LETTER

Spectral Features Based on Local Hu Moments of Gabor Spectrograms for Speech Emotion Recognition

Huawei TAO\textsuperscript{(a)}, Nonmember, Ruiyu LIANG\textsuperscript{(†)}, Member, Cheng ZHA\textsuperscript{1}, Xinran ZHANG\textsuperscript{†}, and Li ZHAO\textsuperscript{†}, Nonmembers

SUMMARY To improve the recognition rate of the speech emotion, new spectral features based on local Hu moments of Gabor spectrograms are proposed, denoted by GSLHu-PCA. Firstly, the logarithmic energy spectrum of the emotional speech is computed. Secondly, the Gabor spectrograms are obtained by convoluting logarithmic energy spectrum with Gabor wavelet. Thirdly, Gabor local Hu moments(\textit{GLHu}) spectrograms are obtained through block Hu strategy, then discrete cosine transform (DCT) is used to eliminate correlation among components of \textit{GLHu} spectrograms. Fourthly, statistical features are extracted from cepstral coefficients of \textit{GLHu} spectrograms, then all the statistical features form a feature vector. Finally, principal component analysis (PCA) is used to reduce redundancy of features. The experimental results on EmoDB and ABC databases validate the effectiveness of GSLHu-PCA.

\textbf{key words:} Gabor wavelet, Hu invariant moment, speech emotion recognition, spectral features

1. Introduction

To make human-computer interaction more intelligent, speech emotion recognition (SER) becomes more important\cite{1}. In addition, people often produce some negative emotions, effective identification of these emotions can help to improve the individual cognition and work efficiency \cite{2}. So, SER is of great practical significance.

To effectively recognize the emotion in speech, many kinds of speech features have been presented. Among them, spectral features are the most common features, such as MFCC and PLP \cite{3}, \cite{4}. Although these two features can accurately recognize speech content, they have poor performance in SER. Recently, research has shown that image feature extraction algorithms can extract emotional features from speech spectrogram. Some work (e.g. \cite{5}) uses Hu moments to describe local energy distribution of spectrogram which is related to emotion types, while others\cite{6}, \cite{7} use texture feature extraction algorithms to extract local texture features from spectrogram. Features mentioned above achieve better recognition results than MFCC. Despite these contributions, those image features still cannot characterize the emotional information in spectrogram adequately, and the spectrogram cannot provide more information for those algorithms, so further study is needed.

According to the above analysis, new spectral features based on local Hu moments of Gabor spectrograms are proposed, denoted by GSLHu-PCA. GSLHu-PCA uses Gabor wavelet to highlight local energy changes of spectrogram, and adopts block Hu strategy to describe local energy distribution of Gabor spectrograms, then DCT is used to eliminate correlation among components of Gabor local Hu moments spectrograms. Statistical features are computed for further processing. Finally, principal component analysis (PCA) is used to reduce redundancy of features. The experimental results on the EmoDB and ABC databases show that GSLHu-PCA not only achieves the best SER performance, but also the best fusion performance.

2. Spectral Features Based on Local Hu Moments of Gabor Spectrograms

2.1 The Principle of GSLHu-PCA

Figure 1 shows the flowcharts of the proposed GSLHu-PCA and HuWSF \cite{5}. As described in Fig. 1, The main differences between the two features lie in three points. First of all, Gabor wavelet is used to process logarithmic energy spectrogram \( E \) in GSLHu-PCA. Secondly, only \( L_{\mu,v} \) are re-
maintained in GSLHu-PCA, while $Q$ and $TQ$ are remained in HuWSF. Thirdly, PCA is used to reduce redundancy of features in GSLHu-PCA. GSLHu-PCA contains 8 steps.

1) The emotional speech signal $s$ is framed into short-term segments $S(n)$.

2) STFT coefficient $X(f)$ is generated by short-time Fourier transform (STFT) from $S(n)$.

