A Multilingual Database of Natural Stress Emotion

Xin Zuo, Tian Li, and Pascale Fung

Human Language Technology Center
HKUST, Hong Kong
xinzuo@ust.hk, tli@ust.hk, pascale@ee.ust.hk

Abstract

In this paper, we describe an ongoing effort in collecting and annotating a multilingual speech database of natural stress emotion from university students. The goal is to detect natural stress emotions and study the stress expression differences in different languages, which may help psychologists in the future. We designed a common questionnaire of stress-inducing and non-stress-inducing questions in English, Mandarin and Cantonese and collected a first ever, multilingual corpus of natural stress emotion. All of the students are native speakers of the corresponding language. We asked native language speakers to annotate recordings according to the participants' self-label states and obtained a very good kappa inter labeler agreement. We carried out human perception tests where listeners who do not understand Chinese were asked to detect stress emotion from the Mandarin Chinese database. Compared to the annotation labels, these human perceived emotions are of low accuracy, which shows a great necessity for natural stress detection research.

Keywords: database, stress emotion, multilingual

1. Introduction

Emotion is an important factor in human communication. We have different ways to express our feelings, such as the tone or energy of speech, emotional keywords in writing or speech, and facial expressions and gestures. Speech is one of the fundamental means of expressing of emotions. Humans are especially capable of expressing their feelings using speech because it is audible to others. In recent years, interest in automatic detection of emotions from speech has grown. There is an increasing demand for emotion identification systems for call centers, the gaming industry, and medical and psychological health care organizations, to name just a few.

Our objective is to design and collect speech for a multilingual natural stress emotion database. Most speech emotion databases contain simulated emotion by professional actors or others faking emotions on demand, which is not suitable for real-life applications. Other databases contain natural speech with universal emotions, such as happiness, anger, sadness, surprise, disgust and fear, but not the emotion of stress (Darwin, 1956). Yet other stress speech is collected under extreme conditions, such as in military settings. There is a need for a stress emotion speech database with natural emotion for real-life applications that can be used by civilians.

Mandarin, Cantonese and English are the three major languages in Hong Kong which is why we have collected speech for our database in these languages. Unlike English, Mandarin and Cantonese are tonal languages with four tones and nine tones respectively. As an important feature in emotion identification, pitch is syllable-dependent in a tonal language, which raises the question whether tones will affect stress emotion detection or not. A multilingual database will help to investigate whether there are any cross lingual differences, in terms of acoustic features, for stress emotion detection in speech.

1.1. Stress Emotion

According to (Selye, 1975), there are two types of psychological stress, eustress and distress. Eustress is the kind of stress which can enhance function, either physical or mental. In contrast, distress can lead to depression or anxiety and cannot be resolved through coping or adaptation. Distress is an aversive state in which a person is unable to adapt completely to things that have occurred in their life. The situation can be worsened by emotions like anger, anxiety or depression which all can cause stress and interfere with the ability to reason.

In this paper, stress is defined as distress. Stress has become one of the major psychological problems among university students in Hong Kong. Therefore, it is important to study stress emotion to allow us to recognize symptoms before they become serious. We propose to detect stress from speech signals.

1.2. Previous Emotional Speech Databases

Human-machine interface technology has been investigated for several decades. There has been increasing interest in emotion analysis to improve the capabilities of current speech technologies, such as speech synthesis, speech detection and spoken dialog systems. Only limited corpora (especially in Mandarin Chinese) are available. There are 32 emotional speech databases recording the most common emotions. These are summarized in Table 1. Of these only three databases are related to emotional stress (Ververidis and Kotropoulos, 2003).

A summary of 32 emotional speech databases is found in (Ververidis and Kotropoulos, 2003). The sizes of the databases vary from half an hour to roughly 24 hours, and there is only one emotional speech corpus for Chinese.

Emotions	Number of databases
Anger	26
Sadness	22
Happiness	13
Fear	13
Disgust	10
Joy	9
Surprise	6
Boredom	5
Stress	**3**
Contempt	2
Dissatisfaction	2
Shame, pride, worry, startle, elation, despair, humor	1

Table 1: Common emotions recorded in previous databases.

Type of Emotion	Number of Databases
Simulated	21
Natural	**9**
Half recordings natural, Half recordings simulated	2

Table 2: Common emotions in previous work.

