SPEECH EMOTIONAL FEATURES MEASURED BY POWER-LAW DISTRIBUTION BASED ON ELECTROGLOTTOGRAPHY

Lijiang Chen¹, Xia Mao¹, Yuli Xue¹ and Mitsuru Ishizuka²

¹School of Electronic and Information Engineering, Beihang University, 100191, Beijing, China
²Department of Information and Communication Engineering, University of Tokyo, Tokyo, Japan

Keywords: Speech emotional features, Power-law distribution, Electroglottography.

Abstract: This study was designed to introduce a kind of novel speech emotional features extracted from Electroglottography (EGG). These features were obtained from the power-law distribution coefficient (PLDC) of fundamental frequency ($F_0$) and duration parameters. First, the segments of silence, voiced and unvoiced (SUV) were distinguished by combining the EGG and speech information. Second, the $F_0$ of voiced segment and the first-order differential of $F_0$ was obtained by a cepstrum method. Third, PLDC of voiced segment as well as the pitch rise and pitch down duration were calculated. Simulation results show that the proposed features are closely connected with emotions. Experiments based on Support Vector Machine (SVM) are carried out. The results show that proposed features are better than those commonly used in the case of speaker independent emotion recognition.

1 INTRODUCTION

As an important channel of human communication, voice contains information of the content of speech, speaker identification and speaker emotion. This paper focuses on the problems existing in the emotional speech processing, which is to recognize the user’s emotional state by analyzing speech patterns. There are already a number of systems that are capable of emotional recognition. However, both speaker dependent and speaker independent speech emotion recognition are far from satisfying. It has always been a dream to give computers speech emotion recognition ability close to or even beyond humans. Looking for new efficient features is one of the effective directions of this study. Several researchers have studied the acoustic correlates of emotion affect in speech signals (Ververidis and C., 2006; Yang and Lugger, 2010). Prosody features such as pitch variables and speech rate were analyzed through pattern recognition (Cowie and Cornelius, 2003) (Borchert and Dusterhoft, 2005). Zhao combined the two kinds of features to recognize Mandarin emotions (Zhao et al., 2005). Shami make use of the segment-level features (Shami and Kamel, 2005). Some other researchers put emphasis on the integration of acoustic and linguistic information (Schuller et al., 2004). There are two basic difficulties of emotion recognition on prosodic features. First, it’s very difficult to extract the fundamental frequency because of the influence of vocal tract and aerodynamic noise. Second, the change of $F_0$ over time is more associated with emotion than the $F_0$ itself. It is challenging to extract the information of $F_0$ distribution over time associated with emotion. Electroglottograph or EGG is a system which gives an information on the vocal folds vibration by measuring the electrical resistance between two electrodes placed around the throat. It gives a very useful information about speech, especially because it’s the source of the phonation (no influence of vocal tract and no aerodynamic noise). EGG has been used to extract $F_0$ in the field of medical and rehabilitation (Kania et al., 2006). In this contradiction, EGG is firstly used for distinguishing the segments of silence, voiced and unvoiced. Then the $F_0$ is obtained from EGG with cepstrum method for emotion recognition.

A power law is a special kind of mathematical relationship between two quantities. When the frequency of an event varies as a power of some attribute of that event (e.g. its size), the frequency is said to follow a power law. As we know, a large number of independent small events meet the normal distribution which is the ideal situation. In reality, events are often dependent and not "small" enough. Therefore, events often meet the power-law distribution rather than normal distribution. For instance, the number of cities...
having a certain population size is found to vary as a power of the size of the population, and hence follows a power law.

The paper is structured as follows: section 2 gives the description of the database used, which includes speech and EGG data; section 3 explains the process of feature extraction, including SUV distinguishing and PLDC calculation; section 4 introduces the experiments based on SVM; and finally conclusions.

2 DATABASE DESCRIPTION

To evaluate the new features proposed, Beihang University Database of Emotional Speech (BHURES) was set up to provide speech utterances. All the utterances were stereo whose left channel contains the acoustic data and right channel contains the EGG data.