3) Mel spectrum $Mel(f)$ and logarithmic energy spectrum $E(m)$ are computed through formula (1) and (2).

$$Mel(f) = 2595 \cdot \log\left(\frac{1 + f/f_0}{2}\right)$$

$$E(m) = \ln\left(\sum_{f=f_i}^{f_h} |Mel(f)|^2 H_m(f)\right), 0 \leq m < M$$

where $f_i, f_h$ are the low and high boundaries of the computed speech frequency. $M$ is the number of Mel filters. $H_m(f)$ is the triangular filter.

4) Gabor spectrograms $P_{\mu,v}$ are obtained through formula (3). Here, $\mu \in \{0, 1, 2, 3, 4\}$ is the Kernel direction of Gabor and $v \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ is the Kernel scale of Gabor.

$$P_{\mu,v} = conv\left(E, \psi_{\mu,v}\right)$$

where $conv(\cdot)$ is convolution operation, $E$ is logarithmic energy spectrogram. The Gabor wavelet $\psi_{\mu,v}$ is defined as following:

$$\psi_{\mu,v} = \begin{bmatrix} k_{\mu,v} \cdot \frac{\alpha^2}{\sigma^2} e^{-\frac{|y| \cdot \alpha}{\sigma^2}} e^{jk_{\mu,v}xz} - e^{-\frac{z^2}{\sigma^2}} \\ k_{\mu,v} \cdot \cos \varphi_{\mu,v} + jk_{\mu,v} \sin \varphi_{\mu,v} \end{bmatrix}$$

where $z$ is the spatial position of pixel; $\sigma$ is the radius of the Gaussian’s function. In formula (5), $k_{\mu,v} = k_{\text{max}}/f_0$, $k_{\text{max}} = \pi/2$, $f_0 = \sqrt{2}, \varphi_{\mu,v} = \pi\mu/8$.

5) Block Hu strategy is used to calculate local Hu moments of each Gabor spectrogram. Firstly, $P_{\mu,v}$ is divided into $(K - w + 1) \times (M/w)$ blocks $B_{ij}$ by (6).

$$B_{ij} = \begin{bmatrix} P_{\mu,v}(i, wj) & \cdots & P_{\mu,v}(i, wj + w - 1) \\ \vdots & \vdots & \vdots \\ P_{\mu,v}(i + w - 1, wj) & \cdots & P_{\mu,v}(i + w - 1, wj + w - 1) \end{bmatrix}$$

where $w$ is the size of block, $wj$ means $w \cdot j$. $K$ is the number of emotional speech frames.

Secondly, Hu moments $I$ of $B_{ij}$ is acquired by (7). Here, $\beta_{20}$ and $\beta_{02}$ are calculated by (8), (9), (10) and (11).

$$I = \beta_{20} + \beta_{02}$$

$$\beta_{pq} = \alpha_{pq} \left(\frac{1}{C_{pq}}\right) r = \frac{(p+q)}{2+1}$$

$$\alpha_{pq} = \sum_{x=1}^{U} \sum_{y=1}^{V} (x-x)(y-y)g(x, y)$$

$$\epsilon_{pq} = \sum_{x=1}^{U} \sum_{y=1}^{V} x^pg^qg(x, y)$$

where $U, V$ is the width and the height of block respectively. $\alpha_{pq}$ is the $(p+q)$-th order normalized center moment, $\beta_{pq}$ is the $(p+q)$-th order center moment, $\epsilon_{pq}$ is the $(p+q)$-th order moment, and $(\bar{x}, \bar{y})$ is the center of energy gravity.

After above processing, Gabor local Hu moments (GLHu) spectrograms $Y_{\mu,v}$ are generated.

6) To eliminate correlation among components of GLHu spectrogram $Y_{\mu,v}$, DCT is used to process $Y_{\mu,v}$ then cepstral coefficients $C_{\mu,v}$ are obtained. As the high cepstral coefficients may be distorted by noise, only the value from the 2-nd to the 13-th coefficients of $C_{\mu,v}$ are selected, which is defined as $L_{\mu,v}$.

