1K1-IOS-1a-5

Discovering Emotion Features in Symbolic Music

Rafael Cabredo, Paul Salvador Inventado, Roberto Legaspi, Masayuki Numao

The Institute of Scientific and Industrial Research, Osaka University

Current music recommender systems only use basic information for recommending music to its listeners. These usually include artist, album, genre, tempo and other song information. Online recommender systems would include ratings and annotation tags by other people as well. We propose a recommender system that recommends music depending on how the listener wants to feel while listening to the music. The user-specific model we use is derived by analyzing brain waves of the subject while he was actively listening to emotion-inducing music. The brain waves are analyzed in order to derive the emotional state of the listener for different segments of the music using an emotion spectral analysis method. The emotional state is used to label segments of music that are fed into a supervised machine learning technique to build an emotion model. This emotion model is used to identify the different music features that are important for recognizing specific emotional states.

1. Introduction

Music induces different kinds of emotions. From the research of [Gabrielsson 03, Kim 10, Livingstone 10] it is known that specific music features causes these changes in emotion. For example, songs with a fast tempo, in a major key, has simple harmony and high pitch, generally make people happy and feel excited, while songs having opposite features, such as, having low tempo, in a minor key, low pitch, and complicated harmony are considered songs that can elicit sadness, despair, or melancholy.

By recognizing these music features, it can be used to anticipate or even change the emotion or mood of a listener. This can be done automatically by using machine learning techniques to learn the dependencies in the music features given a ground truth of emotion labels.

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. [Lin 11] 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 [Trohidis 08] can be used to have richer emotion annotations. These annotations however are still discrete labels.

In this work, emotion changes in the entire song are recorded and analyzed to learn how the music features affect these changes. Instead of using discrete labels, a continuous annotation is used to give a fine-grained description of emotion changes. One method to acquire continuous emotion labels is to use brain waves similar to the work used to develop Constructive Adaptive User Interface (CAUI), which can arrange [Legaspi 07, Numao 97] and compose [Numao 02] music based on one's impressions of music.

2. Data Collection Methodology

A user specific model is built 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, a 29-year old female participated to select and annotate 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 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, the 10 most relaxing songs and 10 most stressful songs with varying levels of familiarity to the subject were chosen. 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, specifically, relaxing music. Relaxing music was chosen 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 a electroencephalograph (EEG) was used to measure brain wave activity while the subject lis-

Contact: Rafael Cabredo, The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan, tel: +81-6-6879-8426, fax:+81-6-6879-8428, cabredo@ai.sanken.osaka-u.ac.jp

tened 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. Using the EEG, electric potential differences were recorded with a reference electrode on the right earlobe.

The subject was advised 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 using the EEG helmet. It was important 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, a 10 second white noise was introduced 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) [Musha 97] 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). The values of the cross-correlation coefficients for the three components on 45 channel pairs were evaluated every 0.64 seconds resulting in 135 variables. This set of variables forms the input vector Y. 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 with a 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 [McKay 06] 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 multidimensional 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 1021-dimension feature vector, where the last feature is the emotion label. 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.

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

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

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

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

3. Emotion Model

C4.5 was 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, it was observed that higher values for the window step size significantly decreased the training data. As such, all features were extracted using the smallest size of $s = 1$.

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 C4.5

C4.5 [Quinlan 93] is a learning technique that builds a decision tree from the set of training data using the concept of information entropy. The implementation of WEKA [Hall 09], J48 with default parameters, was used to build the classifier.

Since C4.5 requires nominal class values, the emotion labels are first discretized into five bins. Initial work used larger bin sizes, but it yielded poorer performance.

