Music Mood Annotator Design and Integration

Cyril Laurier, Owen Meyers, Joan Serrà, Martín Blech, Perfecto Herrera
Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
{cyril.laurier,owen.meyers,joan.serraj,martin.blech,perfecto.herrera}@upf.edu

Abstract

A robust and efficient technique for automatic music mood annotation is presented. A song's mood is expressed by a supervised machine learning approach based on musical features extracted from the raw audio signal. A ground truth, used for training, is created using both social network information systems and individual experts. Tests of 7 different classification configurations have been performed, showing that Support Vector Machines perform best for the task at hand. Moreover, we evaluate the algorithm robustness to different audio compression schemes. This fact, often neglected, is fundamental to build a system that is usable in real conditions. In addition, the integration of a fast and scalable version of this technique with the European Project PHAROS is discussed.

1. Introduction

Psychological studies have shown that emotions conveyed by music are not so subjective that they are invalid for mathematical modeling [1, 7, 9, 18]. Moreover, Vieillard et al. [25] demonstrated that within the same culture, the emotional responses to music can be highly consistent. These results indicate that modeling emotion or mood in music is feasible.

In the past few years, to deal with the huge amount of digital music available, research in Music Information Retrieval has been very active in a wide variety of topics [16]. Recently, mood classification of music has become a topic of interest. It has been shown that, if we restrict the problem to simple categories or dimensions, we can achieve satisfying results. The aim of this work is to create a robust mood annotator that estimates the mood of a music directly from the raw audio data. We use supervised learning classification methods to achieve this goal. In Section 2, we report on the related works in music mood classification. In Section 3, we detail the method and the results we achieve. In Section 4 we expose the integration of this technique in the PHAROS project. Finally, in Section 5, we discuss future work, to obtain better results and user experience.

2. Scientific Background

Although there exist several studies dealing with audio mood classification (like [15, 14, 21, 29]), almost every work differs in the way that it represents the mood concepts. Similarly to psychological studies, there is no real agreement on a common model. Some consider a categorical representation based on basic emotions [10, 23, 20, 13], while others prefer a multi-labeling approach like Wieczorkowska [26]. The latter is more difficult to evaluate since they consider many categories. The basic emotion approach gives simple but relatively satisfying results with accuracies around 80-90%, depending on the data and the number of categories chosen (usually between 3 and 5). Li and Ogihara [11] extract timbre, pitch and rhythm features to train Support Vector Machines (SVM). They consider 13 categories, 11 from Farnsworth [3] and 2 additional ones. However, the results are not that convincing, with low precision (around 0.32) and recall (around 0.54). This might be due to a small dataset labeled by only one person and to the large number of categories they chose. Alternatively, we prefer to use few categories and a ground truth annotated by hundreds of people.

Other works use the dimensional representation (modeling emotions in a space), like Yang [28]. They model the problem with Thayers arousal-valence[1] emotion plane [24] and use a regression approach (Support Vector Regression), to learn each of the two dimensions. They extract mainly spectral and tonal descriptors together with loudness features. The overall results are very encouraging and demonstrate that a dimensional approach is also feasible. In another work, Mandel et al. [15] designed an active learning

[1] In psychology, the term valence describes the attractiveness or aversiveness of an event, object or situation. For instance happy and joy have a positive valence and anger and fear a negative valence.

system using Mel-Frequency Cepstral Coefficients (MFCCs [12]) and SVM, which learns according to the feedback given by the user. Moreover, the algorithm chooses the examples to be labeled in a smart manner, reducing the amount of data needed to build a model with an accuracy comparable to standard methods.

The comparison of these different techniques is very difficult. With the objective to evaluate different algorithms within the same framework, MIREX (Music Information Retrieval Evaluation eXchange) [2] organized a first task on Audio Mood Classification in 2007. MIREX is a reference in the Music Information Retrieval community that provides a solid evaluation of current algorithms in different tasks. The MIREX approach is similar to the Text Retrieval Conference (TREC)[2] approach to the evaluation of text retrieval systems. For the Audio Mood Classification task, it was decided to model the problem with a categorical representation in mood clusters (a word set defining the category). The clusters were chosen to be mutually exclusive (one instance could only belong to one mood cluster). In that aspect, it is similar to a basic emotion approach. There were five mood clusters and the best results achieved were around 60% of accuracy [10, 6]. The results from the MIREX participants are lower than for the approaches described previously. This might be due to a semantic overlap between the different clusters [6]. Indeed, if the categories are mutually exclusive, the category labels have to be chosen carefully.

