School of Big Data & Software Engineering, Chongqing University, Chongqing 400044, China
[e] Faculty of Materials and Energy, Southwest University, Chongqing 400715, China

ARTICLE INFO

Article history:
Received 29 January 2019
Revised 23 February 2020
Accepted 23 February 2020
Available online 25 February 2020

Communicated by Prof. R. Capobianco Guido

Keywords:
Affective computing
Speech signal
Wavelet packet
Wavelet packet coefficient

ABSTRACT

Extracting effective features from speech signals is essential to recognize different emotions. Recent studies have demonstrated that wavelet analysis is a useful technique in signal processing. In this study, we extract emotion features using wavelet packet analysis from speech signals for speaker-independent emotion recognition. We explore and evaluate these features from two databases, i.e., EMODB and EESDB. It is found that the extracted features are effective for recognizing various speech emotions. Furthermore, compared with common features such as Mel-Frequency Cepstral Coefficients (MFCC), these features can improve the recognition rates by 14.9 and 4.3 percentages on EMODB and EESDB, respectively.

1. Introduction

Speech emotion recognition plays an important role in many applications, such as human-computer intelligent assistance [1], man-machine interaction [2,3], daily living activity, activity recognition [4,5] and complex activity recognition [6]. Most speech features for emotion recognition are prosodic, spectral and qualitative features [3]. Exisiting studies [3] have shown that features such as pitch, energy, duration, perceptual linear predictive coefficients (PLPs), linear predictive cepstral coefficients (LPCCs), Mel-Frequency Cepstral Coefficients (MFCCs) and combinations of these are important for speech emotion recognition.

Recently, some new approaches using wavelet analysis to analyze speech signals [7,8] have been proposed. The method of wavelet analysis is based on multi-resolution in order to reflect nonlinear vortex-flow interactions. It has been applied to noising, detection, compression, classification and so on [8–15].

Wavelet packet analysis has been studied in different domains [8,12,16–22]. In [16], it analyzed the dynamic emotional responses of the participants to self-selected music and implemented the multi-resolution analysis algorithm using wavelet packet decomposition for music-induced emotion. In [19], wavelet-based pH time-frequency vocal source features together with the MFCC and the Teager-Energy-Operator (TEO) based features were extracted for emotion classification. The amplitude modulation spectrogram (AMS) and the Gaussian mixture models (GMM) (AMS-GMM) acoustic mask were also proposed to improve the classification. In [20], it employed biorthogonal wavelet entropy to extract features for facial emotion recognition. In [21], it proposed new features based on the energy content of wavelet-based time–frequency (TF) representations, including (1) the continuous wavelet transform (CWT), (2) the bionic wavelet transform (BWT), and (3) the synchro squeezed wavelet transform (SSWT), to model emotional speech. In [22], it extracted PLP, MFCCs, LPCCs, stationary wavelet transform features, wavelet packet energy and entropy features for emotion recognition. In [23], Tuned Q-factor Wavelet Transform (TQWT) and Wavelet Packet Transform (WPT) methods were used to predict the emotion of stroke patients. In [24], it proposed a new feature set using wavelet packet transform based energy and non-linear entropies for infant cry classification. These works mentioned above addressed the optimization of wavelet packets and showed effectiveness for different research domains.

In this study, we propose a new method for recognizing the emotion, which is based on wavelet packet coefficient (WPC) features. In order to evaluate the WPC features, we explore the efficiency of these features for emotion classification. Our contribution includes (1) proposing new WPC features for emotion recognition, (2) further improving speaker-independent speech emotion

* Corresponding authors.
E-mail addresses: kxwang@ahjzu.edu.cn (K. Wang), guoxin@uow.edu.au (G. Su), dcsliuli@cqu.edu.cn (L. Liu), shuwang@swu.edu.cn (S. Wang).

https://doi.org/10.1016/j.neucom.2020.02.085
0925-2312/

recognition by using the Sequential Floating Forward Search (SFFS) method, and (3) carrying out the effectivity of these methods on two language databases. Compared with a preliminary version of this work [25], we have made significant enhancements by presenting more thorough analysis on the background knowledge and experimental results. Especially we have optimized the features and expanded the experimental validation. Finally, we also have presented some additional examples, which are generated from our test.

The remainder of this paper is organized as follows. The wavelet packet coefficient model, which is based on the wavelet packet transformation is presented in Section 2. The evaluations of WPC features for speech emotion on two databases are analyzed in Section 3. The experimental results for recognizing emotion are analyzed Section 4. Finally, conclusions are drawn in the final section.

2. Wavelet packet coefficient model

Wavelets can be defined as [17]. The wavelet transform of signal $f \in L^2(R)$ at time u and scale s is computed by correlating f with a wavelet atom as Eq. (1).

$$Wf(u,s) =< f, \psi_{u,s} >= \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)\mathrm{d}t. \quad (1)$$

The wavelet transform can be rewritten as a convolution product as Eq. (2).

$$Wf(u,s) = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\psi^*\left(\frac{t-u}{s}\right)\mathrm{d}t = f * \bar{\psi}_s(u)\ . \quad (2)$$

Wavelet packet transform (WPT) [17] applies the transform step to both the low pass and the high pass result. It can be implemented using a multistage filter bank which offers a richer range of possibilities for signal analysis. It is calculated through a time-domain filtering with a sub-signal representation obtained from frequency components within each subband [18]. In Eq. (3), it is defined as the decomposition of the wavelet packet.

$$\psi^i_{j,k}(t) = 2^{\frac{j}{2}}\psi^j(2^j t - k), i = 1, 2, 3, \ldots \quad (3)$$

In Eq. (3), integers i is the modulation parameters, j is the scale parameters, k is the translation parameters and ψ^j is the wavelet function. Eqs. (4) and (5) show the recursive ones.