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

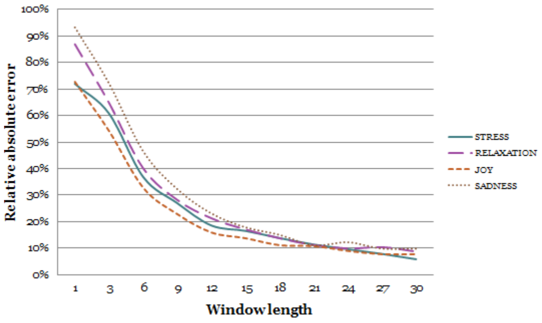

Figure 1: Relative absolute error using $1 \leq n \leq 30$

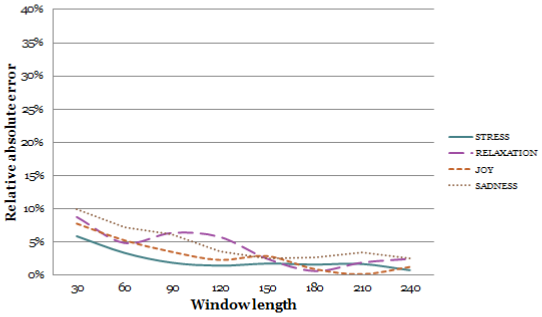

Figure 2: Relative absolute error using $30 \leq n \leq 240$

3.2 Testing and Evaluation

In order to test and evaluate the models, 10-fold cross-validation was used. The models were also built using music features extracted using different values for the window length. Window length values were varied from 1 to 30 samples (i.e. 0.64 seconds to 19.2 seconds of music) to see how the length of music used for feature extraction affected the classifier accuracy.

Relative absolute error was the main criteria for evaluating the performance of the classifiers. Accuracy in terms of F-measure was unreliable for analysis because of the nature of the dataset. This is further discussed in the next section. Figure 1 shows the change in relative absolute error as the window length is increased.

4. Analysis

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.

Experiments were expanded to include window sizes upto 240 samples. Results of these are shown in Figure 2. When $n = 60$, it is observed that the model already has an average relative absolute error of 5.1% with an average root mean squared error of 0.0871, and average Kappa statistic of 0.9530. The Kappa statistic describes the chance-corrected measure of agreement between the classifications and the true classes. This means that using a window length that captures around 40 seconds of music is enough to generate instances for the classifier.

Category	Stress	Relaxation	Sadness	Joy	Average
Rhythm	40.4%	32.4%	32.8%	34.0%	34.9%
Pitch	21.3%	29.7%	28.4%	32.0%	27.8%
Melody	10.6%	16.2%	19.4%	20.0%	16.6%
Instrument	17.0%	10.8%	10.4%	8.0%	11.6%
Texture	8.5%	5.4%	4.5%	2.0%	5.1%
Dynamics	0.0%	2.7%	1.5%	0.0%	1.0%
Others	2.1%	2.7%	3.0%	4.0%	3.0%

Table 2: Distribution of features used by classifier

Depth	Features for Stress
d = 0	R: Rhythmic Looseness
d = 1	M: Melodic Interval Histogram-28
	R: Beat Histogram-88
d = 2	R: Beat Histogram-86
	R: Beat Histogram-68
	R: Beat Histogram-53
	R: Beat Histogram-41
d = 3	M: Melodic Interval Histogram-6
	R: Beat Histogram-4
	R: Beat Histogram-37
	R: Beat Histogram-132
d = 4	I: Variability of Note Prevalence of Unpitched Instruments
	I: Time Prevalence of Pitched Instruments-6
	O: Duration
	R: Beat Histogram-85
	R: Beat Histogram-150

Table 3: First 5 levels of the decision tree for Stress; letters before the features denote the feature's category (R: Rhythm, M: Melody, P: Pitch, I: Instrumentation, O: Others)

When $n \geq 120$, 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.

4.1 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 2 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.

