LETTER

A Salient Feature Extraction Algorithm for Speech Emotion Recognition

Ruiyu LIANG[†a)], ***Member***, **Huawei TAO**[††], **Guichen TANG**[†], **Qingyun WANG**[†], ***and*** **Li ZHAO**[††], ***Nonmembers***

SUMMARY A salient feature extraction algorithm is proposed to improve the recognition rate of the speech emotion. Firstly, the spectrogram of the emotional speech is calculated. Secondly, imitating the selective attention mechanism, the color, direction and brightness map of the spectrogram is computed. Each map is normalized and down-sampled to form the low resolution feature matrix. Then, each feature matrix is converted to the row vector and the principal component analysis (PCA) is used to reduce features redundancy to make the subsequent classification algorithm more practical. Finally, the speech emotion is classified with the support vector machine. Compared with the tradition features, the improved recognition rate reaches 15%.

key words: *speech emotion recognition, spectrogram, selective attention, support vector machine*

1. Introduction

With the development of human-computer interaction technology, speech emotion recognition (SER) has been one of the key technologies [1]. The selection and construction of speech emotion features has great influence on the recognition performance [2]. The common features for SER can be summarized in three types [3]: prosodic feature, spectral feature and quality feature. But these three features are either the time domain features or the frequency domain features. There has been very little research on the time-frequency features for SER. As a visual expression of time-frequency distribution of the speech energy [4], the spectrogram includes some speech features, such as energy, formant, fundamental frequency, tone and so on [5]. The studies on the spectrogram include sound classification, sound recognition [6], and sound enhancement [7], but there have been few studies for SER [8], [9].

Based on the analysis of the spectrogram, a SER algorithm based on the selective attention mechanism is proposed. Here, two improved strategies are introduced: 1) the selective attention mechanism is used to extract the effective time-frequency components to calculate emotional features; 2) imitating the selective attention mechanism, the color, direction and brightness features of the spectrogram are calculated and normalized to form the feature matrix. Then, the principal component analysis (PCA) is adopted to reduce the feature dimensionality to form emotional feature vector. Finally, the support vector machine (SVM) is used to classify the speech emotion. From the experimental results, compared with the SER algorithms based on the traditional features, the recognition rate of the proposed algorithm improves about 15%.

Manuscript received April 18, 2015.
Manuscript publicized May 29, 2015.
[†]The authors are with School of Communication Engineering, Nanjing Institute of Technology, Jiangsu Nanjing, 211167, China..
[††]The authors are with School of Information Science and Engineering, Southeast University, Jiangsu Nanjing, 210096, China.
a) E-mail: lly1711@163.com
DOI: 10.1587/transinf.2015EDL8091

2. Spectrogram Feature Extraction

The SER algorithm based on the selective attention mechanism can be divided into three parts: the extraction of the feature based on the selective attention mechanism, the establishment of classification model and the emotion recognition. Here, the emotional feature extraction is the key part. Compared with the previous studies, the proposed algorithm has two differences: 1) the spectrogram is used as the feature carrier of emotion recognition; 2) the salient region of the spectrogram is extracted to calculate the emotional feature.

The spectrograms of different emotions have some differences. For example, if a person is angry or happy, his mood is strong and the deep color area of the spectrogram is larger; if a person is surprised, the tone is changed frequently and the fluctuations of horizontal stripes in the spectrogram are more. To efficiently extract the feature of the spectrogram, the selective attention mechanism is used. The selective attention mechanism is the high-level cognitive ability, which is helpful to pay attention to the target information and eliminate the interference [10]. The study on the auditory selective attention mechanism is very few and mainly focused on the research field of physiology and cognition [11]. The model of feature extraction is shown as Fig. 1. The steps are as follows:

1) The speech signal is framed and fast Fourier transform (FFT) is used to compute the spectrogram.

2) The spectrogram is decomposed. The convolution of the spectrogram and linear decomposition Gauss kernel (6x6 Gauss kernel [1, 5, 10, 10, 5, 1]/32) is done. Then, based on the color channels, the brightness channels and the direction channels, the image is decomposed into multi-channel and multi-scale image sets. The relationship between the images of different scale σ [12] is:

$$P^{\sigma} = P^{(\sigma-1)}/2,\ \sigma = \{2, 3, 4, 5\} \tag{1}$$

Here, P^{σ} is the decomposed image with the scale of σ and P^{1} is the original image.