Among these 32 databases, there are both simulated and natural emotional speech databases. Simulated emotional speech databases collected data from professional actors, drama students and others faking emotions on demand. The majority of the databases are simulated emotion with only variations at the prosodic level since the linguistic content has been controlled. The following table lists the number of natural and simulated emotional speech occurrences in the reviewed databases (Ververidis and Kotropoulos, 2003; Devillers et al., 2005).

A typical corpus of spontaneous stressed speech was constructed at MIT labs (Fernandez and Picard, 2003). Drivers were asked to add two numbers while driving a car. Another database is Speech Under Simulated and Actual Stress ("SUSAS") [1]. This corpus was half natural, half simulated. However, neither corpus contains natural or continuous speech utterances.

1.3. Stress Emotion Detection

Conversation content is a direct way to detect emotional states. Words such as happy, great, funny etc. infer a positive emotional state. Therefore, linguistic information as features set has already been studied in emotion detection (Devillers and Vidrascu, 2006). However, in most conversations, people can sense the emotional state even when the speaker is talking about something else. Sound pronunciation system forms a special tonic activation in emotional states, and emotional speech is therefore different from neutral speech and can be perceived by human beings. The dominant research interest for emotion detection in the last decade has been to discover the acoustic information that conveys emotional states in order to allow computers to understand speakers as human beings do (Devillers and Vidrascu, 2006). Researchers have also devoted their efforts to combining linguistic features with acoustic features (Polzehl et al., 2009).

The definition of emotion used to be a sentimental criterion. Emotion detection makes this criterion computable, and it's possible to discover micro emotions which are not sensitive enough for human beings to detect. People often don't express their feelings of stress as freely as other emotions in daily life. Research on stress emotion detection can help to discover people with slight stress.

2. A Natural Stress Emotion Database

Stress in students is an important issue, which greatly affects their studies and research. Most students are self-controlled and not willing to show stress until the situation becomes serious. Research on a natural stress emotion detection corpus can help psychologists to find stressed students at an early stage.

Students have stress in several aspects, such as social relationship, study, campus life, especially in examination periods. In this study, we have been collecting speech from university students during their examination periods. The stress emotions in this corpus are therefore spontaneous and natural. We held an interview and asked each interviewee some carefully designed questions. The interviewees were told this was a survey on campus life, so that they wouldn't act the stress emotion. They were also asked not to simulate, hide or exaggerate any emotions. The recordings took place in a quiet conference room with high-quality equipment (Creativer Labs, Model No. SB0490). A student interviewer chatted with them first, the purpose being to eliminate the nervousness of the interviewees in relation to being interviewed. Speech was recorded in a lossless format with a sampling rate of 16,000Hz, using a single channel, 16-bit digitization.

Thus far, 61 university students have been asked to contribute to the Mandarin database, 42 university students to the English database and 69 university students to the Cantonese database. All interviewees are native speakers of the corresponding language from the Schools of Engineering, Science and Business of the Hong Kong University of Science and Technology with various levels of academic standing and family backgrounds. In consideration of gender influence on speech emotion classification (Fu et al., 2010), we divided the database by genders. Table 3 lists the details of this natural stress emotion corpus. Each part is labeled by two annotators with a stressed or unstressed label for every reference audio file.

[1] Linguistic Data Consortium, http://www.ldc.upenn.edu/.

Database	Female	Male	Time(hh:mm:ss)
Mandarin	27	34	9:46:04
English	15	27	4:40:35
Cantonese	16	53	12:27:01

Table 3: Detail of the natural stress emotion database.

2.1. The Questionnaire

The questionnaire consists of twelve questions which are designed to record stressed or unstressed emotions from the interviewee. Since the interviewees are university students, topics included personal life, academic pressure, career choice, etc.

The questions were designed in two categories. The first five questions were designed not to induce stress and to put the interviewees at ease. These unstressful questions help the annotators to distinguish a person's neutral status and stress status. Questions 6 to 12 were designed with the expectation that they would induce stressed emotion in some subjects. Since different students will have different problems leading to stress, the wide topic range explores the potential sources of stress in their lives.

Based on previous psychological studies on stressed emotion (Lazarus, 2006), the questions were asked in increasing order from the least stress inducing to the most. This strategy ensures a gradual change in expression of emotion from the subjects, in order to maximize the differentiation degrees of the corpora.

