# Music mood recognition: State of the Art Review

**MUS-15**

Sebastian Napiorkowski

July 10, 2015

## 1 Introduction

There is a saying that goes, music is what feelings sound like. Not only common sense suggests a strong bond between music and emotion but also researchers refer to music as a language of emotion [1] and discuss this relationship in many different disciplines, including philosophy, musicology, psychology etc. [2]. Kim, Youngmoo E. et al. (2010) states in [3] that automatic recognition of emotions (or mood) as a music information retrieval task is in its early stages. Against this background, the central question that motivates this seminar paper is: How automatic music classification is done? The paper will introduce foundations for the classification of mood, show general ideas of competing solutions for automatic audio based mood recognition and dive then into the hybrid approach of [4].

For the purposes of this report, the terms emotion and mood will be taken to be interchangeable, as [3] suggests. Even though [5] indicates that the two terms have slight differences in music psychology and Music Information Retrieval (MIR). In music psychology the term mood "refers to a long lasting and stable emotional state" and the term emotion is used for "very subjective responses to music which can be acute, momentary and fast changing, while the MIR community tries to find the common affective consequences of music that are shared by many people and are less volatile." [5].

## 2 Mood definition and quantification

Emotions are a complex psychological process and therefore measuring them requires some careful considerations. Whether it is measured in form of self-reports or direct biophysical indicators, one must consider the source of emotion(s) being measured, as emotion(s) can be *expressed* or *induced* by music [3, 6]. In this report the presented solutions focus on emotions expressed, rather than induced, by music. Besides that one must consider the possibility of cultural differences in the perception of mood in music. Fortunately cross-cultural studies of musical power show "that there may be universal psychophysical and emotional cues that transcend language and acculturation" [3, 7]. In addition to that Fritz, Thomas, et al. shows in a noteworthy ethnomusicology study that people without exposure to Western music categorized music examples into three categories of emotion in the same way as western people [8]. Also, the representation of emotion (mood) has to be considered. Paper [3, 5] suggests that emotion models can be generally classified in two categories: categorical descriptions and parametric models.

### 2.1 Categorical models

Categorical models organize mood usually in mood spaces/clusters, which consist of a set of discrete mood categories. One of the earliest and best-known attempts to formalize mood was by Hevner [9] which dates back to 1936 [5]. As shown in figure 1 Hevner used 66 adjectives

which are organized in 8 clusters. These clusters are arranged in a circle such that neighboring clusters express similar moods and opposite ones exhibit the most difference.

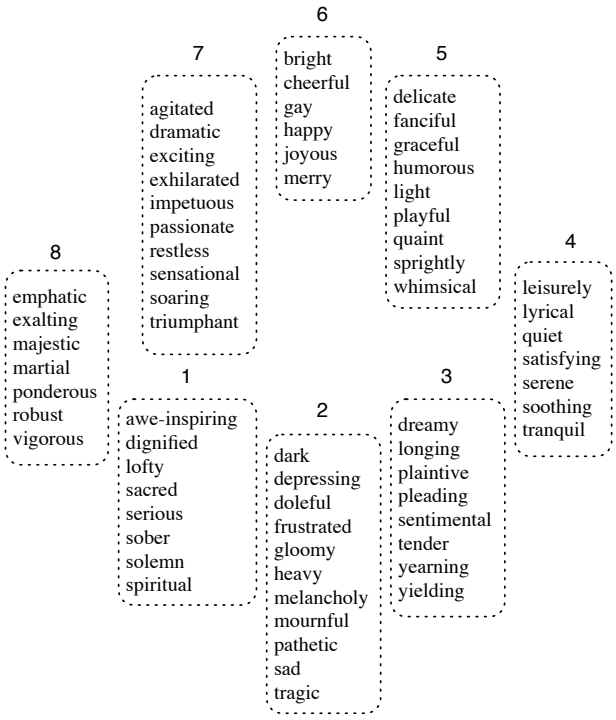


Figure 1: Hevner's adjective cycle

An other categorical model was created for the *Music Information Research Evaluation eXchange* (MIREX), an annual evaluation campaign for MIR algorithms which is hosted by the *International Symposium on Music Information Retrieval* (ISMIR). The first mood classification task was formulated 2007 and Hu, Xiao, et al. [10] proposed the same year the mood clusters shown in Table 1 as a recommendation for future MIREX mood classification tasks. These 5 clusters where "derived by performing clustering on a co-occurrence matrix of mood labels for popular music from the All Music Guide[1]" [3].