$$\psi^{2j}(t) = \sqrt{2}\sum_{-\infty}^{+\infty} h(k)\psi^j(2t-k) \quad (4)$$

$$\psi^{2j+1}(t) = \sqrt{2}\sum_{-\infty}^{+\infty} g(k)\psi^j(2t-k) \quad (5)$$

Eq. (6) defines j level decomposition of $f(t)$. $f^i_j(t)$ can be defined as Eq. (7). $c^i_{j,k}(t)$ is defined as Eq. (8) which is wavelet packet components. So a signal can be computed by a series wavelet packet coefficients combination. In this paper, the wavelet packet coefficients, their first-order difference and their second-order difference are extracted from each frame. Four-level and five-level wavelet packet decomposition are explored. Daubechies wavelets are employed for emotion recognition.

$$f(t) = \sum_{i=1}^{2j} f^i_j(t) \quad (6)$$

$$f^i_j(t) = \sum_{-\infty}^{+\infty} c^i_{j,k}(t)\psi^i_{j,k}(t) \quad (7)$$

$$c^i_{j,k}(t) = \sum_{-\infty}^{+\infty} f(t)\psi^i_{j,k}(t)dt \quad (8)$$

3. The analysis of WPC features

The proposed features as wavelet packet coefficients, their delta and acceleration ones are extracted from two speech emotion databases in two different languages. To explore the effectiveness of the WPC features, we compare these features with other features in terms of classification. The experimental setting and results are described as follows.

3.1. Experimental databases

Two speech emotion databases, EMODB [26] and (EESDB) [29], are considered. EMODB is database using the German language. It includes seven emotional states from daily communication from 10 people. The sentences of this database are predefined. EMODB, as a standard dataset, has been extensively used for emotion research. The EESDB database also covers seven types of emotions derived from 11 people. It is collected from Chinese TV shows about elderly people.

3.2. Wavelet packet coefficient features

The wavelet packet coefficients are estimated by wavelet packet analysis from every frame. As shown in Eq. (8), $C_{j,k}$ is kth coefficient in j-level decomposition of the speech. There are 32 coefficients after 5 level decomposition, which are $C_{5,0}$, $C_{5,1}$, $C_{5,2}$, ... and $C_{5,31}$, respectively.

3.3. Global wavelet packet coefficient features

Studies [1,3,28,30–32] have found that features containing global information are better than local ones for computational efficiency and recognition performance [3]. So, the maximum of the coefficient is calculated as in [3,31,32] from EESDB.

Fig. 1 shows the maximum of all wavelet packet coefficients with different emotions by five-level decomposition on the EESDB database. From Fig. 1 we can see that the maximum of $C_{5,0}$ to $C_{5,31}$ change with the emotions and coefficients. For the same coefficient, the maximum values of different emotions manifest difference. For the same emotion, the maximum values of different coefficients also manifest difference. The values from $C_{5,0}$ to $C_{5,15}$ of each emotion are significantly higher than those from $C_{5,16}$ to $C_{5,31}$. Furthermore, it is found that the fluctuation of the lower WPC of every type of emotion is rapid, while the fluctuation of the higher one is not obvious. Moreover, the angry emotion has highest maximum values of the coefficients among all seven emotions, the fear emotion has lower maximum values among the $C_{5,0}$, $C_{5,1}$, $C_{5,3}$, $C_{5,5}$, $C_{5,8}$, $C_{5,9}$, $C_{5,11}$, $C_{5,12}$ and $C_{5,14}$ coefficients. Except for the angry emotion, the maximum values of the disgust emotion among these $C_{5,1}$, $C_{5,2}$, $C_{5,3}$, $C_{5,7}$, $C_{5,8}$, $C_{5,9}$ and $C_{5,11}$ coefficients are higher than those of the other five emotions. The maximums of the WPCs are also calculated for EMODB. A similar result has been observed that the max value of each type of emotion always exists at a low wavelet packet coefficient.

4. Experiment results of speaker-independent recognition

Speaker-independent means a leave-one speaker-out scheme. It is the latest trend in the affective computing field. Since speaker-independent recognition can handle an unknown speaker, it demonstrates better ability than that of speaker-dependent recognition [27]. To date, a few researchers have conducted speaker-independent recognition [27,33]. Margarita Kotti and Fabio Patern [27] extracted features related to statistics of the pitch, formants, and energy contours, as well as the spectrum, cepstrum, perceptual and temporal features, autocorrelation and others, totaling 2327

