Speech Emotion Feature Analysis Based on Emotion Fingerprints

Yuantao Jiang[1], Kaifa Deng[2*] and Chunxue Wu[1]

[1] School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2] Shanghai University of Engineering Science, Shanghai, China
Corresponding author: Kaifa Deng, Shanghai University of Engineering Science, Shanghai, China

Tel.:+86-176-2160-7346,E-mail:j17621607346@163.com, [*]dengkaifa@126.com, wcx@usst.edu.cn

Abstract: The speech recognition has caught the attention of more and more researchers as the important branch of artificial intelligence. However, speech recognition just can recognize the neutral meaning and ignore the important emotion information. So,this paper proposes an emotion fingerprint extraction algorithm. The speech signal is decomposed by lifting wavelet packets. And combining singular value entropy and sample entropy as the feature vector to characterize the speech time-frequency matrix coefficient. Then adopting the statistical values to extract emotion fingerprints. The experiments show that this algorithm can reflect the emotion information of the speech fully and distinguish several major emotions well.
Keyword: speech recognition, emotion recognition, emotion fingerpringting, lifting wavelet packets, singular value entropy , sample entropy

1.Introduction

Artificial intelligence is one of the hottest technologies recently. Speech recognition has also developed rapidly as an important branch of it. At present, speech recognition is the neutral semantic recognition, which extracts the acoustic feature parameters of human language and recognizes the meaning of the speech according to the recognition models. However, the emotion information which is one of the most important features of language is ignored. One of the founders of artificial intelligence, Prof. Minsky of the Massachusetts Institute of Technology in the United States, pointed out that the question is not whether smart machines can have emotions, but whether intelligent machines can realize intelligence in "*The Society of Mind*"[1].And Damasio's research shows that human intelligence not only shows normal rational thinking and logical reasoning ability, but also shows normal emotional ability[2]. The emotion is an important symbol of human intelligence. So, adding emotion features to speech recognition is an important research topic[3]. Emotion recognition is a classification problem of pattern recognition, which includes three mainaspects: speech emotion feature extraction, selection and speech emotion recognition[4]. The principle of speech emotion recognition is as figure 1.

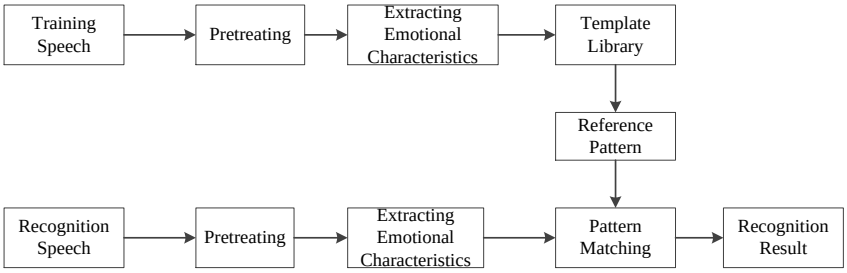


Figure 1. The principle of emotion recognition.

The audio fingerprint technology as the important audio retrieval technology has become more and more important. It is a compact digital signature which is extracted from audio data to characterize the acoustic characteristics of the audio fully. Great audio fingerprint should meet the following four conditions: robustness, compactness, distinguishability, simplicity[5].

2. Speech emotion feature analysis

The parameters of speech can describe emotion features and express human's emotions.And the parameters include acoustic characteristics, time-frequency characteristics and statistical characteristics of speech signals. And the time domain waveforms of different emotions are shown as figure 2.