A closer inspection of the decision tree reveals that each emotion can be classified faster using a different ordering of music features. Tables 3 and 4 show the features found in the first 5 levels of the decision trees for the Stress and Relaxation emotion models. 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

Depth	Features for Relaxation
d = 0	M: Melodic Interval Histogram-6
d = 1	P: Basic Pitch Histogram-54
d = 2	O: Music Position
	R: Beat Histogram-63
d = 3	I: Note Prevalence of Unpitched Instruments-36
	P: Fifths Pitch Histogram-4
	R: Beat Histogram-102
	P: Basic Pitch Histogram-85
d = 4	D: Variation of Dynamics In Each Voice
	R: Beat Histogram-5
	R: Beat Histogram-38
	R: Beat Histogram-135
	P: Basic Pitch Histogram-67

Table 4: First 5 levels of the decision tree for Relaxation

Depth	Features for Joy
d = 0	M: Size of Melodic Arcs
d = 1	I: Note Prevalence of Unpitched Instruments-36
	P: Basic Pitch Histogram-55
d = 2	R: Beat Histogram-4
	P: Basic Pitch Histogram-89
	P: Basic Pitch Histogram-86
d = 3	P: Pitch Class Distribution-12
	I: Number of Unpitched Instruments
	R: Beat Histogram-22
	R: Beat Histogram-12
d = 4	I: Unpitched Instruments Present-15
	O: Duration
	R: Beat Histogram-17
	P: Basic Pitch Histogram-57
	P: Basic Pitch Histogram-31

Table 5: First 5 levels of the decision tree for Joy

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, *Melodic Interval Histogram*, *Basic Pitch Histogram*, and *Music Position* are used for the first 3 levels. These are music features that support the statements of the subject during the interview.

Tables 5 and 6 show a portion of the decision tree for the Joy and Sadness emotion models. 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.

4.2 Accuracy of Emotion labels

Accuracy of the emotion models can also be affected by the quality of the labels used for annotation. It is also important to take note of the class distribution of the datasets. 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 3 shows the steady increase of the values for the class labels. The standard deviation also follows a linear trend and steadily increases from 0.091 – 0.272 for the same window lengths.

The low average values also affected the discretization of the emotion labels. It resulted to having a majority class. Table 7 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. However, 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.

Depth	Features for Sadness
d = 0	P: Basic Pitch Histogram-30
d = 1	I: Number of Unpitched Instruments
	M: Melodic Interval Histogram-33
d = 2	I: Note Prevalence of Unpitched Instruments-4
	P: Importance of High Register
	R: Beat Histogram-73
	R: Beat Histogram-28
d = 3	M: Melodic Interval Histogram-12
	R: Beat Histogram-34
	R: Beat Histogram-29
	R: Beat Histogram-134
	P: Basic Pitch Histogram-51
	P: Basic Pitch Histogram-49
d = 4	M: Melodic Tritones
	R: Beat Histogram-52
	R: Beat Histogram-42
	P: Basic Pitch Histogram-27

Table 6: First 5 levels of the decision tree for Sadness

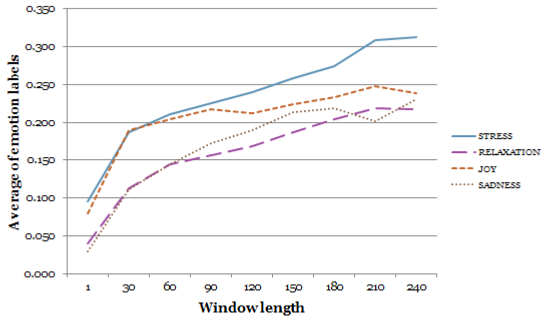

Figure 3: Average of emotion value for different window lengths

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, only 24% of the emotion labels from EEG were the same with the manual annotations, 63% 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, in most cases, 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

			n		
Label	1	30	60	90	120
class 1	84.0%	56.5%	52.3%	51.0%	49.1%
class 2	13.3%	31.6%	28.6%	26.1%	25.7%
class 3	1.9%	8.7%	15.4%	18.4%	20.7%
class 4	0.5%	1.8%	1.8%	2.3%	1.6%
class 5	0.3%	1.4%	1.9%	2.1%	2.9%

Table 7: Class sizes for Stress data after discretization

			n		
Label	1	30	60	90	120
class 1	95.3%	82.2%	80.5%	81.5%	80.5%
class 2	3.8%	9.9%	6.0%	3.7%	3.2%
class 3	0.7%	6.5%	10.3%	11.1%	9.4%
class 4	0.2%	1.0%	2.2%	2.7%	5.7%
class 5	0.0%	0.4%	1.0%	1.1%	1.1%

Table 8: Class sizes for Relaxation data after discretization

turned out not stressful at all. The emotion annotations were examined and analysis was performed to determine if dependency between the values exist.

