Defining the dominance axis of the 3-D emotional model for expressive human audio emotion

Simon Lui*

*The Information Systems Technology and Design Pillar,

Singapore University of Technology and Design,

Singapore

simon_lui@sutd.edu.sg

Abstract

The two-dimensional emotional model by Thayer is widely used for emotional classification. It identifies emotion by arousal and valence. However, the model is not fine enough to classify among the rich vocabularies of emotions. Another problem of the traditional methods is that they don't have a formal definition of the axis value of the emotional model. They either assign the axis value manually or rate them by listening test. We propose to use the PAD (Pleasure, Arousal, Dominance) emotional state model to describe speech emotion in a continuous 3-dimensional scale. We suggest an initial definition of the continuous axis values by observing into the pattern of Log Frequency Power Coefficients (LFPC) fluctuation. We verify the result using a database of German emotional speech. The model clearly separates the average value of the 7 emotions apart (the neutral and big-6 emotion). Experiments show that the classification result of a set of big-6 emotions on average is 81%. We will further refine the definition of the axis formula, in order to reduce the overlapping between different emotions. Our ultimate goal is to find a small set of atomic and orthogonal features that can be used to define emotion in a continuous scale model. This work is the first step to approach this final goal.

Keywords: 3-D emotional model, human speech emotion, speech dominance

1. Introduction

Understanding speech emotion is very important to explore into the world of information retrieval. Audio emotion research is useful for many different applications. For example, to understand the emotion of the speaker on the other side on a phone, to review and improve singers' performance technique by visualizing their expressive performance, or perform semantic music search according to information directly extracted from the audio file, etc. Speech emotion research has been a hot topic in recent years. It is now an age of information explosion, yet most information is presented in the text format. For example, the online search engine seeks for text information, social network platform shares information in text format (or picture and video with text tag), online music store presents music selection according to the text index information. On the other hand, a lot of high-level information is embedded in the text such as the thankfulness in a speech or the anger in a conversation, which are usually related to emotions. Traditionally, the arousal-valence emotional model by Thayer [1] is widely used for expressing emotion. One problem of the Thayer's model is that only using two axes is not enough to classify among different emotions. For example, it is obvious that happy is more positive than angry. However it is hard to compare the valence between disgusting and fear. Also, when comparing very angry with angry, it is hard to define whether very refers to higher energy or enhanced valence. Mehrabian proposed the PAD emotional state model [2]. The PAD model has one additional dominance axis on top of Thayer's model. We

assume that the pleasure dimension in the PAD model is equivalent to Thayer's valence. However, very little work has been done on speech emotion with the PAD model. In this work, we investigate on how to define speech emotion by using the 3-dimensional PAD model. The model should have objective and measurable axis value so that it doesn't require manual input.

2. Implementation

2.1. Data Source

The database of German emotional speech is used for emotion classification [3]. This database consists of speech clips with 7 different emotions, including neutral and the big-6 emotion set (angry, joy, fear, disgust, bored, sad). 10 professional actors record the clips. There are 800 clips. Each lasts for 1-6 seconds.

2.2. The energy axis

The formula of the energy axis is well proven in many other previous works and it is easy to measure. Nwe used energy to classify two groups of emotions [4]. In his work, group 1 consists of anger, surprise, and joy, which refer to high-energy sound clips; group 2 consists of fear, disgust, and sadness, which refers to low-energy sound clips. He achieved accuracy ranging from 70% to 100%. In our work, we use a well-agreed formula of energy - the summation of square of root mean square (RMS) amplitude. We measured the average energy relative to the maximum of 7 different emotions. The result is as shown in Table 1. It is very clear that the majority of joy and angry emotional clips belong to the high-energy class; neutral, fear and disgust emotional clips belong to medium-energy class; sad and bored emotional clips belong to low-energy class.

Table 1: Relative energy of 7 emotions.

