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

Lijiang Chen[1], Xia Mao[1] , Yuli Xue[1] and Mitsuru Ishizuka[2]
[1]*School of Electronic and Information Engineering, Beihang University, 100191, Beijing, China*
[2]*Department 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

Chen L., Mao X., Xue Y. and Ishizuka M..
SPEECH EMOTIONAL FEATURES MEASURED BY POWER-LAW DISTRIBUTION BASED ON ELECTROGLOTTOGRAPHY.
DOI: 10.5220/0003886301310136
In *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing* (BIOSIGNALS-2012), pages 131-136
ISBN: 978-989-8425-89-8

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 (*BHUDES*) 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 *S*70 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 *S*70 subset.

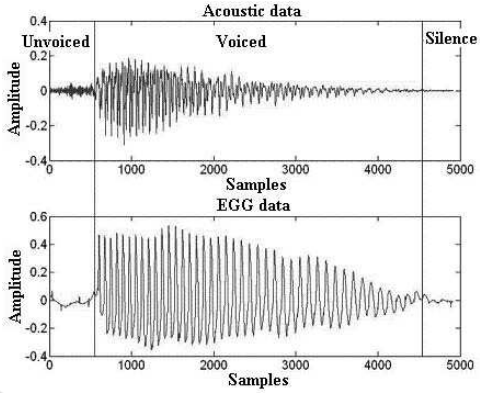

Figure 1: A raw data of acoustic and EGG.

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.

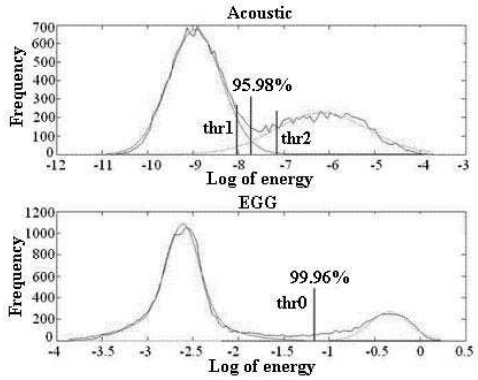

Figure 2: Log energy histograms of acoustic and EGG.

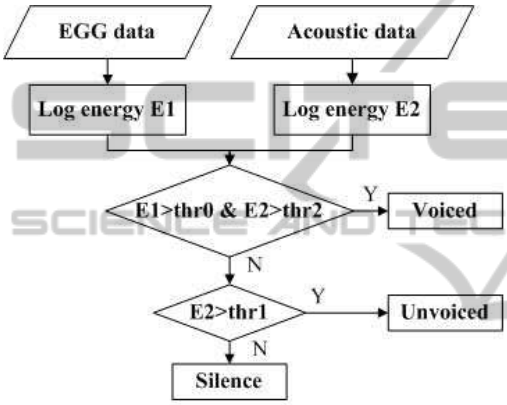

Figure 3: *SUV* division algorithm.

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):

$$E_1 = \log(\sqrt{\sum_{i=1}^{n} x_1(i)^2 \Big/ L}) \tag{1}$$

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.

$$H(z) = 1 - 0.95 \times z^{-1} \tag{2}$$

- Calculate the energy of each frame of filtered acoustic data as equation (3):

$$E_2 = \log(\sqrt{\sum_{i=1}^{n} x_2(i)^2 \Big/ L}) \tag{3}$$

where $x_2(i)$ is the ith frame of acoustic data.

- If the current frame meet equation (4):

$$E_1 > thr_0 \& E_2 > thr_2 \tag{4}$$

this frame belongs to voiced segment.

- Else if the current frame meet equation (5):

$$E_2 > thr_1 \tag{5}$$

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:

$$V(n), n = 1, 2, \cdots N \tag{6}$$

where N means the number of voiced segments.