SUBJECTS
Fifteen healthy volunteers were invited to establish the database, including seven male and eight female. The emotions used resemble the far spread MPEG-4 set, namely joy, anger, disgust, fear, sadness, surprise and added neutrality. The database contains twenty texts with no emotional tendencies. Each sentence was repeated three times for each emotion, thus 6,300 utterances were obtained. All these utterances have a sample-frequency as 11025Hz and mean duration as 1.2s.

INSTRUMENTATION
Acoustic data was obtained by a BE-8800 electret condenser microphone. TIGEX-EGG3 (Tiger DRS, Inc., America) measured the EGG signals. The output of the EGG device was processed by an electronic preamplifier and then by a 16-bit analog-to-digital (A/D) converter that was included in a OPTIPLEX 330 personal computer. Both the EGG and acoustic data were analyzed by MATLAB. A raw data of acoustic and EGG are shown in Fig. 1. The segments of silence, voiced and unvoiced are signed with vertical lines.

EVALUATION
Besides, an emotional speech evaluation system is established to ensure the reliability of the utterances. Emotional speech which is accurately recognized by at least p% of strange listeners is collected into a subset, where \( p \in \{50, 60, 70, 80, 90, 100\} \).

The subset S70 is selected for further experiments because of the appropriate quality and quantity. There are in total 3456 mandarin utterances covering all the emotion categories in the S70 subset.

3 FEATURE EXTRACTION

There are two steps of feature extraction. First, the segment of voiced speech, unvoiced speech and silence were separated using information from both acoustic data and EGG data. Second, we focus on the characteristics distribution of time-domain. The duration distribution of voiced segment, pitch rise segment and pitch down segment were analyzed by the power-law distribution coefficient (PLDC).

SUV DISTINGUISHING
In speech analysis, the SUV decision which is used to divide whether a given segment of a speech signal should be classified as voiced speech, unvoiced speech, or silence, based on measurements made on the signal. The measured parameters include the zero-crossing rate, the speech energy, the correlation between adjacent speech samples, etc. (Atal and Rabiner, 1976). It is usually performed in conjunction with pitch analysis. However, without the information of EGG, the linking of SUV decision to pitch analysis results in unnecessary complexity. Fig. 2 shows the log energy histograms of acoustic and EGG data.

In Fig. 2, both the log energy histograms of acoustic and log energy histograms of EGG have two peaks. The left one represents the unvoiced or silent segments, while the right one represents the voiced segments. We use the maximum a posteriori method to fit the two classes data near the two peaks for both acoustic and EGG. The recognition rates obtained are 95.98% and 99.96% respectively. This indicates that the EGG has shown excellent in the recognition of voice segment.

Based on the above analysis, we designed three threshold which are determined by the result of statistics as shown in Fig. 2. A SUV division algorithm was designed as shown in Fig. 3.
The algorithm to divide SUV is given in detail as follows:

- Use 50Hz high-pass filters to filter the low-frequency noise generated by muscle movement.
- Calculate the energy of each frame of filtered EGG data as equation (1):
  \[ E_1 = \log\left(\sqrt{\frac{\sum_{i=1}^{n} x_1(i)^2}{L}}\right) \]  
  where \( x_1(i) \) is the ith frame of filtered EGG data.
- Use digital filter as equation (2) to enhance the high-frequency components of acoustic signal.
  \[ H(z) = 1 - 0.95 \times z^{-1} \]  
- Calculate the energy of each frame of filtered acoustic data as equation (3):
  \[ E_2 = \log\left(\sqrt{\frac{\sum_{i=1}^{n} x_2(i)^2}{L}}\right) \]  
  where \( x_2(i) \) is the ith frame of acoustic data.
- If the current frame meet equation (4):
  \[ E_1 > \text{thr}_0 \text{ and } E_2 > \text{thr}_2 \]  
  this frame belongs to voiced segment.
- Else if the current frame meet equation (5):
  \[ E_2 > \text{thr}_1 \]  
  this frame belongs to unvoiced segment.
- Else this frame belongs to silence segment.

