AN EMOTION MODEL FOR MUSIC USING BRAIN WAVES

Rafael Cabredo[1,2], Roberto Legaspi[1], Paul Salvador Inventado[1,2], and Masayuki Numao[1]

[1]Institute of Scientific and Industrial Research, Osaka University, Japan,
[2]Center for Empathic Human-Computer Interactions, De La Salle University, Philippines,
{cabredo,roberto,inventado,numao}@ai.sanken.osaka-u.ac.jp

ABSTRACT

Every person reacts differently to music. The task then is to identify a specific set of music features that have a significant effect on emotion for an individual. Previous research have used self-reported emotions or tags to annotate short segments of music using discrete labels. Our approach uses an electroencephalograph to record the subject's reaction to music. Emotion spectrum analysis method is used to analyse the electric potentials and provide continuous-valued annotations of four emotional states for different segments of the music. Music features are obtained by processing music information from the MIDI files which are separated into several segments using a windowing technique. The music features extracted are used in two separate supervised classification algorithms to build the emotion models. Classifiers have a minimum error rate of 5% predicting the emotion labels.

1. INTRODUCTION

Listening to music brings out different kinds of emotions. It can be involuntary and different for every person and primarily caused by musical content. A lot of research has been done identifying music features that are associated with affecting emotion or mood [3, 5, 17]. The work of [9] also investigates music features and discusses how changing these features can affect the emotions the music elicits.

With a good background of how different music features affect emotions, it is possible to automatically classify and predict what kind of emotions a person will experience. A survey of music emotion research by Kim et al. [6] report that the typical approach for classifying music using emotion is to build a database of ground truth of emotion labels by subjective tests. Afterwards, a machine learning technique is used to train a classifier to automatically recognize high-level or low-level music features.

A common problem encountered by previous work is the limitation of the annotation for emotion. It takes a lot of time and resources to annotate music. Lin, et al. [8] reviews various work on music emotion classification and utilize the vast amount of online social tags to improve emotion classification. However, a personalized emotion model for labelling music would still be desirable. Music that is relaxing for some people may be stressful for others.

Songs are also usually annotated with the most prominent emotion (i.e. only one emotion label per song). Multi-label classification [18] can be used to have richer emotion annotations. These annotations however are still discrete-valued emotion labels.

In our work, we are interested in learning how emotion changes throughout the song and identify music features that could have caused these changes. Because of this, continuous-valued emotion annotations are preferred. One method to do this is to use an electroencephalograph (EEG) in recognizing emotions similar to the work used to develop Constructive Adaptive User Interface (CAUI), which can arrange [7, 13] and compose [14] music based on one's impressions of music. In addition to collecting continuous-valued annotations for full-length music, we focus our work on considering individual emotion reactions to music as opposed to building a generalized emotion model.

2. DATA COLLECTION METHODOLOGY

We construct a user specific model by using supervised machine learning techniques to classify songs using music features. As mentioned earlier, this task requires songs that can elicit emotions from a listener and the music features of these songs.

For this research, we had a 29-year old female participant who selected and annotated songs. The music collection is a set of MIDI files comprised of 121 Japanese and Western songs having 33 Folk, 20 Jazz, 44 Pop, and 24 Rock music. By using MIDI files, the music information can be easily extracted to produce high-level features for the classifier. MIDI files also eliminate any additional emotions contributed by lyrics.

2.1 Emotion annotation

Music emotion annotation is performed in 3 stages. First, the subject listened to all songs and manually annotated each one. The subject was instructed to listen to the entire song and was given full control on which parts of the song she wanted to listen to.