- The histogram of $V(n)$ is obtained as (7):

$$[F, xout] = hist(V(n)) \tag{7}$$

where F means the frequency rate and *xout* means the bin locations

- F and *xout* probably follow a power law distribution as (8):

$$F \approx A \cdot xout^B \tag{8}$$

where A and B are constants.

- Taking the logarithm on both sides

$$\ln(F) \approx \ln(A) + B\ln(xout) \tag{9}$$

- The power-law distribution coefficient (*PLDC*) is calculated as equation (10).

$$[P_1, P_2] = polyfit(ln(F), ln(xout), 1) \tag{10}$$

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 $ln(xout)$, in a least squares sense.

Fig. 4 shows the relation of $ln(F)$ and $ln(xout)$. 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* $[P_1, P_2]$.

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

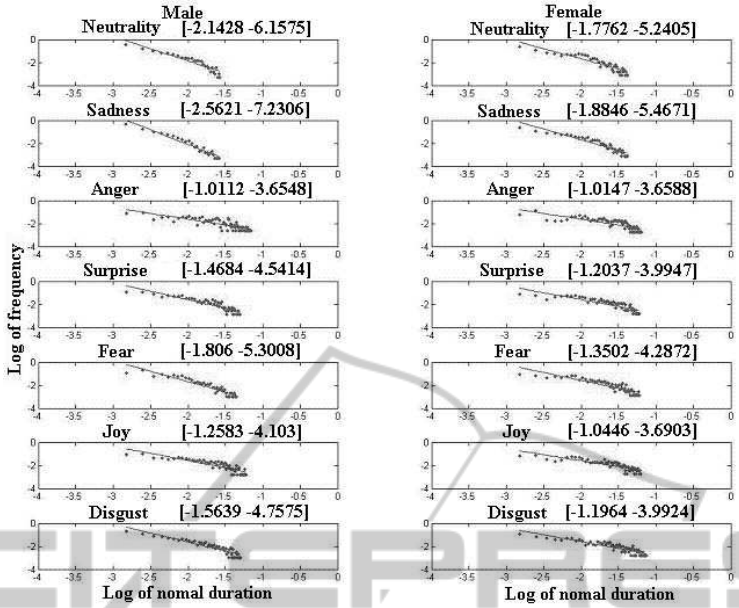

Figure 4: *PLDC* of voiced duration.

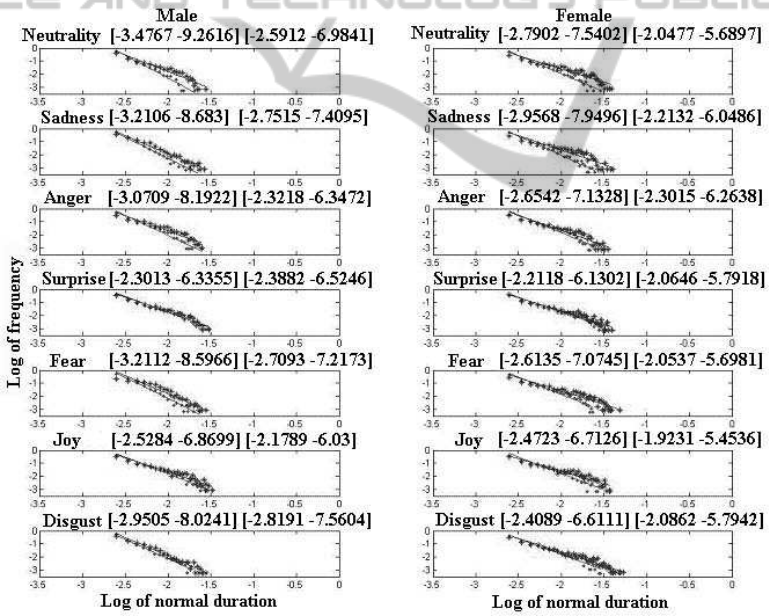

Figure 5: *PLDC* of pitch rise and pitch down duration.