| Nr. | MIREX Mood Clusters |
|---|---|
| 1 | Passionate, Rousing, Confident, Boisterous, Rowdy |
| 2 | Rollicking, Cheerful, Fun, Sweet, Amiable/Good natured |
| 3 | Literate, Poignant, Wistful, Bittersweet, Autumnal, Brooding |
| 4 | Humorous, Silly, Campy, Quirky, Whimsical, Witty, Wry |
| 5 | Aggressive, Fiery, Tense/Anxious, Intense, Volatile, Visceral |

Table 1: MIREX Mood Clusters

## 2.2 Parametric models

In comparison to the categorical models, parametric models follow an entire other approach, as they measure mood in a continuous multidimensional space or simple multidimensional metrics and not in discrete categories. Well-known is the *Valence-Arousal* Model proposed by Russel (1980) [11] shown in figure 2. This model designates one axis to represent the *Arousal* level, which denotes the intensity in form of high *(active)* and low values *(inactive)* and the other axis to represent *Valence*, which is an appraisal of polarity ranging from positive *(happy)* to negative *(unhappy)*.

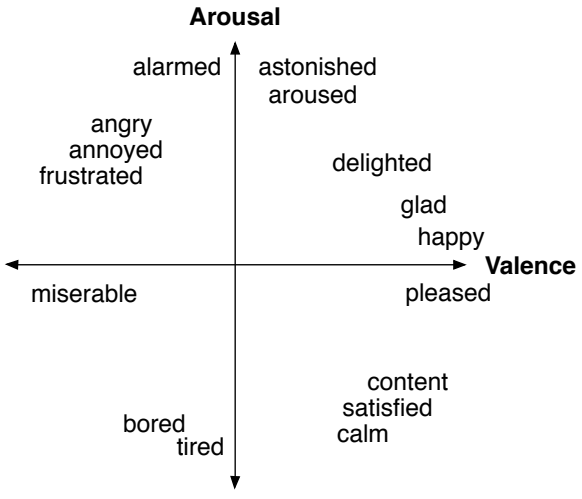


Figure 2: Valence-Arousal Model

Some studies argue that the VA-Model is incomplete and introduce a third dimension like "tension" [12], "kinetics" [13] or "dominance" [14]. Which in essence demonstrates that there is cur-

[1] see section 4.1.1: All Music Guide – http://www.allmusic.com

rently no consensus among researchers regarding the optimal amount of dimensions. A well cited study [15] even lists nine dimensions.

Thayer (1989) constructed in [16] a very similar model where arousal corresponds to energy and valence to stress.

# 3 Music mood recognition

Music mood recognition is usually defined as a regression problem or classification problem where one tries to annotate a music piece with a set of moods. Kim, Youngmoo E., et al. states in [3] that a "*music piece* might be an entire song, a section of a song (e.g., chorus, verse), a fixed-length clip (e.g., 30-second song snipet [*sic*]), or a short-term segment (e.g., 1 second)."

The mood annotation task is approached essentially in three different ways. The first approach is very straightforward as one can use humans to annotate tracks by using surveys, social tagging mechanisms or annotation games. The second approach is to use contextual text information, which means *e.g.*, mining web documents, analyzing lyrics and social tags. And finally the third approach is to evaluate the content itself, thus one can use signal processing to detect and analyze audio features. These methods can also be combined in to hybrid methods, to improve overall accuracy [3].