After listening to each song, the subject gives a general impression on how joyful, sad, relaxing, and stressful each

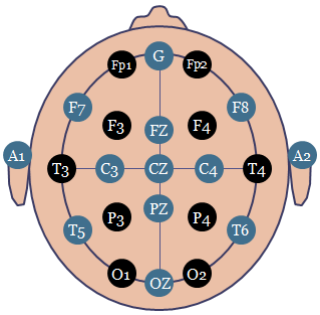

Figure 1. The EEG has 23 electrodes used to record electrical changes on the scalp. Each node is identified by a letter to indicate lobe position: F-Frontal lobe, T-Temporal lobe, C-Central lobe, P-Parietal lobe, O-Occipital lobe. 'Z' refers to an electrode placed on the mid-line

song was using a five-point Likert scale. Aside from the emotions felt, the subject was also asked to rate whether she was familiar with the song or not using the same scale. With this feedback, we chose the 10 most relaxing songs and 10 most stressful songs with varying levels of familiarity to the subject. The manual annotation was done in one session for approximately one and a half hours.

Since collection of the emotion annotations takes a lot of time and effort from the subject, it was decided to concentrate time and resources on a certain type of emotion. We opted to concentrate on relaxing music because these are normally the kind of music people would want to listen to on stressful days. The stressful songs are meant to serve as negative examples for the classifier.

In the second stage an EEG was used to measure brain activity while the subject listened to the 20 songs previously selected. The EEG device is a helmet with electrodes that can be placed on all scalp positions according to the International 10–20 Standard. Figure 1 shows the location of the different electrodes. Using the EEG, electric potential differences were recorded with a reference electrode on the right earlobe.

Work on EEG to recognize emotions find that different mental state produces a distinct pattern of electrical activity [1, 2]. The right hemisphere is responsible for negative emotions (i.e. stress, disgust, sadness) while the left hemisphere is responsible for positive emotions (i.e. happiness, gratitude, amusement).

The EEG device is very sensitive. As such, the subject was instructed to close her eyes and remain still while data was being collected. Listening sessions had to be limited to a maximum of 30 minutes or upto the moment that the subject begins to feel uncomfortable wearing the helmet. We had to ensure that the subject was comfortable and eliminate external factors that may contribute to changes in emotion. On average, EEG readings for 7 songs were recorded per session.

Prior to playing each music, we introduce a 10 second white noise to help the subject focus on the task at hand without stimulating a strong emotional response. After listening to one song, a short interview is conducted to determine if the subject particularly liked or disliked specific parts of the song. The interview also helped confirm the initial manual annotations of the subject.

In the final stage, continuous emotion annotations were obtained using EMonSys. This software [1] uses the emotion spectrum analysis method (ESAM) [12] to convert brain wave readings to emotion readings. Using data from 10 scalp positions at Fp1, Fp2, F3, F4, T3, T4, P3, P4, O1, O2, electric potentials were separated into their θ (5–8 Hz), α (8–13 Hz) and β (13–20 Hz) frequency components by means of fast Fourier transforms (FFT). Cross-correlation coefficients for each pair of channels are computed (i.e., 10 channels * 9 channels/2) and these are evaluated for every time step together with the 3 bands to obtain an input vector Y having 135 variables at each time step. EMonSys can evaluate the EEG readings at different time steps. We used the smallest available: 0.64 seconds.

Using an emotion matrix C, this 135-dimensional vector is linearly transformed into a 4-D emotion vector $E = (e_1, e_2, e_3, e_4)$, where e_i corresponds to the 4 emotional states, namely: stress, joy, sadness, and relaxation. Formally, the emotion vector is obtained by

$$C \cdot Y + d = E, \tag{1}$$

where d is a constant vector. The emotion vector is used to provide a continuous annotation to the music every 0.64 seconds. For example, if one feels joy, the emotion vector would have a value of $E = (0, e_2, 0, 0)$.

2.2 Extracting Music Features

A song having length m is split into several segments using a sliding window technique. Each segment, or now referred to as a window w has a length n, where one unit of length corresponds to one sample of emotion annotation.