3. Method

To classify music by mood, we frame it as an audio classification problem using a supervised learning approach. We consider clear categories to allow a greater agreement between people (human annotators and end-users). We build the ground truth to train our system on both social network knowledge (wisdom of crowds) and experts validation (wisdom of the few). Then we extract a rich set of audio features that we describe in 3.2. We employ standard feature selection and classification techniques and we evaluate them in 3.3. Once the best algorithm is chosen, we evaluate its robustness as reported in 3.4.

3.1. Ground Truth from wisdom of crowds and wisdom of the few

For this study we use a categorical approach to represent the mood. We focus on the following categories: happy, sad, angry, relaxed. We decided on these categories because these moods are related to basic emotions from psychological theories (reviewed in [8]) and they cover the four quadrants of the 2D representation from Russell [19] with valence and arousal dimensions. The Russell's 2D model (called "circumplex model of affect") is a reference widely accepted and cited in psychological studies about emotions. In this space, "happy" and "relaxed" are with positive valence and respectively with high and low arousal. "Angry" and "sad" have negative valence and respectively high and low arousal. As we do not want to be restricted to exclusive categories, we consider the problem as a binary classification for each mood. One song can be "happy" or "not happy", but also independently "angry" or "not angry" and so on.

The main idea is to exploit information extracted from both a social network and several experts validating the data. To do so, we have pre-selected the tracks to be annotated using last.fm[3] tags. Last.fm is a music recommendation website with a large community that is very active in associating labels (tags) with the music they listen to. These labels are then available to all users in the community. For each mood category we have generated a set of close semantic words using Wordnet[4] and looked for the songs frequently tagged with these terms. For instance "joy", "joyous", "cheerful" and "happiness" are grouped under the "happy" category to get a larger result set. Note that the music for the "not" categories was evenly selected using music tagged with antonyms and a random selection to create more diversity. Afterwards, we asked expert listeners to validate this selection. We considered a song to be valid if the tag was confirmed by at least one listener, as the preselection from last.fm granted that the song was likely to deserve that tag. We included this manual tag confirmation in order to exclude songs that could have received the tag by error, to express something else, or by a "following the majority" type of effect. The listeners were exposed to only 30 seconds of the songs, to avoid changes in the mood as much as possible and then to speed up the annotation process. We asked them to concentrate on acoustic information. Consequently only these 30s excerpts have been included in the final dataset. In total, 17 different evaluators participated and an average of 71% of the songs originally selected from last.fm were included in the training set.

We observe that the "happy" and "relaxed" categories have a better validation rate than the "angry" and "sad" categories. This might be due to confusing terms in the tags used in the social networks for these latter categories or to a better agreement between people for "positive" emotions. These results indicate that the validation by experts is a necessary step to ensure the quality of the dataset, otherwise, around 29% of error, on average, would have been introduced. However, we also notice that this method is relevant to pre-selecting a large amount of tracks that potentially belong to one category.

[2] http://trec.nist.gov/

[3] http://www.last.fm

[4] http://wordnet.princeton.edu

At the end of the process, the database was composed of 1000 songs divided between 4 categories of interest plus their complementary categories (“not happy”, “not sad”, “not angry” and “not relaxed”), i.e. 125 songs per categories. We have used an equal distribution of these binary classes.

3.2. Audio Feature Extraction

In order to classify the music from audio content, we automatically extracted several state-of-the-art MIR audio features, including timbral and tonal features based on temporal and spectral representations of the audio signal. Here is a summary of the most relevant features for this algorithm:

- MFCCs [12]
- Barkbands [22]
- Harmonic Pitch Class Profiles (HPCPs) [4]
- Key Strength (HPCPs) [4]
- Spectral Centroid, Skewness, Kurtosis, Flatness [5]
- Mode [4]
- Loudness [5]
- Tempo (Beats Per Minute)
- Spectral Strong Peak [5]
- Dissonance [5]

For each excerpt of the dataset, its 200ms frame-based extracted features were summarized with their component-wise means and variances.

