A PRELIMINARY ANALYSIS OF THE CONTINUOUS AXIS VALUE OF THE THREE-DIMENSIONAL PAD SPEECH EMOTIONAL STATE MODEL

Simon Lui

The Information Systems Technology and Design Pillar,
Singapore University of Technology and Design,
Singapore
simon_lui@sutd.edu.sg

ABSTRACT

The traditional way of emotional classification involves using the two-dimensional (2D) emotional model by Thayer, which identifies emotion by arousal and valence. The 2D model is not fine enough to classify among the rich vocabularies of emotions, such as distinguish between *disgusting* and *fear*. 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. Experiments show that the classification result of a set of big-6 emotions on average is 81%.

1. INTRODUCTION

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. Hence understanding 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.

Traditionally, the arousal-valence emotional model by Thayer [1] is widely used for expressing emotion. Lu used the Thayer's model to detect the mood of music [2]. There are two axes in the model, the valence axis and the arousal (energy) axis. The axis values are usually assigned manually. For example, Wöllmer assigned the value by performing a listening test [3]. Cowie described speech emotions by the activeness axis and the positivity axis with manual input data [4]. However, database that requires manual input will limit the number of training samples available, which will greatly affect the accuracy of data training.

To classify among different emotions, the traditional way is to define each emotion class with feature parameters. For example, Dimitrios used pitch, Teager energy operator, vocal tract features, formant, speech energy, MFCC, intensity, and speech rate to describe emotion [5]. Cowie used pitch, energy, spectral information such as MFCC and LFPC, zero-crossing rate, formant, and spectral of vocal tract feature to perform emotion identification [6]. Luengo used a small set of features including fundamental frequency, intensity slope, and duration, and obtain a good classification result [7]. Schuller used a multi-level SVM with 33 features to classify among 7 emotions [8]. However, these works are not scalable, since it is not easy to expand the current set of trained emotion. We are interested in a small set of atomic and orthogonal features that can approximate key aspects of emotions as they are communicated vocally.

Another 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 [9]. 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. Zhang has worked on facial expression by using the PAD model [10]. However, very little work has been done on speech emotion with the PAD model. Moreover, the existing works usually use a finite number of values. For example, Mehrabian uses 16 discrete values to define emotion with the PAD model [11]. This limits the number of emotions that can be defined.

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

We use the database of German emotional speech for emotion classification [12]. This database consists of speech clips with 7 different emotions, including *neutral* and the big-6 emotion set (*angry, joy, fear, disgust, bored, sad*). The clips are recorded by 10 professional actors. There are 800 clips. Each clip lasts for 1-6 seconds.

2.2. Language independent components

Ververidis proved that the language dependent emotional component does exist in speech [5]. On the other hand, we believe that there also exist language independent emotional components. We performed a small test to illustrate this. This test was performed with 13 people with various mother tongues including English, Cantonese, Mandarin, Japanese and French. 5 different native speakers (group A) were invited to record expressive speeches of 4 different emotions, including *happy, sad, angry, fear*. They were only allowed to use "EE" and "AH" to pronounce the sentence. As a result, we obtained 20 emotional clips of 1 to 3 seconds. 8 people who all know English but having 4 different native languages (group B) were requested to identify the emotion of the clips produced by group A. The result is as shown in Table 1 with average accuracy of 91.25%. We conclude that some emotion can be expressed without language, and hence there are language independent emotional components in speech. This is the fundamental emotion we aim for. We use a free German library to ensure consistency. We will examine language dependent factor in future work.

Table 1: The language independent factor test result.

Source\Answer	Happy	Sad	Angry	Fear
Happy	90%	0%	5%	0%
Sad	0%	95%	0%	5%
Angry	10%	0%	85%	0%
Fear	0%	5%	10%	95%

2.3. Numerical scale model

Different people might have different feelings on listening to the same piece of speech or music. We performed another simple test to demonstrate the acquired emotional factor of audio information. The test is performed with 34 people, where 17 people are from Hong Kong, and the other 17 people are from US, Japan, Canada, Britain and Taiwan. They are invited to listen to a 10-seconds extract of the Massenet's Meditation from Thais, and rate it either as angry, sad, neutral, calm or delighting. The result is as shown in Table 2. Most Hong Kong people rate the music as sad, while people from the other countries tend to rate it as calm or neutral. In fact, it is a smooth and beautiful piece. The extract is written in D Major key, begin with major I chord, with the majority of melody notes consisting of tonic, median and submediant, which make it sounds positive [13]. This piece of music should present a positive feeling. However, due to the special historical situation in Hong Kong, the Massenet's Meditation from Thais was widely used as the background music for a lot of tragedy TV programs in the 1970s-1990s, many Hong Kong people recall the tragedy picture when listening to this piece of music. This is the acquired emotional factor of audio information.

