Indexing Music by Mood: Design and Integration of an Automatic Content-based Annotator

Cyril Laurier, Owen Meyers, Joan Serrà, Martin Blech, Perfecto Herrera and Xavier Serra
Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

contact: cyril.laurier@upf.edu

Extended version of the CBMI 2009 paper: "Music Mood Annotator Design and Integration"

Abstract In the context of content analysis for indexing and retrieval, a method for creating automatic music mood annotation is presented. The method is based on results from psychological studies and framed into a supervised learning approach using musical features automatically extracted from the raw audio signal. We present here some of the most relevant audio features to solve this problem. A ground truth, used for training, is created using both social network information systems (wisdom of crowds) and individual experts (wisdom of the few). At the experimental level, we evaluate our approach on a database of 1000 songs. Tests of different classification methods, configurations and optimizations have been conducted, showing that Support Vector Machines perform best for the task at hand. Moreover, we evaluate the algorithm robustness against 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. This real world application demonstrates the usability of this tool to annotate large-scale databases. We also report on a user evaluation in the context of the PHAROS search engine, asking people about the utility, interest and innovation of this technology in real world use cases.

Keywords music information retrieval . mood annotation
content-based audio . social networks . user evaluation

1. Introduction

Psychological studies have shown that emotions conveyed by music are objective enough to be valid for mathematical modeling [4, 13, 24, 32]. Moreover, Vieillard et al. [43] demonstrated that within the same culture, the emotional responses to music could be highly consistent. All these results indicate that modeling emotion or mood in music is feasible.

In the past few years, research in content-based techniques has been trying to solve the problem of tedious and time-consuming human indexing of audiovisual data. In particular, Music Information Retrieval (MIR) has been very active in a wide variety of topics such as automatic transcription or genre classification [5, 29, 41]. Recently, classification of music mood has become a matter of interest, mainly because of the close relationship between music and emotions [1, 20].

In the present paper, we present a robust and efficient mood annotator that automatically estimates the mood of a piece of music, directly from the raw audio signal. We achieve this task by using a supervised learning method. In Section 2, we report on related works in classification of music mood. In Section 3, we detail the method and the results we achieved. In Section 4, we describe the integration of this technique in the PHAROS project (Platform for searcHing of Audiovisual Resources across Online Spaces). In Section 5, we present the protocol and results of a user evaluation. Finally, in Section 6, we discuss future works.

2. Scientific Background

Although there exist several studies dealing with automatic content-based mood classification (such as [4, 26, 37, 47]), almost every work differs in the way that it represents the mood concepts. Similar to psychological studies, there is no real agreement on a common model [16]. Some consider a categorical representation based on mutually exclusive basic emotions such as "happiness", "sadness", "anger", "fear" and "tenderness" [19, 26, 36, 39], while others prefer a multi-labeling approach (i.e., using a rich set of adjectives that are not mutually exclusive) like Wieczorkowska [45]. The latter is more difficult to evaluate since they consider many categories. The basic emotion approach gives simple but relatively satisfying results, around 70-90% of correctly classified instances, depending on the data and the number of categories chosen (usually between 3 and 5). Li and Ogihara [22] extract timbre, pitch and rhythm features from the audio content to train Support Vector Machines (SVMs). They consider 13 categories, 11 from the ones proposed in Farnsworth [10] plus 2 additional ones. However, the results are not that convincing, obtaining low average precision (0.32) and moderate recall (0.54). This might be due to the small dataset labeled by only one person and to the large number of categories they chose. Conversely, it is very advisable to use few categories and a ground truth annotated by hundreds of people (see Section 3.1).

Other works use the dimensional representation (modeling emotions in a space), like Yang [47]. They model the problem with Thayer's arousal-valence[1] emotion plane [40] 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. [27] describe an active learning system using timbre features and SVMs, which learns

[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.

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, and has an accuracy equivalent to that of state-of-the-art methods.

