Biomedical Signal Processing and Control 18 (2015) 80–90

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

Weighted spectral features based on local Hu moments for speech emotion recognition

Yaxin Sun, Guihua Wen*, Jiabing Wang

School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

ARTICLE INFO

Article history:
Received 17 February 2014
Received in revised form 23 July 2014
Accepted 17 October 2014
Available online 2 January 2015

Keywords:
Speech emotion recognition
Speech spectral features
Feature extraction
Hu moments

ABSTRACT

Features greatly influence the results of speech emotion recognition, among which Mel-frequency Cepstral Coefficients (MFCC) is the most commonly used in speech emotion. However, MFCC does not consider both the relationship among neighbor coefficients of Mel filters of a frame and the relationship among coefficients of Mel filters of neighbor frames, which possibly leads to lose many useful features from spectrogram. This paper presents novel weighted spectral features based on Local Hu moments. The idea is motivated by that the energy on spectrogram would drastically vary with some emotion types such as angry and happy, while it would slightly change with other emotion types such as sadness and fear. This phenomenon would affect the local energy distribution of spectrogram in both time axis and frequency axis of spectrogram. To describe local energy distribution of spectrogram, Hu moments computed from local regions of spectrogram are used, as Hu moments can evaluate the degree how the energy is concentrated to the center of energy gravity of local region of spectrogram and can significantly vary with the speech emotion types. The conducted experiments validate the proposed features in terms of the effectiveness of the speech emotion recognition.

1. Introduction

Speech is one of the most important communication means of human beings. It is much expected that human emotions can automatically be recognized by machines to improve human machine interaction [1]. With the growth in the electronic and computer technologies, new spoken dialog systems with emotion recognition capability are needed. For example, a nursing robot that can continuously monitor the patients' emotional state could provide better health-care services for patients [2]. However, it is very challenging to achieve this goal for the following reasons. Firstly, it is not clear which speech features are most powerful in distinguishing emotions [3]. Secondly, the acoustic variability introduced by the different sentences, speakers, and speaking styles adds another obstacle [2,3]. In addition, speech signals with noises can also change acoustics' distribution greatly.

To solve the above problems, many kinds of speech feature types have been presented. Among them, Mel-frequency Cepstral Coefficients (MFCC) is the most commonly used feature type [4,8,11,14–23,26,28,35–37]. It reduces some negative impact produced by Discrete Cosine Transform (DCT) for noisy and different sentences, speakers, and speaking styles. However, MFCC does not consider both the relationship among neighbor coefficients of Mel filters of a frame and the relationship among coefficients of Mel filters of neighbor frames, which may lose many useful features from spectrogram.

There are some methods that can characterize both kinds of relationships among coefficients of Mel filters, such as mean, standard, entropy, skewness, kurtosis, and so on. However, these methods have two defects. One is that the differences among emotions could not be well discovered by mean, standard deviation, skewness and kurtosis. The other is that some methods can be easily influenced by the differences among the sentences, speakers, and speaking styles, which is much bad for speech emotion recognition. In this paper, Hu moments [25] are utilized to overcome these two problems.

Hu Moments have been widely used as basic feature descriptors in image analysis, pattern and object recognition, image classification, and template matching [59–61]. Hu moments have two advantages: (1) the first absolute orthogonal invariant of Hu moments can evaluate the degree how the energy is concentrated to the center of energy gravity for two dimensional data; (2) Hu moments are invariant with respect to translation, scaling, as well as rotation [25,59–61]. It is reasonable to conclude that Hu moments could be good to characterize the relationship among coefficients of neighbor Mel filters within a frame, as well

* Corresponding author. Tel.: +86 18998384808.
E-mail addresses: sunyaxin2005@163.com (Y. Sun), crghwen@scut.edu.cn (G. Wen).

http://dx.doi.org/10.1016/j.bspc.2014.10.008
1746-8094/© 2014 Elsevier Ltd. All rights reserved.

as the relationship among coefficients of Mel filters of neighbor frames.

Firstly, Hu moments can be used to extract some differences among different emotions. Under different speech emotions, the pronounced strength, the articulation, the degree of changing pitch frequency, and the pronounced speed would obviously be changed [50–52]. These changes will alter the degree how the energy concentrated to some frequencies in a spectrogram. For example, energy is more concentrated in certain frequency in spectrogram when the utterance owns better articulation and higher pronounced strength. Obviously, if Hu moments are computed in local regions of a spectrogram, they can evaluate the degree how the energy concentrated to some frequencies in a spectrogram. It indicates that Hu moments have good ability to extract the differences among the emotions.

Secondly, Hu moments can reduce the changes introduced by the sentences, the speakers, and the speaking styles, because Hu moments are invariant with respect to translation, scaling, as well as rotation. For example, different words may own different formant frequencies, easily leading to translation and rotation of energy. However, Hu moments computed in local regions of a spectrogram can effectively reduce these negative influences.

The remainder of this paper is organized as follows. Section 2 introduces the proposed features. Section 3 presents the other features that can be combined with the proposed features to reach the better performance for speech emotion recognition. Section 4 presents the speech emotion framework used in our context. Experimental results are presented and discussed in Section 5, while Section 6 gives concluding remarks.

2. Proposed spectral features

This section introduces the proposed weighted spectral features based on Local Hu moments, denoted by HuWSF.

2.1. Background of Hu moments

Hu moments have been used as basic feature descriptors for images, for the reasons that they are invariant with respect to translation, scaling, as well as rotation. Here only the first absolute orthogonal invariant of Hu moments is used as a feature descriptor for speech, because it can also evaluate the degree how the energy is concentrated to the center of energy gravity of two-dimensional data. For a density distribution function $g(u, v)$, the first absolute orthogonal invariant of Hu moments is defined as:

$$\theta = \eta 20 + \eta 02 \tag{1}$$

$$\eta pq = \frac{\mu pq}{(\varphi 00^{\rho})}, \ \rho = \frac{(p+q)}{2+1} \tag{2}$$

$$\mu pq = \sum_{u=1}^{U}\sum_{v=1}^{V}(u-\bar{u})^{p}(v-\bar{v})^{q}g(u,v), \quad p,q = 0,1,2\cdots \tag{3}$$

$$\varphi pq = \sum_{u=1}^{U}\sum_{v=1}^{V}u^{p}v^{q}g(u,v), \quad p,q = 0,1,2\cdots \tag{4}$$

where ηpq is $(p+q)$-th order normalized center moment, μpq is $(p+q)$-th order center moment, φpq is $(p+q)$-th order moment, and $(\bar{u}, \bar{v})$ is the center of energy gravity, where $\bar{u} = \varphi 10/\varphi 00$, $\bar{v} = \varphi 01/\varphi 00$.

2.2. HuWSF algorithm

The flowchart of extracting HuWSF is presented in Fig. 1, where the flowchart of extracting MFCC is also presented with an attempt to make comparison between HuWSF and MFCC. It can be seen from Fig. 1 that the main difference between HuWSF and MFCC lies in that MFCC is calculated directly from E while HuWSF is calculated from Q, where Q is constructed by θ which computed from local blocks on E.

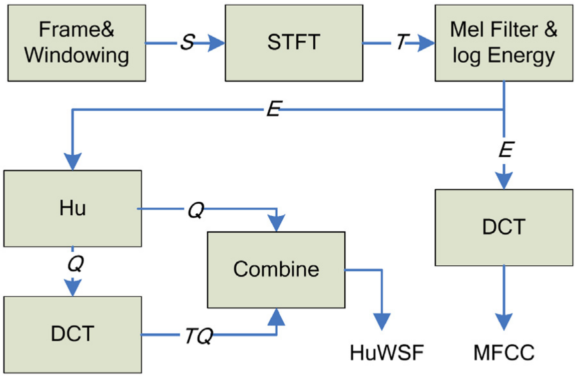


Fig. 1. The flowchart of extracting HuWSF

As described in Fig. 1, HuWSF includes six main steps. *In the first step*, the speech signal S is framed into short-term segments S_k by multiplying w_l samples width Hamming window with w_s samples shift, where k denotes the frame index, $wl = 1.81 * fs/b$, $ws = 1.81/(4*b)$ [33], f_s is the sampling rate of S, and b is the frequency resolution of short time Fourier transform(STFT).

In the second step, T_k is computed by STFT from S_k.