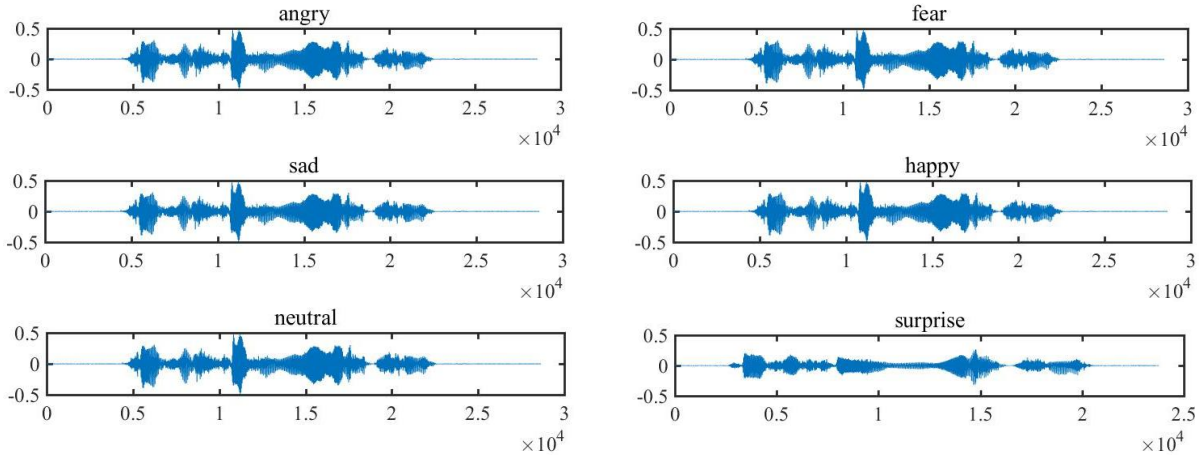


Figure 2. The time domain wave forms of different emotions.

Many literatures have done a lot of research on emotion feature parameters and they mainly focus on amplitude, time, fundamental frequency, formant,energy, etc[6]. At present, speech features can be simply divided into three categories: rhythm features, spectrum signature, and other features.

1). Prosodic features

The prosodic features is the physical property of the speech. Because the prosodic features express more speech emotion information in sentences. Prosodic features include (1).The pitch frequency characteristics;(2).Formant characteristics;(3) Energy characteristics;(4).Time characteristics[7,8].

2). Spectrum features

Compared to the prosodic features, speech does not change significantly in a short time. The classical spectrum features include short-term Fourier transform, linear predictor coefficients (LPC), Mel frequency cepstrum coefficients (MFCC), and perceptual linear predictive cepstral coefficients (perceptual linear predictive cepstral coefficients PLP), line spectrum pair (LSP),short time coherence (SMC).

3) Other features

In addition, there are other speech feature extraction methods in speech emotion recognition, such as teager energy operator (TEO), empirical mode decomposition (EMD),Fractal dimensions, deep

learning, etc.

As for the differences in the language and research environment, the research methods and emphasis are also different. And the majority classification of emotions include angry, happy, sad, neutral, surprised, fear.

3. Emotion fingerprint extraction algorithm

The traditional fingerprint extraction algorithm has limitations which is base on time-frequency analysis. The two-dimensional time-frequency information matrix dimension from the time-frequency transformation of the original signal is too large. It is necessary to process the time-frequency matrix information multiple times to make it effective. Therefore, considering the nonlinear parameter methods to extract the signal features. In this paper, the singular value entropy and sample entropy are introduced in the emotion fingerprint extraction. The feature vector is based on the lifting wavelet packet transform and the combination of singular value entropy and sample entropy is adopted to characterize the time-frequency matrix coefficients of emotions.

3.1 Lifting Wavelet Packet Transform

The lifting wavelet construction method proposed by Sweldens[9] and it's different from the traditional Mallat wavelet construction method. It doesn't depend on the translation and stretching of the mother wavelet, nor does it require spectrum analysis, so it is particularly suitable for structure of wavelets in Finite Area, Surface and Non-uniform Sampling. And the lifting wavelet packet transform not only overcomes the shortcomings of traditionals, such as the heavy computation, the inability to accurately reconstruct the original signal, but also has the advantages of traditionals, such as fast operation speed,no extra space, and the integer wavelet transform, and easy parallel computing. It is suitable for speech recognition which requires higher real-time signal processing.

Constructing biorthogonal wavelet packets with a lifting scheme is simpler and more efficient. The high-frequency components of signal can be obtained by interpolation subdivision, and constructinga scale function to obtain the low-frequency components. The lifiting wavelet packet is divided into: splitting, predicting, updating[9]. And the inverse transform of the lifting algorithm can be implemented by changing the plus or minus sign in the forward calculation formula. The entire lifting wavelet transform process is shown as figure 3.