Comparing evaluations of these different techniques is an arduous task. With the objective to evaluate different algorithms within the same framework, MIREX (Music Information Retrieval Evaluation eXchange) [8] organized a first task on Audio Mood Classification in 2007[2]. MIREX is a reference in the MIR community that provides a solid evaluation of current algorithms in different tasks. The MIREX approach is similar to the Text Retrieval Conference (TREC)[3] approach to the evaluation of text retrieval systems, or TREC-VID[4] for video retrieval. For the Audio Mood Classification task, it was decided to model the mood classification problem with a categorical representation in mood clusters (a word set defining the category). Five mutually exclusive mood clusters were chosen (i.e, one musical excerpt could only belong to one mood cluster). In that aspect, it is similar to a basic emotion approach, because the mood clusters are mutually exclusive. They asked human evaluators to judge a collection of 1250 30-second excerpts (250 in each mood cluster). The resulting human-validated collection consisted of 600 30-second clips in total. The best results approached 60% of accuracy [14, 18]. In Table 1, we show the categories used and the results of different algorithms, including our submitted algorithm [18] (noted CL). One should note that the accuracies from the MIREX participants are lower than those found in most of the existing literature. This is probably due to a semantic overlap between the different clusters [14]. Indeed, if the categories are mutually exclusive, the category labels have to be chosen carefully.

Mood Clusters	CL	GT	TL	ME
rowdy,rousing,confident,boisterous,passionate	45.83%	42.50%	52.50%	51.67%
amiable,good natured,sweet,fun,rollicking,cheerful	50%	53.33%	49.17%	45.83%
literate,wistful,bittersweet,autumnal,brooding,poignant	82.50%	80%	75%	70%
witty,humorous,whimsical,wry,campy,quirky,silly	53.33%	51.67%	52.50%	55%
volatile,fiery,visceral,aggressive,tense/anxious,intense	70.83%	80%	69.17%	66.67%
Mean accuracy	60.50%	61.50%	59.67%	57.83%

Table 1. Extract from the Audio Mood Classification task results, MIREX 2007. Mean accuracies in percentage over a 3-fold Cross Validation. Comparison of our submitted algorithm (CL[18]), with the other top competitors (GT[42], TL[23], ME[28]). We used several of the audio features presented later in this paper and SVMs.

Performing a statistical analysis on this data with the Tukey-Kramer Honestly Significantly Differently method (TK-HSD) [2], the MIREX organizers found that our algorithm had the first rank across all mood clusters despite its average accuracy being the second highest [14]. Another

[2] http://www.music-ir.org/mirex2007/index.php/Audio_Music_Mood_Classification
[3] http://trec.nist.gov/
[4] http://www-nlpir.nist.gov/projects/trecvid/

interesting fact from this evaluation is that, looking at all the submissions, the most accurate algorithms were using SVMs. The results of the MIREX task show that our audio feature extraction and classification method are state-of-the-art. Thus, to create a new music mood annotator, even though we tried different classification methods, we focused on the optimization of Support Vector Machines [3]. Moreover, we especially focused on using a relevant taxonomy and on finding an efficient and original method to create a reliable ground truth.

3. Method

To classify music by mood, we frame the problem as an audio classification problem using a supervised learning approach. We consider unambiguous categories to allow for a greater understanding and agreement between people (both 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 Section 3.2. We employ standard feature selection and classification techniques and we evaluate them in Section 3.3. Once the best algorithm is chosen, we evaluate the contribution of each descriptor in 3.5 and the robustness of the system as reported in Section 3.4. In Figure 1, we show a general block diagram of the method.

