

# EmoMusic – Classifying Music According to Emotion

Doris Baum

Department of Software Technology and Interactive Systems  
Vienna University of Technology  
Vienna, Austria

**Abstract** Like artist and genre, mood or emotion is a useful attribute to categorise music and it would be desirable for listeners to be able to automatically organise their music according to emotion and generate mood playlists.

This paper presents a user study on the usefulness of the PANAS-X emotion descriptors as mood labels for music. It describes the attempt to organise and categorise music according to emotions with the help of different machine learning methods, namely Self-organising Maps and Naive Bayes, Random Forest and Support Vector Machine classifiers.

## 1 Introduction

As file sharing networks and online music stores become increasingly popular, there is also an increasing need for databases with meta data on digital music. Customers want to be able to select their music not only according to song name, album and artist but also according to other attributes like genre and mood. Thus an immense amount of music needs to be classified according to these attributes, preferably in an automated way.

This paper describes a study on automatic classifiability of music according to emotions. A selection of music from different genres and artists was given to a number of test subjects, who were to report their emotional connotation with the songs. The music was taken from George Tzanetakis' benchmark collection [1], the emotion categories from the PANAS-X schedule [2].

Then several Machine Learning methods were tried out, to see in how far music is automatically classifiable according to emotion:

Rhythm Pattern features [3] were extracted from the pieces, to transform the music into a form processable by the Machine Learning algorithms.

A Self-Organising Map (SOM) [4] was trained with the features to gain an overview where the emotions come to lie on the map and whether they overlap or can be separated. The details of this specific experiment will be presented in a different paper, however the results will be mentioned here.

Then, several types of classifiers were tested to evaluate the success of automatic classification: Naive Bayes [5], Random Forest [6] and Support Vector Machines [7].

The results are, as were to be expected, that different users associate different feelings with a given piece of music, but that for quite a lot of songs the majority of users agree on the connotated emotions. Also, not all the tested emotions make sense for music classification, or can automatically and reliably be derived from the rhythmic features used, but for some emotions the features give a good idea of which emotional categories the song could be placed in.

This paper describes the study conducted, the methods used and some interesting results derived. Section 2 gives an overview of related work in the subject. Section 3 describes the setup of the user study conducted, the music, emotion labels, users, questionnaire, and methods of analysis. Section 4 gives some statistics on the data, explains the experiments conducted, and the results derived from the study. Section 5 closes this paper and draws a conclusion.

## 2 Related Work

The relation between music and emotions has been addressed psychologically as early as 1936 by Hevner [8], who also created an “adjective circle” with 8 emotion categories.

Since then, a lot of studies have been devoted to the emotional aspects of music, a substantial number of them summed up by Gabrielsson and Juslin in [9]. However, as Gabrielsson and Juslin conclude, there still is no universally accepted set of adjectives or other characterisations for classifying music according to emotion.

Thayer proposed one of the psychological models for emotion: Thayer’s Model of Mood [10]. Therein, emotion is not defined as groups of adjectives but rather as a two-dimensional scale of Energy (calm - energetic) and Stress (happy - anxious).

The work of Tellegen, Watson, and Clark [11] led to the development of the PANAS-X schedule [2], which combines a dimensional scale of Positive and Negative Affect with adjective groups for a number of emotions.

Based on the vast pool of studies on emotional aspects of music, numerous groups have turned to different emotional models in the context of automated analysis of music.

Pampalk et al. [12] have provided emotional interpretations of certain parts of the Rhythm Pattern feature set. The resulting *Weather Charts* were used to describe, i.e. label areas of the Islands of Music maps, clustering pieces of music according to perceived sound similarity. Apart from pure frequency spectra characteristics, such as *low frequencies dominant*, emotional aspects such as *non-aggressiveness*, based on the ratio of low-frequency amplitude modulations in the lower bark bands, were used to characterise certain areas on the map.