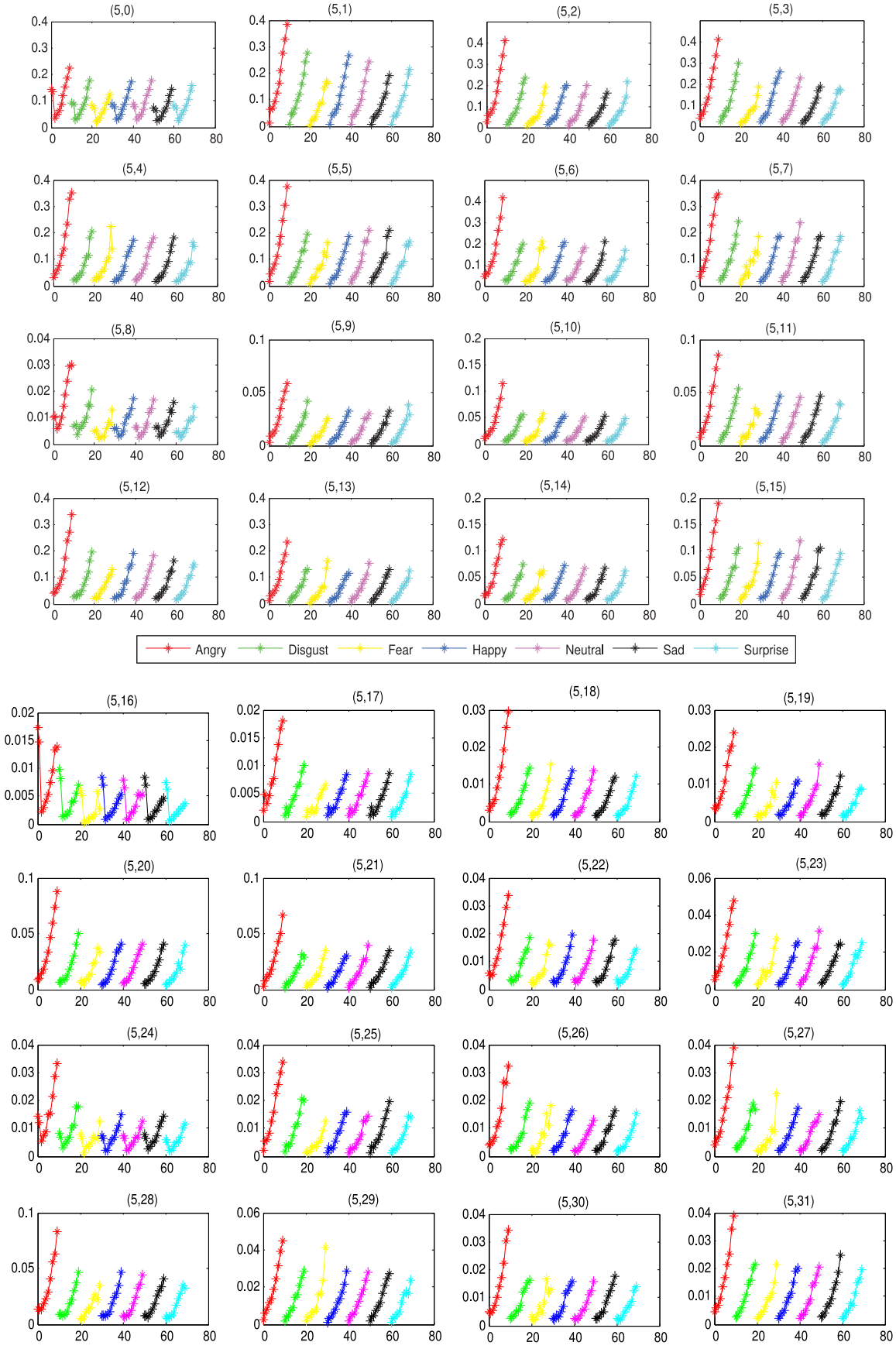


Fig. 1. The maximum of each WPC with different emotions over five-level decomposition on EESDB database.

Table 1
Speaker-independent emotion recognition comparison between level 4 and level 5 (%).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Level-5	59.8	64.8	60.7	61.5	56.9	57.7	62.8	60.2	57.0	60.4	60.0	59.9	55.3	57.3	62.3	56.1
Level-4	57.3	55.4	50.1	52.7	53.5	55.0	47.2	53.3	51.0	52.0	49.4	53.5	54.5	55.7	48.9	55.8
Difference	2.5	9.4	10.6	8.8	3.4	2.7	15.6	6.9	6.0	8.4	10.6	6.4	0.8	1.6	13.4	0.3

features for speaker-independent recognition. Wang [33] extracted features for speaker-independent classification using Energy, Zero-crossing rate, MFCCs and Fourier Parameters. Seven types of emotions are classified in EMODB and EESDB databases. In our experiments, we select Support Vector Machine (SVM) as the classifier. Linear SVM (LSVM) with a linear kernel and SVM with a radial basis function (RBF) kernel (abbreviated as RSVM) are selected as classifiers.

4.1. Feature extraction

In order to test the performance of speaker-independent emotion recognition, we extract MFCC features and WPC features.

4.1.1. MFCC features

Mel Frequency Cepstral Coefficients (MFCCs) were first introduced in [34]. Studies in [3,27,33] adopted MFCC for speech recognition. MFCCs come from the characteristics of audition and feelings of human hearing. In our work, MFCC features are extracted for speech emotion recognition and compared with the proposed wavelet packet coefficient features. The speech signal is firstly filtered with a pre-emphasis coefficient of 0.97. The global features including mean, maximum, minimum, median, and standard deviation of MFCC features are calculated. The original 13 MFCCs, delta- MFCCs and double-delta MFCCs were extracted to 39-MFCCs. Those global features of the 39-MFCCs were further calculated. Therefore, we obtained 195 MFCC feature vectors in total.

4.1.2. Wavelet packet features

In addition to MFCCs, wavelet packet coefficient features were extracted from speech as described in Sections 3.1 and 3.2. Temporal derivative features may enhance the recognition performance [3], so the dynamic features are also extracted. The WPC features include the coefficients C, the first-order difference (ΔC) and the second-order difference ($\Delta\Delta C$). The global features were also computed as they can reduce the number of features. Wavelet packet supports many algorithms such as Daubechies wavelet filter and Gabor filter [10]. Daubechies wavelets have been widely applied to solve signal process problems, especially speech signal. In our studies, the families of Daubechies wavelets [10] were chosen. DB2 was firstly employed for emotion recognition.

Level 4 wavelet packet decomposition produces 16 coefficients, and level 5 wavelet packet decomposition produces 32 coefficients, as described in Eq. (8). The original features, the first-order difference and second-order difference are calculated. This yields 240 features from level 4 wavelet decomposition and 480 features from level 5 wavelet decomposition. These features were explored for speech emotion recognition.

4.2. Comparison between level 4 and level 5 wavelet packet decomposition

We obtain 16 wavelet packet coefficients in level 4 and 32 wavelet packet coefficients in level 5. Table 1 shows the classification results of the former 16 coefficients from both level 4 and level 5 on EESDB. It is observed that the emotion classification rate of level 5 is better than that of level 4. The wavelet packet coefficient in level 5 improves the recognition performance by 6.7% on average. It indicates that wavelet packet coefficient features in level 5 attain better accuracy for recognizing emotion. Hence, we plan to study the WPCs in level 5 as follows. Experiments have been done on emotion databases with different wavelet packet coefficient features.

