A CROSS GENDER AND CROSS LINGUAL STUDY ON ACOUSTIC FEATURES FOR STRESS RECOGNITION IN SPEECH

Xin Zuo & Pascale Fung

Human Language Technology Center, Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Hong Kong
xinzuo@ust.hk; pascale@ee.ust.hk

ABSTRACT

We present a systematic study of the acoustic features for emotional stress in university students across gender and language groups. We design a common questionnaire of stress-inducing and non-stress-inducing questions in Chinese and English, and interviewed 25 native speakers of Mandarin and 31 native speakers of English, of both gender. We extract 560 acoustic features including as low-level descriptors and Teager energy operator (TEO). Our acoustic feature-based classifier recognizes stress in the subjects' speech with 81.28% accuracy on average within the same gender and language group, largely outperforming human perception tests which showed only 39.27%. Moreover, we show for the first time that whereas the emotion detection accuracy decreases by 28.18% across gender, our system maintains the same performance across Mandarin and English. Feature ranking experiments show that the most important stress features are TEO and MFCCs, rather than pitch. This explains the relative language-independence of our model, even though Mandarin is a tonal language. TEO features are also founded to be insensitive to gender difference.

Keywords: emotion recognition, stress detection, feature selections

1. INTRODUCTION

In recent years, interest in automatic detection of emotions from speech has grown. There is an increasing demand for emotion recognition systems for call centers, the gaming industry, medical and psychological health care organizations, to name just a few. Since stressed emotion has become one of the major psychological problems among university students, we are interested in exploring automatic methods to detect stress in this context. However, according to previous study on emotional speech database [10], only limited resources of corpora (especially in Chinese Mandarin) are available.

A typical corpus of spontaneous stressed speech was constructed at MIT labs [3], in which the subjects in this database are drivers who were asked to sum up two numbers while driving a car. Another database called SUSAS (Speech Under Simulated and Actual Stress) was constructed at the University of Colorado Boulder [6]. This corpus was collected from military helicopter pilots during a flight. These databases clearly have limited relevance to daily life. Previous work on stressed speech recognition [7, 12] use TEO features whereas emotion recognition system use low-level descriptors. We propose to study both feature sets for our task. In addition, there has not been any cross language study.

The database collection will be presented in detail in section 2. The feature extraction and dimension reduction are in section 3. Experimental setup and the results of our system, both within and cross gender, within and cross language groups, are compared to human perception tests in Section 4. Section 5 concludes the paper.

2. NATURAL STRESS EMOTION DATABASE

In this study, speech was collected from university students during the examination period. The stressed emotion in this corpus is therefore spontaneous and natural. The interviewees were asked not to simulate, hide or exaggerate any emotions. The recording took place in a quiet conference room with high-quality equipments (Creative® Labs, Model No. SB0490). Speech was recorded in a lossless format with a sampling rate of 16,000Hz, using a single channel 16-bit digitization. The entire database constitutes 2:42:39 hours for Mandarin (13 female (1014 segments) and 12 male (728 segments)) and 3:2:28 hours for English (14 female (697 segments) and 17 male (958 segments)).

Each corpus is labeled by 2 annotators with stressed/unstressed label for every answer audio file. The Kappa inter-labeler agreement is 0.9093 for Mandarin database and 0.7562 for English.

2.1. Human subjects

Twenty-five university students (13 female and 12 male) were asked to contribute to the Mandarin database and thirty one university students (14 female and 17 male) for the English database.

All interviewees were native speakers of the corresponding language specific corpus and they were randomly chosen from the Schools of Engineering, Science and Business of The Hong Kong University of Science and Technology, with various academic standing and family backgrounds.

2.2. The questionnaire

The questionnaire consists of twelve questions which are designed in order to record stressed or unstressed emotions from the interviewee. Topics included personal life, academic pressure, career choice, etc. The first five questions were designed to be non stress-related in order to eliminate the nervousness of the interviewees unrelated to the questions. Questions 6 to 12 were expected to induce stressed emotion in some subjects. All answers of the questionnaire were annotated by 2 professional annotators such that each answer has a label of stressed/unstressed emotion. The labelers are native speaker of the corresponding database. They label the original audio file after listen to the whole answer of the interviewee. Based on previous psychological studies on stressed emotion [5], the questions were asked in the increasing order from the least stress inducing to the most. This strategy ensures the gradual change in expression of the emotion from the subjects in order to maximize the differentiation degrees of the corpora.