MIDI information for each window is read using a module adapted from jSymbolic [10] to extract 109 high-level music features. These features can be loosely grouped into the following categories: Instrumentation, Texture, Dynamics, Rhythm, Pitch Statistics, and Melody. The feature set includes one-dimensional and multi-dimensional features. For example, *Amount of Arpeggiation* is a one-dimensional Melody feature, *Beat Histogram* is a 161-dimensional Rhythm feature, etc. All features available in jSymbolic were used to build a 1023-dimension feature vector. The category distribution of the feature vector is shown in Table 1. The *Others* category refers to the features *Duration* and *Music Position*. *Duration* is a feature from jSymbolic, which describes the length of the song in seconds. *Music Position* refers to the position of the window relative to duration of the song. Although it was known that not all of the features will be used, this approach allows utilization of feature selection techniques to determine which features were the most important in classification.

After extracting the features for one window, the window goes through the data using a step size s until the end

[1] software developed by Brain Functions Laboratory, Inc.

Category	Amount	Percentage
Dynamics	4	0.39%
Instrumentation	493	48.19%
Melody	145	14.17%
Pitch	174	17.01%
Rhythm	191	18.67%
Texture	14	1.37%
Others	2	0.20%

Table 1. Distribution of features used for the instances

of the song is reached. Each window was labelled using the average emotion values within the length of the window. Formally, the label for w_i is the emotion vector

$$E^i = \frac{1}{n}\sum_{j=i}^{i+n} E^j = \frac{1}{n}\sum_{j=i}^{i+n}\left(e_1^j, e_2^j, e_3^j, e_4^j\right), \qquad (2)$$

where $1 \leq j \leq m - n$.

3. EMOTION MODEL

Weka's [4] implementation of linear regression and C4.5 were used to build the emotion models for each emotion. The training examples were derived from the window given one emotion label, which results to four datasets. Each dataset has a maximum of 6156 instances using the smallest values for the sliding window (i.e. $n = 1$ and $s = 1$). The number of instances depends on the parameters used for windowing. During preliminary experiments we observed that the decrease of training data due to larger step sizes had too much of a negative influence on performance. As such, all features were extracted using the smallest size of $s = 1$ for all experiments.

Prior to training, all features that do not change at all or vary too frequently (i.e. varies 99% of the time) are removed. Afterwards, normalization is performed to have all feature values within $[0, 1]$.

3.1 Using Linear Regression

The linear regression used for building the emotion models uses the Akaike criterion for model selection and M5 method [15] to select features. The M5 method steps through the features and removes features with the smallest standardized coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion.

3.2 Using C4.5

C4.5 [16] is a learning technique that builds a decision tree from the set of training data using the concept of information entropy. Since this technique requires nominal class values, the emotion labels are first discretized into five bins. Initial work used larger bin sizes but we observed poorer performance using these.

3.3 Testing and Evaluation

We used 10-fold cross-validation to assess the models generated by the two methods using different values for the

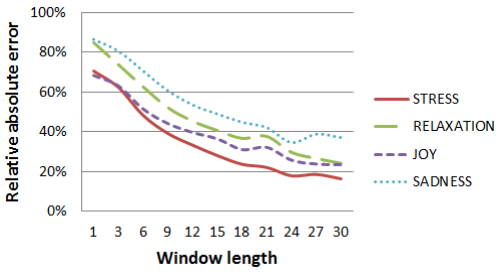

Figure 2. Relative absolute error using linear regression

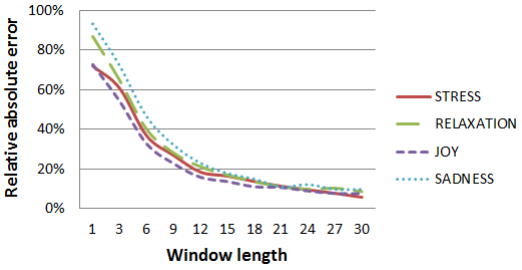

Figure 3. Relative absolute error using C4.5

window length. We use the relative absolute error for evaluating performance of the classifiers. Weka computes this error measure by normalizing with respect to the performance obtained by predicting the classes' prior probabilities as estimated from the training data with a simple Laplace estimator. Figures 2 and 3 show the change in relative absolute error using linear regression and C4.5, respectively. Window length values were varied from 1 to 30 samples (i.e. 0.64 seconds to 19.2 seconds of music).