4.3. Recognition results with single and incremental wavelet packet coefficient features

Fig. 2 shows the results with single and incremental wavelet packet coefficient features in level 5. Fig. 2(a) shows the results for each of the 32 wavelet packet coefficient feature. The results also show that the recognition rate decreases as the coefficient number becomes larger. Moreover, the recognition rates for the first eight WPC coefficients are distinctly higher than those for the last twenty-four. Besides, it shows that the accuracy increases and then decreases on the Chinese database. The highest accuracies are 57.4% and 55.1% on EMODB and EESDB database, respectively, using signal wavelet packet coefficient. Fig. 2(b) shows the results of the incremental wavelet packet coefficient features. It is also observed that the accuracy increases and then decreases on both databases, but this trend is more obvious on EMODB. Therefore, the recognition rate with the WPC features over the former coefficients is better than that over the later ones. The best performance is 64.5% on EMODB and 62.8% on EESDB. Although the performance with multiple WPC features is better than that with single WPC, it is hard to get good results with too many features. In Fig. 2, it shows that the first eight coefficients attain a higher accuracy compared with other coefficients.

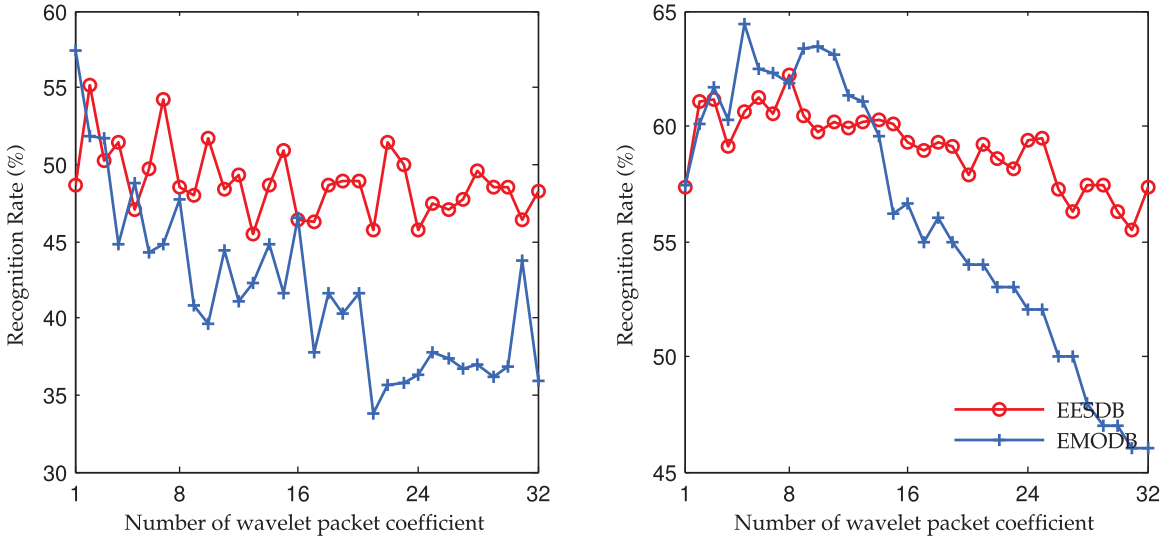
4.4. Recognition results with DB2, DB6 and DB12 wavelet packet

The Daubechies wavelet family has been widely used in speech recognition [35], speech compression [36] and many other signal processing domains [9,37,38]. However, it is rare that Daubechies wavelet family is reported for speech emotion recognition. In this study, in order to evaluate the performance of the Daubechies wavelet function, Daubechies of order two, four and six (db2, db4, db6) are considered. The recognition rate of DB2, DB6 and DB12 are calculated for comparison respectively.

Fig. 3 shows speaker-independent recognition results using DB2, DB6 and DB12 on EESDB database. Fig. 3 (a) shows results using single wavelet packet coefficient features with DB2, DB6 and DB12. The average recognition rates are 48.8%, 43.1% and 46.6% for DB2, DB6 and DB12, respectively. Fig. 3(b) shows the results using incremental wavelet packet coefficient features with DB2, DB6 and DB12. The average recognition rates are 51.9%, 46.9% and 46.7%, respectively. It is also observed that the accuracy increases and then decreases with the incremental order of WPC. Similar results also show that incremental WPC features get better accuracy compared with single ones on the EMODB database. This suggests that the Daubechies wavelet is efficient for speech emotion recognition.

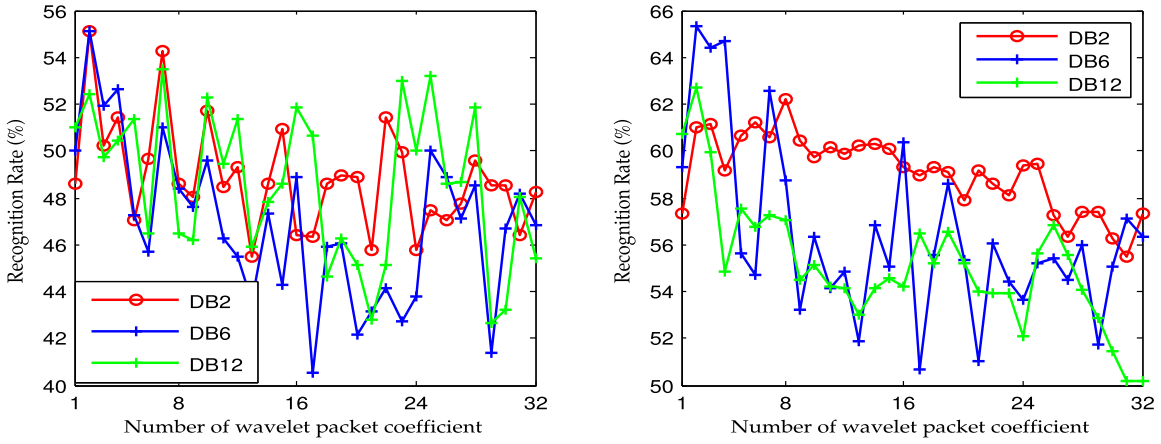
4.5. Recognition results with feature selection

Feature selection, which is important for emotion recognition [39], can be divided into two categories: supervised and unsupervised methods. Recently, a number of new feature selection



(a)Results using single wavelet packet coefficient features (b)Results using multiple wavelet packet coefficient features

Fig. 2. Speaker independent recognition results using single WPC features and multiple WPCs features.