In the third step, Mel frequency Mel_k and the log energy $E_k(m)$ are computed by Eqs. (5) and (6) respectively,

$$Melk = 2595 * lg\left(\frac{1+T_k}{700}\right) \tag{5}$$

$$Ek(m) = \ln\left(\sum_{f=fl}^{fh}\left|Mel_k(f)\right|^{2}H_m(f)\right), \quad 0 \le m < M \tag{6}$$

where f_l and f_h, assigned by f_l = 0HZ and f_h = 4000HZ, are the low and high boundaries of the computed speech frequency, $H_m(f)$ defines a triangular filter by Eq. (7), and M is the number of Mel filters.

$$H_m(f) = \begin{cases} 0, & f < C(m-1) \\ \dfrac{2(f - C(m-1))}{(C(m+1) - C(m-1))(C(m) - C(m-1))}, & C(m-1) \le f < C(m) \\ \dfrac{2(C(m+1) - f)}{(C(m+1) - C(m-1))(C(m+1) - C(m))}, & C(m) \le f < C(m+1) \\ 0, & f \ge C(m+1) \end{cases} \tag{7}$$

where $\sum_{m=0}^{M-1} H_m(n) = 1$, $C(m)$, $m = 1, 2, \cdots M$ is the center frequency of triangular, and the interval between two $C(m)$ increases with m.

The fourth step computes the local Hu moments Q of E. To compute local Hu moments Q, E is firstly divided into $(K-w+1)*(M/w)$ blocks Bij by Eq. (8), where w is the size of block. The blocks are square with the reason that b can control frequency resolution of STFT, which can balance the relationship between time domain and frequency domain. Secondly, θ of Bij is computed by Eq. (1), where $g(u, v)$ in Eqs. (3) and (4) can be defined by Eq. (9). Finally, all θ are used to construct $Q \in R^{(k-w+1)*(M/w)}$.

$$Bij = \begin{bmatrix} Ei(w*j) & \cdots & Ei(w*j+w-1) \\ \vdots & \vdots & \vdots \\ Ei+w-1(w*j) & \cdots & Ei+w-1(w*j+w-1) \end{bmatrix} \tag{8}$$
$$i = 1\cdots K-w+1, \ j = 1\cdots M/w$$

$$g(u, v) = Bij(u, v) \tag{9}$$

In the fifth step, $TQk \in R^{1*(M/w)}$ is computed by Eq. (10) from Q_k. Here only the values from the 2-nd to 13-th coefficients of TQ_k are selected, as the high cepstral coefficients may be distorted by noise.

$$TQk(m) = Dm\sum_{n=1}^{M/w} Qk(n)\cos\frac{(2n-1)m\pi}{2(M/w)} \tag{10}$$

where D_m is defined as

$$Dm = \begin{cases} \sqrt{\dfrac{1}{M/w}} & m = 1 \\ \sqrt{\dfrac{2}{M/w}} & m = 2, \cdots M/w \end{cases} \tag{11}$$

In the sixth step, HuWSF is generated by combining Q *with* TQ, whose dimension is $(K-w+1)*(M/w+12)$.

The algorithm of HuWSF is summarized as Algorithm 1. The code of HuWSF can be downloaded from: http://yunpan.cn/QCKNH3GyMR4wi (password: d5af).

Algorithm 1. HuWSF

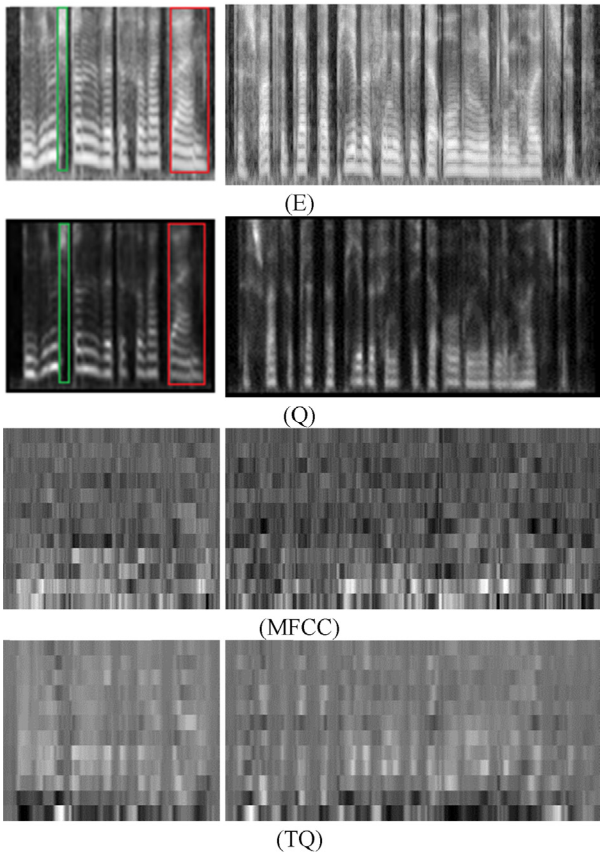
```
Input: a speech signal $S$, the frequency resolution of STFT $b$, the block width $w$,
  and the number of Mel filters $M$.
Output: HuWSF features $F$
Algorithm:
  1: Frame $S$ into short-term segments $S_k k = 1, 2, \cdots K$ by hamming window,
  where the sampling rate $f_s$ can be achieved from $S$.
  2: For $k = 1$ to $K$
       Compute STFT results $T_k$ of all $S_k$.
     End For
  3: For $k = 1$ to $K$
       Compute log energy $E_k(m)$ by Eq. (6).
     End For
  //Compute Q
  4: Divide $E$ into $(K-w+1)*(M/w)$ blocks by Eq. (8) and compute $\theta$ for each
  block by Eq. (1). As a result, $\theta$ of all blocks are constructed $Q$.
  //Computed TQ
  5: For $k = 1$ to $K$-$w$+1
       Compute $TQk \in R^{1*(M/w)}$ from $Q_k$. by Eq. (10)
     End For
  //Get HuWSF
  6: Compute HuWSF features $F$ by connecting $Q$ and $TQ$ in frequency
  dimension as $F = [Q\ TQ]$
```

Similar to the reference [14] which uses the dynamic feature vector of MFCC by appending the first and the second time derivatives of MFCC into the feature vector, we also use the dynamic feature vector of HuWSF, defined by $FG\Delta = [FG, \Delta FG, \Delta\Delta FG]$ where F_G represents HuWSF, ΔFG represents the first-order derivative of F_G, and $\Delta\Delta FG$ represents the second-order derivative of F_G.

2.3. The analysis of HuWSF

It can be seen from Eqs. (1)–(4) that θ can evaluate the degree of how the energy concentrated to the center of gravity, which is proved in appendix. We compute each θ in local blocks of spectrogram, so that θ can evaluate the degree of how energy of frequencies concentrates to some frequencies in a spectrogram, which may significantly vary with the speech emotion types. Furthermore, θ is not only independent of position, size, and orientation, but also independent of the parallel projection [25], so that the differences among the sentences, speakers, and speaking styles can be reduced by θ. Because HuWSF is based on θ, HuWSF share the advantages of θ.

To further illustrate the advantages of HuWSF, we present the visualization results of Q and TQ, which are components of HuWSF. We also give the visualization results of E that are the results of the third step of Algorithm 1. The visualization results of MFCC are also presented for comparison, as HuWSF are improved from MFCC. The visualization results of E, Q, MFCC, and TQ are presented respectively from the top line to the bottom line in Fig. 2. The pictures in each column of Fig. 2 are the experimental results of an utterance that are randomly selected from Berlin Speech Emotion Database (EmoDB), where the red box in the first and second rows of Fig. 2 indicates an example of vowel part, and the green box in the first and second rows of Fig. 2 indicates an example of voiceless part.



(E)

(Q)

(MFCC)

(TQ)

Fig. 2. Visualization results of E, Q, MFCC and TQ

By comparison with the visualization results of E and Q, Q has the following advantages. Firstly, in the vowel part of Q, strong energy peaks are more clearly and weak energy peaks have the opposite effect. This is very significant as the strong energy peaks are often more important for extracting features from speech signals [13,15,50–52]. Secondly, in the voiceless part of Q, the energy distributions of Q is smoother, showing that Q is also the more useful features for the part of voiceless of speech signal. The above advantages are illustrated in more detail in Appendix A.

Compared with the visualization results of MFCC, TQ owns the advantages as follows. Firstly, the boundaries between the part of voice activity and silence are more clear. Secondly, in the part of voice activity, the coefficients among the neighbor frames of TQ are smoother than that of MFCC, which are more in line with the characteristics of the speech signal. These advantages are also illustrated in more details in Appendix A.

As HuWSF is formed by Q and TQ, HuWSF shares the advantages of both Q and TQ.