3. DETECTING STRESS USING ACOUSTIC FEATURES

Each file of average length 35 seconds was divided into 2, 3 or 5 second segments by ‘waveSplit’ [11] for the experiment and the segments’ labels are the same as the original file. Procedures for feature extraction include 1) chunk based feature extraction 2) normalization and 3) dimension reduction. After feature extraction, a SVM classifier (Bioinformatics toolbox version 3.5 in Matlab (R2010a)) with a third order polynomial kernel function was used to distinguish stressed and unstressed emotions from speech segments.

3.1. Acoustic feature extraction

3.1.1. Acoustic Low-level Descriptor Features

A total number of 384 acoustic features were extracted by openSMILE [2]. It was originally used for the Interspeech 2009 Emotion Challenge [9]. The 384-feature set contains 16 low-level descriptors (LLD) and then 12 functionals are applied on a chunk basis [9].

3.1.2. Critical band based TEO autocorrelation envelope features (TEO-CB-Auto-Env)

Speech features such as MFCC are derived from linear speech production model. However, according to Teager’s theory, the true source of sound production is actually the vortex-flow interactions, which are nonlinear. The Teager energy operator (TEO) was shown to be effective in classification of stressed speech in previous study [7, 12]. The Teager energy operator for discrete-time signals [4] is defined as:

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (1)$$

where $x(n)$ is the sampled speech signal.

Table 1: Functionals for TEO-CB-Auto-Env features.

Functionals (11)
max , min ,mean , range
std , variance, kurtosis, skewness
1st ,2nd ,3rd quartile

The critical band based TEO autocorrelation envelop is obtained by first partitioning the entire audible frequency range into 16 critical bands as in [12]. Gabor bandpass filter [8] is implemented with RMS bandwidth corresponding to the critical band for the partition process. Each TEO profile for the 16 critical bands is segmented into 400-sample (25ms) frame with 200-sample overlap. Then the autocorrelation function is calculated and the area under the normalized autocorrelation functions is computed for each frame to get the TEO-CB-Auto-Env frame by frame values. Then 11 functionals are applied on the vectors to get the 176 TEO features. The functionals are listed in Table 1. Therefore we have 560 features which including 384 linear and 176 non-linear acoustic features.

3.2. Feature normalization

Since a speaker independent recognition system is expected, subject based normalization is performed after the feature extraction. The normalization allows us to mitigate the problem of speaker-specific emotional patterns. Particularly,

$$f_{ij} \xrightarrow{Normalization} \frac{2(f_{ij} - f_{mean})}{f_{\max} - f_{\min}} \forall j \qquad (2)$$

where f_{ij} is the ith feature of the jth segment of one subject.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Human perception test

We conduct human perception test to evaluate the performance of our emotion detection system. Ten native English-speakers (5 male, 5 female) were asked to detect the stress/unstressed in the Mandarin speech data set. They have no knowledge of Chinese Mandarin so that the result of the human perception test can be comparable with the machine recognition result since our system only use acoustic features to detect stress emotion. The audio files of the whole answer were used in human perception test. Each audio file was labeled by 2 male and 2 female perceivers with indication of stressed/unstressed emotion. The human perception results are shown in Table 2.

Table 2: Emotion detection accuracy of human perceiver.

Male Data (%)	Female Data (%)
40.07	38.46

Since the interviewees were asked not to simulate or exaggerate any emotion, it is not surprising to see that human performs poorly with no linguistic knowledge. Our system significantly outperforms humans in detecting stress using acoustic features only.

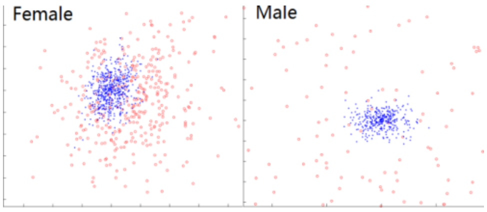
4.2. Segmentation experiment

We use LDA separation rate and PCA plots to select the optimal segmentation length. The separation rate was obtained by calculating the overlap rate of the Kernel smoothing density estimation after dimension reduction by LDA. Five-sec segmentation data set yielded the best result as shown in Table 3 and Figure 1.

