Evaluating Emotional Algorithms using Psychological Scales

Shangfei Wang
School of Computer Science and Technology
University of Science and Technology of China
Hefei, Anhui, P.R.China 230027
sfwang@ustc.edu.cn

Xufa Wang
School of Computer Science and Technology
University of Science and Technology of China
Hefei, Anhui, P.R.China 230027
xfwang@ustc.edu.cn

ABSTRACT

Evaluating the effectiveness of emotional algorithms is a challenge in the research of affective computing. This paper presents an evaluation method using psychological scales. Affective interfaces and artifacts generated by different emotional algorithms are regarded as the external stimuli, while users' feelings are regarded as subjective evaluations. Psychophysical methods, such as the Scheffe method of paired comparison, are adopted to measure users' feelings and obtain a psychological scale. Thus objects generated by different algorithms can be compared on this psychological scale. The higher the value, the better the algorithm is. Empirical studies on three kinds of emotional algorithms, namely emotional image retrieval, emotional speech synthesis, and emotional music generation, are used to illustrate various uses of our approach and how it may be applied to evaluating affective interaction and user-centered design.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Evaluation/methodology

General Terms

Human Factors, Evaluation

Keywords

evaluation method, emotional algorithm, psychological scales

1. INTRODUCTION

Emotions play a very important role in human-human and human-machine interaction. As a result, affective computing [1] has been the focus of much research in recent years with significant progress having been made. However, since human emotion is subjective and diverse, evaluating the effectiveness of emotional algorithms, which are used to generate different affective interfaces and artifacts, such as speech, music and virtual agents, etc, is still a challenge.

AFFINE '09, November 6, 2009, Boston, MA, USA.

The key to evaluating emotional algorithms is assessing users' emotions while they are communicating with affective interfaces and artifacts. A few studies have been reported in the past five years [2] [3] [4]. Höök [5] proposed a two-tiered design and evaluation model, while Wiberg [6] investigated to what extent traditional evaluation methods, used for evaluating traditional usability, are applicable in understanding individuals' experiences of affective systems. Chateau [7] developed AMUSE, an evaluation tool for affective interfaces that gathers and aggregates many sources of data, including eye gaze and physiological data. Fallman [8] introduced the Repertory Grid Technique for affective evaluation. Picard [9] considered behavior-based measures, including both measures of body movements or physiological signals and task-based performance measures. Regan [10] provided a method using four physiological signals for quantifying emotional states continuously during a play experience. Richard [11] used facial electromyography (EMG) measures combined with verbal and performance measures to provide feedback in the software design process on the user's emotional state. Isomursu [12] experimentally evaluated five self-report methods, including SAM [13], Emocards [14], Experience Clip [15], 3E [16], and Mobile Feedback Application [17], for collecting information about emotional responses to mobile applications. Haringer [18] proposed a scalable user state assessment framework that integrates heterogeneous physiologic sensor configurations. In this paper, we propose a new evaluation method using psychological scales.

The difficulty of evaluating emotional algorithms lies in how to use objective criteria to evaluate subjective feelings [19]. Quantitative research on subjective feelings caused by physical stimulation is one of the focuses in psychophysics [20]. Inspired by this, we present an evaluation method using psychological scales. Affective interfaces and artifacts generated by different emotional algorithms can be regarded as the external physical stimuli, while the internal feelings evoked as users are watching, listening to or using these objects, can be seen as subjective evaluations of the algorithms. By measuring users' feelings, that is, by establishing a psychological scale, the method is able to assess whether emotional algorithms are good or bad. Experiments on emotional image retrieval algorithms, emotional speech synthesis algorithms, and emotional music generation algorithms are used to prove the effectiveness of our approach. *Thick* clothes images and *wide view* landscape images retrieved by our algorithm and a random algorithm, *happy*, *angry*, *fearful*, *surprised*, *disgusted*, and *sad* speeches generated by our

algorithm and MIT's Affect Editor, and *happy* and *sad* music generated by our algorithm and that given in [21] are used in the psychological experiments. The test subjects are required to give evaluations according to their own subjective feelings. Scheffe's method of paired comparisons [22] is adopted to produce psychological interval scales. Thus, quantitative evaluations of a variety of emotional algorithms are obtained.