(a) The color channel contains two groups of images.

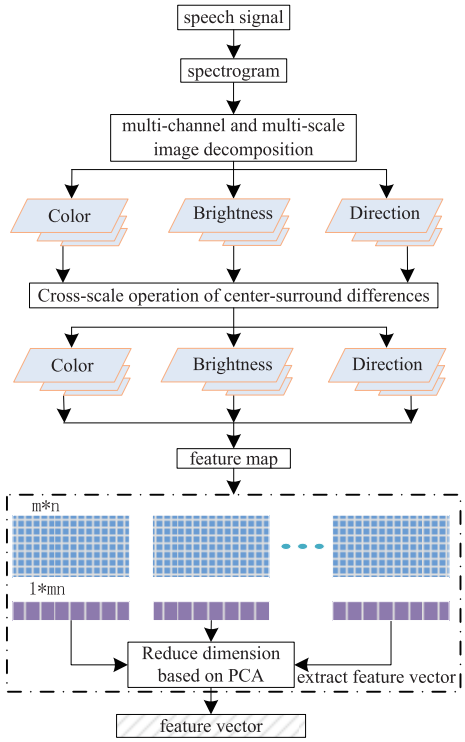

Fig. 1 Features extraction based on selective attention mechanism.

According to the opponent color theory proposed by German physiologist Hearin, the antagonism of $R-G$ and $B-Y$ are used to describe the contribution of color information to the saliency map. These two images are calculated by the following formula:

$$P^{\sigma}_{R-G} = (r^{\sigma} - g^{\sigma}) / \max(r^{\sigma}, g^{\sigma}, b^{\sigma}) \tag{2}$$

$$P^{\sigma}_{B-Y} = (b^{\sigma} - \min(r^{\sigma}, g^{\sigma})) / \max(r^{\sigma}, g^{\sigma}, b^{\sigma}) \tag{3}$$

Here, P^{σ}_{R-G} and P^{σ}_{B-Y} represent the decomposed images of color pairs ($R-G$ and $B-Y$), respectively. r^{σ}, g^{σ} and b^{σ} represent the red, green and blue values of the decomposed image with the scale of σ, respectively.

(b) The image of the brightness channel P^{σ}_{I} is expressed as the mean of r^{σ}, g^{σ} and b^{σ}.

$$P^{\sigma}_{I} = (r^{\sigma} + g^{\sigma} + b^{\sigma})/3 \tag{4}$$

(c) After convoluting with the two-dimensional Gabor filter, the image of different scale can be converted to the decomposed images of direction channel. The exploded pictures of direction features with different scale and angle can be calculated by the following formula:

$$P^{\sigma}_{\theta} = |P^{\sigma}_{I} * G_0(\theta)| + |P^{\sigma}_{I} * G_{\pi/2}(\theta)| \tag{5}$$

Here, $G_0(\theta)$ and $G_{\pi/2}(\theta)$ represent Gabor filters. $\pi/2$ and 0 are phases. θ denotes degree and $\theta = \{0°, 45°, 90°, 135°\}$.

3) The decomposed image with central scale σ_c and surrounding scale σ_s will be done center-surround differences, and then be normalized to get the final feature map.

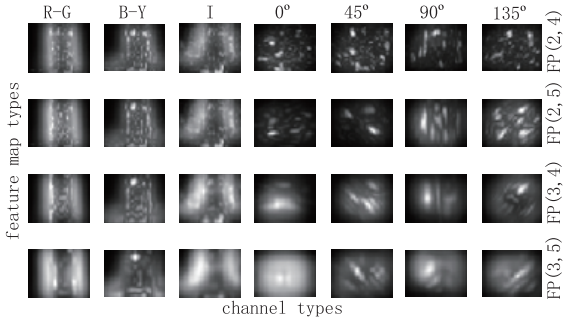

Fig. 2 Decomposed feature maps.