Many studies follow a common pattern, first they decide which kind of approach they want to use (content analysis, social tags, etc.) and which kind of mood model/s. Then they acquire the data, prepare it, extract features and use usually a supervised machine learning technique or a combination of such techniques on these features, to be finally able to recognize mood in the *music piece*.

Reviewing the research on mood recognition displays different often re-occurring supervised machine learning techniques[2]:

- Support Vector Machines (SVM) [4] [17] [18] [19] [20]
- Gaussian Mixture Models (GGM) [21] [22]
- Support Vector Regression (SVR) [20] [23] [24] [25]
- Neural Networks [26]
- Label propagation (Semi-supervised) [27]
- Linear regression: Partial Least Squares [25] [28], Multiple Linear Regression [20] [25]
- Naive Bayes [4]

In addition to the list, all recent submissions to the MIREX mood classification task in 2014 use SVMs [29] [30] [31] [32] [33] except for one approach which uses neural networks [34].

Admittedly, this list is far from complete, but one can see a trend as the common approaches are on the one hand SVMs and GMMs for solving the classification problem and on the other hand SVRs and Linear Regression techniques for solving the regression problem. This conclusion is consistent with Wu, Bin, et al. (2014) [35] findings.

# 4 Hybrid Approach based on Bischoff et al. [4]

As stated before, one major approach in mood recognition is to use SVMs, to narrow this report we will focus now on: Bischoff et al. [4] which makes use of two classifiers and two approaches:

- audio features classified by a SVM
- social tags classified by an Naives Bayes Classifier

And is thus using a hybrid approach to detect mood and themes[3] in music and makes therefore a good example for both techniques. Additional to that, this paper was selected because it presents the task in its eternity, giving a blueprint for the problem and pointing out

[2] This list is not exhaustive.

[3] refers to context or situations which fit best when listening to the track (e.g. party time, christmas, at the beach)

caveats associated with supervised learning solutions.

## 4.1 Datasets

As Bischoff et al. is aiming for a solution using supervised machine learning, they search for a ground truth dataset. After acquiring it, Bischoff et al. searched for tags and audio files for every track in the set. Reducing it to a subset if necessary.

### 4.1.1 Ground Truth Dataset

The term ground truth refers in machine learning to a highly accurate training data set.

Bischoff et al. used collected moods and themes from *AllMusic.com* pages, a website created 1995 for music fans. This site features reviews from music experts with inter alia manually assigned moods and themes. Therefore Bischoff et al. concluded it would form "a good ground truth corpus". They found 178 different moods and 73 Themes and 5770 songs with assigned mood or theme. In detail there were 8158 track-mood assignments (avg. 1.73 moods, max. 12 moods) and 1218 track-theme assignments (avg. 1.21 themes, max. 6 themes).

### 4.1.2 Social Tags

For the acquisition of social tags they used the popular site *Last.fm*, the dataset consisted of the tags and the corresponding frequencies. As not every track had social tags, they reduced the dataset to 4737 tracks. For this set of songs they collected in total 59525 tags. These tags were assigned by users and therefore have to be taken with caution.

Lamere and Celma lists in [36] the caveats associated with social tags: multiple spellings of tags, malicious tagging, sparsity due to the cold-start problem, popularity bias, ad-hoc labeling techniques, etc.

According to a previous study [37] led by Bischoff too, "the majority of the generally accurate and reliable user tags on Last.fm fall into the genre category (60%). Considerably less frequent are tags referring to moods/opinions/qualities (20%) or themes/context/usage (5%) of the music songs" [4].