2. EVALUATION METHODS USING PSYCHOLOGICAL SCALES

Psychophysics investigates the relationships between sensations in the psychological domain and stimuli in the physical domain. Inspired by this, our proposed evaluation method is shown in Fig. 1. Affective artifacts generated by different emotion algorithms, such as emotional speeches, music, avatars, and robots, etc., are regarded as the external physical stimuli. These stimuli evoke the subjects' feelings, which are subjective evaluations of the algorithms. Psychophysical scaling methods are used to collect reactive-level emotional responses to the stimuli, and to analyze the relationships between these responses and the artifacts produced by different algorithms. Though measuring users' feelings, users' subjective evaluations can be quantified on a psychological scale. In other words, objects generated by different emotional algorithms are sorted on the psychological scale. The higher the ranking on the scale, the better is the corresponding algorithm. Thus we can quantitatively evaluate a variety of algorithms.

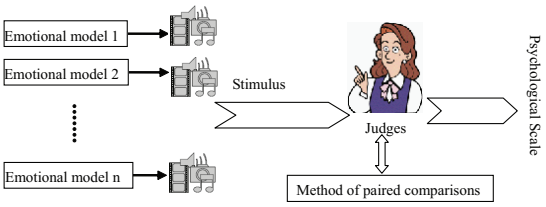


Figure 1: Psychological scale-based evaluation method for emotional algorithms

There are typically four kinds of scales, namely nominal, ordinal, interval, and ratio scales. The nominal scale is qualitative, rather than quantitative, because there is no ordering of the resulting categories. Moreover, since it is difficult to define absolute zero for a user's subjective evaluation, ratio scales are not suitable either in our approach. This leaves the ordinal and interval scales.

There are two main methods for producing ordinal and interval scales: the rank-order method and paired comparison method. In the former method, all items are presented to one or more judges at once. The judge then orders the items according to the specified judgment criterion. In the latter method, items are presented in pairs to one or more judges. For each pair, the judge selects the item that best satisfies the specified judgment criterion. Both these methods can yield an ordinal or interval-scale ordering of items along a dimension such as preference. Although space and time errors may occur in the rank-order method, the paired comparison method can avoid such errors by comparing two items twice using different presentation orders or locations. Therefore, paired comparisons are recommended for our evaluation method.

We adopt Scheffe's paired comparison method [22], since the judges are required not only to select the better item, but also to give a score of the difference between the two items according to the specified judgment criterion. Suppose there are n items, then the number of comparisons is C_n^2. For the pair consisting of item i and item j, the judge k allocates a 5- or 7-step score to the difference between them according to the criterion of how closely the emotion evoked by each item comes to the given emotional requirement. This score is denoted by x_{ijk}. Summarizing the scores of all the judges according to Eq. (1), we obtain the matrix $X = (x_{ij})_{n\times n}$.

$$x_{ij} = \sum_{i=1}^{K} x_{ijk} \tag{1}$$

where K is the number of judges. Then analysis of variance (ANOVA) is used to analyze the matrix to get α_{itemj}, which is regarded as the position of item j on the psychological scale. Thereafter, we may identify the statistically significant pair wise differences. If the distance between two items is larger than the threshold, it means that the difference between the two items is statistically significant at the $\alpha = 0.05$ or 0.01 level. Otherwise, the difference is not statistically significant.