Li and Oghihara [13] use thirteen adjective groups – Farnsworth’s [14] ten groups, which were in turn derived from Hevner’s, plus three their test user created – to classify emotion in music with Support Vector Machines.

Liu, Lu, and Zhang [15], on the other hand, use a two-dimensional scale with the dimensions “Stress” and “Energy” derived from Thayer’s Model of

Mood. They employ three features sets – intensity, timbre, and rhythm – and use Gaussian Mixture Models as classifiers.

Yang and Lee [16] use the Positive and Negative Affect dimensions and the emotion categories of the PANAS-X schedule [2] and the Russel [17] and Tellegen-Watson-Clark [11] models. They extract 12 standard low-level features from the audio sources, generate 12 more features by genetic algorithms, and apply Support Vector Machine regression to them. Also, the lyrics of the songs are taken into account to disambiguate the emotion of the song.

Cunningham, Downie, and Bainbridge in [18] analyse the rather neglected area of dislike, even disgust, for music pieces. From 395 responses to a survey on “the worst song ever”, they extracted, with a grounded theory approach, the most commonly named reasons why songs are “bad”. Among these are the quality of the lyrics or voice, the “earworm effect”, a dislike of the corresponding music video, perceived pretentiousness of the song, over-exposure to the music, and unpleasant personal experience associated with the piece.

### 3 User Study

#### 3.1 Music Collection

The music used in this study was based on George Tzanetakis’ benchmark collection [1]. It consists of mp3s with 30 second extracts of songs from ten different genres: Blues, Classical Music, Country, Disco, Hiphop, Jazz, Metal, Pop, Reggae and Rock. From the full benchmark collection of 1000 songs, 20 from every genre were selected randomly to make up a collection of 200 songs. They were all named uniformly so as to not give any hints concerning the genre to the test subjects. Also, the ordering of the songs was randomised to further prevent any bias when presenting them to the test subjects during a form-based evaluation session.

#### 3.2 Emotions

As previously discussed, there have been a number of studies on music and emotions, but no standard set of emotions seems to have been established. Thus, a set of emotions had to be selected, which should be founded in psychology and would prove useful in the study. The emotional categories were taken from THE PANAS-X Manual for the Positive and Negative Affect Schedule - Expanded Form by Clark and Watson [2]. This may later allow us to bring together adjective groups and two-dimensional scales, as a Positive/Negative Affect axis together with an Arousal / Activation level can form a two-dimensional model similar to Thayer’s.

The categories we use are: fear, hostility, guilt, sadness, joviality, self-assurance, attentiveness, shyness, fatigue, serenity and surprise. Each category was represented in the test questionnaires by two adjectives which the subjects could associate with the music. The adjectives were:

**Fear:** scared / nervous  
**Hostility:** angry / disgusted  
**Guilt:** guilty / ashamed  
**Sadness:** sad / lonely  
**Joviality:** happy / cheerful  
**Self-Assurance:** proud / strong  
**Attentiveness:** determined / alert  
**Shyness:** timid / shy  
**Fatigue:** tired / sleepy  
**Serenity:** calm / relaxed  
**Surprise:** surprised / astonished

There was also a twelfth “don’t know” category the users could attribute. This category, however, was later on not explicitly used, but all songs not attributed to any other category were discarded for the respective overall and per user results.

### **3.3 Test subjects / Users**

Ten subjects took part in the study, all between 20 and 40 years old, 6 male and 4 female, 9 of them with academic background (university students or graduates). Their musical expertise varies from amateurs to experts with theoretical background knowledge, playing instruments or singing, or even writing their own music. The subjects also showed a rather diverse range of preferred musical styles. Obviously, the small number of participants does not allow a purely quantitative evaluation of the study results, but requires a qualitative evaluation of this user study.

### **3.4 Online Questionnaire**

The study was conducted via a form-based on-line questionnaire. While data collection via the Internet form was basically open to the general public, most participants, and specifically the ones that the results in this paper are based upon, were personally asked to participate in the study.