(a)Result using single WPC feature (b)Results using multiple features

Fig. 3. Speaker independent recognition results using DB2, DB6 and DB12 WPCs features (on Chinese database).

strategies, such as Semi-supervised method [40] and paraconsistent feature engineering [41], have been proposed. In this study, we focus on the Sequential Floating Forward Search (SFFS), which is one kind of feature selection methods first proposed in [42]. It has been extensively used in the speech emotion classification domain [21,29]. To decrease the number of feature vectors and improve the classification performance, we choose the SFFS iterative approach to improve recognition performance.

Fig. 4 shows the results of the first eight wavelet packet coefficient (WPC-8) features, WPC+SFFS features and MFCC features. The wavelet packet coefficient was extracted with DB2. The classifiers are LSVM and RSVM, respectively. Fig. 4(a) shows that the WPC8 features can improve the recognition performance by 1.7% on EMODB database and 5.5% on EESDB database compared with MFCC features. It is also observed that the improved performance is significant after SFFS selection, if the RSVM classifier for emotion recognition are used. The improvements of the recognition rates are roughly 14.9% and 4.3% on EMODB and EESDB databases, respectively. In Fig. 4 (b), it shows that the WPC8 features enhance the recognition rates by 1.8% on EMODB and 5.4% on EESDB databases, compared with the recognition rate of MFCC features. After selecting the SFFS method with LSVM the improvements are about 17.4% on EMODB and 10.1% on EESDB. The Unweighted Average Recall(UAR) is necessary for unbalanced distributions of instances among classes in emotion recognition [41]. Fig. 5 shows the recognition results by UAR for MFCC, WPC8 and WPC+SFFS. After using the SFFS method, the best results are 79.2% on EMODB and 71.3% on EESDB. The improvements are 14% and 14.8%, respectively, compared with MFCC features.

Tables 2 and 3 show the confusion matrix of the best emotion recognition results using WPC+SFFS methods on the EMODB and EESDB databases. The average recognition rates are 79.5% and 60.7%, respectively. Table 4 shows the best number of wavelet packet coefficient after feature selection. Most selected features are belong to the first 16 coefficients. The reason is that the global values of the lower coefficients are obviously higher than those of the higher coefficients which are effective to recognize different emotions. These selected coefficient features on the EMODB database are effective for recognizing happy, sad, angry and disgust emotions as shown from Table 2. The selected ones from the EESDB database are effective for recognizing sad, angry and neutral emotions as shown from Table 3. Although fewer of the higher coefficients are selected, some of them are also useful for emotion recognition.

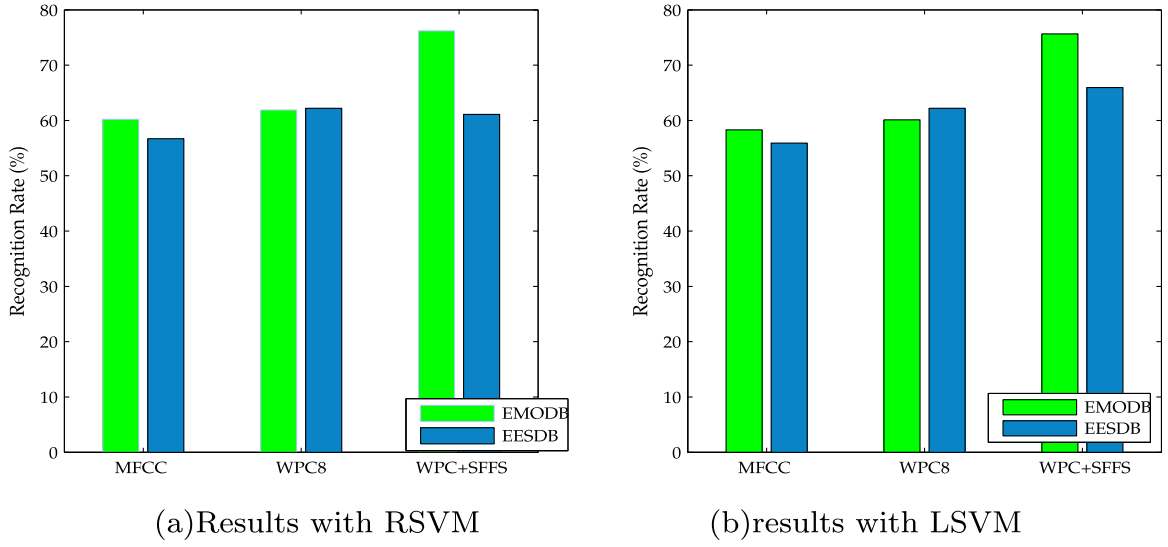


Fig. 4. Emotion classification with MFCC and WPC features on two databases using SVM classifier (%).