3. EXPERIMENTS

Evaluation experiments on emotional image retrieval algorithms, emotional speech synthesis algorithms, and emotional music generation algorithms are used to prove the feasibility of our proposed evaluation method. 20 judges from our university, aged between 22 and 30, participated in the test. Each was requested to give 5- or 7-step scores to the difference between a pair of items according to the criterion of how close the emotion expressed by the items came to the given emotional requirement.

3.1 Evaluation experiments for emotional image retrieval algorithms

The semantic structure of images comprises spatial relationships, objects, events, behavior, and emotion. Of these, the emotional semantics is the most abstract, and is usually described by adjectives [19] such as *happy*, *romantic*, *brilliant* etc. Emotional image retrieval was first studied in Japan in the 1990's. Since then, many retrieval algorithms have been proposed and several prototype systems have been developed. However, to-date there is still no standard image database, adjectives nor evaluation methods in the field. Here, we evaluate our emotional image retrieval algorithm [23] by comparison with a random algorithm using our image database and adjectives.

3.1.1 Experimental design

Two kinds of emotional images, *thick* clothes images and *wide view* landscape images were retrieved and compared. For each adjective, six images were retrieved. Three of these were retrieved by our algorithm using an Interactive Genetic Algorithm (IGA), and labeled A, C, and E. The other three images were selected randomly, and labeled B, D, and F. The number of pairs of images used for the subjective test was $C_6^2 = 15$. For each pair, the two images were displayed together, and then the judges scored the images according to their satisfaction. For example, for *thick* clothes images, we displayed images A and B simultaneously. If the judge considered the clothes in image A to be much thicker than those in image B, a rating of 2 was given. If he/she considered the clothes in image A to a little thicker than those in image B, a rating of 1 was given. If he/she considered the clothes in image A to be the same thickness as those in B, a 0 rating was given. Otherwise, either -2 or -1 was given as the score. We then collected all the judges' ratings and analyzed them using ANOVA.

3.1.2 Experimental analysis

Table 1 and Fig. 2 show the distances between the six images. The horizontal axis in Fig. 2 represents the constructed psychological scale, with better evaluation depicted to the right of the scale. If the distances are larger than the threshold, it means that the difference between the two images is statistically significant.

For the *thick* clothes images, the distance between E and B on the psychological interval scale is larger than the threshold. This means that A, C, and E, which were retrieved by our algorithm, are statistically significantly better than D, B, and F, which were selected randomly, both at the $\alpha = 0.05$ and $\alpha = 0.01$ levels. From this we conclude that our retrieval algorithm is better than the random algorithm. Furthermore, although A is given a better evaluation than C, and C is given a better evaluation than E, the distances between A and C, A and E, and C and E, are not larger than the threshold. So we cannot conclude that A obtains a better evaluation than either C or E with statistical significance. This may be evidence of the stability of our algorithm. For the *wide view* landscape images, C, A, and E are also statistically significantly better than B, D, and F. However, D is statistically significantly better than F. This may be caused by the instability of the random algorithm. Thus, the experimental results show that our retrieval algorithm is better than the random algorithm.

3.2 Evaluation experiments for emotional speech synthesis algorithms

Generating natural and expressive speeches is the goal of the study of speech synthesis. Cahn at MIT first studied emotional speech synthesis and developed the Affect Editor [24]. Whether the synthesized emotional speeches are understood by users and which emotion synthesizer is better are worthy of study. In this section, we assess two emotional speech synthesis algorithms, our algorithm [25], which uses an IGA to optimize prosody parameters, and the algorithm developed at MIT, the Affect Editor.

3.2.1 Experimental design

Six kinds of emotional speeches were synthesized, namely *happy*, *angry*, *fearful*, *surprised*, *disgusted* and *sad* speeches.