Table 2: The acquired emotional factor test result.

Emotion	Hong Kong	Other countries
Angry	0%	0%
Sad	76%	18%
Neutral	18%	29%
Calm	6%	53%
Delighting	0%	0%

Since different people might have different self-definition of emotion due to acquired factor, it is much better to use a numerical scale model to describe emotions rather than training manually labelled audio data. The coordinate of a numerical model is objective, while emotional descriptors are subjective. For a certain coordinate on the model, with fixed numerical value, different people can define different emotional descriptors for it.

Also, a particular emotion descriptor can have several sample points scattered on the emotional model. For example, a high energy and aggressive shout such as "Great! I win the game!" can be regarded as *happy*; on the other hand, a low energy and defensive sentence such as "Good, I am really satisfied with today's dinner…" can also be regarded as *happy*. If we train the two clips with the same descriptor *happy*, the resulting classification library will not be accurate. By using a measureable scale, we can locate scattered point of a same emotion accurately.

2.4. The energy axis

Here we propose the measurable parameters of the three axes of the PAD speech emotional model. 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 [14]. 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 3. 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. There are exceptional cases, but on average, sad clips won't have higher energy than angry clips.

Table 3: 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%

2.5. The valence axis

We believe that 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.

Busso classified between emotional sound clips and neutral sound clips. He achieved 77% accuracy by only using pitch features [15], which shows the existence of the valence axis. Many previous researches obtained good results in classifying emotions valence by spectral shape, MFCC, or LFPC. In this project, we use the LFPC shape to classify among three classes of positive, neutral and negative valence. We choose LFPC instead of MFCC and other spectral representation because it preserves low energy component, which is essential for speech emotion research [16].

2.6. The dominance axis

Frijda first suggested the idea of emotional approach and avoidance but without formal definition [17], which is similar to our concept of dominance. 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:

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

where *LFPC(n,k)* is the LFPC of the n^{th} frame and the k^{th} bin.
S is the RMS Amplitude.

Table 4 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:

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

Table 5 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 4. 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 6 shows some examples of different dominance.

Table 4: 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 5: 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 6: 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

2.7. 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 7. It is found that the three axis-formulas have small positive correlation, but quite independent of each other.

Table 7: covariance of axis value

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

2.8. Semantic meaning of the model axis

The proposed 3-dimensional emotional model describes emotions with 3 fundamental components. The energy axis represents the power of speech. The valence axis is described by spectral shape, which represents the tone of speech. The dominance axis is described by change of normalized spectral value, which shows the fluctuation and hesitation of speech.

3. EXPERIMENT AND DISCUSSION

We performed several experiments 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. 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.

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

where *LFPC_norm(n,k)* is the coefficient of normalized LFPC of the n^{th} 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*.

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 overlapping 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 8 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 9, the average accuracy is 81.5%.

As a comparison, Guven [18] 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 [19] 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 [20] 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 accuracy 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.

Table 8: 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 9: 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%

4. FUTURE WORK

We proposed an initial definition of the three continuous axis of the PAD speech emotional model. This model clearly separates the average value of the 7 emotions apart (the neutral and big-6 emotion), but there is 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. 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.

5. ACKNOWLEDGMENTS

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

6. REFERENCES

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

[2] Lu, L., Liu, D., Zhang, H.J. 2006. “Automatic mood detection and tracking of music audio signals”, IEEE Transactions on Audio, Speech & Language Processing 14: 5-18.

[3] Wöllmer, M., Schuller, B., Eyben, F., Rigoll, G. 2010. "Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening", Selected Topics in Signal Processing, IEEE Journal of , vol.4, no.5, pp.867-881.

[4] Cowie, R., Cornelius, R. R. 2003. “Describing the emotional states that are expressed in speech”, Speech Communication, Volume 40, Issues 1-2, pp 5-32.

[5] Ververidis, D., Kotropoulos, C. 2006. “Emotional speech recognition: Resources, features, and methods”, Speech Communication 48(9): 1162-1181.

[6] Cowie, R., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J. 2001. “Emotion Recognition in Human-Computer Interaction”, IEEE Signal Processing Magazine pp. 32-80.