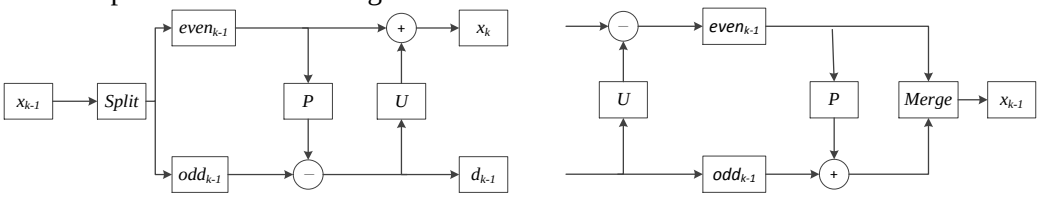


Figure 3. Lifting wavelet packet transform process.

3.2 Singular Value Decomposition and singular value entropy Calculation

Singular values are inherent in the matrix and reflect the information contained in the matrix fully. The singular value entropy of wavelet packet is based on the singular value decomposition theory. The signal after wavelet packet transform is decomposed into a series of singular values and then utilizing statistical properties of information entropy to analyze the uncertainty of a set of singular values and then give a deterministic measure of the complexity of the speech signal. Therefore, using the singular value of the wavelet packet space feature matrix A as the characteristics of the speech signal to construct the feature vector[10].

Assuming A is a $m\times n$ ($m>n$) matrix with a rank r ($r\leq n$), there are $m\times m$ orthogonal matrices U and $n\times n$ orthogonal matrices V to make $U \wedge V^T = A(or\ U \wedge V^H = A)$. And $\wedge$ *is* $m\times n$ non-negative diagonal matrix.

$$A = \begin{bmatrix} R & 0 \\ 0 & 0 \end{bmatrix} \tag{1}$$

In (1), $R = diagram(\sigma_1, \sigma_2, \dots, \sigma_r)$ and the diagonal elements are the singular values of matrix *A*. And if these non-zero singular values are combined into a feature vector *X*($X=(\sigma_1, \sigma_2, \dots, \sigma_r)$. From the properties of the matrix singular value, this feature vector can characterizes the characteristics of the wavelet packet coefficient matrix. The coefficient matrix can reflect the nature and characteristics of the signal. So the feature vector can be used to represent the emotion characteristics.

So, the singular value contains the characteristics of the emotion, and the difference in the singular value reflects the different characteristics of the different frequency bands. Each wavelet coefficient has the different frequency component, and the singular value is also different, normalizateing each component then get:

$$T_i = \frac{\sigma_i^2}{E} \tag{2}$$

In (2), $E = E_1 + E_2 + \cdots + E_n$. Then the singular value entropy of each wavelet coefficient can be obtained.

$$H = -\sum_{i=1}^{n} p_i \, log \, p_i \tag{3}$$

In (3), $p_i = \frac{T_i}{T}, T = \sum_{i=1}^{n} T_i$.

3.3 Sample entropy of speech signal

Add 0.37 seconds sliding time window to calculate the sample entropy of the speech signal. Then calculating the sample entropy of the next 0.37 seconds time window until get the sample entropy of the speech signal of the last time window, thereby the time sequence of the sample entropy is obtained,that is, $X = \{x_{(1)}, x_{(2)}, \dots, x_{(N)}\}$. Sample entropy calculation steps as follows[11][12].

(1). Given a mode dimension m, the original sequence constitutes the m-dimensional vector.

$$X_{(i)} = \{x_{(i)}, x_{(i+1)}, \dots, x_{(i+m-1)}\} \tag{4}$$

In (4), *i=0,1,2,…,N-m+1*.

(2). Defining the distance between $X_{(i)}$ and $X_{(j)}$.

$$d(i,j) = max_{k=1-m-1} |x(i+k) - x(j+k)| \tag{5}$$

In (5), *k=0,1,2,…,m-1*.

(3). Given a threshold *r*, counting the number of *d(i,j)<r* for each *i* and the ratio of the number to the distances total number *N-m+1* and denoted as $B_i^m(r)$.

$$B_i^m(r) = \frac{L[d_{(i,j)<r}]}{N-m+1} \tag{6}$$

In (6), $L[d_{(i,j)}<r]$ is the number of $d_{(i,j)}<r$. $1 \le j \le N-m$ and $j \neq i$.Then calculating the average of all of *i*.

$$B^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} B_i^m(r) \tag{7}$$

(4). Repeat steps (1)~(3) for *m+1* to obtain $B^{m+1}(r)$.