Table 3: Separation rate after dimension deduction by LDA from 384 features set to 1 dimension.

384 Features	Female (%)	Male (%)	Mix (%)
5 sec	86.25	88.01	79.12
3 sec	78.10	73.17	71.07
2 sec	75.2	84.31	72.42

Figure 1: PCA scatter plot after LDA for 5-sec-seg dataset showing clear separation between two classes. (Circle: Stressed, Dot: Unstressed)



4.3. Feature comparison and ranking

4.3.1. Feature comparison

As discussed above, in this approach, we used the **5-second** segmentation data set for evaluation. The cross validation method is speaker independent [1] which means that the same speaker does not appear in both training and testing sets. We held out 30% of the interviewees for testing and the rest for training and the results were compared with that of the human perceiver afterwards.

LDA was used to generate one dimension vector for both 384-feature set and TEO feature set. Then we combine the vectors and use SVM to distinguish the emotions. The correct classification result is based on each **5-sec** audio segment as shown in Table 4. Without any feature selection, the averaged detection rate for stressed and unstressed speech is 51.13%.

The linear acoustic feature outperforms the nonlinear TEO feature except for Female English data. The average outperforming rate is 5.03%. The classification result improves 4.72% by adding TEO features to the 384-feature set. Thus we use the combining feature set to evaluate the later experiments.

Table 4: Correct segment recognition rate for different sets of features, **5-sec** segmentation data set.

Correct Rate (%)	384	TEO Only	384+TEO
Man Male	84.62	73.35	86.91
Man Female	77.10	67.07	80.46
Eng Male	75.27	70.07	79.63
Eng Female	69.28	75.66	78.13

4.3.2. Feature ranking

Sequential forward feature selection (SFFS) method was used in finding the **best subset** of features. In order to speed up the selection process, we use LDA to reduce the feature space to one dimension for TEO and each LLDs feature sets, where MFCC 1-12 vectors are reduce to one vector, thus 11 dimensions were obtained. The feature subsets for individual gender and language groups are listed in Table 5. TEO and MFCC features are clearly more important than others. F0 seems to be more important for male emotion speech whereas RMS frame energy is more important for female speech.

Table 5: Feature subsets and classification results for different gender and language groups.

Mandarin Male (%)		**Mandarin Female (%)**	
TEO	73.35	ΔMFCC	67.80
MFCC	78.83	TEO	73.76
ΔMFCC	**81.76**	MFCC	**76.91**
F0	81.59	RMS Energy	76.82
English Male (%)		**English Female (%)**	
TEO	70.07	TEO	75.66
ΔMFCC	75.79	ΔMFCC	79.76
MFCC	77.93	MFCC	82.70
F0	**78.16**	ΔRMS Energy	**82.79**
ΔZCR	75.86	ΔF0	80.64

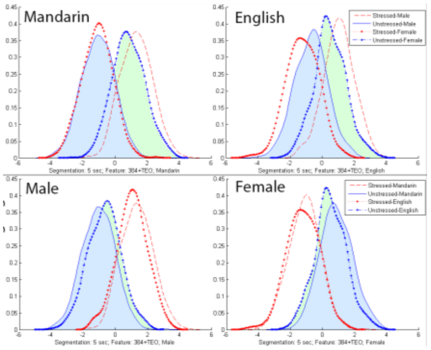
4.4. Cross gender test

We use the same TEO plus 384 features training models and same TEO training models from speaker of one gender to detect stress in the other gender. The test data sets were also exactly the same which were used in the individual gender classifier. Table 6 shows a huge performance difference across gender for both English and Mandarin database whereas for TEO feature only, there is no difference across gender. For stand along 384 feature set, the performance is similar with the TEO plus 384 feature set.

4.5. Cross language test

Similarly, we use the training models from one language to detect stress in the other language. Interestingly, the cross language prediction result is consistent by gender which may imply that the gender difference is more important than language difference for stressed emotion. Figure 2 also shows that the distributions of the density estimation are different across gender by similar across language.