All answers to the questionnaire were annotated by two annotators to get the Kappa inter-labeler agreement [2]. The Kappa inter-labeler agreement measures the agreement between two raters. We exclude any answer that has a Kappa inter-labeler agreement lower than 0.6. It is defined as:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \quad (1)$$

where $Pr(a)$ is the relative observed agreement among raters and $Pr(e)$ is the probability of random agreement.

There are two native speaking annotators for each language database, and each answer has a label of stressed or unstressed emotion. The annotators are from our research lab and did not participate in the recordings. The speakers were required to give a label of stressed or unstressed to each answer themselves. Some students thought the speech to be "a little stressed" and were unable to define the stress level. They might pick a stressed or unstressed label randomly. Labels with great uncertainty were marked in the self-labeling process. Our annotators listened to all of the original audio recordings independently and gained an overall understanding of the stress level according to labels from students. Then the annotators listened to the recordings again and gave each answer a modified label of stressed or unstressed emotion , in order to get the Kappa inter-labeler agreement. The final labels were decided after discussion between the two annotators.

[2] Cohen's_kappa, http://en.wikipedia.org/wiki/Cohen's_kappa.

		Number of labels Annotator A	
		Stressed	Unstressed
Annotator B	Stressed	213	19
	Unstressed	41	231

Table 4: The annotation statistics of the two annotators for the English database.

There are 42 students for the English database and each has 12 recordings, which means 504 recordings in total. Note that there were 213 audio files that were labeled as stressed speech by both annotators, and 231 were labeled as unstressed speech by both readers. Thus, the observed percentage agreement is $Pr(a) = (213+231)/504 = 0.8810$.

To calculate $Pr(e)$ (the probability of random agreement) we note that:

- Annotator A labeled "stressed" to 254 files and "unstressed" to 250 files. Thus annotator A said "stressed" 50.40% of the time;
- Annotator B labeled "stressed" to 232 files and "unstressed" to 272 files. Thus annotator B said "stressed" 46.03% of the time.

Therefore the probability that both of them would say "stressed" randomly is $0.5040 \times 0.4603 = 0.2320$ and the probability that both of them would say "unstressed" is $0.4865 \times 0.5315 = 0.2677$. Thus, the overall probability of random agreement is $Pr(e) = 0.2320 + 0.2677 = 0.4997$.

So now applying equation 1 we get:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} = \frac{0.8810 - 0.4997}{1 - 0.4997} = 0.7621.$$

The Kappa inter-labeler agreement is 0.9021 for the Mandarin database, 0.7621 for English and 0.8545 for Cantonese. Kappa inter-labeler agreement between $0.6 \sim 0.8$ means good agreement and between $0.8 \sim 1.0$ means very good agreement. Although the English annotators and English student interviewees are native English speakers, they are not from the same country. Cultural differences may affect understanding of stressed and unstressed emotion, and accordingly the English database Kappa value is a bit worse than the other two . The overall Kappa value shows very good agreement.

2.2. The Transcription

Linguistic features are expensive to obtain compared with acoustic features since they must be obtained from accurate ASR transcriptions or manual transcriptions. In our research, we are interested in comparing the relative

	female speaker	male speaker	average detecting accuracy
female listener	49.36%	55.71%	52.54%
male listener	46.79%	49.31%	48.05%
average accuracy being perceived	48.08%	52.51%	50.29%

Table 5: Human perception results.

contribution of acoustic features versus linguistic features for stress emotion identification. Therefore we need to transcribe the speech database. So far, we have been carrying out the transcription of the English database using two native speakers. Each document of the transcription contains the entire speech from each answer. Sixteen speakers' speeches (eight female and eight male) have been transcribed. (There are some sample transcriptions in Table 6.)

Label	**Example**
	English
unstress	When I'm free I cook a lot, like **I do like** the food I cook and - - and **I do like** create some new food that others did not cook. Um I like because my mum likes the stuff I cooked and my dad like it as well. I don't like it that much but they like it.
stress	The work **didn't go very well** recently because um I've got a lot to do but for some of the tasks I **have no idea** where to start. Um, I think I kind of wrote a paper last month but it's not like a serious paper it's like a student paper contest. Um I think I didn't have enough time to do that so I'm just kind of finish it up and like not really pay a lot of attention to that.
	Mandarin
unstress	香港大学生的生活我觉得**太美好**了.平时课也不多,都可以做自己想做的事情.然后上的课都是自己挺**喜欢**的.
stress	就业,现在,还没开始找工作呢.然后,觉得,**不怎么好找**.因为自己没有什么特长.学业,这个学期学分还多,学起来**超累**.
	Cantonese
unstress	人际关系啊,都**好好的**.因为嗯我自己同一些partner这些啊之前UG一起都玩的**好嗨**这样啊. 咁大家好多时间一起煮饭啊,一起聊天啊,咁搞到好夜才返校才开始温书这样啊
stress	因为现在还没搵到.我自己又**不清楚**我自己的人生目标是乜嘢,我真是想做乜嘢都**不清楚**,既然不清楚的话,对我来讲比较难设一个目标.