[7] Luengo, I., Navas, E., Hernáez, I. 2010. “Feature Analysis and Evaluation for Automatic Emotion Identification in Speech”, IEEE Transactions on Multimedia 12(6): 490-501.

[8] Schuller, B., Rigoll, G., Lang, M. 2004. "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture", Acoustics, Speech, and Signal Processing,

2004. Proceedings. (ICASSP '04). IEEE International Conference on , vol.1, no., pp. I- 577-80 vol.1, 17-21.

[9] Mehrabian, A. 1996. "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament". Current Psychology. Springer.

[10] Zhang, S. Wu, Meng, H.M., Cai, L. 2007. "Facial expression synthesis using pad emotional parameters for a chinese expressive avatar". Modeling Machine Emotions for Realizing Intelligence, Smart Innovation, Systems and Technologies Vol. 1, 2010, pp 109-132.

[11] Mehrabian, A. 1980. "Basic dimensions for a general psychological theory" ISBN 0-89946-004-6. pp. 39–53.

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

[13] Kemp, I., 1973. "Harmony in Weill: Some Observations", Tempo (New Series), pp 11-15. Cambridge University Press.

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

[15] Busso, C., Lee, S., Narayanan, S. 2009 “Analysis of Emotionally Salient Aspects of Fundamental Frequency for Emotion Detection”, IEEE Transactions on Audio, Speech & Language Processing 17(4): 582-596.

[16] Nwe, T.L., Foo, S.W., De Silva, L.C. 2003. “Detection of Stress and Emotion in Speech Using Traditional and FFT Based Log Energy Features”, In: Proc. Pacific Rim Conference on Multimedia, vol. 3, pp. 1619-1623.

[17] Frijda, N. H., 1986. “The Emotions”. Cambridge, UK: Cambridge Univ. Press.