They chose to use feature vectors that "have as many elements as the total number of distinct tags assigned to the songs belonging to the mood classes. The elements of a vector will have values depending on the frequency of the tags occurring along with the song" [4]. Each element $t_j$ of the feature vector $F_t = \{t_j | t_j \in T\}$ where $T$ denotes the set of tags from all songs in all mood classes, is composed as following:

$$t_j = \begin{cases} log(freq(t_j) + 1) & \text{if } song \text{ has tag } t_j \\ 0 & \text{otherwise.} \end{cases}$$

### 4.1.3 Audio

They used 30 second long audio file excerpts encoded in MP3 with 192 kbps, which corresponds to the in section 3 defined *music piece*. From these tracks they extract different state-of-the-art MIR audio features like timbral, tonal, rhythmic including MFCCs, BPM, chroma features, spectral centroid and others. They refer to [38] for a complete list, besides that Table 2 lists common features used for mood recognition [3].

| Type | Features |
|---|---|
| Dynamics | RMS energy |
| Timbre | Mel-frequency cepstral coefficients (MFCCs), spectral shape, spectral contract |
| Harmony | Roughness, harmonic changes, key clarity, maharanis |
| Register | Chromagram, chroma centroid and deviation |
| Rhythm | Rhythm strength, regularity, tempo, beat histograms |
| Articulation | Event density, attack slope, attack time |

Table 2: Common audio feature types for emotion recognition according to [3]

Each feature was extracted from 200ms frame-windows and then summarized with their component wise means and variances for better suitability with the SVM classifier. This process generates at the end a 240-dimensional feature vector consisting of low-level and mid-level audio features.

### 4.2 Classification

They reduced the dataset a second time to tracks which fall exactly in one mood category, cutting it to 1192 distinct songs. To get a balanced cluster size for the multiclass classifiers, they used 200 instances per cluster.

They experimented with the MIREX mood clusters and the Thayer energy-stress model [16], which is very similar to the Russel Model from section 2.2. However, they didn't use it as a continuous model but created 4 categories corresponding to each quadrant in the graph.

Bischoff et al. experimented with classifiers, different other features and also how to combine the results from classification of social tags and audio features. After several experiments they decided to use SVM classifiers with Radial Basis Function (RBF), as this kernel performed best. Stating that it "outperformed Logistic Regression, Random Forest, GMM, K-NN and Decision Trees" [4]. For tag features they concluded that Naive Bayes Multinomial performed best and for the combination they used a linear combination of the separate classifiers output.

The total number of songs in the training set was split in a set of songs for training and a set of songs for testing the classifiers learned model. Then the already described features where calculated. In the next step the classifiers where trained and tested. Because the classifiers where trained for the whole set of classes (*e.g.* moods) they produced a set of probability distributions and therefore in the last step the highest probability was considered for the assignment of the track to the related class.

### 4.3 Results

Table 3 shows their results on the mood task, using the metrics Precision (P), Recall (R), F1-Measure (F1) and Accuracy (Acc). The $\alpha$-value denotes the weight in which the audio-based classifier results or respectively the socialtags-based results are considered in the linear combination. A $\alpha$-value greater than 0.5 indicates a higher weight for the audio-based results and $\alpha$-values lower than 0.5 a higher weight for the socialtags-based results. After analyzing the influence of the $\alpha$-value on the combined results they concluded that $\alpha = 0.7$ works best for the MIREX part and $\alpha = 0.8$ for the Russel/Thayer Model. The reason for this choice is that Naive Bayes is known for producing probabilities close to 1 for likely cases and for the rest probabilities closer to 0. In contrast to Naive Bayes the SVM produces more even probability distributions, therefore the assigned *alpha*-weights even out the Naive Bayes probabilities.

| Classifier | Class | R | P | F1 | Acc |
|---|---|---|---|---|---|
| SVM (audio) | Mood MIREX | 0.450 | 0.442 | 0.420 | 0.450 |
| NB (tags) | Mood MIREX | 0.565 | 0.566 | 0.564 | 0.565 |
| Comb ($\alpha = 0.7$) | Mood MIREX | 0.575 | 0.573 | 0.572 | 0.575 |
| SVM (audio) | Mood THAYER | 0.517 | 0.515 | 0.515 | 0.517 |
| NB (tags) | Mood THAYER | 0.539 | 0.542 | 0.539 | 0.539 |
| Comb ($\alpha = 0.8$) | Mood THAYER | 0.570 | 0.569 | 0.569 | 0.569 |

Table 3: Results for the MIREX and THAYER part in [4]

One can see that the results from the socialtags-based classifier strictly dominate the results of the pure audio-based classifier. Yet, combining both types results in improved overall results. Therefore it shows the feasibility of hybrid approaches and that they can have advantages over solo approaches.