Emotion	Relative Energy
Angry	43.85%
Joy	33.47%
Neutral	16.69%
Fear	13.12%
Disgust	10.99%
Sad	7.98%
Bored	6.15%

2.3. The valence axis

The valence axis should consist of discrete values. In this work, we worked on negative, neutral and positive valence. The reasons are as follow. First, the energy axis alone should be enough to tell the difference between very angry and angry. These two emotions shouldn't have difference in terms of valence and dominance. Second, the difference between angry and fear should be described by the dominance axis, which they have no difference in valence and can have no difference in energy. However, the valence axis cannot be removed. For example, angry, excited and joy are all having high energy and high dominance, which they have negative, neutral and positive valence respectively. Similarly, sad, sleepy and satisfied are all having low energy and low dominance, and they have negative, neutral and positive valence respectively. In this work, we use the LFPC shape to classify among three classes of positive, neutral and negative valence.

2.4. The dominance axis

We performed a test to proof the existence of the dominance axis. We divided the emotional audio data clips into frames of 32ms and calculate the 12-bins LFPC accordingly. We calculate the normalized LFPC as follow:

$$LFPC\_norm(n,k) = \frac{LFPC(n,k)}{\frac{1}{N}\sum_{i=1}^{N} LFPC(i,k) \times \frac{1}{N}(\sum S^2)} \quad (1)$$

where LFPC(n,k) is the LFPC of the nth frame and the k^{th} bin. S is the RMS Amplitude. Table 2 shows the relative standard deviation (RSD) of the normalized 2nd – 6th LFPC. The RSD refers to the standard deviation divided by mean. It shows that the RSD is high for lower dominance emotion. We also measured the RSD of LFPC in another way. We further normalized the LFPC to be bounded by 0 and 1. The formula is as follow:

$$LFPC\_bounded(n,k) = \frac{LFPC(n,k)}{\max(LFPC)} \quad (2)$$

Table 3 shows the RSD of the normalized 2nd – 6th LFPC bounded within the range of 0 to 1. It presents a similar trend as in Table 2. From the observations of the two experiments above, we suggest that the dominance axis can be described by the RSD of normalized

LFPC. A small RSD refers to very firm and aggressive emotion, while a large RSD refers to high level of hesitation and hence defensive emotion. Table 4 shows some examples of different dominance.

Table 2: The RSD of the normalized 2nd – 6th LFPC.

LFPC	2	3	4	5	6
Fear	23.20%	32.31%	31.27%	29.62%	22.43%
Disgust	21.33%	27.44%	31.06%	25.21%	19.17%
Sad	12.34%	14.48%	20.32%	21.99%	16.42%
Bored	11.13%	12.32%	17.43%	18.54%	15.76%
Neutral	13.64%	13.42%	13.37%	13.32%	13.78%
Joy	11.14%	10.96%	12.11%	17.78%	13.01%
Angry	10.52%	8.42%	13.42%	15.63%	12.24%

Table 3: The RSD of the normalized 2nd – 6th LFPC bounded within the range of 0 to 1.

LFPC	2	3	4	5	6
Fear	28.20%	33.12%	32.12%	28.54%	26.32%
Disgust	27.33%	25.51%	30.23%	24.35%	21.28%
Sad	15.12%	15.83%	21.38%	22.54%	17.94%
Bored	14.43%	13.29%	18.76%	15.20%	16.74%
Neutral	12.54%	12.95%	16.89%	13.19%	14.56%
Joy	11.11%	11.65%	15.22%	14.83%	14.49%
Angry	10.87%	9.33%	14.23%	12.47%	12.03%

Table 4: Example of emotions with different dominance.

Emotion	Valence	Dominance description	Dominance
Angry	Negative	Approaching, present out the feeling	Very aggressive
Jealous	Negative	Approaching, keep the feeling in heart	A bit aggressive
Sad	Negative	No desire	nil
Disgust	Negative	Repelling, keep the feeling in heart	A bit defensive
Fear	Negative	Repelling, present out the feeling	Very defensive

2.5. Orthogonality of the three axes values

We used the Pearson product-moment correlation coefficient (PCC) to evaluate the orthogonality of the three axes. The result is as shown in Table 5. It is found that the three axis-formulas have small positive correlation, but quite independent of each other.