3. More features related to HuSWF

The currently used features in speech emotion recognition can be also applied to combine with HuSWF to further improve the performance. This is because most of the speech emotion recognition methods often use more than one feature extraction method to extract features [1,11,16,26,39,56,58] and the combined features can greatly enhance the effect of speech emotion recognition. These

Table 1
Feature extraction methods and feature statistical methods.

References	Feature types	Feature statistical types
Ref. [1]	Mel filter-bank, pitch, HNR, PWP	Mean, variance, median, first and third quartiles
Ref. [11]	Pitch, formant frequency, jitter, shimmer, MFCC, duration	Mean, std, min, max, some features use local derivative
Ref. [16]	LogEnergy, Pitch, ZCR, MFCC, spectral melspectra0-25, Frequency Band	Mean, variance, skew, kurtosis, maximum, minimum, median, quartiles, range
Ref. [26]	Energy, low pass energy, high pass energy, pitch, MFCC	Mean, maximum, minimum, range, variance, median, first quartile, third quartile, inter-quartile range, mean absolute value of the local derivative, some features use local derivative
Ref. [39]	Pitch, energy, envelop, MFCC1-16, Formant1-5amplitude, HNR, shimmer, jitter Formant1-5 bandwidth, Formant1-5 position	Max, min, range, quartile, kurtosis, centroid. all features use local derivative
Ref. [56]	Duration, energy, pitch, spectrum, MFCC, PLP, NHR, HNR, shimmer, Wavelets, jitter, TRAP, Teager operator	Minimum, maximum, onset, offset, duration, regression
Ref. [58]	ZCR, RMS, pitch logarithmic frame energy HNR, MFCC	Mean, absolute mean, kurtosis, standard deviation skewness, variance, minimum, range, maximum, relative position

features can be divided into two categories: spectral features and prosodic features [3].

3.1. Spectral features

In addition to MFCC and HuWSF, there are some other features belonging to spectral features, such as LPCC (Linear Predictor Cepstral Coefficients) [5,17,22,28,32,35,36], ZCPA(Zero Crossings with Peak Amplitudes) [13,15,38], Harmonic [4], MSF (Modulation Spectral Features) [8], LFPC (Log Frequency Power Coefficients)[6,28,35,40], PLP (Perceptual Linear Predictive) and RASTA-PLP(RASTA Perceptual Linear Predictive) [7,8,40], WMFCC (Weighted Mel-frequency Cepstral Coefficient) [14], and RSS (Ratio of a Spectral flatness measure to a Spectral center) [22]. Among them, some will be compared with HuWSF in the later experiments.

3.2. Prosodic features

Prosodic features are often used together with spectral features in speech emotion recognition, as they have good supplement effectiveness to spectral features. However, it can be seen from the second column of Table 1 that the prosodic features used in different references are very different [1,11,16,26,39,56,58]. To prove that HuWSF has a better supplement effectiveness to prosodic features than the other spectral features, the combined prosodic features should be typical and complete. Fortunately, Björn Schuller proposed five different baseline acoustic feature sets for speech emotion recognition [45–49]. Here INTERSPEECH 2010 [45] feature set is selected, for the reasons that INTERSPEECH 2010 feature set contains most prosodic features that are used in the references [1,11,16,26,39,56–58].

Table 2
Description of the acoustic features based on 38 LLD and their first derivate and 21 feature statistics functions.

Descriptors	Functions
PCM loudness	maxPos, minPos, mean,
MFCC[0–14]	stddev, skewness, kurtosis,
Log Mel freq. band[0–7]	quartile1/2/3
LSP frequency[0–7]	quartile range (2–1)/(3–1)/(3–1)
F0 envelop	lin.regression coeff.1/2
Voicing probability	lin.regression error Q/A
F0final	percentile 1/99
jitterLocal, jitterDDP	percentile range (99–1)
shimmerLocal	up-level time75/90

INTERSPEECH 2010 feature set results from a base of 34 Low-Level Descriptors (LLD) with 34 corresponding delta coefficients appended, and 21 functions applied to each of these 68 LLD contours (1428-features). In addition, 19 functions are applied to the 4 pitch-based LLD and their four delta coefficient contours (152 features), where 19 functions are selected from 21 functions mentioned by removing the minimum value and the range functions. Finally the number of pitch onsets (pseudo syllables) and the total duration of the input are appended (2 features). Table 2 gives an overview of the low-level descriptors and associated feature statistics functions. The details of each item can be seen from reference [45]. This feature set can be obtained by the toolbox named OpenSmile [29].

It can be seen from Table 2 that INTERSPEECH 2010 feature set contains MFCC features and prosodic features. In order to obtain the prosodic features, we remove MFCC features from INTERSPEECH 2010 feature set. The remaining features are prosodic features, which will be used in the later experiments. We name this feature set as PROS.

4. The speech emotion recognition framework used to evaluate HuWSF

To evaluate the HuSWF, typical speech emotion recognition framework is applied, which includes four steps [9,11,15–19,23]: feature extraction, feature statistics, dimensional reduction, and classification. HuSWF is performed for feature extraction. The remaining three steps in the framework will be simply introduced as follows.

4.1. Feature statistics

Feature statistics estimates the global statistics of the extracted features in order to form a feature vector for the speech signal. The global statistics are useful in speech emotion recognition, as they are less sensitive to linguistic information. Currently, many statistical methods have been applied to estimate the global statistics of the extracted features. Some of them are illustrated in the third column of Table 1.

In our framework, the following feature statistics methods are selected: *mean, std, max, min, kurtosis, skewness*, and *median*. They are the most used feature statistics methods in the recent speech emotion recognition [1,11,16,26,39]. The feature statistics of the spectral features are computed by the above feature statistics methods, while the feature statistics of the prosodic features are computed directly by OpenSmile toolbox [29].

4.2. Dimension reduction

There may be high correlation among the global statistics features, which could have negative influence on speech emotion recognition. Currently there are many approaches performing

Table 3
WA of PI experiments on EmoDB with different dimension reduction methods, where the maximum WA for each dimension reduction method is shown in bold font.

	MCFS	MRMR	DISR	LDA
HuWSF	**74.71**	72.46	70.75	39.30
LPCC	**56.58**	53.28	53.00	47.69
MFCC	**66.25**	63.76	62.33	53.36
PLP	**73.43**	71.11	70.56	48.32
ZCPA	**54.59**	51.40	52.75	45.86

Table 4
WA of HuWSF against M and b, where the WA of HuWSF with the selected M and bis shown in bold font.

b	M				
	80	100	120	140	160
40	73.86	72.16	72.42	73.16	73.83
50	75.67	74.19	73.90	73.19	74.33
60	**74.71**	74.33	74.01	76.29	73.20
70	75.74	73.76	73.76	74.27	72.84
80	74.27	74.79	72.52	73.38	71.32

Table 5
WA of MFCC against M and b, where the WA of MFCC with the selected M and b is shown in bold font.

b	M				
	20	40	**60**	80	100
40	64.84	66.02	63.55	64.42	64.38
50	63.44	65.48	**66.25**	65.08	65.18
60	60.37	64.64	64.99	65.24	65.10
70	59.65	62.54	61.26	63.91	63.68
80	63.09	63.45	65.10	64.38	64.64

dimensionality reduction to solve the problem. For example, PCA (Principal Component Analysis) [38] and LDA (Linear Discriminant Analysis) [8] can be used to perform feature extraction. mRMR (Minimum Redundancy Maximum Relevance Feature Selection) [27,30,31], SFS (Sequential Forward Feature Selection) [3,8,9,17], SFFS (Sequential Forward float Feature Selection) [3], MCFS(multi-cluster feature selection) [44], and DISR(Double Input Symmetrical Relevance) [43] can be applied to perform the feature selection. Which one is the best in the framework will be discussed in the later experiments.

4.3. Classification

Classification aims to obtain the emotion type for the input speech feature vector. Typical classification approaches include HMM [19,20,28], GMM [18–23,35,37], ANN [18,38], K-NN [17,18,21,26,42], SVM [8,10,11,15,16,18,23,26,38,54], and Bayesian classifier [12,15].

In our framework, SVM is used for Classification, where the toolbox of SVM named LIBSVM is used [24] with the polynomial kernel, as polynomial kernel nearly owns the same results as that of RBF kernel.

5. Experiments

5.1. Speech emotion databases

To validate HuWSF, experiments are conducted on three speech databases.

Berlin emotional speech database (EmoDB) [34]. It is a German database and one of the most popular databases used for emotion recognition, so that it can facilitate comparisons with the related works. The numbers of speech files for each emotion category are: anger (127), anxiety fear (69), boredom (81), disgust (46), happiness (71), neutral (79), and sadness (62). Ten actors (5 male and 5 female) speak 49, 58, 43, 38, 55, 35, 61, 69, 56, 71 utterances respectively in the final database.