Table 9 shows that the subject treated the emotion Stress to be the bipolar opposite of Relaxation due to the high negative correlation value. Using ESAM, a similar situation can be observed but there is only a moderate negative correlation between the two as shown in Table 10. When the other emotions are examined, Joy labels are found to have a correlation with Relaxation and a negative correlation with Stress labels. This is consistently reported for both manual annotations and annotations using ESAM.

Finally, the amount of discrepancy between manual and automated annotations were compared with the subject's familiarity with the song. 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, 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.

C4.5 was 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.

	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 9: 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 10: Correlation of annotations using ESAM

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, other metrics to summarize the emotion values for each window need to be studied.

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. Instead of high-level features, accuracy could be improved by using low-level features or a combination of both.

The window length greatly affects model accuracy. There is still need to investigate if there is a relationship between the average tempo of the song with window length. It can be hypothesized 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.

References

[Gabrielsson 03] Gabrielsson, A., Juslin, P. N: 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),

[Hall 09] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H.: The WEKA Data Mining Software: An Update, SIGKDD Explorations, vol. 11, no. 1 , pp. 10–18 (2009),

[Kim 10] Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J.,Speck, J. A., Turnbull, D.: Music emotion recognition: A state of the art review, Proceedings of the 11th International Society for Music Information Retrieval Conference, pp. 255–266 (2010),

[Legaspi 07] Legaspi, R., Hashimoto, Y., Moriyama, K., Kurihara, S., Numao, M.: Music Compositional Intelligence with an Affective Flavor, Proceedings of the 12th International Conference on Intelligent User Interfaces, pp. 216–224 (2007),

[Lin 11] Lin, Y.-C., Yang, Y.-H., Chen, H. H.: Exploiting online music tags for music emotion classification, ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 7S, no. 1, pp. 1–16 (2011),

[Livingstone 10] Livingstone, S. R., Muhlberger, R., Brown, A. R., and Thompson, W. F.: Changing musical emotion: A computational rule system for modifying score and performance, Computer Music Journal, vol. 34, no.1, pp. 41–64 (2010),

[McKay 06] McKay, C. and Fujinaga, I.: jSymbolic: A feature extractor for MIDI files, Proceedings of the International Computer Music Conference, pp. 302–305 (2006),

[Miranda 05] Miranda, E. R. and Brouse, A.: Toward direct brain-computer musical interfaces, New Interfaces for Musical Expression (2005),

[Musha 97] Musha, T., Terasaki, Y., Haque, H. A. and Ivanitsky, G. A.:Feature extraction from EEGs associated with emotions, Journal of Artificial Life and Robotics, Vol. 1, No. 1, pp. 15–19 (1997),

[Numao 97] Numao, M., Kobayashi, M. and Sakaniwa, K.: Aquisition of human feelings in music arrangement, Proceedings of IJCAI '97, pp. 268–273 (1997),

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

[Quinlan 92] Quinlan, J. R.: Learning with continuous classes, Proceedings of AI92, 5th Australian Joint Conference on Artificial Intelligence, Adams & Sterling (eds.), World Scientific, Singapore, pp. 343–348 (1992),

[Quinlan 93] Quinlan, J. R.: C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers (1993),

[Trohidis 08] Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.: Multilabel classification of music into emotions, Proceedings of 9th International Conference on Music Information Retrieval, pp. 325–330 (2008)