Table 5: covariance of axis value

Axes	PCC
Energy vs Valance	0.1213
Energy vs Dominance	0.1632
Valance vs Dominance	0.2754

3. Experiments and Discussion

Several experiments are performed to present the identification ability of the proposed 3-dimensional emotional model.

In the first experiment, we plot the mean axis values of 7 different emotions we used. The result is as shown in Figure 1. The 7 emotions are clearly apart from each other, although sad and bored seems rather closed to each other.

In the second experiment, we demonstrate the emotion identification ability of the dominance axis. We use the product of sequence of the normalized LFPC (LFPC_ps) as shown in Equation 3 to calculate the dominance of each sound clip:

$$LFPC\_ps(p,q) = \frac{1}{N}\sum_{n=1}^{N}\prod_{k=p}^{q} LFPC\_norm(n,k) \quad (3)$$

where LFPC_norm(n,k) is the coefficient of normalized LFPC of the nth frame and the k^{th} bin. LFPC_ps(p,q) is the product of sequence of the normalized LFPC from bin p to bin q. Figure 2 shows the distribution of LFPC_ps(2,6) of 184 negative emotions clips, with 46 clips for each of the angry, fear, disgust and sad emotion. The order of average dominance from descending order is fear, disgust, sad and angry. We observed that there are several cases such that the dominance of sad clip is higher than fear clip. This shows that subjective emotional descriptor can scatter in an objective numerical scale. Also, our axis value formula is not finalized. We expect to narrow down the overlapping area with formula refinement.

In the third experiment, we performed two linear classification tests. The first test runs with a single class SVM. A linear classification is enough for working with linear axis value. We use SVM to ensure that it is computational efficient with over-lapping data. We use a sample size of 46 clips per emotion, for 7 emotions. Each clip has a length of 1-6 second. For each clip, we calculate the average energy, valence and dominance value respectively, which form a 3-dimension vector data. Then we setup SVM machines to train the target emotion data against the other 6 emotions' data. This is done by a 10-fold cross validation

where 90% data are used for training and 10% data are used for testing. Table 6 shows the experiment result. It is found that the average accuracy is 85.81%. We also performed another a linear classification with k-NN, k=7. The result is as shown in Table 7, the average accuracy is 81.5%.

Table 6: 10-fold cross validation SVM linear classification result.

Emotion	Accuracy
Angry	87.95%
Joy	93.47%
Disgust	78.46%
Fear	81.06%
Neutral	89.01%
Sad	81.22%
Bored	89.54%

Table 7: 10-fold cross validation k-NN (k=7) linear classification result. Left: sample class. Top: target class.

	Angry	Joy	Disgust	Fear	Neutral	Sad	Bored
Angry	78.6%	0.5%	6.7%	9.9%	4.3%	0.0%	0.0%
Joy	9.2%	87.7%	0.9%	2.3%	0.0%	0.0%	0.0%
Disgust	6.4%	0.0%	71.4%	14.3%	3.3%	3.5%	1.2%
Fear	8.2%	0.0%	13.7%	69.4%	4.3%	2.1%	2.3%
Neutral	0.0%	0.0%	3.7%	5.2%	78.4%	3.5%	9.3%
Sad	0.3%	0.0%	2.3%	1.7%	1.8%	84.5%	9.4%
Bored	0.0%	0.0%	3.2%	2.1%	5.2%	7.6%	81.9%