(5). Theoretically, the sample entropy of this sequence is

$$SampEn(m,r) = lim_{n\to\infty}[-ln\frac{B^{m+1}(r)}{B^m(r)}] \tag{8}$$

When *N* takes a finite value, the estimated value of the sample entropy can be obtained.

$$SampEn(m,r,N) = -ln\frac{B^{m+1}(r)}{B^m(r)} \tag{9}$$

The parameters *m*, *r*, and *N* in the formula are referenced in [9]. And in this paper, *m=2*, *r=0.2SD*, *N=1024* and *SD* is the standard deviation of the original data.

From the above theoretical analysis, it can be found that the singular value entropy of each wavelet packet coefficient contains more comprehensive characteristics, and the difference of singular values reflects different characteristics of different frequency bands. The stable estimates can be obtainde from the sample entropy analysis and it can represent signal complexity and irregularity and has good noise immunity and anti-jamming capability. So, the emotion fingerprintings which is extracted by the combination of singular value entropy and sample entropy can fully reflect the emotion characteristics of speech.

4. Experiment and result analysis

4.1 Emotion fingerprinting extraction process

The emotion fingerprinting extraction process is shown as the figure 4:

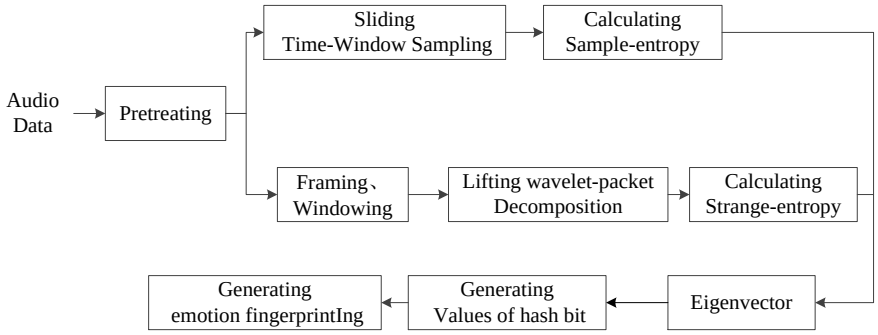


Figure 4. Emotion fingerprinting extraction process.

(1). Converting speech to 16 bit/sample and sample rate is 44. 1 kHz mono signal, then adopting endpoint detection to distinguish speech signal and non-voice signal[13]. The endpoint detection results shown in figure 5.

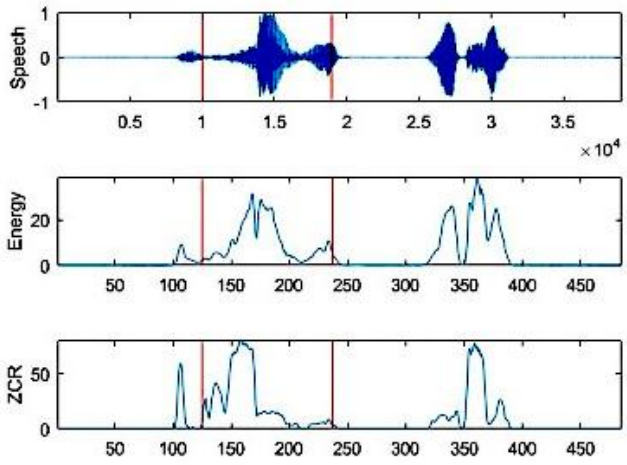


Figure 5. Endpoint detection processing

(2). Calculating sample entropy $SampE_{n0}$, $SampE_{n1}$, …,$SampE_{nn}$ according to formula (8) for all sample points.

(3). In the experiment, the frame length is 2048 and adopting the Hanning window to smooth the frame edge. The overlap factor $P = 28/32$.The Hanning window formula is as follows:

$$W_{(n)} = \begin{cases} 0.5 \times [1 - cos(2\pi n/(N-1))], 0 \le n \le N-1 \\ 0 \qquad\qquad , \quad else \end{cases} \tag{10}$$

(4). Adopting the lifting wavelet packet to perform 5-layers decomposition on each frame,then extracting the decomposition coefficients of the fifth layer from low-frequency to high-frequency bands. Then reconstructing the obtained wavelet packet coefficients for the feature matrix A of the lifting wavelet packet.