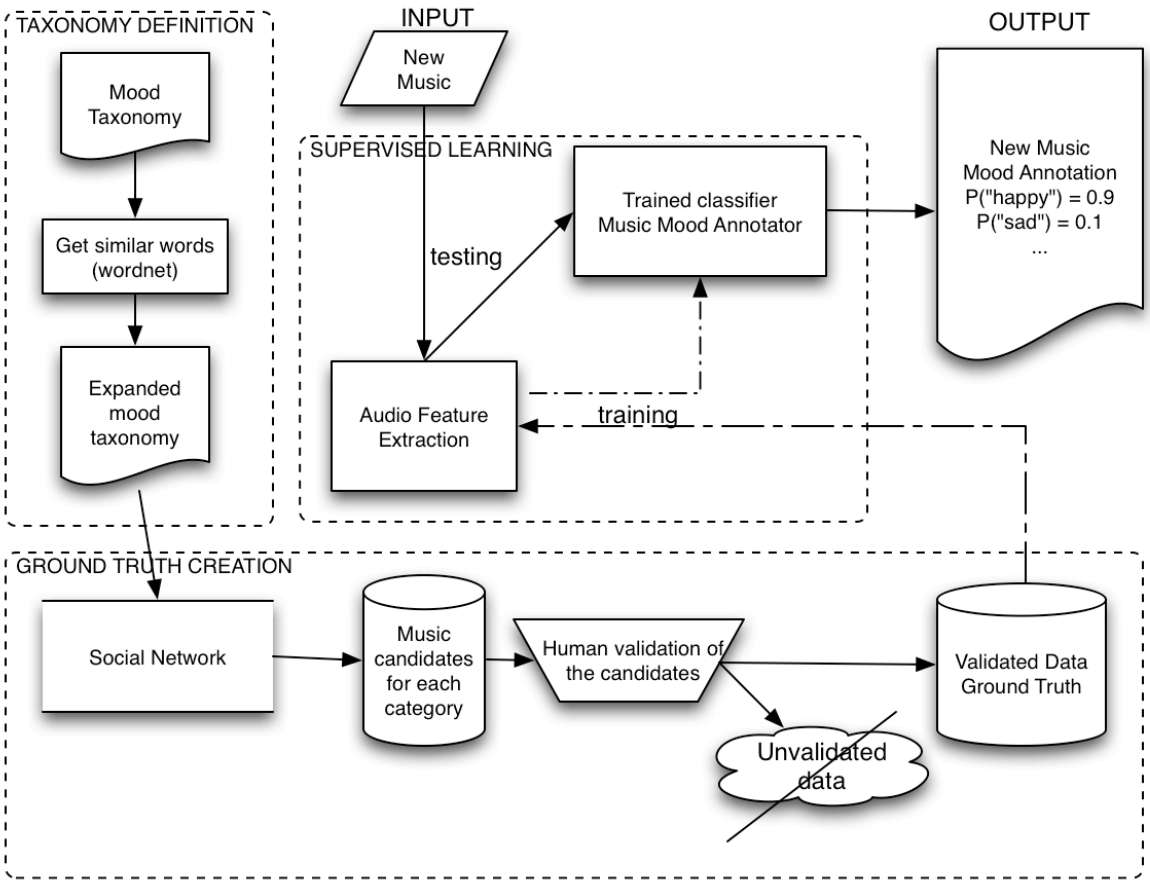

Figure 1. Schema of the method employed to create the ground truth, validate it and design the music mood annotator.

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,* and *relaxed*. We decided to use these categories because these moods are related to basic emotions from psychological theories (reviewed in [15]) and they cover the four quadrants of the 2D representation from Russell [34] with valence and arousal dimensions (see Figure 2).

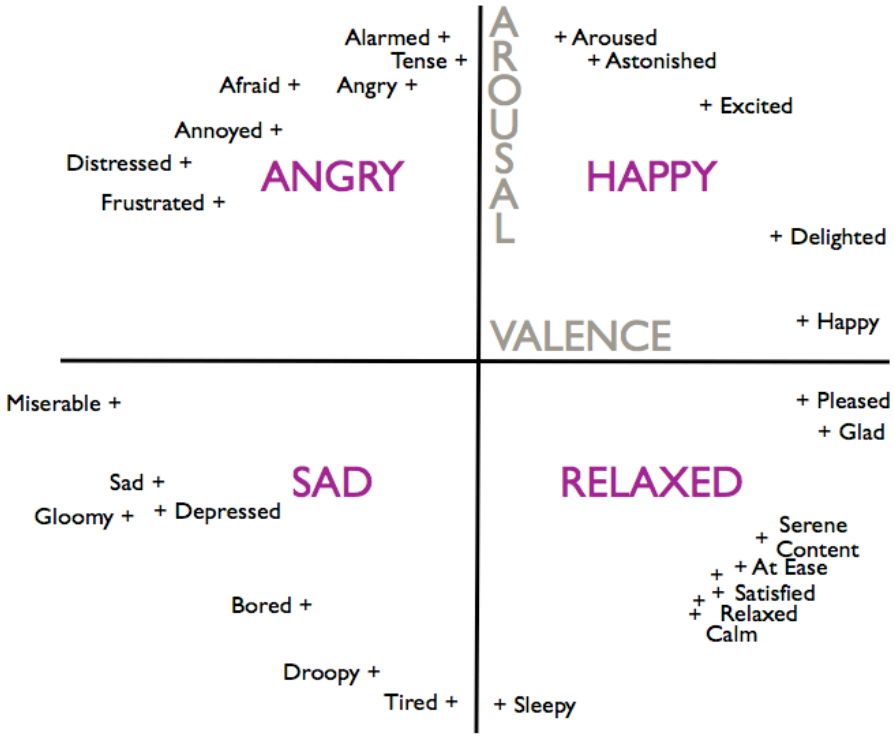

Figure 2. "Circumplex model of affect" (adapted from Russel [34]).