4. RESULTS AND ANALYSIS

Increasing the window size increases accuracy of the classifiers. Further experiments were done to include window sizes upto 240 samples. Results of these are shown in Figures 4 and 5. From these results, we find the value of n which minimizes the average relative absolute error over $n = [1..20]$. For linear regression, using $n = 90$ gives the minimum average relative absolute error of 7.6% with a correlation coefficient of 0.8532 and root mean squared error of 0.1233. The average is taken from values for the four emotion model results.

Using C4.5, a smaller window length is necessary to obtain similar results. Using $n = 60$, the average relative absolute error is 5.1%, average root mean squared error is 0.0871, and average Kappa statistic is 0.9530. The Kappa statistic describes the chance-corrected measure of agreement between the classifications and the true classes.

When $n \geq 120$, we notice that some songs are no longer included in the training data as the window length becomes greater than the song length. As such, results using these window lengths may not be accurate.

	$n = 1$		$n = 30$		$n = 60$		$n = 90$		$n = 120$	
Class No.	S	R	S	R	S	R	S	R	S	R
1	84.0%	95.3%	56.5%	82.2%	52.3%	80.5%	51.0%	81.5%	49.1%	80.5%
2	13.3%	3.8%	31.6%	9.9%	28.6%	6.0%	26.1%	3.7%	25.7%	3.2%
3	1.9%	0.7%	8.7%	6.5%	15.4%	10.3%	18.4%	11.1%	20.7%	9.4%
4	0.5%	0.2%	1.8%	1.0%	1.8%	2.2%	2.3%	2.7%	1.6%	5.7%
5	0.3%	0.0%	1.4%	0.4%	1.9%	1.0%	2.1%	1.1%	2.9%	1.1%

Table 2. Class sizes for Stress (S) and Relaxation (R) data after discretization

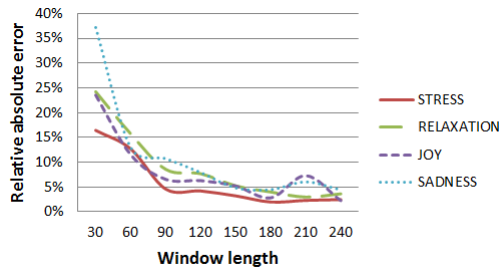

Figure 4. Relative absolute error using linear regression

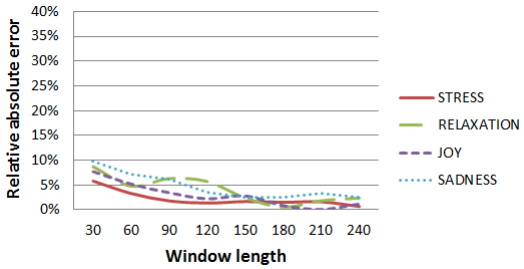

Figure 5. Relative absolute error using C4.5

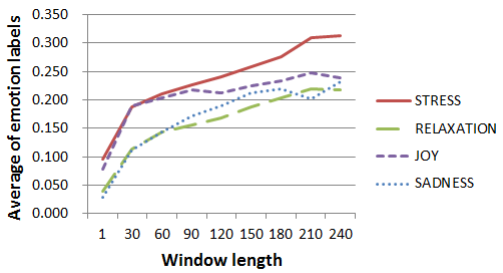

Figure 6. Average of emotion value for different window lengths

Category	Stress	Relaxation	Sadness	Joy
Rhythm	40.4%	32.4%	32.8%	34.0%
Pitch	21.3%	29.7%	28.4%	32.0%
Melody	10.6%	16.2%	19.4%	20.0%
Instrumentation	17.0%	10.8%	10.4%	8.0%
Texture	8.5%	5.4%	4.5%	2.0%
Dynamics	0.0%	2.7%	1.5%	0.0%
Others	2.1%	2.7%	3.0%	4.0%

Table 3. Distribution of features used in C4.5

4.1 Influence of window length

Model accuracy is highly dependent on the parameters of the windowing technique. Increasing the window length allows more music information to be included in the instances making each more distinguishable from instances of other classes.