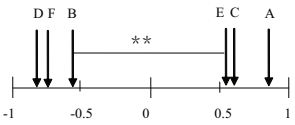
Table 1: Distances between retrieved images

thick clothes images

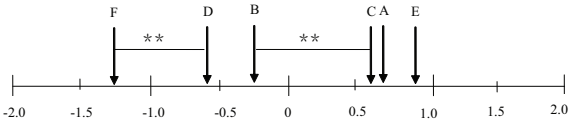
distance	$\alpha = 0.01$		$\alpha = 0.05$	
AC	$0.267 < 0.439$		$0.267 < 0.369$	
CE	$0.063 < 0.439$		$0.063 < 0.369$	
EB	$1.008 > 0.439$	**	$1.008 > 0.369$	**
BF	$0.131 < 0.439$		$0.131 < 0.369$	
FD	$0.045 < 0.439$		$0.045 < 0.369$	

wide view Landscape images

distance	$\alpha = 0.01$		$\alpha = 0.05$	
EA	$0.225 < 0.454$		$0.225 < 0.382$	
AC	$0.088 < 0.454$		$0.088 < 0.382$	
CB	$0.845 > 0.454$	**	$0.845 > 0.382$	**
BD	$0.305 < 0.454$		$0.305 < 0.382$	
DF	$0.705 > 0.454$	**	$0.705 > 0.382$	**



(a) *thick* clothes images



(b) *wide view* landscape images

Figure 2: Retrieved images on the psychological scale

For each emotion, four samples were used as objects in the evaluation experiments, including emotional speeches synthesized by our algorithm, neutral speeches, random speeches with the prosody parameters set randomly, and the emotional speeches downloaded from the MIT website for the Affect Editor. These samples were labeled B, D, C, and A, respectively. The number of pairs of speeches used for the subjective test was $C_4^2 = 6$. Each pair of speeches was played in turn, and then the judges scored the samples based on their satisfaction. The procedure was similar to that described in Section 3.1.1. Having collated all the reports, we analyzed them using ANOVA.

3.2.2 Experimental analysis

Table 2 and Fig. 3 show the distances between four samples of the speeches. For all kinds of emotional speeches, except *anger*, both A and B are significantly better than C and D at both the $\alpha = 0.05$ and $\alpha = 0.01$ levels. This means that the speeches synthesized by the MIT algorithm and our algorithm are better than those synthesized by the random algorithm and neutral speech. For the *happy* speech, B is significantly better than A, while for *fear*, A is significantly better than B. Furthermore, for the *angry* and *surprised* speeches, A is given a superior evaluation to B, but this is not significant. For the *disgusted* and *sad* speeches, although

Table 2: Distances between synthesized emotional speeches

happy speech for *I am almost finished*

distance	$\alpha = 0.01$		$\alpha = 0.05$	
BA	$0.17 > 0.15$	**	$0.17 > 0.13$	**
AD	$1.11 > 0.15$	**	$1.11 > 0.13$	**
DC	$1.07 > 0.15$	**	$1.07 > 0.13$	**

angry speech for *I am almost finished*

distance	$\alpha = 0.01$		$\alpha = 0.05$	
AB	$0.09 < 0.27$		$0.09 < 0.25$	
BC	$1.51 > 0.27$	**	$1.51 > 0.25$	**
CD	$0.27 = 0.27$	**	$0.27 > 0.25$	**

fearful speech for *I am going to the city*

distance	$\alpha = 0.01$		$\alpha = 0.05$	
AB	$0.76 > 0.39$	**	$0.76 > 0.31$	**
BD	$0.28 < 0.39$		$0.28 < 0.31$	
DC	$1.02 > 0.39$	**	$1.02 > 0.31$	**

surprised speech for *I am going to the city*

distance	$\alpha = 0.01$		$\alpha = 0.05$	
AB	$0.06 < 0.25$		$0.06 < 0.21$	
BC	$1.42 > 0.25$	**	$1.42 > 0.21$	**
CD	$0.62 > 0.25$	**	$0.62 > 0.21$	**