(5). Performing singular value decomposition on the feature matrix A, and the feature vector $\sigma = (\sigma_0, \sigma_1, \dots, \sigma_n)$ formed by the singular value can be obtained. Then calculating singular value entropy according to formula (3).

(6). Combining the singular value entropy and sample entropy as the feature vector

$[SampE_n(\sigma_1), SampE_n(\sigma_2), \ldots, SampE_n(\sigma_n)]$ of the emotion fingerprint.

(7) Extracting the hash value to generate the emotion fingerprint. The statistics is an effective method for extracting relevant feature values. This algorithm calculates the following entropy statistics.

$$H_i = \sum_{n=1}^{j} |SampE_n(\sigma_n)|^2 \tag{11}$$

$$H_{sumk} = \sum_{i=1}^{m} H_i \tag{12}$$

$$H_{avg} = \frac{1}{t}\sum_{k=1}^{t} H_{sumk} \tag{13}$$

In the formula (11),(12),(13), $SampE_n(\sigma_n)$ is the value of singular value entropy and sample entropy of the *n*th wavelet packet coefficient of the *k*th frame's *i*th subspace. H_i represents the entropy value of the *k*th frame's *i*th subspace. H_{sumk} represents the entropy value of the *k*th frame , H_{avg} represents the average entropy value of all frames.

(5). Extracting audio fingerprints by hash bit values. Comparing the energy value E_{sumk} of each frame with the average of energy value E_{avg} of all frames, one bit of hash value of each frame is generated according to the formula (12).The splicing of all generated hash bit values constitutes the emotion fingerprint.

$$H_{(k)} = \begin{cases} 1, H_{sumk} > H_{avg} \\ 0, else \end{cases} \tag{14}$$

Then the emotion fingerprint of the audio signal which length is *N* can be expressed as:

$$S_{(t)} = \sum_{t=1}^{N}(k_1 h_1(t) + k_2 h_2(t) + \cdots + k_n h_n(t)) \tag{15}$$

In (15), h_i is the hash value of the emotion fingerprint extracted for each frame signal, and K_i (i=1, 2, 3, 4, 5) is the synthesis coefficient.

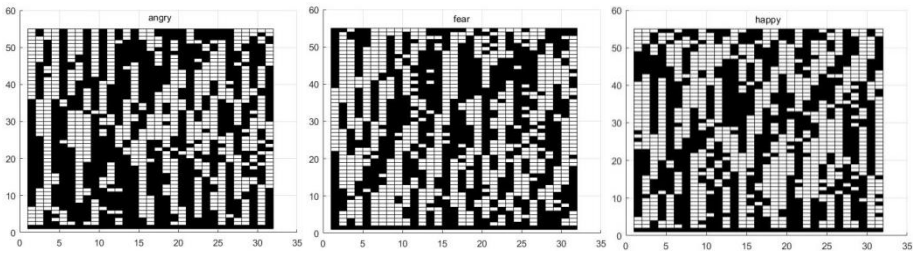
4.2. Experiment Analysis

This paper take the casia's Chinese emotion database for experiments[14]. There are four speechers include two men and two women. And the number of speeches are 50 with angry, fear, happy, neutral, sad, and surprise respectively and the time of each speech is 1 second.

The emotion fingerprint is extracted with the algorithm of this paper for these speeches. The fingerprint distance is an important indicator to distinguish different emotions. If the speech with different emotion has similar emotion fingerprint, it will cause higher bit error rate(*BER*) in emotion recognition. And the emotion fingerprint of angry, happy, sad, fear, neutral, surprise are shown as figure 6. And taking the *BER* to represent the fingerprint distance and the result is shown as table 1. Assuming that the length of the two fingerprints is *B* bits, the number of mismatched bits is *n*, and the *BER* presses as the formula (16).

$$BER = \frac{100}{B}\sum_{n=0}^{B-1} \begin{cases} 1, h'(n) \neq h(n) \\ 0, h'(n) = h(n) \end{cases} \tag{16}$$

In (16), $h'(n)$ and $h(n)$ represent different fingerprints respectively.