Table 6
The parameter setting for HuWSF, MFCC, PLP, LPCC, and ZCPA.

	HuWSF			MFCC		PLP	LPCC		ZCPA
	w	M	b	M	b	p	p	b	M
EmoDB	5	80	60	60	50	12	14	60	10
SAVEE	5	140	70	40	40	12	12	60	10
CASIA	3	160	60	60	70	12	14	60	10

Table 7
WA of SI experiments for SPEC#, PROS# and OTHER# on three database, where the maximum and second maximum WA for SPEC#, PROS# and OTHER# on three database are shown in bold and italics fonts.

Features	WA(EmoDB)	WA(SAVEE)	WA(CASIA)
HuWSF	**74.71**	**45.42**	**41.92**
LPCC	56.58	41.04	31.42
MFCC	66.25	37.71	36.17
PLP	*73.43*	*43.96*	*39.33*
ZCPA	54.59	34.17	29.33
PROS + HuWSF	**81.01**	46.88	**43.50**
PROS + LPCC	75.67	47.08	41.50
PROS + MFCC	78.41	**47.92**	41.67
PROS + PLP	*79.67*	*47.08*	40.75
PROS + ZCPA	77.02	46.67	39.83
PROS	76.23	47.08	*41.83*
OTHER + HuWSF	**81.74**	*48.75*	**43.17**
OTHER + MFCC	79.24	47.50	41.58
OTHER	*81.44*	47.50	42.33
OTHER + MFCC + HuWSF	80.26	**50.00**	*42.75*

Surrey Audio-Visual Expressed Emotion Database (SAVEE) [41]. It is an English database that consists of recordings from 4 male actors in 7 different emotions. The numbers of speech files for each emotion category are: anger (60), disgust (60), fear (60), happiness (60), sadness (60), surprise (60), and neutral (120). Each speaker says 120 utterances. The sentences were selected from the standard TIMIT corpus and phonetically balanced for each emotion.

Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA) [53]. It is a Chinese database that consists of recordings from 4 actors in 6 different emotions. The numbers of speech files for each emotion category are: anger (200), fear (200), happiness (200), sadness (200), surprise (200), and neutral (200). Each speaker says 300 utterances.

5.2. Performance evaluation criteria

Two experimental strategies are used. They are Speaker-independent (SI) [6,21] and Speaker-dependent (SD) [6,21]. In SI strategy, for each fold, all utterances from one of the speakers were sequentially assigned to the hold-out set used for testing data and the utterances of the remaining speakers were used as the labeled data for training data. In SD strategy, all utterances of each emotion are randomly divided into five equal parts, among which four parts are taken as the training data and the remained one is taken as the testing data. This procedure is repeated ten times, and the average classification results across all trials were computed.

The Weighted Average Recall (WA) is employed to evaluate all the approaches. WA is the total number of correctly classified test samples of all classes averaged by the total number of test samples [16,20,55].

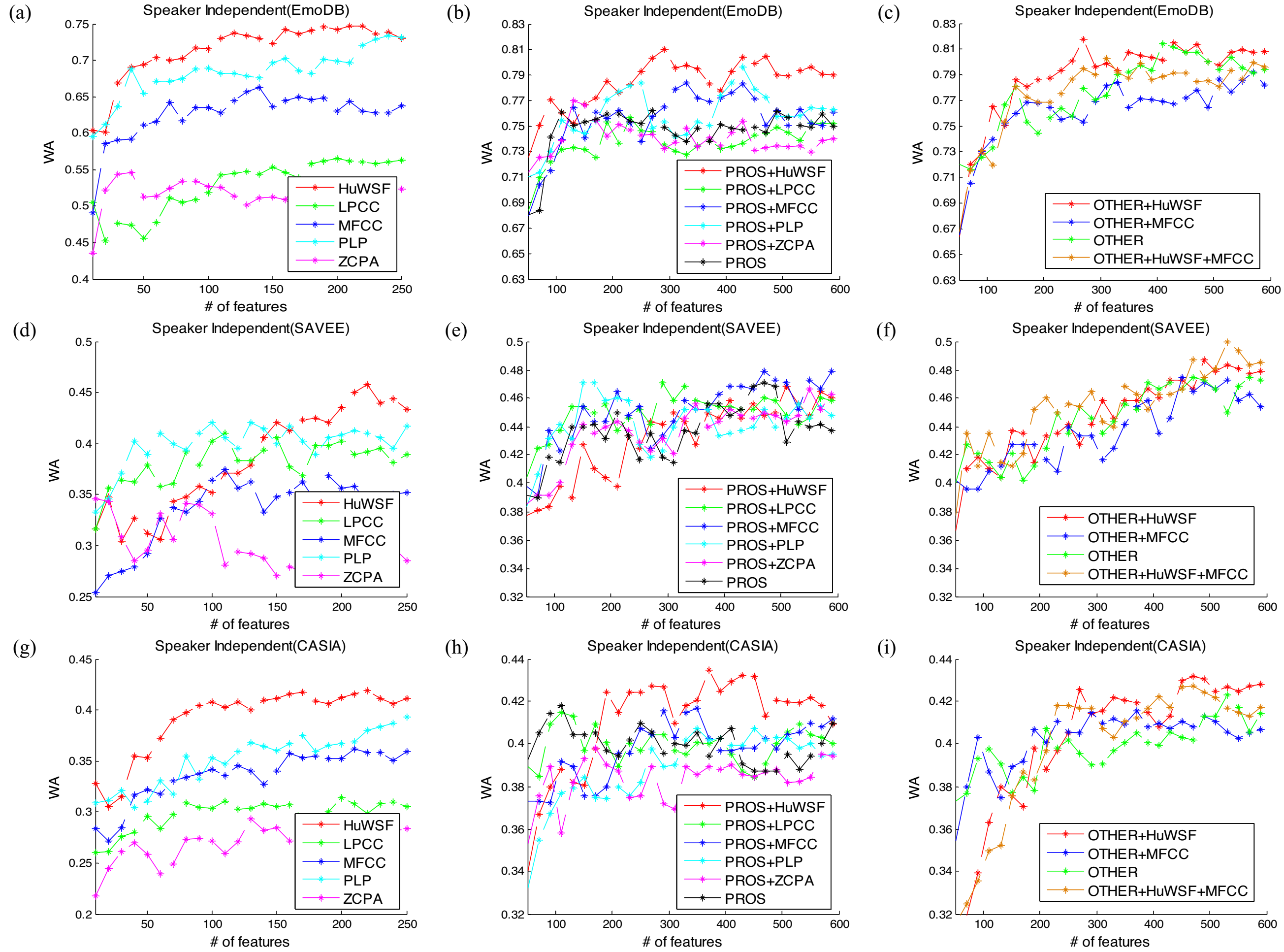


Fig. 3. The performances against the dimensions of SI experiments, where (a)–(c) are the performances on EmoDB, (d)–(f) are the performances on SAVEE, and (g)–(i) are the performances on CASIA.

5.3. The compared features

HuWSF are the spectral features, so that it would be compared with the mostly used spectral features: ZCPA [13], PLP [40], MFCC [8], and LPCC [5], where PLP represents the features combined by PLP and RASTA-PLP. Similar to HuWSF, all these spectral features contain the first and the second time derivatives into the final feature vectors. The feature set composed of all these features is marked as SPEC#.

Furthermore, to compare the supplement effectiveness of the spectral features to prosodic features, we make comparison between the following features: PROS + HuWSF, PROS + ZCPA, PROS + PLP, PROS + MFCC, PROS + LPCC, and PROS, where PROS represents prosodic features, PLP represents the features combined by PLP and RASTA-PLP. The feature set composed of all these features is marked as PROS#.

To prove HuWSF that has better supplement effectiveness to both prosodic features and other spectral features together than MFCC, we also make comparison between the following features: OTHER + HuWSF, OTHER + MFCC, OTHER + HuWSF + MFCC, and OTHER. The OTHER is combined by PROS, ZCPA, PLP, and LPCC. The feature set formed by these features is marked as OTHER#.

5.4. Selection of a suitable dimension reduction method

To select a suitable feature selection method, we take some of PI experiments on EmoDB with all spectral features, where the compared dimension reduction methods include mRMR [30], LDA [8], DISR [43], and MCFS [44]. The SFS and SFFS are not compared, as the dimensions of the features are too high while SFS and SFFS are very slow on high dimensional features. The best performances among all dimensions of these feature selection methods are given in Table 3. It can be seen that MCFS has the best WA for all features, so that MCFS is used in the following experiments.