3.3. Classification and Evaluation

Once the ground truth was created and the features extracted, we performed a series of tests with many different classifiers and configurations. We evaluated the classifiers using Weka [27] implementations with 10 runs of 10-fold cross-validation (See Table 1 for the mean accuracies) . We list here the different configurations with the reference name in Table 1:

- SVM default: SVM with normalized features, default parameters and a polynomial kernel
- SVM std: Same configuration but standardizing the features
- SVM Cfs: Same configuration as the first one, but with a pre-processing step of correlation feature selection
- SVM RBF: Same configuration as the first one, but with a Radial Basis Function as kernel
- k-NN: Nearest-neighbor algorithm with k=10
- DecTree: Decision tree, J48 (default parameters)
- LogisticReg: Logistic Regression (default parameters)

For comparison purposes, we show the accuracies obtained for each mood category separately. The best results were achieved by the SVM algorithm (SMO in Weka) with linearly normalized features between 0 and 1, default parameters and a polynomial kernel.

The accuracies we obtained using audio-based classifiers are quite satisfying and even exceptional when looking at the “angry” category with 98%. All four categories reached classification accuracies above 80%, and two categories (“angry” and “relaxed”) above 90%. Even though these results might seem surprisingly high, this is coherent with similar studies [21]. Moreover, as we deal with binary comparisons on a balanced dataset, the random baseline is 50%. Also, the training examples were selected and validated only when they clearly belong to the category or its complementary. This can bias the database towards very clear between-class differences.

3.4. Audio encoding robustness

The cross-validation evaluation previously described gives relatively satisfying results in general. However, since the goal is to integrate this technology into a working platform, we tested the stability and robustness of the mood classification to low quality encodings. The original encodings of the training set were mp3 at 128 kbps (kilobits per second). We generated two modified versions of the dataset, modifying the bit rate : one at 64 kbps and one at 32kbps. In Figure 1, we represent the degradation of the accuracy of the classifier trained with the entire dataset, tested on the same dataset with different encodings. We decided to train and test with full datasets, as this classifier model would be the one to be used in the final integrated version.

We observe a degradation due to lower bit rate encoding. However, in all cases, this does not seem to have a strong impact. In Table 2, we report the degradation, in percentage, compared to the original version at 128 kpbs. For instance, we observe that for the “angry” category, at 32 kbps, only 0.7% of the dataset is no longer correctly classified as before. We notice that the highest percentage of degradation is 3.6% obtained for the “relaxed” category (with 32 kbps). Even if there is a drop in the accuracy, the classification still gives satisfying results.

Table 1. Mean classification accuracy with 10 runs of 10-fold cross-validation, for each category against its complementary (mp3 128 kbps)

	SVM default	SVM Cfs	k-NN	SVM std	RandForest	SVM RBF	DecTree	LogisticReg
Angry	**98.1%**	97.3%	96.3%	97.4%	95.4%	96.4%	92.71	95.9%
Happy	**82.6%**	81.3%	79.7%	81.5%	77.7%	72.1%	71.42	74.8%
Relaxed	**91.4%**	91.2%	90.9%	87.3%	91.2%	90.6%	91.1	80.9%
Sad	**87.7%**	86.2%	85.9%	87.0%	86.2%	87.0%	83.89	85.9%
Mean	**90.0%**	89.1%	88.2%	88.3%	87.6%	86.4%	84.8	84.4%

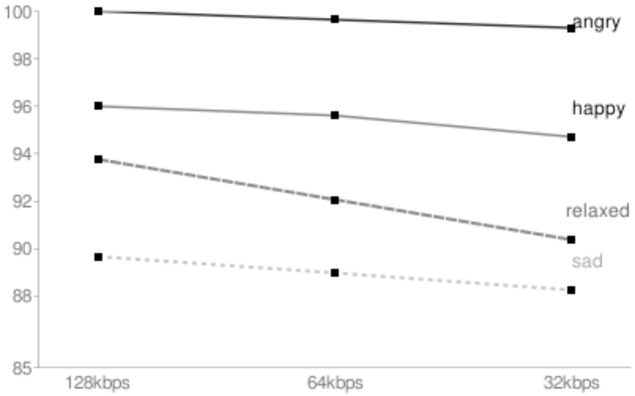

Figure 1. Effect of the audio bitrate reduction on the accuracy (in %) for the entire dataset.