Increasing the window length also affects the emotion annotations. ESAM was configured to produce emotion vectors having positive values. Since most of the emotion values are near zero, the average emotion values for the windows are also low. Figure 6 shows the steady increase of the values for the class labels as the window length is increased. The standard deviation also follows a linear trend and steadily increases from 0.091 to 0.272 for the same window lengths. Using larger window lengths diversifies the emotion labels as well which, in turn, contributes to better accuracy.

The low average values also affected the discretization of the emotion labels for C4.5. It resulted to having a majority class. Table 2 shows that class 1 is consistently the majority class for the data set. With a small window length, more instances are labelled with emotion value close to 0. We note, however that as window length is increased, the number of classes steadily balances out. For example, at $n = 1$, 84% of the data is labelled as class 1, but when $n = 90$, it is only 51%. This is the general trend for all the emotion models. At $n = 90$, the instances labelled as class 1 for the other emotion labels are as follows: 62.2% for Joy, 78.8% for Sadness, and 81.5% for Relaxation.

4.2 Important features used in C4.5

C4.5 builds a decision tree by finding features in the data that most effectively splits the data into subsets enriched in one class or the other. This causes a side effect of identifying music features that are most beneficial for classifying emotions.

Table 3 summarizes the features included in the trees generated by the algorithm using $n = 60$. The items are ordered according to the number of features present in the decision trees. A big portion of the features included are rhythmic features averaging 34.9% of the feature set. Features related to instrumentation also play a big part in identifying Stress unlike the other emotions. On the other hand, melody features are more important for Relaxation, Stress and Joy.

A closer inspection of the decision tree reveals that each emotion can be classified faster using a different ordering of music features. Table 4 shows the distribution of features found in the first 5 levels of the different decision

Category	Stress	Relaxation	Sadness	Joy
Rhythm	23.4%	13.5%	6.0%	14.0%
Pitch	0.0%	10.8%	9.0%	10.0%
Melody	4.3%	2.7%	1.5%	6.0%
Instrumentation	4.3%	2.7%	4.5%	4.0%
Texture	0.0%	0.0%	0.0%	0.0%
Dynamics	2.1%	2.7%	1.5%	0.0%
Others	0.0%	2.7%	0.0%	0.0%

Table 4. Distribution of features found in the first 5 levels of the decision trees of C4.5

trees. The Stress model mostly uses rhythmic features and 2 melodic features for the first 4 levels and uses Instrumentation for the 5th level. During the interview with the subject, when asked which parts of the songs are stressful, she explains that songs with electric guitar and rock songs in general are very stressful for her. Rock songs used in the dataset had a fast tempo and may be a factor as to the construction of the decision tree.

For relaxing music, the subject mentioned that there are specific parts of the songs that made her feel relaxed. These include introductory parts, transitions between chorus and verses, piano and harp instrumentals, and climactic parts of the song (i.e. last verse-chorus or bridge). Examining the decision tree for relaxation, we find that *Melodic Interval Histogram*, *Basic Pitch Histogram*, and *Music Position* are used for the first 3 levels, which are features that support the statements of the subject. Although emotion models for Joy and Sadness are available, a complete analysis of these cannot be done since the dataset was primarily focused on relaxing and stressful music.