The Russell 2D model (called "circumplex model of affect") is a reference widely accepted and cited in psychological studies on emotion. In this space, "happy" and "relaxed" have positive valence and, respectively, 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 task for each mood. One song can be "happy" or "not happy", but also independently "angry" or "not angry" and so on.

The main idea of the present method 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[5] tags (textual labels). Last.fm is a music recommendation website with a large community of users (30 million active users based in more than 200 countries) that is very active in associating tags with the music they listen to. These tags are then available to all users in the community. In Figure 3, we show an example of a "tag cloud", which is a visualization of the tags assigned to one song with the font size weighted by the popularity of the tag for this particular song.

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

Figure 3. Tag cloud of the song "Here comes the sun" from the Beatles. The tags recognized as mood tags are underlined. The bigger the tag is, more people have used it to define that song.

In the example shown in Figure 3, we can see that "happy" is present and quite highly weighted (which means that many people have used this tag to describe the song). In addition to "happy", we also have "cheerful", "joy", "fun" and "upbeat". To gather more data, we need to extend our query made to last.fm with more words related to mood. For the four chosen mood categories, we generated a set of related semantic words using Wordnet[6] and looked for the songs frequently tagged with these terms. For instance "joy", "joyous", "cheerful" and "happiness" are grouped under the "happy" category to generate a larger result set. We query the social network to acquire songs tagged with these words and apply a popularity threshold to select the best instances (we keep the songs that have been tagged by many users).

Note that the music for the "not" categories (like "not happy") was evenly selected using both music tagged with antonyms and a random selection to create more diversity. Afterwards, we asked listeners to validate this selection. We considered a song to be valid if the tag was confirmed by, at least, one listener, as the pre-selection 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 to speed up the annotation process. Consequently, only these 30 second 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 was 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 errors, on average, would have been introduced. This method is relevant to pre-selecting a large number of tracks that potentially belong to one category.

[6] Wordnet is a large lexical database of English words with sets of synonyms http://wordnet.princeton.edu/

At the end of the song selection process, the database was composed of 1000 songs divided between the 4 categories of interest plus their complementary categories ("not happy", "not sad", "not angry" and "not relaxed"), i.e. 125 songs per category. The audio files were 30-second stereo clips at 44khz in a 128kbps mp3 format.

3.2 Audio Feature Extraction

In order to classify the music from audio content, we extracted a rich set of audio features based on temporal and spectral representations of the audio signal. For each excerpt, we merged the stereo channels into a mono mixture and its 200ms frame-based extracted features were summarized with their component-wise statistics across the whole song. In Table 2, we present an overview of the extracted features by category.

Timbre	Bark bands, MFCCs, pitch salience, hfc, loudness, spectral: flatness, flux, rolloff, complexity, centroid, kurtosis, skewness, crest, decrease, spread
Tonal	dissonance, chords change rate, mode, key strength, tuning diatonic strength, tristimulus
Rhythm	bpm, bpm confidence, zero-crossing rate, silence rate, onset rate, danceability

Table 2. Overview of the audio features extracted by category. See [31], [12] and [25] for a detailed description of the features.

For each excerpt we obtained a total of 200 feature statistics (minimum, maximum, mean, variance and derivatives), and we standardized each of them across the whole music collection values. In the next paragraphs, we describe some of the most relevant features for this mood classification task, with results and figures based on the training data.

3.2.1 Mel Frequency Cepstral Coefficients (MFCCs)

MFCCs [25] are widely used in audio analysis, and especially for speech research and music classification tasks. The method employed is to divide the signal into frames. For each frame, we take the logarithm of the amplitude spectrum. Then we divide it into bands and convert it to the perceptually-based Mel spectrum. Finally we take the discrete cosine transform (DCT). The number of output coefficients of the DCT is variable, and is often set to 13, as we did in the present study. Intuitively, lower coefficients represent spectral envelope, while higher ones represent finer details of the spectrum. In Figure 4, we show the mean values of the MFCCs for the "sad" and "not sad" categories. We note from Figure 4 a difference in the shape of the MFCCs. This indicates a potential usefulness to discriminate between the two categories.