5.5. The parameter setting

In this subsection, we describe in detail the parameter setting process for HuWSF and MFCC on EmoDB.

It can be seen from Algorithm 1 that three parameters for HuWSF need to be set: the block width w, the number of Mel filters M, and the frequency resolution of STFT b. In this experiment, w takes the value from 3, 5, 7 respectively, $M = 160$, and $b = 60$, leading to the corresponding WA as 71.85%, 74.00%, and 73.05%. It can be easily seen that WA of HuWSF is the biggest when w takes 5, so that we set $w = 5$ in the following experiments on EmoDB. As to M and b of HuWSF, we set M from 80 to 160 and set b from 40 to 80. The achieved experimental results are presented in Table 4. It can be seen from Table 4 that WA of HuWSF does not vary significantly when M takes the value around 80 and b takes the value around 60. Consequently we set $M = 80$ *and* $b = 60$ for HuWSF in the following experiments on EmoDB.

For MFCC, it can be seen from Fig. 1 that there are two parameters that need to be set: M and b. We set M from 20 to 100 and set

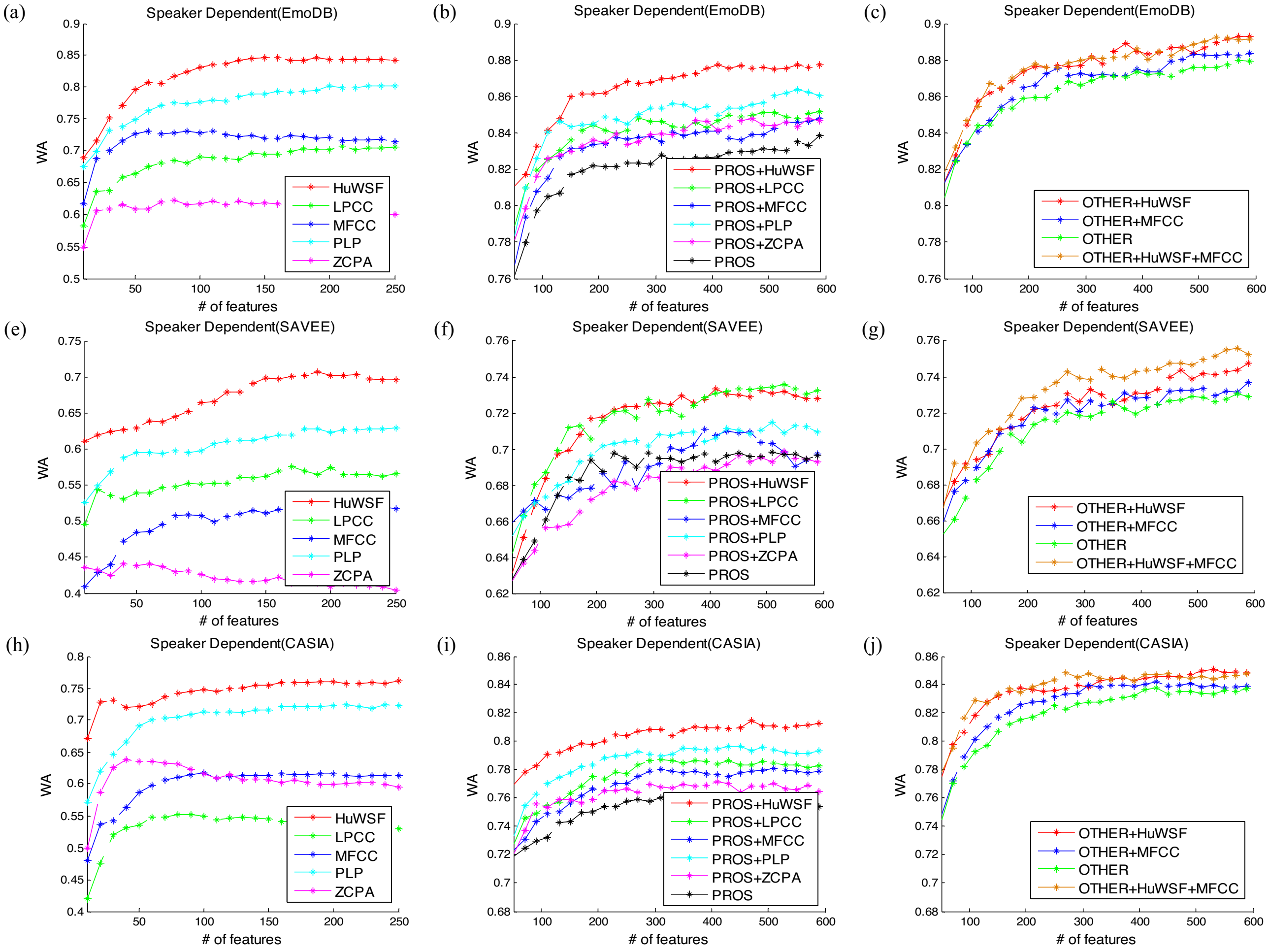


Fig. 4. The performances against the dimensions of SD experiments, where (a)–(c) are the performances on EmoDB, (d)–(f) are the performances on SAVEE, and (g)–(i) are the performances on CASIA.

b from 40 to 80. The experimental results are presented in Table 5. It can be seen from Table 5 that WA of HuWSF does not vary significantly when M takes the value around 60 and b takes the value around 50, so that we set $M = 60$ and $b = 50$ for MFCC in the following experiments on EmoDB.

We take the same method to set the parameters for the other features. The parameter that needs to be set for PLP is the order of linear predictor p. The testing values of p are 8, 10, 12, 14, and 16. The parameters need to be set for LPCC are Hamming window width w_l, window shift w_s, and linear predictor p, where w_l, w_s can be computed from b with testing values from 40, 50, 60, 70, and 80. The testing values of p are 8, 10, 12, 14, and 16. The parameter that needs to be set to ZCPA is the number of filters M, which takes the default value as 10. The final parameter settings for all features are summarized as Table 6.

5.6. The results for speaker-independent (SI) experiments

In most real application of speech emotion recognition, the labeled samples of testing speakers are hardly obtained, so that speaker independent experiments should be used to evaluate HuWSF. These speaker independent experiments would be used to answer the following questions:

(1) Is HuWSF better than other spectral features?
(2) Has HuWSF better supplement effectiveness to prosodic features than any other spectral feature set?
(3) Has HuWSF better supplement effectiveness to prosodic features together with other spectral features than MFCC?
(4) Should HuWSF be used for speech emotion recognition?

The experiments for SPEC#, PROS# and OTHER# are performed on all three speech emotion databases. The performances against the dimensions are given in Fig. 3, and the best performances among all dimensions are given in Table 7.

As to the first problem, it can be seen from (a), (d), (g) subfigures of Fig. 3 and Table 7 that HuWSF has the highest WA among all features in SPEC# on all databases, while the WA of HuWSF are 8.5%, 7.7%, and 5.8% respectively higher than that of MFCC on these three databases. It illustrates that HuWSF is better than the compared spectral features, and much better than MFCC in SI experiments.

As to the second problem, it can be seen from (b), (e), (h) subfigures of Fig. 3 and Table 7 that WA of PROS + HuWSF is highest among all feature sets in PROS# on EmoDB and CASIA, showing that HuWSF has the best supplement effectiveness to prosodic features among the compared spectral features in most SI experiments.

As to the third problem, it can be seen from (c), (f), (i) subfigures of Fig. 3 and Table 7 that WA of OTHER + HuWSF are higher than that of OTHER + MFCC on all database, illustrating that HuWSF has better supplement effectiveness to prosodic features together with other spectral features than MFCC. Furthermore, WA of OTHER + HuWSF + MFCC is much higher than that

Table 8
WA of SD experiments for SPEC#, PROS# and OTHER# on three databases, where the maximum and second maximum WA for SPEC#, PROS# and OTHER# on three database are shown in bold and italics fonts.

Features	WA(EmoDB)	WA(SAVEE)	WA(CASIA)
HuWSF	**84.72**	**70.63**	**76.14**
LPCC	70.72	57.52	55.21
MFCC	73.18	52.02	61.68
PLP	*80.23*	*62.94*	*72.44*
ZCPA	62.28	44.10	63.79
PROS + HuWSF	**87.76**	*73.33*	**81.44**
PROS + LPCC	85.18	**73.58**	78.69
PROS + MFCC	84.77	71.13	78.13
PROS + PLP	*86.36*	71.50	*79.65*
PROS + ZCPA	84.78	69.88	77.15
PROS	83.87	69.85	75.10
OTHER + HuWSF	**89.32**	*74.48*	**85.08**
OTHER + MFCC	88.37	73.67	84.23
OTHER	87.97	73.06	83.80
OTHER + MFCC + HuWSF	*89.27*	**75.60**	*84.87*

of OTHER + HuWSF and OTHER + MFCC on SAVEE, indicating that HuWSF has good supplement effectiveness to MFCC in some cases.