disgusted speech for *I thought you really meant it*

distance	$\alpha = 0.01$		$\alpha = 0.05$	
BA	$0.15 < 0.22$		$0.15 < 0.17$	
AD	$0.35 > 0.22$	**	$0.35 > 0.17$	**
DC	$1.25 > 0.22$	**	$1.25 > 0.17$	**

sad speech for *I thought you really meant it*

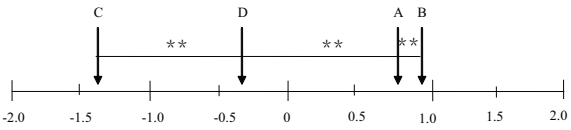
Distance	$\alpha = 0.01$		$\alpha = 0.05$	
BA	$0.19 < 0.37$		$0.19 < 0.33$	
AC	$0.88 > 0.37$	**	$0.88 > 0.33$	**
CD	$1.03 > 0.37$	**	$1.03 > 0.33$	**

B is given a superior evaluation to A, it is not significant. From the above, we may conclude that MIT's algorithm is as good as our algorithm, and far better than the random algorithm. Both MIT's algorithm and our algorithm are effective, since the synthesized speeches are superior to the neutral speeches with significance in most cases.

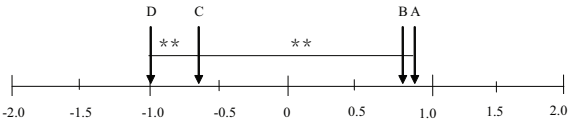
For the *angry*, *surprised* and *sad* speeches, C is significantly better than D, whereas for the other three emotions, D is significantly better than C. This may be evidence of the instability of the random algorithm.

3.3 Evaluation experiments on algorithms to generate emotion music

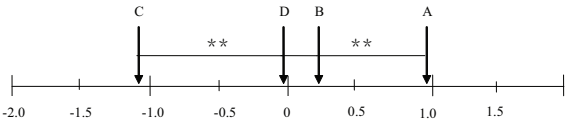
Emotion is a key factor in music. In this section, we assess our algorithm to generate emotional music [26] using a modified KTH rule system [27], whose parameters are optimized by an IGA. We compare music generated by our algorithm, a random algorithm, in which the parameters of



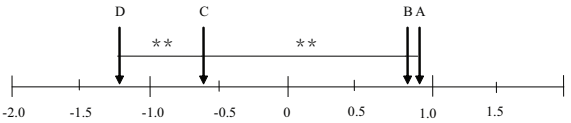
(a) *happy* speech for *I am almost finished*



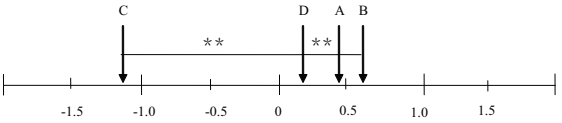
(b) *angry* speech for *I am almost finished*



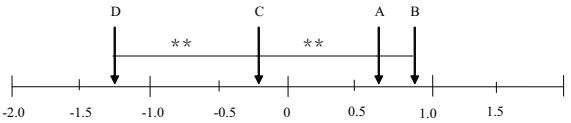
(c) *fearful* speech for *I am going to the city*



(d) *surprised* speech for *I am going to the city*



(e) *disgusted* speech for *I thought you really meant it*



(f) *sad* speech for *I thought you really meant it*

Figure 3: Synthesized emotional speeches on the psychological scale

Table 3: Distances between generated emotional music

happy music

distance	$\alpha = 0.01$		$\alpha = 0.05$	
BA	$0.836 > 0.338$	**	$0.836 > 0.277$	**
DA	$0.175 < 0.338$		$0.175 < 0.277$	
AC	$1.341 > 0.338$	**	$1.341 > 0.277$	**

sad music

distance	$\alpha = 0.01$		$\alpha = 0.05$	
BC	$0.338 > 0.22$	**	$0.338 > 0.18$	**
CA	$0.736 > 0.22$	**	$0.736 > 0.18$	**
BA	$1.074 > 0.22$	**	$1.074 > 0.18$	**

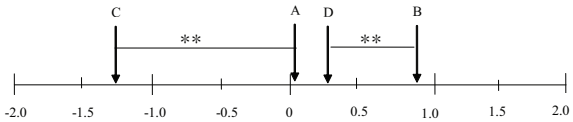
the KTH system are set randomly, and the original KTH system whose parameters are set by the authors in [21].