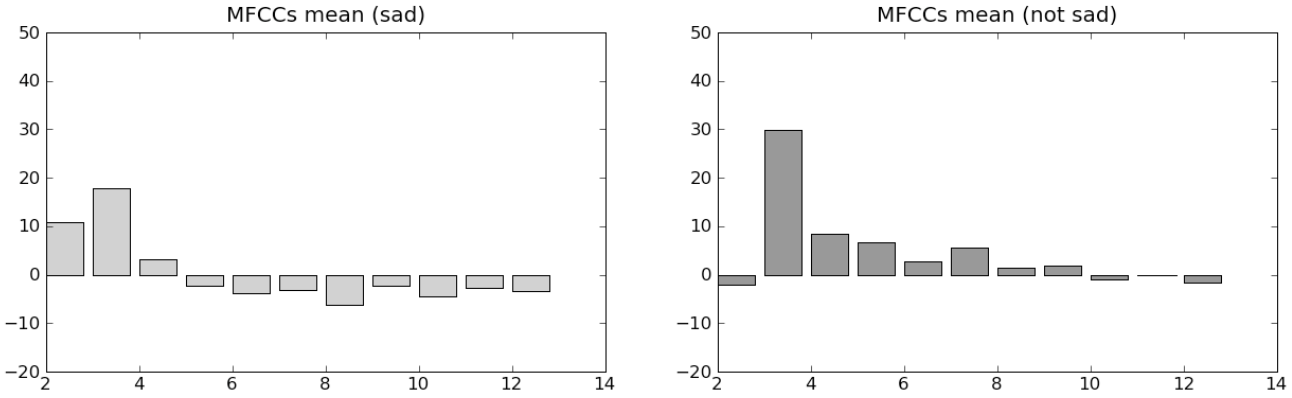

Figure 4. MFCC mean values for coefficients between 2 and 13 for the "sad" and "not sad" categories of our annotated dataset.

3.2.2 Bark bands

The Bark band algorithm computes the spectral energy contained in a given number of bands, which corresponds to an extrapolation of the Bark band scale [31, 38]. For each Bark band (27 in total) the power-spectrum is summed. In Figure 5, we show an example of the Bark bands for the "sad" category.

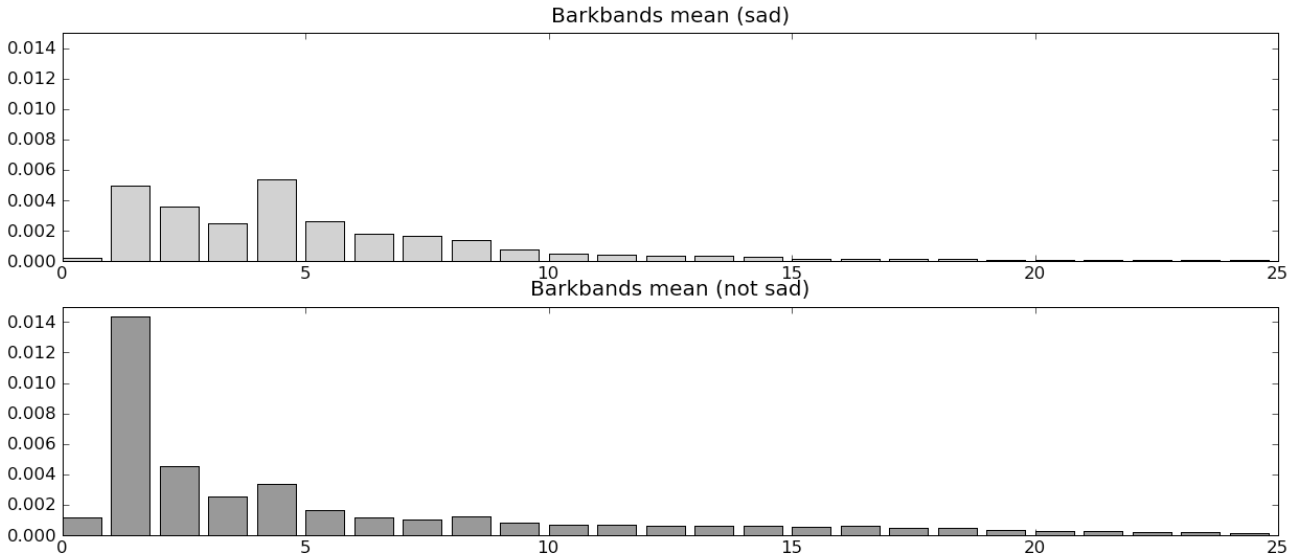

Figure 5. Bark band mean values for coefficients between 1 and 25 for the "sad" and "not sad "categories of our annotated dataset.