As to the fourth problem, the answer can be illustrated in two ways. Firstly, it can be seen from Table 7 that WA of PROS + HuWSF is only a little lower than the highest WA on EmoDB, and WA of PROS + HuWSF is higher than WA on CASIA. Furthermore, the complexity of the algorithm of computing HuWSF is not very high, so that if the system of speech emotion recognition needs a fast speed, the system can only use HuWSF together with prosodic features. Secondly, it can be seen from Table 7 that the feature sets for the best WA all contain HuWSF, so that if a better result of speech emotion recognition is needed, HuWSF should be used.

Furthermore, it is noted that WA of PROS + HuWSF, WA of PROS + LPCC, and WA of PROS + ZCPA are equal to or lower than WA of PROS on SAVEE. WA of PROS + HuWSF is also lower than WA of PROS + MFCC on SAVEE. WA of OTHER + MFCC is lower than WA of OTHER on EmoDB. Besides, WA of OTHER + HuWSF are higher than WA of OTHER + MFCC + HuWSF on EmoDB and CASIA. These phenomena may result from two aspects. One is that the used feature selection methods cannot select the best subset of features. The other is that the number of features is much higher than that of training samples. In such a case, the selected features may be good for recognizing emotion from training samples but not good for recognizing emotion of a new testing sample. This problem is typically called the curse of dimensionality.

5.7. *The results for speaker-dependent (SD) experiments*

In some real application of speech emotion recognition, some labeled samples of testing speakers could be used for training data, so that SD experiments are used to evaluate the HuWSF.

The experiments for SPEC#, PROS# and OTHER# are also performed on all three databases with aims to answer the same four questions presented in Section 5.6. The performances against the dimensions are given in Fig. 4 while the best performances among all dimensions are presented in Table 8.

As to the first problem, it can be seen from (a), (d), (g) subfigures of Fig. 4 and Table 8 that HuWSF is also much better than the second best features in SPEC+.

As to the second problem, it can be seen from (b), (e), (h) subfigures of Fig. 4 and Table 8 that WA of PROS + HuWSF is much higher than WA of other feature set in PROS+ on EmoDB and CASIA, indicating that HuWSF also has the best supplement effectiveness to prosodic features among the compared spectral features in most SD experiments.

As to the third problem, it can be seen from (c), (f), (i) subfigures of Fig. 4 and Table 8 that WA of OTHER + HuWSF is much higher than WA of OTHER + MFCC on all database, illustrating that HuWSF has better supplement effectiveness to prosodic features together with other spectral features than MFCC. Furthermore, WA of OTHER + HuWSF + MFCC is much higher than both WA of OTHER + HuWSF and WA of OTHER + MFCC on SAVEE. It means that HuWSF also could have good supplement effectiveness to MFCC on some databases.

The fourth problem can be also illustrated in two ways. Firstly, WA of PROS + HuWSF is nearly the same as WA of OTHER + MFCC on EmoDB and SAVEE, where the features in OTHER + MFCC are much more than the features PROS + HuWSF, so that we can only use prosodic features and HuWSF when a system of speech emotion recognition needs a fast speed. Secondly, it can be seen from Table 8 that the feature sets for the best WA all contain HuWSF, so that if a better result of speech emotion recognition is needed, HuWSF should be used.

6. Conclusion

This paper presents a new spectral feature type named HuWSF, which uses local Hu moments. It has the following advantages. Firstly, HuWSF considers the relationships among neighbor coefficients of Mel filters of a frame and the relationship among coefficients of Mel filters of neighbor frames. These relationships are ignored by MFCC. Secondly, HuWSF evaluates the degrees how the energy concentrated to some frequencies in a spectrogram, which vary greatly with the speech emotion types. Thirdly, HuWSF can reduce the changes brought by the differences among sentences, speakers, and speaking styles.

These advantages have been validated by the experiments on three speech emotion database with two experimental strategies. From these experiments, we can get following conclusion. (1) HuWSF is better than all other typical spectral features, because WA of HuWSF is the highest in both SI and SD experiments on all databases when each feature type is used solely. (2) HuWSF has better supplement effectiveness to prosodic features than the compared typical spectral features, as WA of HuWSF with prosodic features is higher than WA of all compared typical spectral features with prosodic features on EmoDB and CASIA. (3) HuWSF has better supplement effectiveness to prosodic features together with other spectral features than MFCC, because the WA of HuWSF are all higher than that of MFCC when they are used with prosodic features together with other spectral features. It can be easily concluded from the analysis of experimental results that HuWSF has good effectiveness for speech emotion recognition.

Besides the degrees of how the energy concentrated to some frequencies in a spectrogram are sensitive to emotions, there are other kinds of local relationships in a spectrogram also varying with emotions, which are not considered here. In the future, these local relationships will be investigated and combined with HuWSF to further improve the performance of speech emotion recognition.

Acknowledgment

This work was supported by China National Science Foundation under Grants 60973083, 61273363, State Key Laboratory of Brain and Cognitive Science under grants 08B12.

Appendix A.

Proposition 1. θ can evaluate the degree of how the energy concentrated to the center of gravity

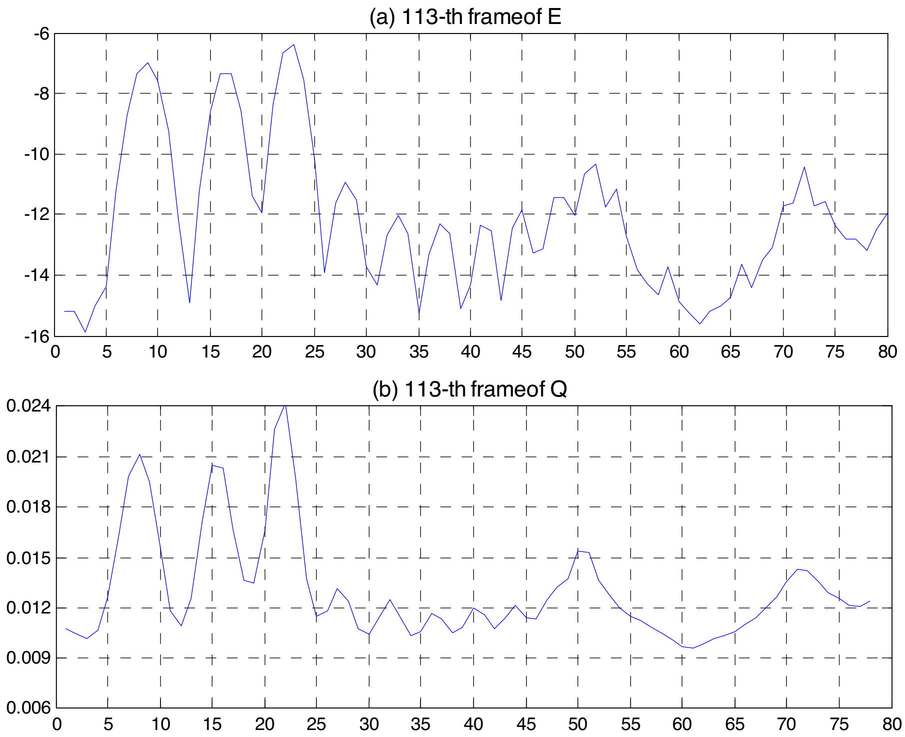


Fig. A.1. Visualization results of 113-th frame of *E*, Q, where *E* and *Q* are computed from a utterance chosen from EmoDB