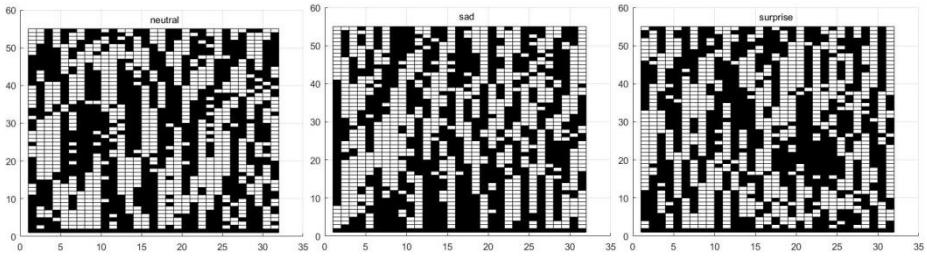


Figure 6. The emotion fingerprinting of the different emotions

Table 1. The distance of the different emotions

Emotion	Angry	Fear	Happy	Neutral	Sad	Surprise
Angry	0	0.5431	0.6553	0.6089	0.5645	0.6643
Fear	0.5431	0	0.6479	0.6728	0.6469	0.7213
Happy	0.6553	0.6479	0	0.7070	0.6145	0.6329
Neutral	0.6089	0.6728	0.7070	0	0.6935	0.7142
Sad	0.5645	0.6469	0.6145	0.6935	0	0.6534
Surprise	0.6643	0.7213	0.6329	0.7142	0.6534	0

Then the emotion fingerprint of different people with the same emotion is shown as figure 7, the *BER* is also adopted to represent the fingerprint distance and the result is shown in table 2.

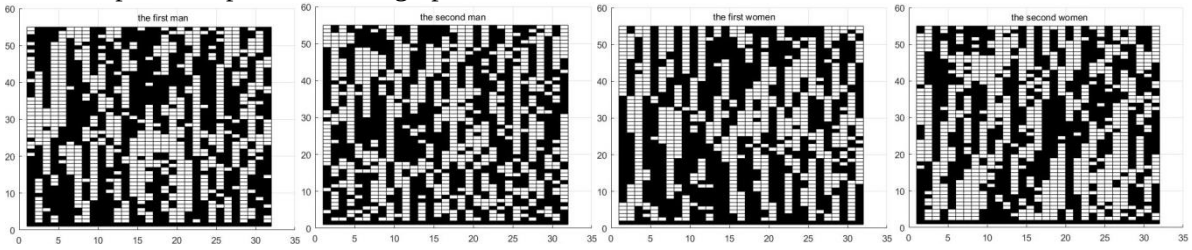


Figure 7. The emotion fingerprinting of different person.

Table 2. The *BRE* of emotion fingerprinting by different person.

Persons	Woman1	Woman2	Man1	Man2
Woman1	0	0.2416	0.3024	0.2977
Woman2	0.2416	0	0.2843	0.2861
Man1	0.3024	0.2843	0	0.3146
Man2	0.2977	0.2861	0.3146	0

The results show that the bigger the *BER*, the greater the difference between different emotions, the better it can distinguish between different emotion. From figure 6 and table 2, it is evident that the *BER* of the fingerprint comparison between the same speecher's different emotion is between 0.5431 and 0.7213, beyond the threshold 0.35 proposed in [15], which shows that this algorithm has a strong distinguishability in different emotions.

And from figure 7 and table 2, the *BER* of fingerprints comparison between the same emotion of different speechers is between 0.2416 and 0.3146, and the distance between the fingerprintings is small, so judging these are the same emotion basically. And the *BER* is that the emotion fingerprint includes the semantics characteristic parameter and the acoustic characteristics of men and women. Therefore,

the emotion fingerprint also provides a good theoretical basis for speech recognition which combine the semantics characteristic and emotion characteristic.