4.3 Accuracy of Emotion labels

The manual emotion labels were also compared to the emotion values from ESAM. The average emotion value for each song was calculated and transformed into a 5-point scale. Comparing the manual annotations with the discretized continuous annotations, we find that only 25% of the emotion labels from EEG were the same with the manual annotations, 62% of the emotion labels from EEG slightly differed from the manual annotations, and 13% were completely opposite from what was originally reported. It is difficult to attribute error for the discrepancy. One possible cause could be the methodology for manual annotations. While the subject was doing the manual annotations, we observed that usually, she would only listen to the first 30 seconds of the song and in some cases skip to the middle of the song. It is possible that the manual annotation incompletely represents the emotion of the entire song.

It is also possible that the subject experienced a different kind of emotion unconsciously while listening to the music. For example some songs that were reported to be stressful turned out not stressful at all. We examined the emotion annotations and checked if there was any dependency between the values.

In Table 5 we can see that the subject treated the emotion Stress to be the bipolar opposite of Relaxation due to

	Joy	Sadness	Relaxation	Stress
Sadness	-0.5638			
Relaxation	0.5870	0.0733		
Stress	-0.6221	-0.0555	-0.9791	
Familiarity	0.7190	-0.2501	0.5644	-0.6252

Table 5. Correlation of manual annotations

	Joy	Sadness	Relaxation	Stress
Sadness	-0.1187			
Relaxation	0.4598	-0.2338		
Stress	-0.4450	0.3100	-0.4223	
Familiarity	-0.0579	0.2956	-0.2343	0.5731

Table 6. Correlation of annotations using ESAM

the high negative correlation value. Using ESAM, we find a similar situation but there is only a moderate negative correlation between the two as shown in Table 6. If we examine the other emotions, we find that Joy has a correlation with Relaxation and a negative correlation with Stress. This is consistently reported for both manual annotations and annotations using ESAM.

Finally, we compared the amount of discrepancy between manual and automated annotations against the subject's familiarity with the song. We found that the discrepancy values for joyful and relaxing songs have a high correlation with familiarity : 0.6061 for Joy and 0.69551 for Relaxation. This implies that measurements of ESAM for Joy and Relaxation become more accurate when the subject is not familiar with the songs. It is possible that unfamiliar songs will help induce stronger emotions as compared to familiar music. This may be an important factor when using psychophysiological devices in measuring emotion.

5. CONCLUSION

This research focuses on building an emotion model for relaxing and stressful music. The model was built by extracting high-level music features from MIDI files using a windowing technique. The features were labelled using emotion values generated using EEG and ESAM. These values were also compared against manual emotion annotations. With the help of interviews conducted with the subject, we observe that EEG and ESAM can be used for annotating emotion in music especially when the subject experiences a strong intensity of that emotion. Familiarity of the subject with the song can affect genuine emotions.

Linear regression and C4.5 were used to build the different emotion models. Using a 10-fold cross-validation for evaluating the models, high accuracy with low relative absolute errors was obtained by using large window lengths encompassing between 38.4 seconds ($n = 60$) to 57.6 seconds ($n = 90$) of music.

6. FUTURE WORK

The current work involves one subject and it would be interesting to see if the model can be generalized using more

subjects or, at the least, to verify if the current methodology will yield similar results when used with another subject.

Instead of using the average value for the emotion label, we intend to explore other metrics to summarize the emotion values for each window.

Further study on the music features is also needed. The current model uses both one-dimensional and multidimensional features. Experiments using only one set of the features will be performed. We also wish to explore the accuracy of the classification if low-level features were used instead of high-level features.

The window length greatly affects model accuracy. We have yet to investigate if there is a relationship between the average tempo of the song with window length. We hypothesize that slower songs would require longer window lengths to capture the same amount of information needed for fast songs. On the other hand, songs with fast tempo would need shorter window lengths.

Finally, this model will be integrated to a music recommendation system that can recommend songs which can induce similar emotions to the songs the user is currently listening to.