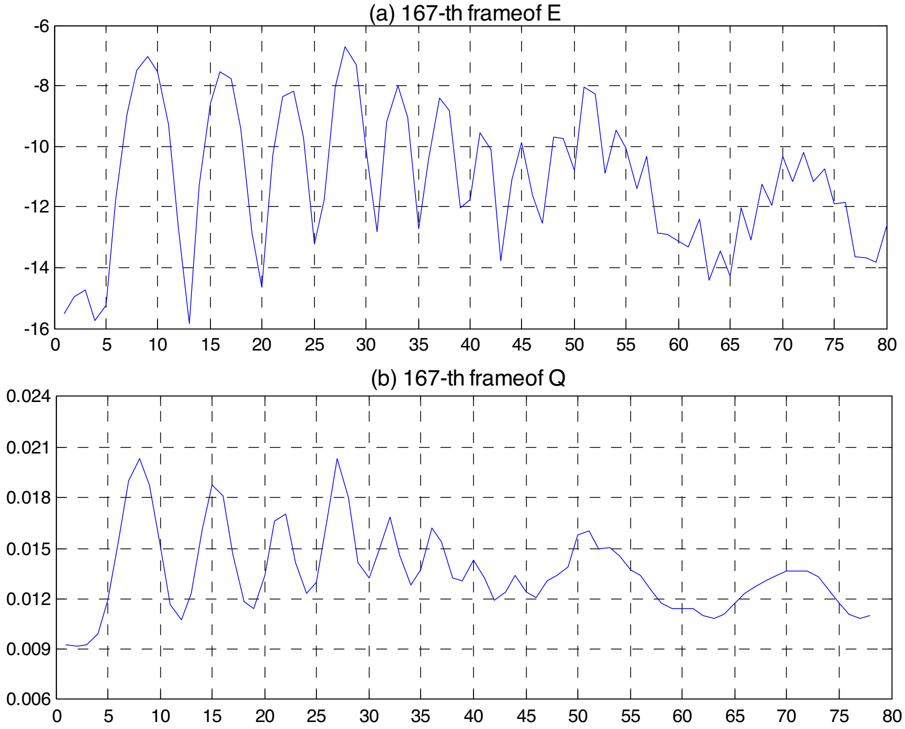


Fig. A.2. Visualization results of 167-th frame of *E*, Q, where *E* and *Q* are computed from a utterance chosen from EmoDB

Proof. To prove this proposition, we only need analyze how $\eta 20$ is changed with the distribution of energy. From Eq. (2), the normalized center moment $\eta 20$ of a block can be computed by Eq. (A.1), where $\bar{u}$ is the center of energy gravity. Obviously, under certain sum of energy in the block, the higher the degree of deviation from the center of gravity of energy, the bigger $\left|\eta 20\right|$, and thus proposition 1 has been proved. □

$$\eta_{20} = \frac{\sum_{u=1}^{w}\sum_{v=1}^{w}(u-\bar{u})^2 g(u,v)}{\left(\sum_{u=1}^{w}\sum_{v=1}^{w} g(u,v)\right)^2} \tag{A.1}$$

Proposition 2. Strong energy peaks are more clear and weak energy peaks have the opposite effect.

Proof. To prove this proposition, we first analyze how η_{20} is changed with the changes of energy. After defining $ai = (u-\bar{u})^2, xi = g^H(u,v), i = 1, 2\cdots k, k = w^2$, the normalized center moment $\eta 20$ can be rewritten as Eq. (A.2), and $\partial\eta_{20}/\partial xi$ can be computed by Eq. (A.3). It can be seen from Eq. (A.3) that $\eta 20$ changed quickly when $|x1+x2+\cdots xk|$ is small, and changed slowly when $|x1+x2+\cdots xk|$ is large. What's more, as can be seen from Fig. A.1(a), $|x1+x2+\cdots xk|$ of strong energy peaks are much smaller than that of weak energy peaks. As a result, $\eta 20$ of strong energy peaks would change more quickly than $\eta 20$ of weak energy peaks, and thus proposition 2 has been proved. □

$$\eta_{20} = \frac{a1x1 + a2x2 + \cdots akxk}{(x1 + x2 + \cdots xk)^2} \tag{A.2}$$

$$\frac{\partial\eta_{20}}{\partial xi} = \frac{ai}{2(x1 + x2 + \cdots xk)} \tag{A.3}$$

Proposition 3. In the part of voiceless of Q, the energy distributions of Q are smoother.

Proof. It can be seen from Fig. A.2(a) that the range of a energy peak in the voiceless part of Q is 2 or 3, which is smaller than the size of block, where the size of block is set to 5 in EmoDB. So that a $\eta 20$ in a energy peak would be computed from all energy of this energy peak, and $\bar{u}$ used to computed this $\eta 20$ is affected by more than one energy peak. As a result, the differences among all $\eta 20$ in this energy peak would be very small, and thus proposition 3 has been proved. □

Proposition 4. The boundaries between the part of voice activity and silence of *TQ* are more clear.

Proof. The energy would be decreased quickly from the part of voice activity to silence, and then $|x1+x2+\cdots xk|$ of blocks would be increased quickly from the part of voice activity to silence. As a result, the difference between the $\eta 20$ in the part of voice activity and silence would be much larger than the difference between the energy in the part of voice activity and silence, and then the boundaries in *TQ* would be more clearenergy peaks would. □

Proposition 5. In the part of voice activity, the coefficients among the neighbor frames of *TQ* are smoother than that of MFCC.

Proof. It can be concluded from Propositions 2 and 3 that the weaker peaks are smoother and the stronger peaks are clearer. It means that θ between neighbor frames have been smoothed, and as a result, the cepstral components of *TQ* between neighbor frames would be varied less drastically than the cepstral components of MFCC between neighbor frames, and thus Proposition 5 has been proved. □

References

[1] S. Ntalampiras, N. Fakotakis, Modeling the temporal evolution of acoustic parameters for speech emotion recognition, IEEE Trans. Affect. Comput. 3 (1) (2012) 116–125.

[2] J.-S. Park, J.-H. Kim, Y.-H. Oh, Feature vector classification based speech emotion recognition for service robots, IEEE Trans. Consum. Electron. 55 (3) (2009) 1590–1596.

[3] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: features, classification schemes, and databases, Pattern Recogn. 44 (3) (2011) 572–587.

[4] B. Yang, M. Lugger, Emotion recognition from speech signals using new harmony features, Signal Process. 90 (5) (2011) 1223–1415.

[5] B.S. Atal, Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification, J. Acoust. Soc. Am. 55 (6) (1974) 1304–1312.

[6] T.L. New, S.W. Foo, L.C. De Silva, Classification of stress in speech using linear and nonlinear features, in: ICASSP, 2003, pp. 9–12.

[7] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, J. Acoust. Soc. Am. 87 (4) (1990) 1738–1752.

[8] S. Wu, T.H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech Commun. 24 (7) (2011) 768–785.

[9] Z. Xiao, E. Dellandrea, L. Chen, Recognition of emotions in speech by a hierarchical approach, in: ACII, Amsterdam, 2009, pp. 1–8.

[10] A. Miltona, S. Tamil Selvib, Class-specific multiple classifiers scheme to recognize emotions from speech signals, Comput. Speech Lang. 28 (3) (2014) 727–742.

[11] D. Bitouk, R. Verma, A. Nenkova, Class-level spectral features for emotion recognition, Speech Commun. 52 (7) (2010) 613–625.

[12] S. Yun, C.D. Yoo, Loss-scaled large-margin gaussian mixture models for speech emotion classification, IEEE Trans. Audio Speech Lang. Process. 20 (2) (2012) 585–598.

[13] D.-S. Kim, J.-H. Jeong, J.-W. Kim, S.-Y. Lee, Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments, in: ICASSP, 1996, pp. 61–64.

[14] E. Bozkurt, E. Erzin, Ç. Eroğlu Erdem, A. Tanju Erdem, Formant position based weighted spectral features for emotion recognition, Speech Commun. 53 (9-10) (2011) 1186–1197.

[15] C.-C. Lee, E. Mower, C. Busso, S. Lee, S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, Speech Commun. 53 (9–10) (2011) 1162–1171.

[16] A. Hassan, R.I. Damper, Classification of emotional speech using 3DEC hierarchical classifier, Speech Commun. 54 (7) (2012) 903–916.

[17] J.-H. Yeh, T.-L. Pao, C.-Y. Lin, Y.-W. Tsai, Y.-T. Chen, Segment-based emotion recognition from continuous Mandarin Chinese speech, Comput. Hum. Behav. 27 (5) (2011) 1545–1552.

[18] A.I. Iliev, M.S. Scordilis, J.P. Papab, A.X. Falcão, Spoken emotion recognition through optimum-path forest classification using glottal features, Comput. Speech Lang. 24 (3) (2010) 445–460.

[19] E.M. Albornoz, D.H. Milone, H.L. Rufiner, Spoken emotion recognition using hierarchical classifiers, Comput. Speech Lang. 25 (3) (2011) 556–570.

[20] V. Bogdan, P. Dmytro, B. Ronald, W. Andreas, Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications, Comput. Speech Lang. 28 (2) (2013) 483–500.