The home page contained an explanation of the study for the subjects and provided a possibility to listen interactively to the individual pieces of music while ticking the check boxes of the appropriate emotional characteristics.

PHP scripts were used to automatically generate an individual music rating questionnaire for each test subject. A different randomised ordering of the songs was produced to prevent any potential bias introduced by the sequence of the pieces of music as well as the duration of the evaluation session.

### **3.5 Methods of Analysis**

The results of the study were analysed in several different manners.

**Please rate these mp3s according to the emotion(s) they convey:**

**1. emo0057.mp3**  
 scared / nervous    angry / disgusted    guilty / ashamed    sad / lonely  
 happy / cheerful    proud / strong    determined / alert    timid / shy  
 tired / sleepy    calm / relaxed    surprised / astonished    don't know / doesn't really touch me

**2. emo0007.mp3**  
 scared / nervous    angry / disgusted    guilty / ashamed    sad / lonely  
 happy / cheerful    proud / strong    determined / alert    timid / shy  
 tired / sleepy    calm / relaxed    surprised / astonished    don't know / doesn't really touch me

**3. emo0031.mp3**  
 scared / nervous    angry / disgusted    guilty / ashamed    sad / lonely  
 happy / cheerful    proud / strong    determined / alert    timid / shy  
 tired / sleepy    calm / relaxed    surprised / astonished    don't know / doesn't really touch me

**Figure 1.** Part of the questionnaire the users were asked to fill in, containing checkboxes to assign adjectives to the songs.

First of all, the variation of emotions assigned to the various titles was analysed statistically to obtain a feeling for inter-indexer consistency, i.e. in how far the test subjects agreed with each other when assigning emotions to music.

Then, Rhythm Pattern features [3] were used in the production of different Machine Learning evaluations of the data. The features are psycho-acoustically weighted descriptions of the loudness amplitude modulation in different frequency bands. Thus, they describe fluctuations in frequency regions all over the audible frequency range, not only rhythm (as the name may suggest). It seems likely that the information that induces emotion in humans is contained in this; also the Rhythm Pattern features already showed to work well for genre classification.

The pieces of music were clustered on a self-organising map according to the Rhythm Pattern features. The SOM groups the music according to sound similarity so that similar pieces of music are located next to each other. The resulting map was then labelled with the emotional labels in order to see where the emotions came to lie and if consistent regions could be identified. The SOMLib software [19] was used for this.

Finally, different classifiers were trained on the Rhythm Pattern features, to obtain an estimate of how well the emotions in question could be predicted with different machine learning methods. Naive Bayes [5], Random Forest [6] and Support Vector Machine Classifiers [7] were used. The WEKA Data Mining Software [20] was employed to do this. Because emotions may overlap and users may assign more than one emotion to a piece of music, not one classifier for all emotions was used. Instead there was one classifier per emotion, to decide whether a piece of music could be attributed this emotion or not. To get a more realistic estimate, cross-validation was employed. However, due to the small number of examples and counter-examples for the emotion classes (160 instances

at the most), further studies (possibly relying on user relevance feedback) must be done to confirm the validity of current experiments.

## 4 Results

### 4.1 Statistics

To get an overview, a small statistical evaluation was done on the collected data for all users.

The music pieces were put into one or more of the emotion categories: A song was assigned an emotion if the emotion in question had at least 50% of the votes (that is 5 or more users agreed on the respective emotion). More than one emotion could be assigned to the same song.

All in all, 153 songs were labelled with one, two, or three emotions. 97 (63%) of the songs were assigned one emotion, 50 (33%) were assigned two emotions, and 6 (4%) were assigned three emotions.

This shows that at least 50% of the users agreed on at least one emotion category for 77% (153 of 200) of the rated songs.