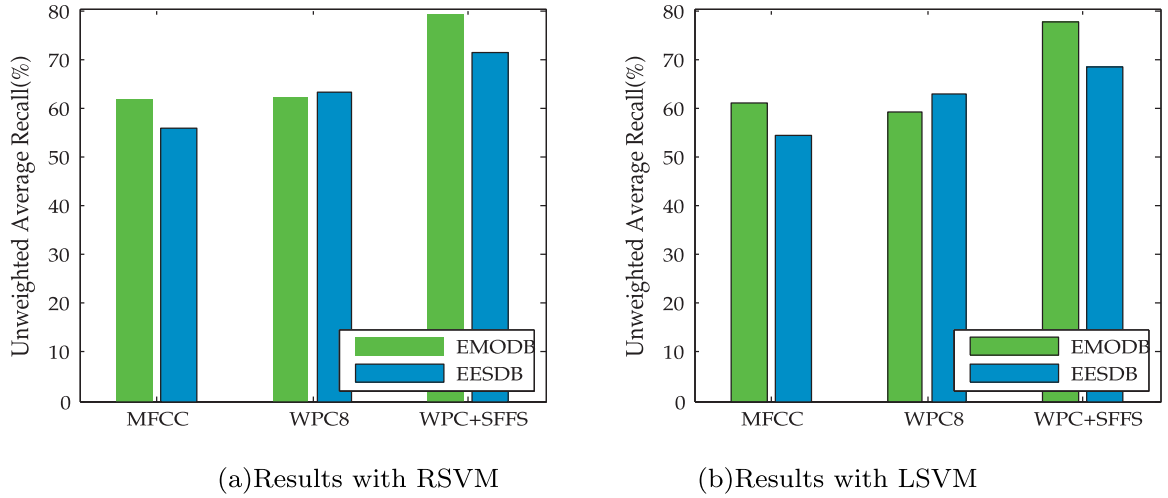


Fig. 5. Recognition result by unweighted average recall with MFCC, WPC8 and WPC+SFFS features on two databases using SVM classifier (%).

Table 2
Confusion matrix result on EMODB after feature selection using RSVM (%).

	Happiness.	Sadness.	Anger.	Neutrality.	Boredom.	Anxiety.	Disgust.
Happiness.	98.6	0	0	1.4	0	0	0
Sadness.	0	96.8	0	1.6	0	1.6	0
Anger.	1.6	0	93.7	0	0	2.4	2.3
Neutrality.	1.26	7.59	3.8	56.96	18.99	8.86	2.53
Boredom.	1.2	8.6	2.5	4.93	75.3	6.17	1.2
Anxiety.	0	7.25	8.7	1.45	17.39	55.07	10.14
Disgust.	6.52	0	6.52	0	2.17	4.35	80.4

Table 3
Confusion matrix result on EESDB after feature selection using RSVM (%).

	Happiness	Sadness	Anger	Neutrality	Surprise	Fear	Disgust.
Happiness	27.9	18.6	16.28	30.2	2.33	0	4.65
Sadness	0	69.7	11.2	11.2	5.6	0	2.3
Anger	0	2.3	93.1	2.3	2.3	0	0
Neutrality	1.7	7.0	7.8	80.9	1.7	0	0.9
Surprise	5.6	0	12.2	22.2	60	0	0
Fear	0	0	25.0	25.0	0	50.0	0
Disgust	0	21.2	30.3	5.1	0	0	43.4

Table 4
The best feature set of WPC selected by SFFS methods on two database (%).

Database	Number of wavelet packet coefficient
EESDB	1, 2, 3, 4, 7, 9, 10, 12, 14, 29
EMODB	1, 2, 3, 4, 5, 7, 9, 10, 26, 31

Studies [19,21,27,43,44,44–47] demonstrated their research on the Berlin emotional speech database EMODB. In [27], they extracted 2327 features for recognizing emotions from the speech signals, and the recognition rates of happiness, sadness, anger and disgust are 89.7%, 88.6%, 90.1% and 47.5%, respectively. While using WPC features, the corresponding accuracy rates are 98.6%, 96.8%, 93.7% and 80.4%, respectively. These WPC features attain an improvement of 8.9%, 8.2%, 3.6% and 32.9%, respectively. In [19], the average accuracy obtained with the pH feature is 68.1% (in particular, happiness is 48%, sadness is 82%, anger is 86%, disgust 67%, boredom is 61%, anxiety is 62%, and neutrality is 71%). Moreover, the pH feature using the amplitude modulation spectrogram (AMS) and the Gaussian mixture models (GMM) (AMS-GMM) acoustic mask improved classification performance as happiness is 70.5%, sadness is 88%, anger is 89%, disgust 78%, boredom is 83%, anxiety is 70%, and neutrality is 89%. These WPC features attain an improvement of 28.1%, 16.8%, 4.7% and 2.4%, respectively, for happy, sad, angry and disgust emotions, compared with the results of [19]. In [21], features based on the energy content of wavelet-based time-frequency (TF) representations were extracted for classification, with the highest UAR being 69.3%. While the WPC features achieve an average UAR 79.2%. In [46], the modulation spectral features (MSFs) are proposed for the automatic emotion recognition. The best recognition rates of happiness, sadness, anger and disgust are 70.4%, 91.9%, 90.6% and 76.1%, respectively, for speaker-dependent classification. MSFs achieve 78.4% and 77.0% on average for speaker-independent classification using four different feature sets of MSFs and two different feature selection method. In [47], Deep Neural Network-Hidden Markov Model (DNN-HMM) was proposed and widely used in speech recognition. Their approach obtained the highest accuracy up to 77.92% on the Berlin database, while our results attain a boost of 1.58%. The Wavelet packet coefficient features are effective for recognizing happy, sad, angry and disgust emotions but ineffective for recognizing neutral, boredom and anxiety emotions. Classifying seven emotions is more difficult than classifying six emotions for EMODB database [33,48,49]. Since EESDB is about the elderly emotion, the emotions expressed by the elderly are usually more difficult to identify [50]. Although our approach is not the best, overall it is affordable for practical usage in speaker-independent emotion recognition.