3.3.1 Experimental analysis

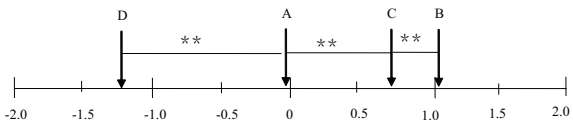
Two types of emotional music, *happy* and *sad*, were generated. For each emotion, we obtained four pieces of music, one generated by our algorithm, one generated by the algorithm in [21], and two generated randomly. These were labeled B, A, C, and D, respectively. The number of pairs used for the subjective test was $C_4^2 = 6$. The experimental procedure was similar to that in Section 3.1.1. In this section, the test subjects used 7-step scores to report their feelings. All reports were subsequently collated and analyzed by ANOVA.

3.3.2 Experimental analysis

Table 3 and Fig. 4 show the statistical results. For both *happy* and *sad* music, B is significantly better than A, C, and D at the $\alpha = 0.05$ and $\alpha = 0.01$ levels. This leads us to conclude that our algorithm is better than both that in [27] and the random algorithm. One piece of randomly generated music is better than that generated by the algorithm in [27]. This may be a result of the instability of the random algorithm.



(a) *happy* music



(b) *sad* music

Figure 4: Generated emotional music on the psychological scale

4. CONCLUSION

The evaluation method is very important in the research field of affective computing. The difficulty lies in how to use objective criteria to evaluate subjective feelings. Psychophysics investigates the relationships between sensations in the psychological domain and stimuli in the physical domain. Inspired by this, we proposed an evaluation method using psychological scales in this paper. Objects generated by different emotional algorithms are regarded as the external stimuli, while users' feelings are regarded as subjective evaluations. Psychophysical methods, such as the Scheffe method of paired comparison, were adopted to measure users' feelings and obtain the psychological scale. Objects generated by different algorithms were then compared on this psychological scale, where the higher the value, the better the algorithm is. Case studies on three kinds of emotional algorithms, emotional image retrieval, emotional speech synthesis and emotional music generation, were used to illustrate various uses of our approach and how it may be applied in evaluating affective interaction and user centered design.

Since subject evaluations are mainly relative, our proposed method can only be given a relative evaluation by comparing it with other algorithms. This is reasonable. How to define an absolute evaluation parameter remains to be considered in future research. Furthermore, a standard image database, sentences for synthesis, and music scores for generation are also envisaged in further research.

Emotion assessment is important in evaluation, which falls into three categories: self-reporting, physiological measurement, and emotion inference from observation [28]. Each method has its advantages and disadvantages. A hybrid evaluation method combining the advantages of the different assessment techniques may be a worthwhile research topic.

Acknowledgement

This paper is supported by National 863 Program(2008AA01Z122), Anhui Provincial Natural Science Foundation (No.070412056) and SRF for ROCS, SEM.

5. REFERENCES

[1] R. W. Picard. *Affective Computing*. MIT Press, 1997.

[2] K. Isbister and K. Höök. Evaluating affective interfaces: innovative approaches. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 2119–2119, New York, NY, USA, 2005. ACM.

[3] K. Isbister and K. Höök. Evaluating affective interactions. *International Journal of Human-Computer Studies*, 65(4):273–274, 2007.

[4] N. S. Shami, J. T. Hancock, C. Peter, M. Muller, and R. Mandryk. Measuring affect in hci: going beyond the individual. In *CHI '08: CHI '08 extended abstracts on Human factors in computing systems*, pages 3901–3904, New York, NY, USA, 2008. ACM.