7) $L_{\mu,v}$ must be transformed into appropriate formats for further processing. Statistical features are extracted from $L_{\mu,v}$ and its first and second derivatives, which are defined as $R_{\mu,v}, \Delta R_{\mu,v}$ and $\Delta \Delta R_{\mu,v}$ respectively. Then feature $R = [R_{1,1}, \Delta R_{1,1}, \Delta \Delta R_{1,1}, \cdots, \Delta R_{4,7}, \Delta \Delta R_{4,7}]$ is gained.

8) PCA is used to reduce features redundancy, and GSLHu-PCA is generated.

2.2 The Analysis and Evaluation Framework of GSLHu-PCA

It has been reported that energy on spectrogram varies with emotion types and local energy distribution of spectrogram is related to speech emotion[5]–[7]. Gabor wavelet [8] is a multi-scales and multi-directions analytical tool, which highlights the local energy changes of signal and also retains the information of signal. Figure 2 shows logarithmic energy spectrograms $E$ and partial Gabor spectrograms $P_{\mu,v}$. It can be seen from Fig. 2 that each Gabor spectrogram has a similar global distribution of voiceprint, which means that each Gabor spectrogram retains partial distribution of logarithmic energy spectrogram. It also can be seen from Fig. 2 that local energy distribution of each Gabor spectrogram is different from each other. This is because Gabor wavelet describes local energy changes of spectrogram under different angles. Gabor wavelet with 5 scales and 8 directions is used in this letter. There are totally 40 Gabor spectrograms for one utterance, which means that Gabor spectrograms have
clearer local energy changes than energy spectrogram. So Gabor spectrograms are used instead of logarithmic energy spectrogram. In addition, block Hu strategy can evaluate the degree of how energy of frequencies concentrates to some frequencies in a block of spectrogram, which describes local energy distribution of spectrogram [5]. So block Hu strategy is used to describe local energy distribution of Gabor spectrograms. After Gabor and block Hu processing, DCT is used to eliminate correlation among components of GLHu spectrograms and PCA is used to reduce redundancy of statistical features.

SER framework includes 3 steps: feature extraction, feature selection, classification. Similar to the experimental setting of Ref. [5], GSLHu-PCA is performed for feature extraction and the following feature statistics methods are selected: mean, std, max, min, kurtosis, skewness, range and median. MCFS algorithm [9] is used for feature selection and SVM with polynomial kernel is selected as classifier.

3. Experiments

3.1 Database and Evaluation Criteria

EmoDB [10] and ABC corpuses [11] are selected as simulation databases. EmoDB corpus is composed of 535 utterances from ten professional actors, which contains seven kinds of emotional speeches: anger, boredom, disgust, fear, happiness, neutral and sadness. ABC corpus is composed of 430 utterances from 8 actors which contains six kinds of emotional speeches: aggressive, cheerful, intoxicated, nervous, neutral, and tired. All the utterances are randomly divided into 5 equal parts, 4 parts act as training set and the remaining 1 part acts as testing set. The above steps are repeated 10 times to eliminate the random impact. The weighted average recall (WA) is used as evaluation criteria. WA is the total number of correctly classified test samples of all classes averaged by the total number of test samples [12].

3.2 Single Feature Recognition Experiment

Some of the most popular spectral features are compared, which contain MFCC, HuWSF and PLP. As described in Sect. 2.2, all spectral features are computed by the same statistical methods. Meanwhile, the first and the second derivatives of these features are also computed. The dimensions of MFCC, HuWSF and PLP are 288, 672 and 1008 respectively. Figure 3 shows the WA of different features after PCA processing, where SVM with polynomial kernel is selected as classifier. It can be seen from Fig. 3 that 250-dimensions GSLHu-PCA achieves better results in both databases. Meanwhile, 250-dimensions features are near to the dimensions of MFCC. So 250-dimensions GSLHu-PCA is used in this letter. It is worth noting that GSLHu-PCA achieves better results than HuWSF after PCA processing. As described in Fig. 1, only $L_{\mu,v}$ are remained in GSLHu-PCA, while $Q$ and $TQ$ are remained in HuWSF, which means that it is easier to extract emotional features from $L_{\mu,v}$ and the introduction of Gabor makes GSLHu-PCA achieving better recognition performance.