In summary, both LSVM and RSVM are used as classifiers in our method. The proposed WPC features can improve speaker-independent emotion recognition by approximately 1.8% and 5.5% on the EMODB and EESDB databases, respectively, compared with MFCC features. The recognition rates can be further enhanced by 16.2% and 7.2% by using SFFS feature selection on the two databases.

5. Conclusion

In this study, we described the wavelet packet features for examining the impact on the performance of speech emotion recognition. We showed that the performance of wavelet packet coefficient features is comparable with MFCC features. To reduce the feature space, we carried out a SFFS feature selection approach while achieving better recognition accuracy. Results demonstrate that the WPC features together with SFFS can achieve the best recognition rate. Additionally, a combination of MFCC and WPC sets further improves recognition rates. By using the leave-one-speaker-out scheme, we achieve a better recognition rate of 98.6%, 96.8%, 93.7% and 80.4% for happiness, sadness, anger and disgust respectively on the EMODB database. This demonstrates that the extracted WPC features successfully capture the vocal characteristics of emotional data. In our future work, we plan to extend the proposed features and evaluate their performance on naturalistic speech data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Kunxia Wang: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Funding acquisition. **Guoxin Su:** Writing - review & editing. **Li Liu:** Conceptualization, Methodology, Writing - review & editing, Supervision, Project administration, Funding acquisition. **Shu Wang:** Validation, Investigation, Supervision.

Acknowledgment

This work was supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (grant nos. 201700014), Anhui Provincial Natural Science Foundation (grant nos. 1708085MF167), the National Major Science and Technology Projects of China (grant nos. 2018AAA0100703, 2018AAA0100700), the National Natural Science Foundation of China (grant no. 61977012), the Chongqing Provincial Human Resource and Social Security Department (grant no. cx2017092), the Central Universities in China (grant nos. 2019CDJGFD-SJ001, CQU0225001104447 and 2018CDXYRJ0030). Any correspondence should be made to Li Liu.

References

[1] N. Fragopanagos, J.G. Taylor, Emotion recognition in human computer interaction, Neural Netw. 18 (4) (2005) 389–405.
[2] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, Int. J. Speech Technol. 21 (1) (2018) 93–120.
[3] M.E. Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recogn. 44 (3) (2011) 572–587.
[4] L. Liu, Y.X. Peng, Z.G. Huang, M. Liu, Sensor-based human activity recognition system with a multilayered model using time series shapelets, Knowl.-Based Syst. 90 (2015) 138–152.
[5] L. Liu, S. Wang, B. Hu, Q.Y. Xiong, J.H. Wen, D.S. Rosenblum, Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition, Pattern Recogn. 81 (2018) 545–561.
[6] L. Liu, S. Wang, G.X. Su, B. Hu, Y.X. Peng, Q.Y. Xiong, J.H. Wen, A framework of mining semantic-based probabilistic event relations for complex activity recognition, Inf. Sci. 418–419 (2017) 13–33.
[7] J. Silva, S.S. Narayanan, Discriminative wavelet packet filter bank selection for pattern recognition, IEEE Trans. Signal Process. 57 (5) (2009) 1796–1810.
[8] Y. Ghanbari, M.R. Karami-Mollaei, A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets, Speech Commun. 48 (8) (2006) 927–940.
[9] K.D. Rao, M.N.S. Swamy, Discrete wavelet transforms, in: Digital Signal Processing, Springer, Singapore, 2018, pp. 619–691.
[10] I. Daubechies, The wavelet transform, time-frequency localization and signal analysis, IEEE Trans. Inf. Theory 36 (5) (1990) 961–1005.
[11] M.T. Islam, C. Shahnaz, W.P. Zhu, Speech enhancement based on student t modeling of teager energy operated perceptual wavelet packet coefficients and a custom thresholding function, IEEE/ACM Trans. Audio Speech Lang. Process. (TASLP) 23 (11) (2015) 1800–1811.
[12] M.T. Islam, C. Shahnaz, W.P. Zhu, Rayleigh modeling of teager energy operated perceptual wavelet packet coefficients for enhancing noisy speech, Speech Commun. 86 (2017) 64–74.
[13] M. Hariharan, R. Sindhu, V. Vijean, Improved binary dragonfly optimization algorithm and wavelet packet based non-linear features for infant cry classification, Comput. Methods Progr. Biomed. 155 (2018) 39–51.