**TIME-DOMAIN DISTRIBUTION**

In this section, we focus on the characteristics distribution of time-domain. First, the duration distribution of voiced segment was processed as follows:

- Assume the length of each voiced segment is:
  \[ V(n), n = 1, 2, \ldots, N \]  
  where \( N \) means the number of voiced segments.
- The histogram of \( V(n) \) is obtained as (7):
  \[ [F, x_{out}] = \text{hist}(V(n)) \]  
  where \( F \) means the frequency rate and \( x_{out} \) means the bin locations.
- \( F \) and \( x_{out} \) probably follow a power law distribution as (8):
  \[ F \approx A \cdot x_{out}^B \]  
  where \( A \) and \( B \) are constants.
- Taking the logarithm on both sides
  \[ \ln(F) \approx \ln(A) + B \ln(x_{out}) \]  
- The power-law distribution coefficient (PLDC) is calculated as equation (10).
  \[ [P_1, P_2] = \text{polyfit}([\ln(F)], [\ln(x_{out})], 1) \]  
  where function polyfit is used to find the coefficients of a polynomial \( p(x) = P_1 \cdot x + P_2 \) of degree 1 that fits the data, \( \ln(F) \) to \( \ln(x_{out}) \), in a least squares sense.

Fig. 4 shows the relation of \( \ln(F) \) and \( \ln(x_{out}) \). The left column is six basic emotions and neutrality data for male. The right column is the same emotions but for female. The data followed each title of the subfigure means the PLDC \([P_1, P_2]\].

From Fig. 4 we can indicate that, for both male and female data, PLDC of the voiced segment duration are closely related with emotions. Especially in the anger situation, the \( |P_1| \) is significantly higher than neutrality and sadness situation. Other emotions with...
high arousal have the similar trends, such as surprise and joy.

The duration distribution of pitch rise segment and pitch down segment were analyzed by the PLDC method as the same process above. F0 of each frame are obtained by a cepstrum method (Noll, 1967). Fig. 5 shows the power law distribution of the pitch rise duration and pitch down duration. The left column is six basic emotions and neutrality data for male. The right column is the same emotions data but for female. The data followed each title of the subfigure means the PLDC of pitch rise duration and pitch down duration.

From Fig. 5 we can indicate that, for both male and female data, PLDC of pitch rise duration are closely related with emotions. For example, both in male and female surprise subfigure, [P1] of pitch rise duration are higher than other emotions. The relations between PLDC of pitch down duration and emotion is not obvious.
4 EXPERIMENTS

In order to evaluate the efficiency of PLDC emotional features, comparative experiments are designed. All these experiments are based on the s70 corpus described in section 2.

4.1 Sort of Speech Features

We apply a comparative test of the proposed PLDC features and traditional features based Sequential Forward Floating Search (SFS) and Sequential Backward Floating Search (SBS). SFS and SBS are known for their high performance as shown in (Pudil et al., 1994). In addition to fundamental frequency ($F_0$), the first three formant ($F_1,F_2,F_3$), the energy ($E$) and zero crossing rate ($Z$) are extracted. We calculated the statistic of the frame-level feature, including maximum, minimum, mean and standard deviation. These statistic features are compared with the proposed PLDC features include PLDC of voiced segment duration ($PLDC_{V_1}$ and $PLDC_{V_2}$), PLDC of pitch rise duration ($PLDC_{P_1}$ and $PLDC_{D_2}$) and PLDC of pitch down duration ($PLDC_{P_1}$ and $PLDC_{D_2}$).