Based on SER framework of Sect. 2.2, Fig. 4 shows the recognition results of different features and Table 1 shows the maximum WA of different features in Fig. 4. Compared with HuWSF, PLP and MFCC, the maximum WA of GSLHu-PCA is improved by 1.6%, 5.53% and 12.68% respectively in EmoDB, and is improved by 1.35%, 4.97% and 7.47% respectively in ABC.

Pairwise t-tests are conducted to further evaluate performance of GSLHu-PCA. To ensure the testing set containing sufficient samples, database is randomly divided into 2 equal parts. One part acts as training set and the remained part is divided into 5 equal datasets. Each dataset acts as a testing set. Based on SER framework of Sect. 2.2, Table 2 shows the maximum WA of HuWSF and GSLHu-PCA for one experiment. GSLHu-PCA is statistical significance in both corpuses ($p_{EmoDB} = 0.031 < 0.05$, $p_{ABC} = 0.024 < 0.05$). After the experiment is repeated 10 times, GSLHu-PCA is also statistical significance because the average p in two corpuses are $p_{EmoDB} = 0.0373$ and $p_{ABC} = 0.0294$.

3.3 Fusion Feature Recognition Experiment

In order to detect the fusion performance of GSLHu-PCA,
Table 2 The maximum WA of HuWSF and GSLHu-PCA

<table>
<thead>
<tr>
<th>EmoDB</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuWSF</td>
<td>84.18%</td>
<td>84.91%</td>
<td>82.69%</td>
<td>81.82%</td>
<td>79.25%</td>
</tr>
<tr>
<td>GSLHu-PCA</td>
<td>87.27%</td>
<td>84.91%</td>
<td>86.54%</td>
<td>83.64%</td>
<td>81.13%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ABC</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>HuWSF</td>
<td>63.79%</td>
<td>71.43%</td>
<td>69.05%</td>
<td>69.44%</td>
<td>69.18%</td>
</tr>
<tr>
<td>GSLHu-PCA</td>
<td>65.12%</td>
<td>73.14%</td>
<td>71.43%</td>
<td>69.77%</td>
<td>69.87%</td>
</tr>
</tbody>
</table>

GSLHu-PCA is fused with acoustic feature set PROS [5]. The PROS feature set is INTERSPEECH2010 feature set [13] of removing MFCC and contains: loudness, logarithmic power of Mel-frequency bands, the 8 line spectral pair frequencies, the envelope of the smoothed fundamental frequency contour, the voicing probability of the final fundamental frequency candidate. The PROS feature set totally includes 952 features. The comparison fusion features include: PROS + GSLHu-PCA, PROS + PLP, PROS + MFCC, PROS + HuWSF.

Based on SER framework of Sect. 2.2, Fig. 5 shows the recognition results of different fusion features and Table 3 shows that the maximum WA of different fusion features in Fig. 5. Compared with PROS + HuWSF, PROS + PLP and PROS + MFCC, the maximum WA of PROS + GSLHu-PCA is improved by 1.15%, 3.17% and 3.31% respectively in EmoDB, and is improved by 4.45%, 6.38% and 4.42% respectively in ABC. The experimental results show that GSLHu-PCA not only achieves the best SER performance, but also the best fusion performance.

4. Conclusions

In this letter, new spectral features based on local Hu moments of Gabor spectrograms are proposed, denoted by GSLHu-PCA. GSLHu-PCA uses Gabor wavelet to highlight local energy changes of spectrogram and adopts block Hu strategy to describe local energy distribution of Gabor spectrograms. GSLHu-PCA has the following advantages: (1) SER performance of GSLHu-PCA is better than that of the most classic spectral features. (2) GSLHu-PCA shows the best fusion performance, which can be used as an effective complement to improve the recognition performance of SER system.

Acknowledgments

This research project was founded in part by National Natural Science Foundation (No.61301219, 61375028, 61273266), Natural Science Foundation of Jiangsu Province (No. BK20130241).

References