$$FP_e(\sigma_c, \sigma_s) = N(|P^{\sigma_c}_{e} - P^{\sigma_s}_{e}|) \tag{6}$$

Here, $e \in \{R-G, B-Y, I, 0°, 45°, 90°, 135°\}$ means the sets of the color channel, brightness channel and direction channel; $\sigma_c = \{2, 3\}$, $\sigma_s = \sigma_c + d$ and $d = \{2\}$; N denotes normalized operation. So every decomposed image (seven types) contains four center-surround subtraction $FP_e(2,4)$, $FP_e(3,4)$, $FP_e(2,5)$, $FP_e(3,5)$, which will get four feature maps and total is 28. The extracted feature maps of the spectrogram are shown as Fig. 2.

4) The feature matrix is extracted. One feature map is divided into m rows and n columns, totally $m * n$ sub-areas. Every sub-area is replaced by the average value. The image is normalized to the $m * n$-dimensional matrix [11]. The formula of the feature matrix is shown as:

$$FD_i(p,q) = \frac{mn}{vh} \sum_{g=\frac{pv}{n}}^{\frac{(p+1)v}{n}-1} \sum_{f=\frac{qh}{m}}^{\frac{(q+1)h}{m}-1} FP_i(g,f) \tag{7}$$

Here, $p \in [0, n-1]$, $q \in [0, m-1]$. v and h denote the width and height of the feature map, respectively. $FP_i, i = 1, 2, \cdots, 28$ is the feature map, which is equal to $FP_e(\sigma_c, \sigma_s)$. FD_i is the corresponding feature matrix. Here, m is 4 and n is 5.

5) The features are done dimensionality reduction and restructured. The feature matrix of every feature map is restructured to the $1 * mn$-dimensional vector. Then, PCA is used to reduce the features dimension. By calculating the eigenvectors of the covariance matrix of the original inputs, PCA can linearly transform a high-dimensional input vector into a low-dimensional one whose components are uncorrelated [13]. Here, 25% of the variance is retained.

3. Analyses of Emotional Features

Berlin speech emotion database (EMO-DB) are adopted in the experiments. Here, five persons (03, 08, 09, 10, 11) of the EMO-DB are training sets and others are testing sets. The comparisons of selective features based on the EMO-DB are shown in Fig. 3. Here, Fig. 3 (c) is the feature matrix of single emotion sample, and Fig. 3 (d) is the mean value of 535 emotion samples. From Fig. 3 (d), there are some differences for different emotions. For example, the amplitude range of the disgust is max, while that of the anger is min.

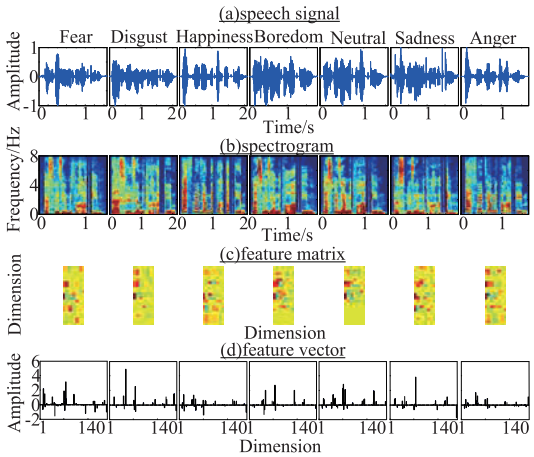

Fig. 3 Feature comparisons for seven emotions.

In addition, the amplitude fluctuation of the disgust and the sadness is smoother. Here, the proposed features are the statistical features of spectrogram, which are different from the traditional features. In order to remove the influence of invalid spectrogram components, the selective attention mechanism is introduced to obtain the effective time-frequency components to compute the emotional features more precisely.

4. Experiments for Emotion Recognition

In order to verify the efficiency of the proposed feature, the SER performance based on the traditional features and other spectrogram features are compared. The selected traditional features include acoustic feature, prosodic feature and chaotic feature [3]. The spectrogram feature includes three types: 1) Spectral pattern features (SPs) and harmonic energy features (HEs) [8]. Here, the spectral pattern feature contains the average value of spectrogram in each frame of sub-band and the relative value of spectrogram in each sub-band, totally 204 dimensions. The harmonic energy feature is the statistical feature of harmonic energy envelope, totally 260 dimensions. 2) Rotationally invariant texture energy measurements (RITEM), totally 42 dimensions [9]. 3) Proposed features (140 dimensions). Considering that the current emotion recognition features are all the superposition of traditional features and new features, every compared feature is composed of traditional features and different spectrogram features.