Table 2. Degradation in percentage of the original accuracy (compared to mp3 at 128 kbps)

	mp3 64 kbps	mp3 32 kbps
Angry	0.4%	0.7%
Happy	0.4%	1.4%
Relaxed	1.8%	3.6%
Sad	1.9%	2.8%
Mean	1.1%	2.1%

4. Integration in the PHAROS Project

4.1. The PHAROS project

PHAROS[5] (Platform for searcHing of Audiovisual Resources across Online Spaces) is an Integrated Project co-financed by the European Union under the Information Society Technologies Programme (6th Framework Programme) - Strategic Objective 'Search Engines for Audiovisual Content'. PHAROS aims to advance audiovisual search from a point-solution search engine paradigm to an integrated search platform paradigm. One of the main goals of this project is to define a new generation of search engine, developing a scalable and open search framework which lets users search, explore, discover, and analyze contextually relevant data. Part of the core technology includes automatic annotation of content using integrated components of different kinds (visual classification, speech recognition, audio and music annotations, etc...). In our case, we implemented the automatic music mood annotation model computed as above.

4.2. Integration of the mood annotator

As a search engine, PHAROS uses automatic content annotation to index its content. However, there is a clear need to make the content analysis as efficient as possible (in terms of accuracy and time). To integrate mood annotation into the platform, we first created a fast implementation in C++ with proprietary code for audio feature extraction and dataset management together with the libsvm [6] library for Support Vector Machines. The SVMs were trained with full ground truth datasets to achieve the best results. Using a standard representation format defined in the project, we wrapped this binary implementation into a webservice which could be accessed by other elements of the PHAROS platform. Exploiting the probability output of the SVM algorithm, we provided a confidence value for each of the mood classifiers. This added a real value that is used for ranking the results of a query by the annotation. The resulting component extracted audio features and predicted the music's mood at a much higher speed (more than twice realtime) than the previous version. This meant that with an average computer, the algorithm could predict all the mood categories in less than half the duration of the audio excerpt. This annotator contributes to the overall system by allowing a flexible and distributed usage. The mood annotation is used to filter automatically the content according

[5] http://www.pharos-audiovisual-search.eu

[6] Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm

to the user's needs and helps them to find the content they are looking for. This integrated technology can lead to an extensive set of new tools to interact with music, enabling users to find new pieces similar to a given one, providing recommendations of new pieces, automatically organizing and visualizing music collections, creating playlists or personalizing radio streams. Indeed, the commercial success of large music catalogs nowadays is based on the possibility for people to find the music they want to listen to.

5. Conclusions

We presented an approach for music mood annotation introducing a method to exploit both the wisdom of crowds and the wisdom of the few. We reported the results in terms of accuracy and robustness. We explained how the technology was integrated, useful and used in the search engine European Project PHAROS. Preliminary results from user evaluations showed that the mood annotation is considered as innovative and very useful, but the study of user feedback will be part of future work. One may argue that this approach might be too simple to model the complexity of human perception. So what could be done to improve this automatic annotator? First, we can add more categories. Although there might be a semantic overlap, it can be interesting to annotate music moods with a larger vocabulary, if we can still have high accuracies. Then, we can try to make better predictions using a larger ground truth or designing new audio descriptors especially relevant for this task. Another option would be to generate analytical features [17], to try to increase the accuracy of the system. Finally, the mood annotation could be personalized, learning from the user's feedback and his own perception of mood. This would add a great value, although it might require much more processing time per user, thus making the annotation much less scalable. Nevertheless, it would dramatically enhance the user experience.