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. F_0 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, $[P_1]$ 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 *s*70 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 ($PLDC-V_{s1}$ and $PLDC-V_{s2}$), *PLDC* of pitch rise duration ($PLDC-P_{r1}$ and $PLDC-P_{r2}$) and *PLDC* of pitch down duration ($PLDC-P_{d1}$ and $PLDC-P_{d2}$).

Table 1: The first 10 Features in *SFS* and *SBS*.

Order	*SFS*	*SBS*
1th	$PLDC-V_{s1}$	$PLDC-V_{s1}$
2th	$mean-F_0$	$max-F_0$
3th	$max-E$	$PLDC-P_{r1}$
4th	$max-F_1$	$PLDC-P_{r2}$
5th	$max-F_2$	$mean-F_0$
6th	$max-F_0$	$PLDC-V_{s2}$
7th	$PLDC-P_{r1}$	$max-E$
8th	$max-Z$	$PLDC-P_{d1}$
9th	$mean-E$	$PLDC-P_{d2}$
10th	$mean-Z$	$max-F_1$

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 *S*70 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 2: Results using traditional features.

Emo.	Sad.	Ang.	Sur.	Fea.	Hap.	Dis.
Sad.	**0.539**	0	0	0.289	0	0.172
Ang.	0	**0.717**	0.111	0.033	0.128	0.011
Sur.	0.022	0.072	**0.444**	0.083	0.211	0.167
Fea.	0.267	0	0.033	**0.367**	0.044	0.289
Hap.	0.011	0.039	0.311	0.017	**0.522**	0.101
Dis.	0.117	0.017	0.067	0.156	0.217	**0.428**

Table 3: Results using the proposed *PLDC* features.

Emo.	Sad.	Ang.	Sur.	Fea.	Hap.	Dis.
Sad.	**0.717**	0	0	0.117	0	0.167
Ang.	0	**0.822**	0.056	0.033	0.083	0.006
Sur.	0.011	0.022	**0.678**	0.133	0.094	0.061
Fea.	0.226	0	0.028	**0.4**	0.044	0.272
Hap.	0.006	0.039	0.167	0.022	**0.672**	0.094
Dis.	0.106	0.028	0.067	0.156	0.172	**0.472**

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-

tion with applications to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 24(3):201–212.

Borchert, M. and Dusterhoft, A. (2005). Emotions in speech-experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on*, pages 147–151. IEEE.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cowie, R. and Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1-2):5–32.

Kania, R., Hartl, D., Hans, S., Maeda, S., Vaissiere, J., and Brasnu, D. (2006). Fundamental frequency histograms measured by electroglottography during speech: a pilot study for standardization. *Journal of Voice*, 20(1):18–24.

Lin, Y. and Wei, G. (2005). Speech emotion recognition based on HMM and SVM. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 8, pages 4898–4901. IEEE.

Noll, A. (1967). Cepstrum pitch determination. *The journal of the acoustical society of America*, 41:293.

Pudil, P., Novovicová, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters*, 15(11):1119–1125.

Schuller, B., Reiter, S., Muller, R., Al-Hames, M., Lang, M., and Rigoll, G. (2005). Speaker independent speech emotion recognition by ensemble classification. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 864–867. IEEE.

Schuller, B., Rigoll, G., and Lang, M. (2004). Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I577–I580. IEEE.

Shami, M. and Kamel, M. (2005). Segment-based approach to the recognition of emotions in speech. In *2005 IEEE International Conference on Multimedia and Expo*, pages 1–4. IEEE.

Ververidis, D. and C., K. (2006). Emotional speech recognition-resources features and methods. *Speech Communication*, 48:1162–1181.

Yang, B. and Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing*, 90(5):1415–1423.

Zhao, L., Cao, Y., Wang, Z., and Zou, C. (2005). Speech emotional recognition using global and time sequence structure features with mmd. *Affective Computing and Intelligent Interaction*, pages 311–318.