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

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

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

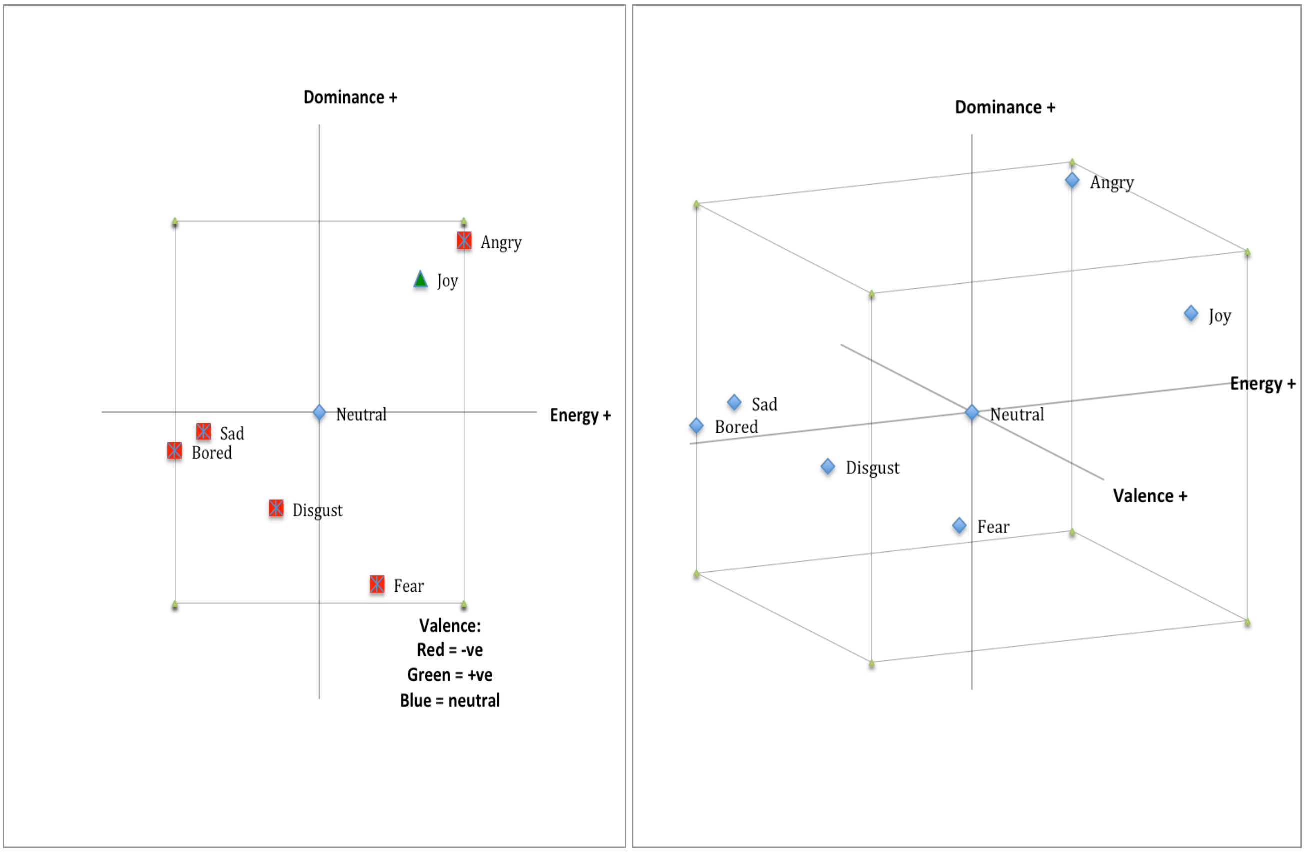

Figure 1: Plot of the mean values of 7 emotions.

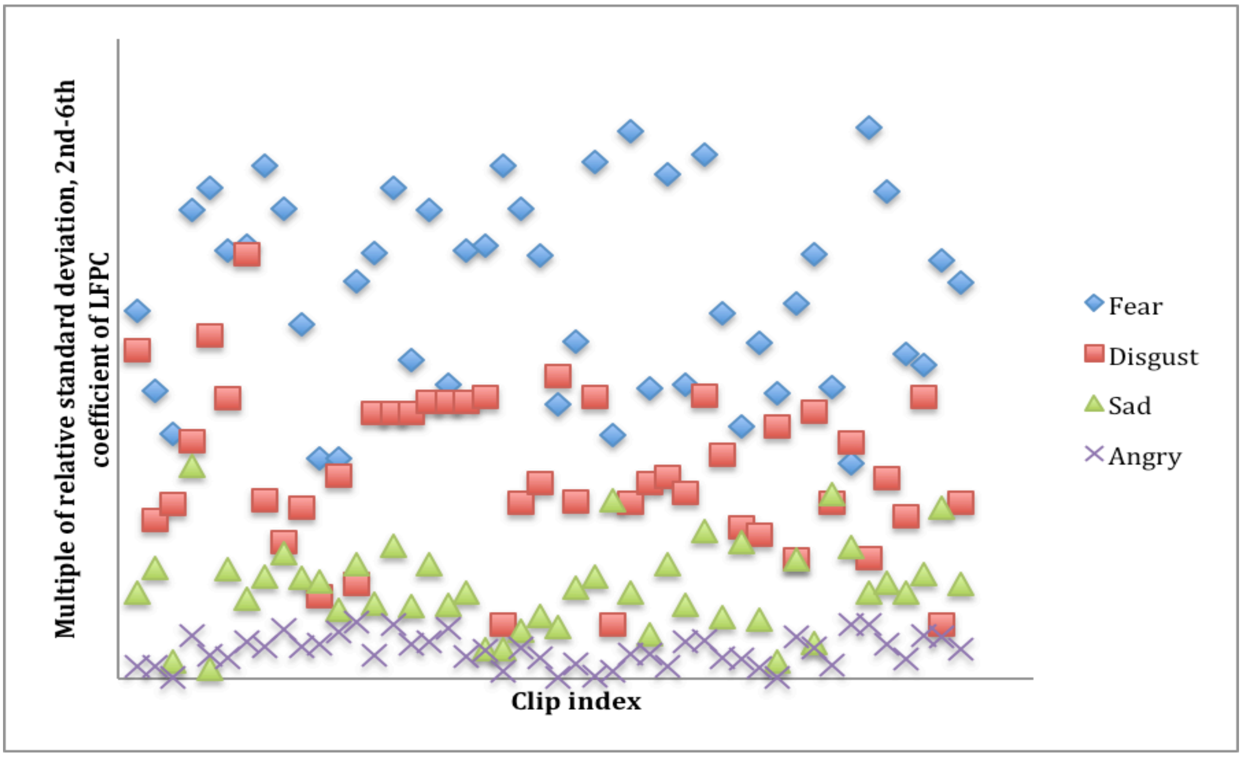

Figure 2: Distribution of product of sequence of the normalized LFPC, 2nd – 6th bin, for 184 clips of four different negative valence emotions.