As a comparison, Guven [5] used the same German database and performed speech emotion recognition using SFTF as features and classified with SVM. He obtained an average accuracy of 68% in identifying 7 emotions. The bottleneck lies on the classifying disgust (50% accuracy) and bored (60% accuracy). Iliou [6] used MFCC features and obtained an average accuracy of 94% of identifying 7 emotions with neural network, where the speakers are known to the classifier (speaker dependent). In the case of speaker independent, the overall accuracy is 78% by classifying with SVM, with 55% accuracy for bored and 54.5% for disgust. In these two works, disgust and bored are both negative and low energy emotions. The two emotions can be classified by the 3rd axis in our proposed model effectively. Lugger [7] used 25 audio features including pitch, formant, harmonic, and MFCC. He performed classification with an iterative sequential floating forward selection algorithm. He obtained an average ac-curacy of 88.8%. However, he only worked on 6 emotions of the German database. He didn't work on disgust and fear, which is the main bottleneck for emotion classification.

4. Future Work

An initial definition of the three continuous axis of the PAD speech emotional model is proposed. This model separates the average value of the 7 emotions apart with some overlapping among individual samples since the current axis value formula is not ultimately defined. This initial definition is subject to refinement since it is not totally orthogonal. We will further refine the definition of the axis formula, in order to reduce the overlapping between different emotions. This work is the first step to approach this final goal. Our ultimate goal is to find a small set of atomic and orthogonal features that can be used to define emotion in a continuous scale model.

5. Acknowledgment

This work is supported by the SUTD-MIT International Design Center Grant (IDG31200107 / IDD11200105 / IDD61200103).

References

[1] Thayer, R. E. 1989. "The Biopsychology of Mood and Arousal", New York: Oxford Univ. Press.

[2] Mehrabian, A. 1996. "Pleasure-arousal-dominance: A gen-eral framework for describing and measuring individual dif-ferences in temperament". Current Psychology. Springer.

[3] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B. 2005. "A Database of German Emotional Speech", Proc. Interspeech.

[4] Nwe, T. L., Foo, S. W., Silva, L. C. D. 2003. "Speech emo-tion recognition using hidden Markov models", Speech Communication 41(4): 603-623.

[5] Guven, E. 2010. "Speech Emotion Recognition using a Backward Context". Applied Imagery Pattern Recognition Workshop. pp1-5.

[6] Iliou, T., Anagnostopoulos, C. 2010. "Classification on Speech Emotion Recognition - A Comparative Study", In-ternational Journal On Advances in Life Sciences, volume 2, pp 18-28.

[7] Lugger, M., Yang, B. 2008. "Cascaded emotion classifica-tion via psychological emotion dimensions using a large set of voice quality parameters", Proceedings of the IEEE ICASSP, pp. 4945–4948.