6. Acknowledgments

We are very grateful to all the human annotators that helped to create our ground truth. We also want to thank all the people contributing to the MTG technology and in particular Nicolas Wack, Eduard Aylon and Robert Toscano. This research has been partially funded by the EU Project PHAROS IST-2006-045035.

References

[1] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet. Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition & Emotion*, 19(8):1113–1139, December 2005.

[2] J. S. Downie. The music information retrieval evaluation exchange (mirex). *D-Lib Magazine*, 12(12), 2006.

[3] P. R. Farnsworth. A study of the hevner adjective list. *The Journal of Aesthetics and Art Criticism*, 13(1):97–103, 1954.

[4] E. Gómez. *Tonal description of music audio signals*. PhD thesis, Universitat Pompeu Fabra, 2006.

[5] F. Gouyon, P. Herrera, E. Gomez, P. Cano, J. Bonada, A. Loscos, X. Amatriain, and X. Serra. *Content Processing of Music Audio Signals*, chapter 3, pages 83–160. Logos Verlag Berlin GmbH, Berlin, 2008.

[6] X. Hu, S. J. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 462–467, Philadelphia, PA, USA, 2008.

[7] P. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, September 2004.

[8] P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3), 2004.

[9] C. L. Krumhansl. An exploratory study of musical emotions and psychophysiology. *Canadian journal of experimental psychology*, 51(4):336–353, December 1997.

[10] C. Laurier and P. Herrera. Audio music mood classification using support vector machine. In *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[11] T. Li and M. Ogihara. Detecting emotion in music. In *Proceedings of the 4th International Conference on Music Information Retrieval*, pages 239–240, Baltimore, MD, USA, 2003.

[12] B. Logan. Mel frequency cepstral coefcients for music modeling. In *Proceeding of the 1st International Symposium on Music Information Retrieval*, Plymouth, MA, USA, 2000.

[13] D. Lu, L. Liu, and H. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on audio, speech, and language processing*, 14(1):5–18, 2006.

[14] 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*, 14(1):5–18, Jan. 2006.

[15] M. I. Mandel, G. E. Poliner, and D. P. W. Ellis. Support vector machine active learning for music retrieval. *Multimedia Systems*, 12(1), 2006.

[16] N. Orio. Music retrieval: a tutorial and review. *Found. Trends Inf. Retr.*, 1(1):1–96, 2006.

[17] F. Pachet and P. Roy. Analytical features: a knowledge-based approach to audio feature generation. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009(1), February 2009.

[18] I. Peretz, L. Gagnon, and B. Bouchard. Music and emotion: perceptual determinants, immediacy, and isolation after brain damage. *Cognition*, 68(2):111–141, August 1998.

[19] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[20] Y.-Y. Shi, X. Zhu, H.-G. Kim, and K.-W. Eom. A tempo feature via modulation spectrum analysis and its application to music emotion classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1085–1088, Toronto, Canada, 2006.

[21] J. Skowronek, M. F. McKinney, and S. van de Par. A demonstrator for automatic music mood estimation. In *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[22] Smith and J. S. Abel. Bark and erb bilinear transforms. *Speech and Audio Processing, IEEE Transactions on*, 7(6):697–708, 1999.

[23] M. Sordo, C. Laurier, and O. Celma. Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of the 8th International Conference on Music Information Retrieval*, pages 531–534, Vienna, Austria, 2007.

[24] R. E. Thayer. *The Origin of Everyday Moods: Managing Energy, Tension, and Stress*. Oxford University Press, Oxford, November 1996.

[25] S. Vieillard, I. Peretz, N. Gosselin, S. Khalfa, L. Gagnon, and B. Bouchard. Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4):720–752, 2008.

[26] A. Wieczorkowska, P. Synak, R. Lewis, and Raś. Extracting emotions from music data. In *Foundations of Intelligent Systems*, pages 456–465. Springer-Verlag, 2005.

[27] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco, 1999.

[28] Y. H. Yang, Y. C. Lin, H. T. Cheng, and H. H. Chen. Mr.emo: Music retrieval in the emotion plane. In *Proceedings of the ACM International Conference on Multimedia*, Vancouver, BC, Canada, 2008.

[29] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, Feb. 2008.