7. ACKNOWLEDGEMENTS

This research is supported in part by the Management Expenses Grants for National Universities Corporations through the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, by the Global COE (Centers of Excellence) Program of MEXT, and by KAKENHI 23300059.

8. REFERENCES

[1] K. Ansari-Asl, G. Chanel, T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood," *EUSIPCO 2007,15th Eur. Signal Proc. Conf.*, p. 1241–1245, 2007.

[2] G. Chanel, J. Kronegg, D. Grandjean, T. Pun: "Emotion assessment: Arousal evaluation using EEGs and peripheral physiological signals, *Lecture Notes in Computer Science*, Vol. 4105, p. 530, 2006.

[3] A. Gabrielsson, P.N. Juslin,: "Emotional expression in music." In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, editors, *Handbook of affective sciences*, New York: Oxford University Press, pp. 503-534, 2003.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten: "The WEKA Data Mining Software: An Update, SIGKDD Explorations," Vol. 11, No. 1, pp. 10–18, 2009.

[5] P.N. Juslin, J.A. Sloboda: "Handbook of music and emotion: theory, research, applications," Oxford University Press, 2010.

[6] Y.E. Kim, E.M. Schmidt, R. Migneco, B.G. Morton, P. Richardson, J. Scott, J.A. Speck, D. Turnbull: "Music Emotion Recognition: A State of the Art Review," *Proc. of the 11th ISMIR Conf.*, pp. 255–266, 2010.

[7] R. Legaspi, Y. Hashimoto, K. Moriyama, S. Kurihara, M. Numao: "Music Compositional Intelligence with an Affective Flavor," *Proc. of the 12th International Conference on Intelligent User Interfaces*, pp. 216–224, 2007.

[8] Y.-C. Lin, Y.-H. Yang, and H. H. Chen: "Exploiting online music tags for music emotion classification, *ACM Transactions on Multimedia Computing*, Communications, and Applications, Vol. 7S, No. 1, pp. 1–16, 2011.

[9] S.R. Livingstone, R. Muhlberger, A.R. Brown, and W.F. Thompson: "Changing musical emotion: A computational rule system for modifying score and performance," *Computer Music Journal*, Vol. 34, No. 1, pp. 41–64, 2010.

[10] C. McKay, and I. Fujinaga: "jSymbolic: A feature extractor for MIDI files," *Proc. of the International Computer Music Conference*, pp. 302–305, 2006.

[11] E.R. Miranda, and A. Brouse: "Toward direct brain-computer musical interfaces," *New Interfaces for Musical Expression*, 2005.

[12] T. Musha, Y. Terasaki, H.A. Haque, and G.A. Ivanitsky: "Feature extraction from EEGs associated with emotions," *Journal of Artificial Life and Robotics*, Vol. 1, No. 1, pp. 15–19,1997.

[13] M. Numao, M. Kobayashi, and K. Sakaniwa: "Aquisition of human feelings in music arrangement," *Proc. of IJCAI '97*, pp. 268–273, 1997.

[14] M. Numao, S. Takagi, and K. Nakamura: Constructive adaptive user interfaces - Composing music based on human feelings," *Proc. of AAAI '02*, pp. 193–198, 2002.

[15] J.R. Quinlan: "Learning with continuous classes, *Proc. AI92, 5th Australian Joint Conference on Artificial Intelligence*, Adams & Sterling (eds.), World Scientific, Singapore, pp. 343–348, 1992.

[16] J.R. Quinlan: "C4.5: Programs for Machine Learning," Morgan Kaufmann Publishers, 1993.

[17] E. Schubert: "Affective, Evaluative, and Collative Responses to Hated and Loved Music," *Psychology of Aesthetics Creativity and the Arts*, Vol. 4, No. 1, pp. 36–46, 2010.

[18] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas: "Multilabel classification of music into emotions, in *Proc. of the 9th International Conference on Music Information Retrieval*, pp. 325–330, 2008.