[14] P.S. Addison, The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine And finance, CRC Press, 2017.
[15] G. Balasubramanian, A. Kanagasabai, J. Mohan, Music induced emotion using wavelet packet decomposition an EEG study, Biomed. Signal Process. Control 42 (2018) 115–128.
[16] G. Balasubramanian, A. Kanagasabai, J. Mohan, et al., Music induced emotion using wavelet packet decomposition an EEG study, Biomed. Signal Process. Control 42 (2018) 115–128.
[17] S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (7) (1989) 674–693.
[18] S. Mallat, S. Zhong, Characterization of signals from multiscale edges, IEEE Trans. Pattern Anal. Mach. Intell. 16 (1992) 710C732.
[19] L. Zao, D. Cavalcante, R. Coelho, Time-frequency feature and AMS-GMM mask for acoustic emotion classification, IEEE Signal Process. Lett. 21 (5) (2014) 620–624.
[20] Y.D. Zhang, Z.J. Yang, H.M. Lu, et al., Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation, IEEE Access. 4 (2016) 8375–8385.
[21] J.C. Vasquez-Correa, T. Arias-Vergara, J.R. Orozco-Arroyave, J.F. Vargas-Bonilla, E. Noeth, Wavelet-based time-frequency representations for automatic recognition of emotions from speech, in: Proceedings of the ITG Symposium, VDE Speech Communication, 12, 2016, pp. 1–5.
[22] H. Muthusamy, K. Polat, S. Yaacob, Particle swarm optimization based feature enhancement and feature selection for improved emotion recognition in speech and glottal signals, PloS One 10 (3) (2015).
[23] B.S. Zheng, W. Khairunizam, S.A.M. Murugappan, et al., Effectiveness of tuned q-factor wavelet transform in emotion recognition among left-brain damaged stroke patients, Int. J. Simul.-Syst. Sci. Technol. 19 (3) (2018).
[24] M. Hariharan, R. Sindhu, V. Vijean, Improved binary dragonfly optimization algorithm and wavelet packet based non-linear features for infant cry classification, Comput. Methods Progr. Biomed. 155 (2018) 39–51.
[25] K. Wang, N. An, L. Li, Speech emotion recognition based on wavelet packet coefficient model, Proceeding of the 2014 Ninth International Symposium on Chinese Spoken Language Processing (ISCSLP) IEEE, 2014.
[26] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of german emotional speech, in: Proceeding of the INTERSPEECH2005: 1517C1520, 2005.
[27] M. Kotti, F. Patern, Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema, Int. J. Speech Technol. (2012) 131–150.
[28] G.F. Choueiter, J.R. Glass, An implementation of rational wavelets and filter design for phonetic classification, audio, speech, and language processing, IEEE Trans. 15 (3) (2007) 939–948.
[29] K.X. Wang, Q.L. Zhang, S.Y. Liao, A database of elderly emotional speech, in: Proceeding of the International Symposium on Signal Processing Biomedical Engineering, and Informatics (SPBEI), 2014, pp. 549–553.
[30] C. Busso, S. Lee, S. Narayanan, Analysis of emotionally salient aspects of fundamental frequency for emotion detection, audio, speech, and language processing, IEEE Trans. 17 (4) (2009) 582–596.
[31] M. Grimm, K. Kroschel, E. Mower, S. Narayanan, Primitivesbased evaluation and estimation of emotions in speech, Speech Commun. 49 (2007) 787–800.
[32] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette, Feartype emotion recognition for future audio-based surveillance systems, Speech Commun. 50 (2008) 487–503.
[33] K. Wang, N. An, B. Li, Y. Zhang, Speech emotion recognition using fourier parameters affective computing, IEEE Trans. 6 (1) (2015) 69–75.
[34] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Audio Speech Lang. Process. 28 (1980) 357–366.
[35] Q. Li, J. Zheng, A. Tsai, Robust endpoint detection and energy normalization for real-time speech and speaker recognition, Speech Audio Process. IEEE Trans. 10 (3) (2002) 146–157.
[36] A. Kumar, The optimized wavelet filters for speech compression, Int. J. Speech Technol. 16 (2) (2013) 171–179.
[37] F. Germain, The Wavelet Transform Applications in Music Information Retrieval, McGill University, Canada, 2009.
[38] S. Fateri, N. V. Boulgouris, A. Wilkinson, W. Balachandran, T.H. Gan, Frequency-sweep examination for wavelet mode identification in multimodal ultrasonic guided wavelet signal. ultrasonics, ferroelectrics, and frequency control, IEEE Trans. 61 (9) (2014) 1515–1524.
[39] C. N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, Artif. Intell. Rev. (2012) 1–23.
[40] X. Chen, G. Yuan, F. Nie, et al., Semi–supervised feature, in: Selection via Rescaled Linear Regression IJCAI. 2017, 2017, pp. 1525–1531.
[41] R.C. Guido, Paraconsistent feature engineering [lecture notes], IEEE Signal Process. Mag. 36 (1) (2018) 154–158.
[42] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (11) (1994) 1119–1125.
[43] B. Schuller, A. Batliner, S. Steidl, D. Seppi, Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge, Speech Commun. 53 (9/10) (2011) 1062–1087.
[44] F. Eyben, A. Batliner, B. Schuller, Towards a standard set of acoustic features for the processing of emotion in speech, in: Proceedings of Meetings on Acoustics,volume 9, 2010, pp. 1–12.
[45] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller, Deep neural networks for acoustic emotion recognition: raising the benchmarks, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), 2011, pp. 5688–5691.
[46] S. Wu, T.H. Falk, W.Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Commun. 53 (5) (2011) 768–785.
[47] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, H. Sahli, Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition, in: Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 312–317.
[48] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech Commun. 52 (7–8) (2010) 613–625.
[49] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: a review, Int. J. Speech Technol. 21 (1) (2018) 93–120.
[50] I.S. Engberg, A.V. Hansen, Documentation of the Danish Emotional Speech Database (DES), Department of Communication Technology, Institute of Electronic System, Aalborg University, Denmark, 1996.

Kunxia Wang works in the Department of Electronic Engineering, Anhui University of Architecture, Hefei, China. She is currently an associate professor. She received her Ph.D. in Computer Science from the School of Computer and Information, Hefei University of Technology in 2008. Her research interests are in affective computing and their applications. She has published widely in conferences and journals with more than 40 peer-reviewed publications.

Guoxin Su is a lecturer at the University of Wollongong (UOW) in Australia. He received his Ph.D. from the University of Technology Sydney in 2013. Before joining UOW, he was a research fellow and a senior research fellow with the School of Computing at the National University of Singapore. His research areas include software engineering, formal methods and big data. He has published a series of papers on the top-ranked journals and conferences

Li Liu is currently an associate professor at Chongqing University. He received his Ph.D. in Computer Science from the Université Paris-sud XI in 2008. He had served as an associate professor at Lanzhou University in China and also a Senior Research Fellow of School of Computing at the National University of Singapore. His research interests are in pattern recognition, data analysis, and their applications on human behaviors. He aims to contribute in interdisciplinary research of computer science and human related disciplines. Li has published widely in conferences and journals with more than 100 peer-reviewed publications. Li has been the Principal Investigator of several funded projects from government and industry.

Shu Wang is currently a lecture at Southwest University. She received her Ph.D. from Lanzhou University. She had severed as an associate professor at Lanzhou University of Technology. She has great research interests in sensor development techniques and electronic materials for sensors.