[21] L. He, M. Lech, N.C. Maddage, N.B. Allen, Study of empirical mode decomposition and spectral analysis for stress and emotion classification in natural speech, Biomed. Signal Process. Control 6 (2) (2011) 139–146.

[22] E.H. Kim, K.H. Hyun, S.H. Kim, K. Yoon Keun, Improved emotion recognition with a novel speaker-independent feature, IEEE Trans. Mechatron. 14 (3) (2009) 317–325.

[23] C.-H. Wu, W.-B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels, IEEE Trans. Affect. Comput. 2 (1) (2011) 10–21.

[24] C.-C. Chang, C.-J. Lin, LIBSVM – a library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 1–27.

[25] M.-K. Hu, Visual pattern recognition by moment invariants, IRE Trans. Inf. Theory 8 (2) (1962) 179–187.

[26] M. Shami, W. Verhelst, An evaluation of the robustness of existing supervised machine learning approaches to the classification of emotions in speech, Speech Commun. 49 (3) (2007) 201–212.

[27] I. Luengo, E. Navas, I. Hernáez, Feature analysis and evaluation for automatic emotion identification in speech, IEEE Trans. Multimed. 12 (6) (2010) 490–501.

[28] N. Tin Lay, F. Say Wei, L.C. De Silva, Speech emotion recognition using hidden Markov models, Speech Commun. 41 (4) (2003) 603–623.

[29] F. Eyben, M. Wöllmer, B. Schuller, openSMILE-The Munich Versatile and Fast Open-Source Audio Feature Extractor, in: ACM Multimedia (MM), Florence, 2010, pp. 1459–1462.

[30] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, IEEE Trans. Pattern Anal. Mach. Intell. 27 (8) (2005) 1226–1238.

[31] G. Brown, A. Pocock, M.-J. Zhao, M. Lujan, Conditional likelihood maximisation: a unifying framework for information theoretic feature selection, J. Mach. Learn. Res. 13 (2012) 27–66.

[32] G. Shashidhar, K. Koolagudi, S. Rao, Emotion recognition from speech: a review, Int. J. Speech Technol. 15 (2) (2012) 99–117.

[33] VOICEBOX. Speech Processing Toolbox for MATLAB, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
[34] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, W. Benjamin, A database of German emotional speech, in: INTERSPEECH, Lisbon, 2005, pp. 1517–1520.
[35] L.-C. Ramón, S. Jan, K. Martin, Enhancement of emotion detection in spoken dialogue systems by combining several information sources, Speech Commun. 27 (9–10) (2011) 1210–1228.
[36] E. Humberto Pérez, A. Carlos, G. Reyes, P. Luis Villaseñor, Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model, Biomed. Signal Process. Control 7 (1) (2012) 79–87.
[37] K. Marcel, B. Lukáš, Č. Jan Honza, Application of speaker and language identification state-of-the-art techniques for emotion recognition, Speech Commun. 53 (9–10) (2011) 1172–1185.
[38] L. Chen, X. Mao, Y. Xue, C. Lee Lung, Speech emotion recognition: features and classification models, Digital Signal Process. 22 (6) (2012) 1154–1160.
[39] S. Bjorn, V. Bogdan, E. Florian, W. Martin, S. Andre, W. Andreas, R. Gerhard, Cross-corpus acoustic emotion recognition: variances and strategies, IEEE Trans. Affect. Comput. 1 (2) (2012) 119–131.
[40] S.G. Koolagudi, K.S. Rao, Emotion recognition from speech using source, system, and prosodic features, Int. J. Speech Technol. 15 (2) (2012) 265–289.
[41] H. Sanaul, P.J.B. Jackson, Speaker-dependent audio-visual emotion recognition, in: AVSP, 2009, pp. 53–58.
[42] Y. Mitani, Y. Hamamoto, A local mean-based nonparametric classifier, Pattern Recogn. Lett. 27 (10) (2006) 1151–1159.
[43] P. Meyer, G. Bontempi, On the use of variable complementarity for feature selection in cancer classification, Appl. Evol. Comput. 3907 (2006) 91–102.
[44] C. Dend, Z. ChiYuan, H. XiaoFei, Unsupervised feature selection for multi-cluster data, in: International conference on Knowledge Discovery and Data Mining, 2010, pp. 333–342.
[45] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2010 paralinguistic challenge, in: INTERSPEECH, 2010, pp. 2794–2797.
[46] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2009 Emotion Challenge feature set, in: INTERSPEECH, 2009, pp. 983–986.
[47] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2011 Speaker State Challenge feature set, in: INTERSPEECH, 2011.
[48] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2012 Speaker Trait Challenge feature set, in: INTERSPEECH, 2012.
[49] B. Schuller, S. Steidl, A. Batliner, The INTERSPEECH 2013 Computational Paralinguistics Challenge feature set, in: INTERSPEECH, 2013, pp. 148–152.
[50] C. Pereira, Dimensions of emotional meaning in speech, in: Workshop on Speech and Emotion: A Conceptual Framework for Research, Belfast, 2000, pp. 25–28.
[51] R. Tato, R. Santos, R. Kompe, J.M. Pardo, Emotion space improves emotion recognition, in: Int. Conf. Spoken Lang. Process., Denver, 2002, pp. 2029–2032.
[52] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, J.-J. Lu, Detecting emotions in mandarin speech, in: ROCLING, 2004, pp. 365–373.
[53] The selected Speech Emotion Database of Institute of Automation Chinese Academy of Sciences (CASIA), http://www.datatang.com/data/39277
[54] A. Yazid, D. Pierre, Anchor models for emotion recognition from speech, IEEE Trans. Affect. Comput. 4 (3) (2013) 280–290.
[55] A. Hassan, R. Damper, M. Niranjan, On acoustic emotion recognition: compensating for covariate shift, IEEE Trans. Audio Speech Lang. Process. 21 (7) (2013) 1458–1468.
[56] A. Batliner, B. Schuller, et al., The automatic recognition of emotions in speech, in: R. Cowie, C. Pelachaud, P. Petta (Eds.), Emotion-Oriented Systems, Springer, Berlin, Heidelberg, 2010, pp. 71–94.
[57] B. Felix, You seem aggressive! monitoring anger in a practical application, in: LREC, 2012, pp. 1221–1225.
[58] B. Schuller, B. Felix, Learning with synthesized speech for automatic emotion recognition, in: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), Dallas, 2010, pp. 5150–5153.
[59] Ž. Dragiša, Ž. Joviša, Shape ellipticity from Hu moment invariants, Appl. Math. Comput. 226 (1) (2014) 406–414.
[60] L. Zhang, F. Xiang, et al., Application of improved HU moments in object recognition, in: IEEE International Conference on Automation and Logistics (ICAL), Zhengzhou, 2014, pp. 224–228.
[61] Z. Huang, J. Leng, Analysis of Hu's moment invariants on image scaling and rotation, in: International Conference on Computer Engineering and Technology (ICCET), Chengdou, 2010, V7-476–V7-480.

Yaxin Sun is a Ph.D. candidate at the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He received his Masco. degree in Computer Application Technology from Guangzhou University of Technology in 2012. His current research interests include speech emotion recognition, pattern recognition and machine learning.

Guihua Wen, Ph.D., is a professor and doctor supervisor. In 2005–2006, he did visiting research on machine learning and semantic web in School of Electronics and Computer, University of Southampton,UK. His main research interests are computational creativity, data mining and knowledge discovery, machine learning, and cognitive geometry. Since 2006, he proposed some original methods based on the computation of cognitive laws, which can effectively solve difficult problems in information science. The research results have been published in the international journals, including Pattern Recognition Neurocomputing, Journal of Software, Journal of computer Research and Development. He also published some papers in the international conferences such as IJCAI. Since 2006, he directed the projects from the China National Natural Science Foundation, State Key Laboratory of Brain and Cognitive Science, the Ministry of Education Scientific Research Foundation for returned overseas students, Guangdong Provincial Science and Technology research project, the Fundamental Research Funds for the Central Universities, SCUT. He also directed many projects from enterprises, with applications of his research results to the practical problems. He has ever been a Council Member of Chinese Association for Artificial Intelligence and a program committee member of many international conferences. He is also a reviewer for China National Natural Science Foundation.

Jiabing Wang received the PhD degree in computer science from Huazhong University of Science and Technology (HUST), China, in 2003. He joined the School of Computer Science and Engineering at South China University of Technology (SCUT) in 2004, and now he is an associate professor. His research interests are machine learning, data mining, and pattern analysis. In these areas, he has published more than 20 technical papers in refereed journals or international conference proceedings.