The classification algorithm for every feature is SVM. SVM is an efficient machine learning algorithm and is widely used for pattern recognition and classification problems. Here, SVM is selected based on two considerations [14]: 1) If the feature dimension is large, the number of samples has little effect on the performance of SVM; 2) The over-fitting problem between the model and data can be avoided.

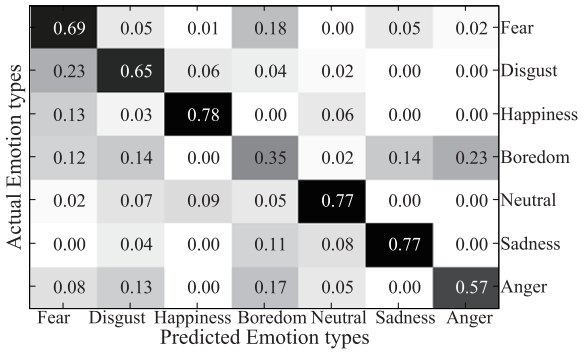

Fig. 4 SER results based on the traditional features.

Fig. 5 SER results based on the selective attention features.

4.1 Validation Experiments of Spectrogram Feature

The SER results based on the tradition features and the proposed features are shown in Figs. 4 and 5, respectively. From Fig. 4, the recognition rate of happiness, neutral and sadness is better, and the recognition rate of anger and boredom is worse. From Fig. 5, the mean recognition rate is 80.4%, which improves about 15% than the tradition features. Moreover, for the emotion of anger and boredom which the arousal effect is close, the recognition rate is obviously improved. From Fig. 3, the feature difference of these two emotions is obvious. So the recognition rate increases 21% and 38%, respectively. For others five emotions, the improvement of happiness is the smallest and increases about 2%. The misjudgment of anger and boredom is still bigger and exceeds 10%, so these two emotions need be further improved. Relatively speaking, the misjudgment of happiness and neutral is the lowest. From the arousal-valence space theory [15], the distance between these two emotions for the arousal effect and the valence effect are both far. It also can be seen from Fig. 3 that the feature differences of these two emotions are also obvious.

4.2 Comparisons of Different Spectrogram Features

The SER results based on four kinds of features are shown in Fig. 6. Here, SPS+HEs [8] and RITEM [9] include the traditional features. From the figure, when different spectrogram

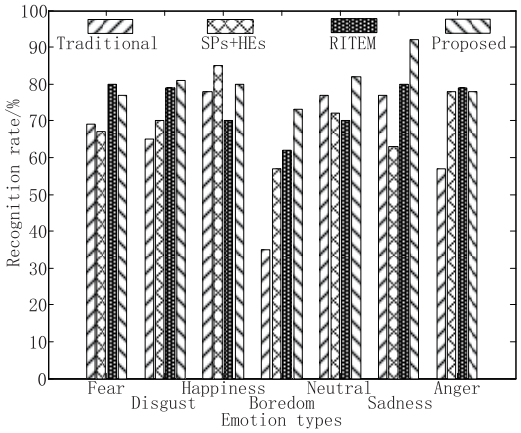

Fig. 6 Comparisons of SER based on different features.

features are added, there are different changes for the SER. The SER rates with four features are 65.4%, 70.3%, 74.3% and 80.4%, respectively. The recognition rates of SPS+HEs and RITEM are almost the same, but lower than that of the proposed features. For the tradition features, the recognition rate of each emotion is not always the lowest. Although the recognition rate of the traditional features for the disgust, boredom and anger emotion is the lowest, the recognition rate for the neutral emotion is higher than those of SPS+HEs and RITEM features. For the happiness emotion, the recognition rate of the SPS+Hes feature is the highest. For the fear and anger emotion, the recognition rate of the RITEM feature is the highest. The proposed features are not only the simple spectrogram feature, but also combined with selective attention mechanism to effectively extract features. So except happiness, fear and anger, the recognition rates of the proposed feature for the other emotions are the highest. To verify the significance of the experimental results, the Friedman test is adopted. Results show that the recognition rates for different methods are statistically significant ($p = 0.0488 < 0.05$). Then, results of multiple comparison show that the proposed features and the tradition features are statistically significant.