5. Conclusion

In this paper, the singular value entropy and sample entropy are introduced into the emotion fingerprint, adopting a feature vector combine the singular value entropy and sample entropy to characterize the time-frequency matrix coefficients of the speech emotion based on the lifting wavelet packet transform. extracting the emotion fingerprints by statistical value calculations. Experiment analysis shows that there are significant differences in emotion fingerprints of different emotion, so emotion fingerprints can characterize the emotion information in the speech well and distinguish different emotion clearly. Emotion fingerprints with emotion information are smaller than original signals, occupy less memory, and are easier to process. It provides a good theoretical basis for speech recognition combined with semantic and emotion recognition. So considering the speech recognition algorithm recognise emotion and semantic information synchronously according to the audio fingerprint algorithm is the future work.

Acknowledgments

The authors would like to appreciate all anonymous reviewers for their insightful comments and constructive suggestions to polish this paper in high quality. This research was supported by the ShanghaiScience and Technology Innovation Action Plan Project (16111107502, 17511107203),the National Natural Science Foundation of China (61502220),the Zhejiang Provincial Natural Science Foundation of China(No.LY14F020044),the program for tackling key problems in Henan science and technology (No.172102310636) and the Nanjing Leading Science and Technology Entrepreneurial Talents Introduction Program Funded Project (2014A090002).

Reference

[1] Minsky M. The Society of Mind[J]. Personalist Forum, 1987, 3(1):19-32.
[2] Zimmerman, Corinne. Descartes' Error: Emotion, Reason and the Human Brain[J]. Psychosomatics, 1996, 310(36):151–153.
[3] Ververidis, D., Kotropoulos, C., & Pitas, I. (2004). Automatic emotional speech classification. *Proc.ieee Int.conf.acoustics Speech & Signal Processing Montreal Canada, 1*(1), I-593-6 vol.1.
[4] Nwe, T. L., Foo, S. W., & Silva, L. C. D. (2003). Speech emotion recognition using hidden markov models. *Speech Communication, 41*(4), 603-623.
[5] Cano, P., Batle, E., Kalker, T., & Haitsma, J. (2002). *A review of algorithms for audio fingerprinting*.
[6] Mencattini, A., Martinelli, E., Costantini, G., Todisco, M., Basile, B., & Bozzali, M., et al. (2014). Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure. *Knowledge-Based Systems, 63*(3), 68-81.
[7] Zhao, X., Zhang, S., & Lei, B. (2014). Robust emotion recognition in noisy speech via sparse representation. Neural Computing & Applications, 24(7-8), 1539-1553.
[8] Ooi, C. S., Seng, K. P., Ang, L. M., & Li, W. C. (2014). A new approach of audio emotion recognition. Expert Systems with Applications, 41(13), 5858-5869.
[9] Sweldens, W. (1998). The lifting scheme: a construction of second generation wavelets. *Siam J.math.anal, 29*(2), 511-546.
[10] Xie, H. B., Zheng, Y. P., & Guo, J. Y. (2009). Classification of the mechanomyogram signal using a wavelet packet transform and singular value decomposition for multifunction prosthesis control. *Physiological Measuremsent, 30*(5), 441-457.
[11] Zurek, S., Guzik, P., Pawlak, S., Kosmider, M., & Piskorski, J. (2012). On the relation between correlation dimension, approximate entropy and sample entropy parameters, and a fast algorithm for their calculation. *Physica A Statistical Mechanics & Its Applications, 391*(24), 6601-6610.
[12] Liang, J., & Yang, Z. (2015). A Novel Wavelet Transform – Empirical Mode Decomposition

Based Sample Entropy and SVD Approach for Acoustic Signal Fault Diagnosis. *International Conference in Swarm Intelligence* (Vol.9142, pp.232-241). Springer International Publishing.

[13] Li, Q., Zheng, J., Tsai, A., & Zhou, Q. (2002). Robust endpoint detection and energy normalization for real-time speech and speaker recognition. *Speech & Audio Processing IEEE Transactions on, 10*(3), 146-157.

[14] MINGYU YOU, GUO-ZHENG LI, JACK Y. YANG, & MARY QU YANG. (2010). An enhanced lipschitz embedding classifier for multi-emotion speech analysis. Internationoal Journal of Pattern Recgnition & Artificial Intelligence, 23(08), 1685-1700.

[15] Cotton, C. V., & Ellis, D. P. W. (2010). Audio fingerprinting to identify multiple videos of an event. *IEEE International Conference on Acoustics Speech and Signal Processing* (pp.2386-2389). IEEE.