They state that it is difficult to compare their results with other related work, as each paper uses it's own models. In comparison to the submitted work to MIREX, they achieve lower results. Nonetheless, they knew that a paper [39] submitted 2007 used the same algorithm and performed better, achieving 60.5% accuracy. Despite the fact that they obtained lower results, through this incident they had an insight about the ground truth dataset. Their hypothesis is

that the difference in accuracy came from not filtering the training and test instances using listeners.

## 5 Conclusion

*1.* Emotions are fuzzy, therefore there exist no straightforward model to represent them. Hence, many paper use their own model and this makes it difficult for comparisons among researchers in this field.

*2.* While reviewing papers one can notice that the vast majority of approaches assume one mood per track, which does not corresponds to the reality as Wu, Bin, et al. [35] stated.

*3.* Also, the vast majority is using SVMs or GMMs for the classification problem and SVRs and Linear regression for the regression problem, and these methods are according to [35] only useful on a small number (four to eight) of music emotion categories. Wu, Bin, et al. presents an other solution by applying Multi-label Multi-layer Multi-instance Multi-view Learning[35].

*4.* My hypothesis is that many of these solutions were originally created for genre classification tasks, in this case the assumption to have one track in one genre category seems more appropriate.

*5.* In case of the hybrid approach presented in section 4 one can notice that they did not emphasize the selection of audio features, which could mean that they followed the premise – the more the better. It would be interesting if careful considerations of which audio-features to use would result in better accuracy.

Furthermore, a future task could be to use audio features based on the predicted genre of the given track.

## References


[1] Pratt, C. C. *Music as the language of emotion.* The Library of Congress, (1952)

[2] Yang, Yi-Hsuan, and Homer H. Chen. *Music emotion recognition.* CRC Press, 2011.

[3] Kim, Youngmoo E., et al. *Music emotion recognition: A state of the art review.* Proc. ISMIR. 2010.

[4] Bischoff, Kerstin, et al. *Music Mood and Theme Classification - a Hybrid Approach.* ISMIR. 2009.

[5] Hu, Xiao. *Music and mood: Where theory and reality meet.* (2010).

[6] Lee, Jin Ha, Trent Hill, and Lauren Work. *What does music mood mean for real users?.* Proceedings of the 2012 iConference. ACM, 2012.

[7] C. McKay, *Emotion and music: Inherent responses and the importance of empirical cross-cultural research* Course Paper. McGill University, 2002.

[8] Fritz, Thomas, et al. *Universal recognition of three basic emotions in music.* Current biology 19.7 (2009): 573-576.

[9] Hevner, Kate. *Experimental studies of the elements of expression in music.* The American Journal of Psychology (1936): 246-268.

[10] Hu, Xiao, and J. Stephen Downie. *Exploring Mood Metadata: Relationships with Genre, Artist and Usage Metadata.* ISMIR. 2007.

[11] J. A. Russell, *A circumspect model of affect* Journal of Psychology and Social Psychology, vol. 39, no. 6, p. 1161, 1980.

[12] Eerola, Tuomas, Olivier Lartillot, and Petri Toiviainen. *Prediction of Multidimensional Emotional Ratings in Music from Audio Using Multivariate Regression Models.* ISMIR. 2009.

[13] Mion, Luca, and Giovanni De Poli. *Score-independent audio features for description of music expression.* Audio, Speech, and Language Processing, IEEE Transactions on 16.2 (2008): 458-466.

[14] Mehrabian, Albert, and James A. Russell. *An approach to environmental psychology.* the MIT Press, 1974.