**Table 1.** Number of songs in the emotion categories, with 50% and 70% user agreement required

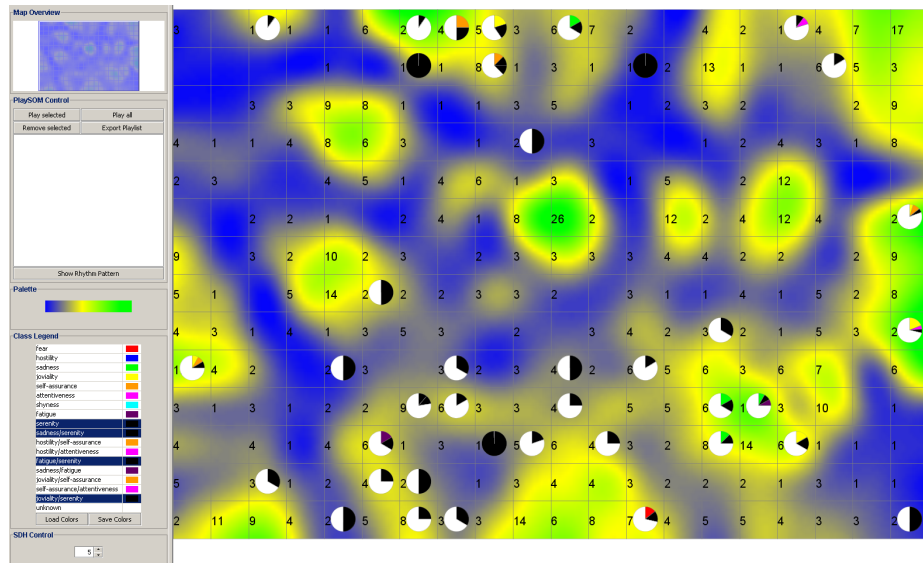
Emotion category	50%	70%
fear	4	0
hostility	22	7
guilt	0	0
sadness	25	3
joviality	41	18
self-assurance	44	20
attentiveness	14	3
shyness	2	0
fatigue	10	3
serenity	53	19
surprise	0	0

A list of how many songs were placed in each category is presented in Table 1. The first column lists the number of songs assigned to each emotional category with 50% user agreement (5 or more participants rated the song as conveying this emotion). If one requires that at least 70% of the test users agree on an emotion, 35% (70) of the songs can be labelled – the second column lists the number of songs in each category for 70% user agreement. It seems that in such a (relatively small) group, there always will be some disagreement on the emotions

connected with a song but that the majority of users can agree on at least one emotion for a high number of songs.

Taking a look at correlation of classes, one can observe that 56 songs are assigned two or three emotions; of these, 16 (29%) are assigned both sadness and serenity; 13 (23%) are assigned both hostility and self-assurance; and 9 (16%) are assigned both joviality and self-assurance.

## 4.2 SOM



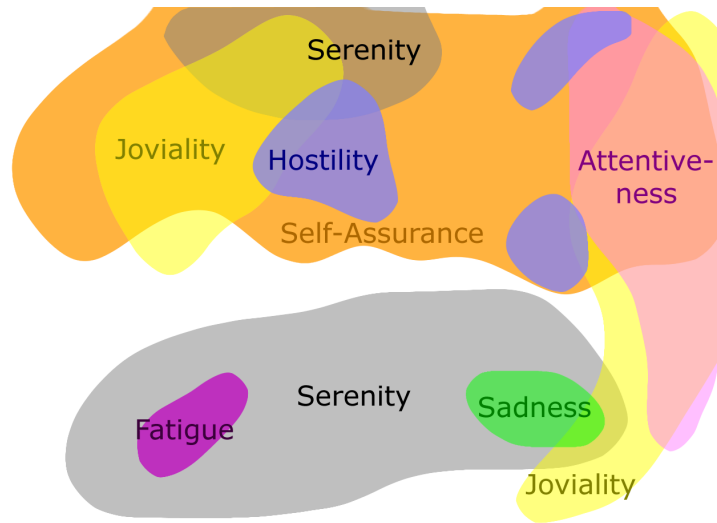
**Figure 2.** Screenshot of the SOM Viewer Software displaying serenity for all users (serenity is marked in black).