5. Conclusions

Inspired by the selective attention mechanism, a salient feature extraction algorithm is introduced into the SER. Firstly, the extraction model of selective attention feature is built, and then the selective attention feature is obtained from the spectrogram by multiscale decomposition and PCA. Finally, SVM is used for speech emotion classification. From the experiments of SER, the mean recognition rate of the proposed features improves about 15% than the tradition features. The improvement means that the proposed features are comprehensive embodiment of time-frequency features of speech signal, which has higher theoretical research value and practical value. For the later work, the speech emotion features should be further extracted and optimized, and the better combination of emotion features should be found to get more improvement. At the same time, the more effective classification algorithm should be studied, e.g., deep learning network.

Acknowledgments

The work was supported by the National Natural Science Foundation of China (No.61301219, 61375028, 61273266, 61301295) and the Natural Science Foundation of Jiangsu Province (No.BK20130241).

References

[1] Y. Attabi and P. Dumouchel, "Anchor models for emotion recognition from speech," IEEE Trans. Affect. Comput., vol.4, no.3, pp.280–290, Aug. 2013.
[2] A. Mencattini, E. Martinelli, G. Costantini, M. Todisco, B. Basile, M. Bozzali, and C. Di Natale, "Speech emotion recognition using amplitude modulation parameters and a combined feature selection procedure," Knowl.-Based. Syst., vol.63, pp.68–81, June 2014.
[3] X. Zhang, C. Huang, L. Zhao, and C. Zou, "Recognition of practical speech emotion using improved shuffled frog leaping algorithm," Acta Acustica, vol.39, no.2, pp.271–280, March 2014.
[4] T.A. Lampert and S.E.M. O'Keefe, "On the detection of tracks in spectrogram images," Pattern Recognit., vol.46, no.5, pp.1396–1408, May 2013.
[5] P.K. Ajmera, D.V. Jadhav, and R.S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram," Pattern Recognit., vol.44, no.10-11, pp.2749–2759, Oct. 2011.
[6] J. Dennis, H.D. Tran, and E.S. Chng, "Overlapping sound event recognition using local spectrogram features and the generalised hough transform," Pattern Recognit. Lett, vol.34, no.9, pp.1085–1093, July 2013.
[7] K. Lee, "Application of non-negative spectrogram decomposition with sparsity constraints to single-channel speech enhancement," Speech Commun., vol.58, pp.69–80, March 2014.
[8] A. Shahzadi, A. Ahmadyfard, K. Yaghmaie, and A. Harimi, "Recognition of emotion in speech using spectral patterns," Malays. J. Comput. Sci., vol.26, no.2, pp.140–158, 2013.
[9] K.C. Wang, "The feature extraction based on texture image information for emotion sensing in speech," Sensors, vol.14, no.9, pp.16692–16714, Sept. 2014.
[10] K.P. Walsh, E.G. Pasanen, and D. McFadden, "Selective attention reduces physiological noise in the external ear canals of humans. I: Auditory attention," Hearing Res., vol.312, pp.143–159, June 2014.
[11] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," IEEE Trans. Audio, Speech, Language Process., vol.17, no.5, pp.1009–1024, July 2009.
[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.20, no.11, pp.1254–1259, Nov. 1998.
[13] A.M. Martinez and A.C. Kak, "PCA versus LDA," IEEE Trans. Pattern Anal. Mach. Intell., vol.23, no.2, pp.228–233, Feb. 2001.
[14] C.-W. Hsu, and C.-J. Lin, "A comparison of methods for multiclass support vector machines," IEEE Trans. Neural Netw., vol.13, no.2, pp.415–425, March 2002.
[15] W.J. Han, H.F. Li, H.B. Ruan, and L. Ma, "Review on speech emotion recognition," J. Software, vol.25, no.1, pp.37–50, Jan. 2014.