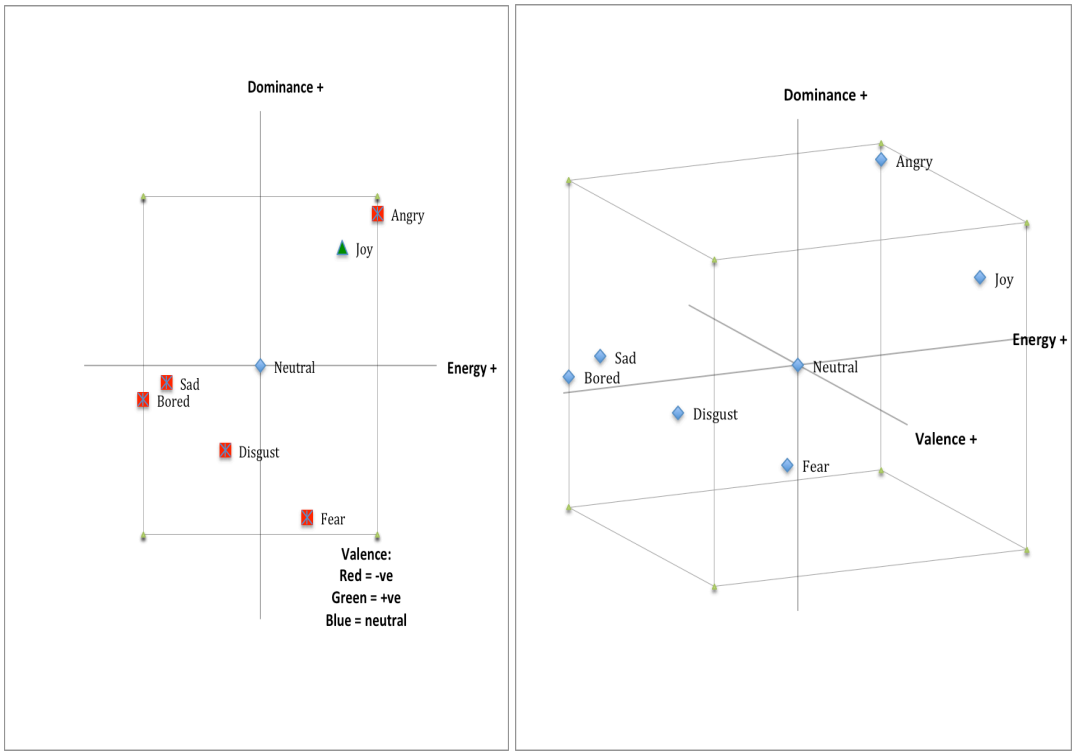

Figure 1: Plot of the mean values of 7 emotions.

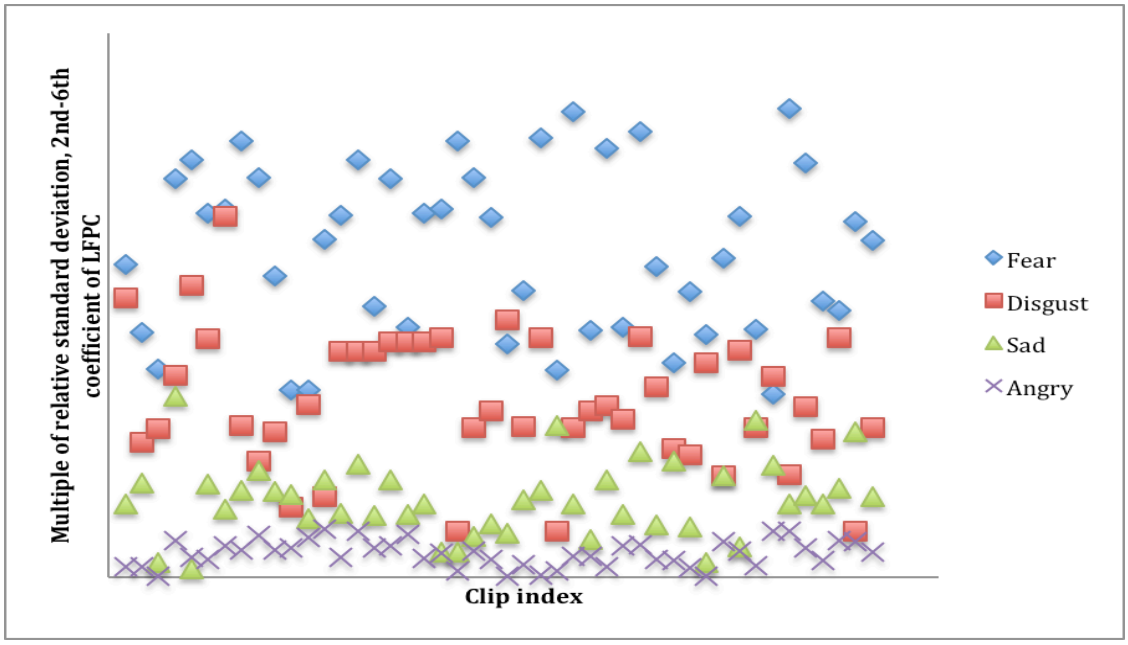

Figure 2: Distribution of product of sequence of the normalized LFPC, 2nd – 6th bin, for 184 clips of four different negative valence emotions.