To get a picture of the results of the study, a SOM was trained on the Rhythm Pattern features of the whole music collection (1000 songs).<sup>1</sup> The SOM was then visualised with the SOMLib Viewer software, displaying the SDH Visualisation [12] which shows the sound similarity cluster structure as Islands of Music. On top of this, pie charts show for each SOM unit how many songs from the different emotion categories have been mapped onto this unit.

Figure 2 provides an example visualisation for the emotion class “serenity” for all users (serenity is marked in black).

When going through the visualisations for the emotion categories, one can detect clusters and areas of higher density of the categories. A detailed explanation of this will be published in future. As an overview of the results, a –

<sup>1</sup> The SOMs had 20 x 14 units, a learn rate of 0.75, and were trained for 2000 iterations.



**Figure 3.** Rough map of the distribution of the emotions on the SOM visualisation.

manually produced – overview of these clusters is given in Figure 3. It is just a rough sketch, rather than a density estimation, but it nonetheless permits some conclusions:

The emotions cannot be separated completely, but they overlap – as is to be expected if one takes into account that feelings mix and music carries a lot of emotional information.

However, it seems that the calmer and quieter emotions come to lie in the lower half of the visualisation and the more aroused feelings can be found in the upper half – with the exception of the “serenity” cluster in the upper half.

On the right-hand edge there seems to be an area of strong and cheerful music – the overlapping clusters of the right “joviality” cluster and the attentiveness area.

The lower “serenity” cluster encompasses “fatigue” and “sadness”, and “self-assurance” overlaps with or encloses “joviality”, “serenity”, “hostility”, and “attentiveness”.

Thus it may indeed be possible to generally derive the emotion connoted with a piece of music. To go into this further, experiments with different classifiers were conducted, as described in the next section.

### 4.3 Classifiers

To see, whether automatic emotion classification is possible with the data from the study, several test classifiers were trained in WEKA with Rhythm Pattern features of examples for the emotion categories. Naive Bayes, Random Forest

and Support Vector Machine (with Radial Basis Function Kernel) classifiers were used.

In order to be able to automatically assign several emotions to one song, a binary classifier for each emotion was built to decide whether a song belonged to this emotion or not. To obtain a complete emotion classification, a song would be run through all binary classifiers to establish all emotion categories it belongs to.

For this, the data from the user study was preprocessed into separate data-files for each emotion, each file containing examples and counter-examples for the respective category.

**Table 2.** Number of examples and counter-examples for the emotion categories.

Category	Pro-examples	Counter-examples
<i>fear</i>	4	123
hostility	22	138
<i>guilt</i>	0	134
sadness	25	109
joviality	41	65
self-assurance	44	54
attentiveness	14	52
<i>shyness</i>	2	145
fatigue	10	113
serenity	53	56
<i>surprise</i>	0	131

A piece of music was selected as an example for an emotion category if 5 or more users had associated it with this emotion. A song was selected as a *counter*-example for an emotion category, if none of the users had rated it as this emotion.

Table 2 shows an account of how many examples and counter-examples each of the emotion category data files contained. From it, it can be seen, that the categories “fear”, “guilt”, “shyness”, and “surprise” contain under 5 pro-examples each. This can mean several things: a) These emotions are not (generally) applicable to music; b) There is a very low user agreement as to which songs represent these categories; c) There were almost no songs representing these emotions in the music collection used.

Through communication with the test subjects, I got the impression that some think a) to be case (“What is a guilty song supposed to be?”). Also, when listening to the music myself, I found very few examples for (in my opinion) fearful music.

However, low user agreement seems to be the case for all four categories: When counting the number of songs attributed with one of these categories by at least *one* user, each of the four is assigned to more than 50 songs. (Fear gets 77 songs, guilt 66, shyness 55, and surprise 69.)

In any case, it is quite clear that with less than 5 examples for a category, a classifier cannot be sensibly built. For the sake of completeness, classifiers were built for these emotions as well, but it is later taken into account that they skew the estimate for the overall precision and recall.