As with the MFCCs, the Bark bands appear to have quite different shapes for the two categories, indicating a probable utility for classification purposes.

3.2.3 Spectral Complexity

The spectral complexity descriptor [31] is based on the number of peaks in the input spectrum. We apply peak detection on the spectrum (between 100Hz and 5Khz) and we count the number of peaks. This feature describes the complexity of the audio signal in terms of frequency components. In Figures 6 and 7, we show the box-and-whisker plots of the spectral complexity descriptor's standardized means for the "relaxed", "not relaxed", "happy" and "not happy" categories. These results are based on the entire training dataset. These plots illustrate the intuitive result that a

relaxed song should be less "complex" than a non-relaxing song. Moreover, Figure 7 tells us that happy songs are on average spectrally more complex.

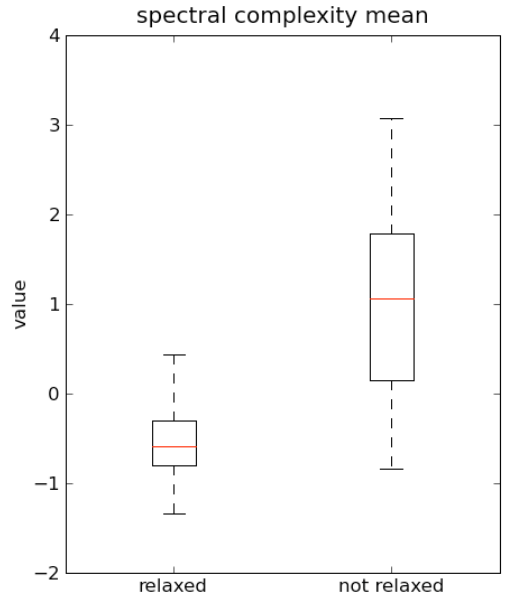

Figure 6. Box-and-whisker plot of the standardized spectral complexity mean feature for "relaxed" and "not relaxed".

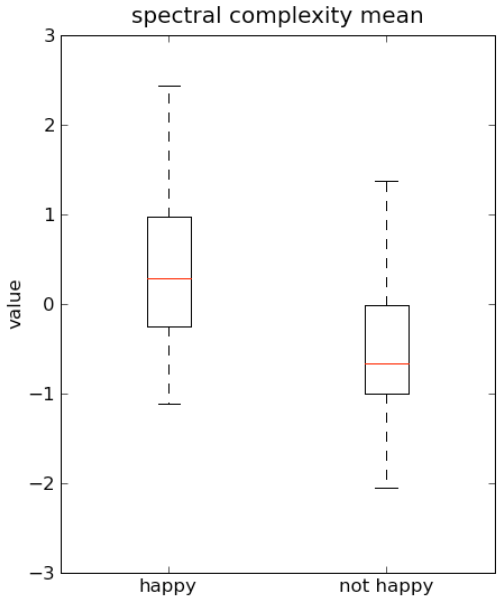

Figure 7. Box-and-whisker plot of the standardized spectral complexity mean feature for "happy" and "not happy".

3.2.4 Spectral Centroid and Skewness

The spectral centroid and skewness descriptors [31] (as well as spread, kurtosis, rolloff and decrease) are descriptions of the spectral shape. The spectral centroid is the barycenter of the spectrum, which considers the spectrum as a distribution of frequencies. The spectral skewness measures the asymmetry of the spectrum's distribution around its mean value. The lower the value, the more energy exists on the right-hand side of the distribution, while more energy on the left side indicates a higher spectral skewness value. In Figure 8 we show the spectral centroid's box-and-whisker plot for "angry" and in Figure 9 the spectral skewness for "sad".

Figure 8 shows a higher spectral centroid mean value for "angry" than "not angry", which intuitively means more energy in higher frequencies. For the spectral skewness, the range of mean values for the "sad" instances is bigger than for the "not sad" ones. This probably means that there is a less specific value for the centroid. In any case, it seems to have on average a lower value for the "not sad" instances.