<table>
<thead>
<tr>
<th>Order</th>
<th>SFS</th>
<th>SBS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1th</td>
<td>$PLDC_{V_1}$</td>
<td>$PLDC_{V_1}$</td>
</tr>
<tr>
<td>2th</td>
<td>$mean - F_0$</td>
<td>$max - F_0$</td>
</tr>
<tr>
<td>3th</td>
<td>$max - E$</td>
<td>$PLDC_{P_1}$</td>
</tr>
<tr>
<td>4th</td>
<td>$max - F_1$</td>
<td>$PLDC_{P_2}$</td>
</tr>
<tr>
<td>5th</td>
<td>$max - F_2$</td>
<td>$mean - F_0$</td>
</tr>
<tr>
<td>6th</td>
<td>$max - F_0$</td>
<td>$PLDC_{V_2}$</td>
</tr>
<tr>
<td>7th</td>
<td>$PLDC_{P_1}$</td>
<td>$max - E$</td>
</tr>
<tr>
<td>8th</td>
<td>$max - Z$</td>
<td>$PLDC_{P_1}$</td>
</tr>
<tr>
<td>9th</td>
<td>$mean - E$</td>
<td>$PLDC_{P_2}$</td>
</tr>
<tr>
<td>10th</td>
<td>$mean - Z$</td>
<td>$max - F_1$</td>
</tr>
</tbody>
</table>

From table 1, we can see that, under the same conditions, the proposed PLDC features is sorted in the head of other features both in SFS and SBS. That means the PLDC features are closely linked to emotion expression.

4.2 Comparative Emotion Recognition

SVM is a novel type of learning machine, which bases on statistical learning theory (SLT) (Cortes and Vapnik, 1995). Some studies suggest that SVM classifier is more effective than others in speech emotion recognition (Lin and Wei, 2005; Schuller et al., 2005). For these reasons, we use SVM as classifiers for comparative emotion recognition. The Sigmoid function is selected for kernel function. Half of the samples in S70 are selected randomly for five times to train classifiers, while the utterances left are objects to be recognized. The average recognition results using traditional features and using the proposed PLDC features are shown in table 2 and table 3.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sad.</td>
<td>0.539</td>
<td>0.111</td>
<td>0.033</td>
<td>0.128</td>
<td>0.033</td>
<td>0.172</td>
</tr>
<tr>
<td>Ang.</td>
<td>0.717</td>
<td>0.111</td>
<td>0.033</td>
<td>0.128</td>
<td>0.033</td>
<td>0.172</td>
</tr>
<tr>
<td>Sur.</td>
<td>0.022</td>
<td>0.072</td>
<td>0.444</td>
<td>0.083</td>
<td>0.211</td>
<td>0.167</td>
</tr>
<tr>
<td>Fea.</td>
<td>0.267</td>
<td>0.033</td>
<td>0.367</td>
<td>0.044</td>
<td>0.289</td>
<td>0.522</td>
</tr>
<tr>
<td>Hap.</td>
<td>0.011</td>
<td>0.039</td>
<td>0.311</td>
<td>0.017</td>
<td>0.522</td>
<td>0.101</td>
</tr>
<tr>
<td>Dis.</td>
<td>0.117</td>
<td>0.017</td>
<td>0.067</td>
<td>0.156</td>
<td>0.217</td>
<td>0.428</td>
</tr>
</tbody>
</table>

Data from table 2 and table 3 shows that the average emotion recognition rate are increased by using the proposed PLDC features. The recognition rate for the emotions of surprise, anger, happiness and sadness are increased most significantly.

5 CONCLUSIONS

In this paper, we propose a kind of novel speech emotion features, PLDC, to recognize emotional states contained in spoken language. This kind of features are obtained based on EGG signal which avoid the influence of vocal tract and aerodynamic noise. Comparative experiments based on SVM proved that the proposed features have high relativity with the speech emotion.

ACKNOWLEDGEMENTS

This research is supported by the International Science and Technology Cooperation Program of China (No.2010DFA11990) and the National Nature Science Foundation of China (No. 61103097).

REFERENCES

Atal, B. and Rabiner, L. (1976). A pattern recognition approach to voiced-unvoiced-silence classifica-