**Table 3.** Results of 10-fold cross-validations, according to emotions and types of classifiers.

	<b>NaiveBayes</b>		<b>RandomForest</b>		<b>SVM</b>	
	Precision	Recall	Precision	Recall	Precision	Recall
fear	0.500	0.250	0.333	0.250	0.000	0.000
no-fear	0.976	0.992	0.976	0.984	0.969	1.000
hostility	0.391	0.409	0.500	0.182	0.000	0.000
no-hostility	0.905	0.899	0.882	0.971	0.863	1.000
guilt	0.000	0.000	0.000	0.000	0.000	0.000
no-guilt	1.000	1.000	1.000	1.000	1.000	1.000
sadness	0.452	0.760	0.571	0.480	0.000	0.000
no-sadness	0.935	0.789	0.885	0.917	0.813	1.000
joviality	0.692	0.659	0.659	0.707	0.000	0.000
no-joviality	0.791	0.815	0.806	0.769	0.613	1.000
self-assurance	0.739	0.773	0.796	0.886	0.885	0.523
no-self-assurance	0.808	0.778	0.898	0.815	0.708	0.944
attentiveness	0.571	0.571	0.667	0.429	0.000	0.000
no-attentiveness	0.885	0.885	0.860	0.942	0.788	1.000
shyness	0.000	0.000	0.000	0.000	0.000	0.000
no-shyness	0.986	1.000	0.986	0.993	0.986	1.000
fatigue	0.467	0.700	0.667	0.600	0.000	0.000
no-fatigue	0.972	0.929	0.965	0.973	0.919	1.000
serenity	0.737	0.792	0.796	0.811	0.784	0.755
no-serenity	0.788	0.732	0.818	0.804	0.776	0.804
surprise	0.000	0.000	0.000	0.000	0.000	0.000
no-surprise	1.000	1.000	1.000	1.000	1.000	1.000
<b>Macro-average</b>	0.663	0.670	0.685	0.660	0.505	0.547
<b>Un-skewed macro-average</b>	0.724	0.749	0.769	0.735	0.511	0.573

A ten-fold cross-validation was done for each emotion category and each of the types of classifier: Naive Bayes, Random Forest and Support Vector Machine. Table 3 contains the resulting precision and recall measures for each category and anti-category (e.g. “fear” and “no-fear”), as well as a macro-average of precision and recall for each type of classifier. The last row shows an “un-skewed” macro-average, for the calculation of which precision and recall of the underrepresented categories (“fear”, “guilt”, “shyness”, and “surprise” and their respective anti-categories) were not taken into account.

This shows that emotion classification – at least on the current data – can indeed be quite successful, especially when regarding that the majority of test users also “only” agreed on one or more emotions for about 75% of the songs. Of the three types, a Naive Bayes or Random Forest classifier seems best suited for the data at hand, with un-skewed precision of 0.724 and 0.769 respectively and un-skewed recall of 0.749 and 0.735 respectively. One may prefer Naive Bayes because it takes less computation time to build the model.

The Support Vector Machine classifier seems not to work well for the datasets at hand – which may be improved upon by better pre-processing, fine-tuning, and choosing another kernel type. But considering that an SVM model takes much longer to build than a Naive Bayes classifier and that the latter seems to work quite well, one might want to stick with the Bayes classifier.

To give more significance to these results, more annotated training data (pro- and counter-examples) would have been desirable, but this data has first to be collected in further user studies – possibly by using a relevance feedback system fed with predictions of the current classifiers. One might for example run the Naive Bayes classifier on as-yet un-annotated songs and present those pieces to users which are assigned an emotion with a high probability by the classifier. The user could then be asked to confirm or reject the automatic classification.

## 5 Conclusions

From this study, it can be concluded that emotions may very well be derivable in an automatic way, although the procedure certainly can be refined further.