Figure 2: All features are sensitive to gender but not to language difference. (Unshaded: Stressed, Shaded: Unstressed)



But for TEO feature only, the distribution of the density estimation across gender and language are almost the same (Figure 3). Consequently, the cross gender and cross language classification by TEO features are similar, which average is 73.23% (Table 6).

Table 6: Results for cross gender/language classification.

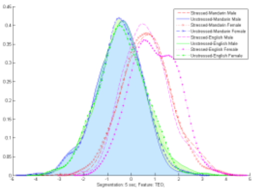
TEO Plus 384 Feature Set (%)				
Train/Test	Man M	Man F	Eng M	Eng F
Man M	86.91	59.48	74.74	73.45
Man F	44.17	80.46	51.89	81.30
Eng M	22.53	42.74	79.63	54.91
Eng F	40.78	79.55	53.84	78.13
TEO Feature Set (%)				
Train/Test	Man M	Man F	Eng M	Eng F
Man M	73.35	71.23	71.04	76.64
Man F	74.71	67.07	71.03	76.03
Eng M	75.31	72.00	70.07	76.88
Eng F	71.45	70.81	70.64	75.66

5. CONCLUSIONS

We present a thorough study of the acoustic features for stress recognition from interview speech across gender and language groups. A corpus has been specifically designed to investigate the stressed emotion among university students. The database currently contains material from a total of 56 university students, with stressed emotions and comparatively neutral emotions for each, of both genders, giving a total of 5:45:7 hours speech. Furthermore, the recording of the corpus and rating procedure have been kept constant or as close as possible over all recordings.

We have proposed a system that is able to distinguish stressed emotion based **only** on acoustic features with up to 86% accuracy. Changes in acoustic features of speech signal have shown to be a reliable indicator of stressed emotion of a person. Our system outperforms human significantly by 44.42%. The classification accuracy improves for gender dependent system. The language difference for stressed emotions is comparably small. Feature ranking experiments show that the most important stress features are TEO and MFCCs, rather than pitch. This explains the relative language-independence of our model, even though Mandarin is a tonal language.

Figure 3: TEO feature is insensitive to gender and language difference. (Unshaded: Stressed; Shaded: Unstressed)



6. REFERENCES

[1] Brendel, M., Zaccarelli, R., Devillers, L. 2010. Building a system for emotion detection from speech to control an affective Avatar. *Proc. LREC 2010* ELRA, Valetta, Malta.

[2] Eyben, F., et.al. 2010. OpenSMILE - the Munich versatile and fast open-source audio feature extractor. *Proc. ACM Multimedia (MM), ACM* Florenze, Italy, 1459-1462.

[3] Fernandez, R., et.al. 2002. Modeling drivers' speech under stress. *Proc. ISCA Workshop (ITRW) on Speech and Emotion: A Conceptual Framework for Research* Belfast.

[4] Kaiser, J.F. 1990. On a simple algorithm to calculate the energy of a signal. *Proc. Int. Conf. Acoustic, Speech, Signal Processing '90*, 381-384.

[5] Lazarus, R.S. 1999. *Stress and Emotion: A New Synthesis*. New York: Springer Pub.

[6] Linguistic Data Consortium. *http://www.ldc.upenn.edu/*

[7] Low, L.A., et.al. 2010. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. *ICASSP* Dallas, Texas, USA.

[8] Maragos, P., Kaiser, J.F., Quatieri, T.F. 1993. Energy separate-ion in signal modulations with application to speech analysis. *IEEE Transactions, Signal Processing* 41, 3024-3051.

[9] Schuller, et.al. 2009. The interspeech 2009 emotion challenge. *Interspeech 2009 ISCA* Brighton.

[10] Ververidis, D., Kotropoulos, C. 2003. A state of the art review on emotional speech databases. *Proc. 1st Richmedia Conference* Laussane, Oktober, 109-119.

[11] waveSplit. *http://wavesplit.sourceforge.net/*

[12] Zhou, G., Hansen, J.H.L., Kaiser, J.F. 2001. Nonlinear feature based classification of speech under stress. *IEEE Transactions, Speech and Audio Processing* 9, 201-216.