[5] K. Höök. User-centered design and evaluation of affective interfaces. pages 127–160, 2004.

[6] C. Wiberg. Affective computing vs. usability? insights of using traditional usability evaluation methods. April 2005.

[7] N. Chateau and M. Mersiol. Amuse: A tool for evaluating affective interfaces. April 2005.

[8] D. Fallman and J. Waterworth. Dealing with user experience and affective evaluation in hci design: A repertory grid approach. April 2005.

[9] R. W. Picard and S. B. Daily. Evaluating affective interactions: Alternatives to asking what users feel. April 2005.

[10] R. L. Mandryk and M. S. Atkins. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, 65(4):329–347, 2007.

[11] R. L. Hazlett and J.Benedek. Measuring emotional valence to understand the user's experience of software. *International Journal of Human-Computer Studies*, 65(4):306–314, 2007.

[12] M. Isomursu, M. Tahti, S. Vainamo, and K. Kuutti. Experimental evaluation of five methods for collecting emotions in field settings with mobile applications. *International Journal of Human-Computer Studies*, 65(4):404–418, 2007.

[13] P.J. Lang. *Behavioral treatment and bio-behavioral assessment: computer applications*. Albex, Norwood, NJ, 1980.

[14] P. M. A. Desmet, C. J. Overbeeke, and S. J. E. T. Tax. Designing products with added emotional value: development and application of an approach for research through design. *The Design Journal*, 4(1):32–47, 2001.

[15] M. Isomursu, K. Kuutti, and S. Vainamo. Experience clip: method for user participation and evaluation of mobile concepts. In *Proceedings of the Participatory Design Conference*, pages 83–92, 2004.

[16] M. Tahti and L. Arhippainen. A proposal of collecting emotions and experiences. *Interactive Experiences in HCI*, 2:195–198, 2004.

[17] L. A. Rantakokko and T. M. Tahti. Mobile feedback application for emotion and user experience collection. In *Proceedings of PROW 2004*, pages 77–81. Helsinki University Press, 2004.

[18] M. Haringer and S. Beckhaus. Framework for the measurement of affect in interactive experiences and games. 2008.

[19] S. F. Wang. Emotion semantics image retrieval: An brief overview. In *Proceeding of 1st International Conference on Affective Computing and Intelligent Interaction*, pages 490–497. Springer, 2005.

[20] Zh. L. Yang. *Experimental Psychology*. Zhejiang Education Publishing Company, 2002.

[21] A. Friberg. pdm: An expressive sequencer with real-time control of the kth music-performance rules. *Comput. Music J.*, 30(1):37–48, 2006.

[22] H. Scheffe. An analysis of variance for paired comparisons. *Journal of American Statistical Association*, 147(1):381–400, 1952.

[23] S. F. Wang. *Research on Kansei Information Processing and its Application in Image Retrieval*. Doctor dissertation, Universtiy of Science and Technology of China, Hefei, Anhui, P.R.China, May 2002.

[24] J. Cahn. Generating expression in synthesized speech. Master thesis, M.I.T., 1989.

[25] S. L. Lv, S. F. Wang, and X. F. Wang. Emotional speech synthesis by xml file using interactive genetic algorithms. In *GEC Summit*, pages 907–910. ACM, 2009.

[26] H. Zhu, S. F. Wang, and Zh. Wang. Emotional music generation using interactive genetic algorithm. In *2008 International Conference onComputer Science and Software Engineering*, pages 345–348, 2008.

[27] A. Friberg, R. Bresin, and J. Sundberg. Overview of the kth rule system for music performance. *Advances in Cognitive Psychology*, 2(1):145–161, 2006.

[28] M. Wong. Emotion assessment in evaluation of affective interfaces. master thesis, University of Waterloo, 2006.