Though different users associate different feelings with a given piece of music, the majority of users agree on the connoted emotions for quite a lot of songs.

Clusters of different emotions appear on the SOM visualisations, showing sensible correlation between emotions for music in the various regions. Emotions perceived as related are also located in neighbouring map regions.

Naive Bayes and Random Forest classifiers can – for example – be used to predict the emotion of a piece of music with reasonable success.

Still, the emotion categories could to be reconsidered – and unnecessary or unused categories, such as “guilt”, “shyness”, “surprise” and maybe even “fear” may be omitted in further studies. This, however, would move the categories away from the PANAS-X schedule and contribute to the problem, that almost

every study on emotion and mood in music uses its own set of emotion descriptors.

Also, with something so subjective as emotion, it might be more promising to build individual emotion classifiers for each listener than to try and derive a general notion of what song belongs to which emotional class.

## References

1. Tzanetakis, G.: Manipulation, Analysis and Retrieval Systems for Audio Signals. PhD thesis, Princeton University, Department of Computer Science (2002) <http://www.cs.uvic.ca/~gtzan/work/pubs/thesis02gtzan.pdf>.
2. Watson, D., Clark, L.A.: The PANAS-X Manual for the Positive and Negative Affect Schedule - Expanded Form. The University of Iowa (1994) <http://www.psychology.uiowa.edu/faculty/Clark/PANAS-X.pdf>.
3. Rauber, A., Pampalk, E., Merkl, D.: Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by musical styles. In: Proceedings of the 3rd International Symposium on Music Information Retrieval, Paris, France (2002) 71–80 <http://www.ifs.tuwien.ac.at/ifs/research/publications.html>.
4. Kohonen, T.: Self-Organizing Maps. 3rd edn. Volume 30 of Springer Series in Information Sciences. Springer, Berlin (2001)
5. Elkan, C.: (Naive bayesian learning)
6. Breiman, L.: Random forests. In: Machine Learning. Volume 45. (2001) 5–32
7. Boser, B.E., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Computational Learning Theory. (1992) 144–152
8. Hevner, K.: Experimental studies of the elements of expression in music. In: American Journal of Psychology. Volume 48. (1936) 246–268
9. Gabrielsson, A., Juslin, P.N.: Emotional expression in music. In Davidson, R.J., Scherer, K.R., Goldsmith, H.H., eds.: Handbook of Affective Sciences, Oxford University Press (2002) 503–534
10. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford University Press (1989)
11. Tellegen, A., Watson, D., Clark, L.A.: On the dimensional and hierarchical structure of affect. In: Psychological Science. Volume 10. (1999) 297–303
12. Pampalk, E., Rauber, A., Merkl, D.: Content-based organization and visualization of music archives. In: Proceedings of ACM Multimedia 2002, Juan-les-Pins, France, ACM (2002) 570–579 <http://www.ifs.tuwien.ac.at/ifs/research/publications.html>.
13. Li, T., Ogihara, M.: Detecting emotion in music. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003), Baltimore, Maryland (USA) (2003)
14. Farnsworth, P.R.: The social psychology of music. Dryden Press (1958)
15. Liu, D., Lu, L., Zhang, H.J.: Automatic mood detection from acoustic music data. In: Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003), Baltimore, Maryland (USA) (2003)
16. Yang, D., Lee, W.: Disambiguating music emotion using software agents. In: Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004), Barcelona, Spain (2004)

17. Russell, J.A.: Core affect and the psychological construction of emotion. In: *Psychological Review*. Volume 110. (2003) 145–172
18. Cunningham, S.J., Downie, J.S., Bainbridge, D.: "the pain, the pain": Modelling music information behavior and the songs we hate. In: *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK (2005) 474–477
19. Neumayer, R., Dittenbach, M., Rauber, A.: Playsom and pocketsomplayer: Alternative interfaces to large music collections. In: *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK (2005) 618–623
20. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*. 2nd edn. Morgan Kaufmann, San Francisco (2005)