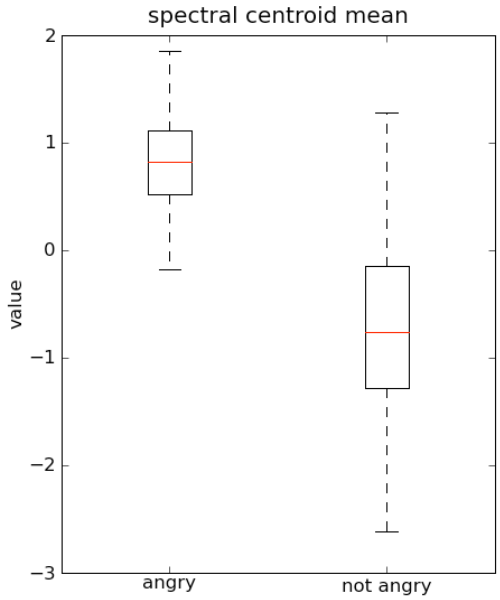

Figure 8. Box-and-whisker plot of the standardized spectral centroid mean for "angry" and "not angry".

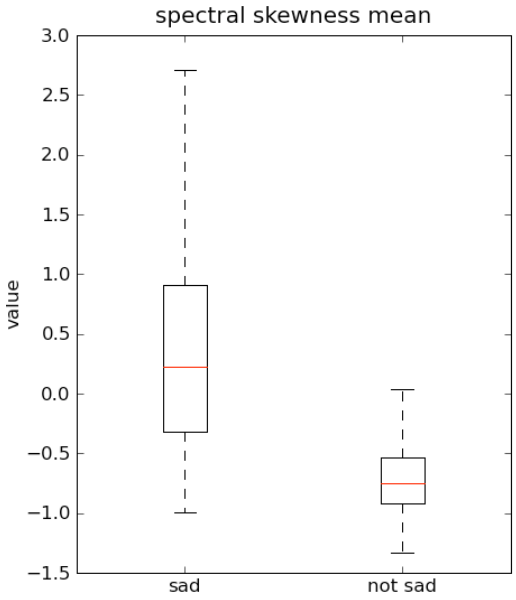

Figure 9. Box-and-whisker plot of the standardized spectral skewness mean for "sad" and "not sad".

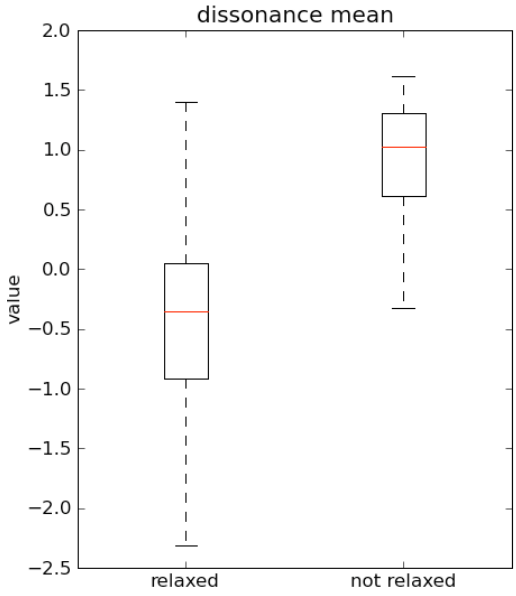

Figure 10. Box-and-whisker plot of the standardized dissonance mean for the "relaxed" and "not relaxed" categories.

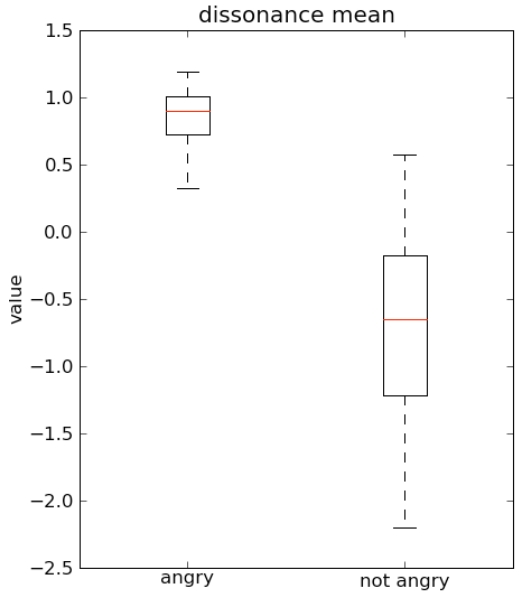

Figure 11. Box-and-whisker plot of the dissonance mean for the "angry" and "not angry" categories.