Table 6: Sample transcriptions from English, Mandarin and Cantonese database.

For male transcriptions, there are 1066 identical words in total, and 987 for female transcriptions. There are 39 POS tags for both male and female transcriptions.

3. Human Perception Test

Recognizing emotions in speech requires classifying voice quality and prosodic information in speech. Previous research on stressed speech detection has shown that acoustic features help to detect stress, and lots of work has been conducted on pitch or spectral features (Zhou et al., 2001; He et al., 2009).

We conducted a human perception test to evaluate the performance of our emotion detection system without linguistic features. Ten non-Chinese-speakers (5 male, 5 female) were asked to detect the stress/unstressed emotion in the Mandarin speech data set. Each answer was labeled by 2 male and 2 female perceivers，which means 4 times in total. They did not use any linguistic knowledge to judge whether the speaker was stressed or not since they have no understanding of Mandarin Chinese. Therefore, the results of the human perception test can be used to compare to the machine detection results based only on acoustic features to detect stress emotion.

The perception labels were compared with given labels, and the average accuracies of the human perception results are shown in Table 5. It is interesting to note that female listeners are better at detecting stress emotion (52.54% vs. 48.05%), while female speakers are more easily mistaken in stress emotion detection (48.08% vs. 52.51%). Since the interviewees were asked not to simulate or exaggerate any emotion, it is not surprising to see that humans perform poorly with no linguistic knowledge and that the accuracy is almost balanced between stressed and unstressed emotion (50.29%).

4. Conclusion and Future Work

Stress detection applications in counseling, psychotherapy, driver safety and call center application are on demand in modern life, while such research requires a natural stress database. In multilingual areas, such as Hong Kong, stress detection is applicable if stress is language independent, and research on this aspect requires a multilingual stress database. Currently, there are not yet any suitable databases.

We have collected the first ever multilingual corpus of natural stress emotion in English, Mandarin and Cantonese. Through a common questionnaire of stress-inducing and non-stress-inducing questions in Chinese and English, we

interviewed 61 Mandarin, 42 English and 69 Cantonese native speakers. We collected 27 hours of speech, recording stressed emotions and comparatively neutral emotions for each of both genders.

In future, we will continue adding more speakers to the database, as well as more languages, such as French. It is also possible to extend the modality of the database to include video images, for example.

5. References

C. Darwin. 1956. The expression of the emotions in man and animals. *The Journal of Nervous and Mental Disease*, 123(1):90.

L. Devillers and L. Vidrascu. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In *Ninth International Conference on Spoken Language Processing*.

L. Devillers, L. Vidrascu, and L. Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

R. Fernandez and R.W. Picard. 2003. Modeling drivers' speech under stress. *Speech Communication*, 40(1-2):145–159. ISBN 0167-6393.

L. Fu, C. Wang, and Y. Zhang. 2010. A study on influence of gender on speech emotion classification. *Proceedings of the second IEEE International Conference on Signal Processing Systems*, 1:534–537.

L. He, M. Lech, N. Maddage, and N. Allen. 2009. Stress and emotion recognition using log-gabor filter analysis of speech spectrograms. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference*, pages 1–6. IEEE.

R.S. Lazarus. 2006. *Stress and emotion: A new synthesis*. Springer Publishing Company, New York. ISBN 0826102611.

T. Polzehl, S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze. 2009. Emotion classification in children's speech using fusion of acoustic and linguistic features. In *Tenth Annual Conference of the International Speech Communication Association*.

H. Selye. 1975. Confusion and controversy in the stress field. *Journal of Human Stress*, 1(2):37–44.

D. Ververidis and C. Kotropoulos. 2003. A state of the art review on emotional speech databases. *Proceedings of the first Richmedia Conference*, pages 109–119.

G. Zhou, J.H.L. Hansen, and J.F. Kaiser. 2001. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3):201–216.