[15] Asmus, Edward P. *The development of a multidimensional instrument for the measurement of affective responses to music.* Psychology of Music 13.1 (1985): 19-30.

[16] R. E. Thayer *The biopsychology of mood and arousal.* Oxford University Press, 1989.

[17] T. Li and M. Ogihara, *Detecting emotion in music.* in Proc. of the Intl. Conf. on Music Information Retrieval, Baltimore, MD, October 2003.

[18] G. Tzanetakis, *Marsyas submissions to MIREX 2007.* MIREX 2007.

[19] C. Cao and M. Li, *Thinkit's submissions for MIREX2009 audio music classification and similarity tasks.* ISMIR, MIREX 2009.

[20] E. M. Schmidt, D. Turnbull, and Y. E. Kim, *Feature selection for content-based, time-varying musical emotion regression.* in MIR 10: Proc. of the Intl. Conf. on Multimedia Information Retrieval, Philadelphia, PA, 2010, pp. 267274.

[21] L. Lu, D. Liu, and H. J. Zhang, *Automatic mood detection and tracking of music audio signals.* IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 1, pp. 518, 2006.

[22] G. Peeters, *A generic training and classification system for MIREX08 classification tasks: Audio music mood, audio genre, audio artist and audio tag.* MIREX 2008.

[23] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, I.-B. Liao, Y.-C. Ho, and H. Chen, *Advances in Multimedia Information Processing.* - PCM 2008, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, December 2008, ch. 8, pp. 7079.

[24] B.Han, S.Rho, R.B.Dannenberg, and E.Hwang, *SMERS: Music emotion recognition using support vector regression.* in Proc. of the Intl. Society for Music Information Conf., Kobe, Japan, 2009.

[25] E. M. Schmidt and Y. E. Kim, *Prediction of time-varying musical mood distributions from audio.* in Proc. of the Int. Society for Music Information Conf., Utrecht, Netherlands, August 2010.

[26] Y. Feng, Y. Zhuang, and Y. Pan: *Popular music retrieval by detecting mood.* SIGIR, 2003.

[27] Wu, Bin, et al. *SMART: Semi-supervised music emotion recognition with social tagging.* SIAM International Conference on Data Mining. 2013.

[28] T.Eerola, O.Lartillot, and P.Toiviainen, *Prediction of multidimensional emotional ratings in music from audio using multivariate regression models.* in Proc. of the Intl. Society for Music Information Conf., Kobe, Japan, 2009.

[29] Seung-Ryoel Baek and Moo Young Kim, *Music Genre Classification MIREX 2014 submissions* MIREX 2014.

[30] Shumin Xu and Yating Gu, *Music Genre/Mood/Composer Classification: MIREX 2014 submissions.* MIREX 2014.

[31] R. Panda et al. *MIREX 2014: Mood Classification tasks submission.* MIREX 2014.

[32] Ming-Ju Wu and Jyh-Shing Roger Jang, *Confidence-based late Fusion for Music Genre Classification.* MIREX 2014.

[33] Klaus Seyerlehner and Markus Schedl *MIREX 2014: Optimizing the fluctuation pattern extraction process.* MIREX 2014.

[34] Qiuqiang Kong and Xiaohui Feng *Music Genre/Mood/Composer Classification: MIREX 2014 submissions* MIREX 2014.

[35] Wu, Bin, et al. *Music emotion recognition by multi-label multi-layer multi-instance multi-view learning.* Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[36] P. Lamere and O. Celma, *Music recommendation tutorial notes,* ISMIR Tutorial, September 2007.

[37] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu *Can all tags be used for search?,* CIKM, pp. 193202, 2008.

[38] C. Laurier and P. Herrera *Automatic detection of emotion in music: Interaction with emotionally sensitive machines.* Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence, pp. 932, 2009.

[39] C. Laurier and P. Herrera: *Audio music mood classification using support vector machine.* MIREX Audio Music Mood Classification contest, ISMIR, 2007.