3.2.5 Dissonance

The dissonance feature (also known as "roughness" [35]) is defined by computing the peaks of the spectrum and measuring the spacing of these peaks. Consonant sounds have more evenly spaced spectral peaks and, on the contrary, dissonant sounds have more sporadically spaced spectral peaks. In Figures 10 and 11, we compare the dissonance distributions for the "relaxed" and "angry"

categories. These figures show that "angry" is clearly more dissonant than "not angry". Listening to the excerpts from the training data, we noticed many examples with distorted sounds like electric guitar in the "angry" category, which seems to be captured by this descriptor. This also relates to psychological studies stating that dissonant harmony may be associated with anger, excitement and unpleasantness [13,44].

3.2.6 Onset rate, Chords change rate

From psychological results, one important musical feature when expressing different mood types is rhythm (generally, faster means more arousal) [15]. The basic measure/element of rhythm is the onset, which is defined as an event in the music (any note, drum, etc.). The onset times are estimated by looking for peaks in the amplitude envelope. The onset rate is the number of onsets in one second. This gives us the number of events per second, which is related to a perception of the speed. The chords change rate is a rough estimator of the number of chords change per second.

In Figure 12, we compare the onset rate values for the "happy" and "not happy" categories. It shows that "happy" songs have higher values for the onset rate, which confirms the psychological results that "happy" music is fast [15]. In Figure 13, we look at the chords change rate, which is higher for "angry" than "not angry". This is also a confirmation of the studies previously mentioned, associating higher arousal with faster music.

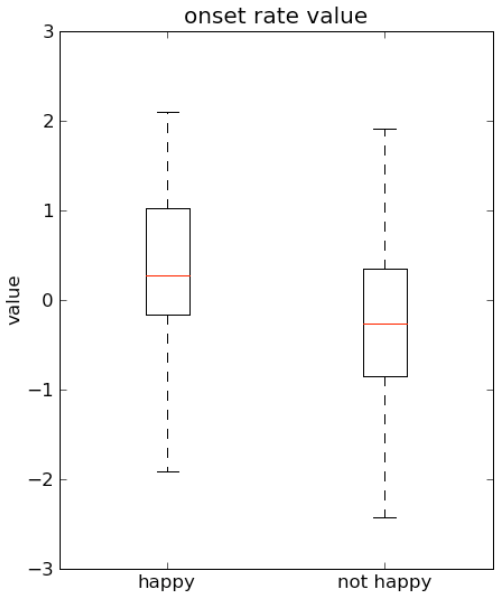

Figure 12. Box-and-whisker plot of the standardized onset rate value mean for the "happy" and "not happy" categories.

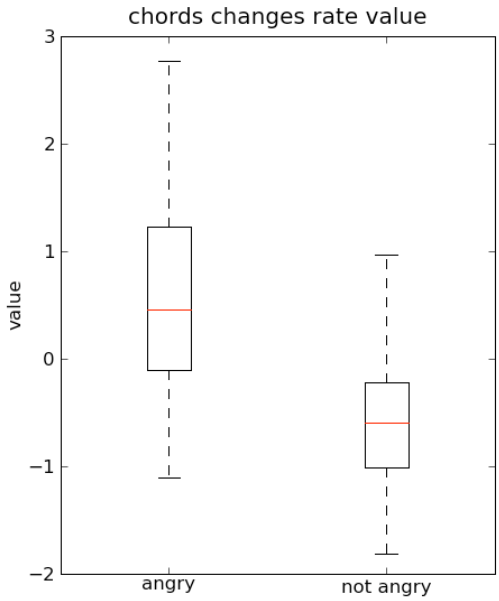

Figure 13. Box-and-whisker plot of the chords change mean for the "angry" and "not angry" categories.

3.2.7 Mode

In Western music theory, there are two basic modes: major and minor. Each of them has different musical characteristics regarding the position of tones and semitones within their respective musical scales. Gómez [11] explains how to compute an estimation of the mode from raw audio data.

The signal is first pre-processed using the direct Fourier transform (DFT), filtering frequencies between 100 Hz and 5000 Hz and locating spectral peaks. The reference frequency (tuning

frequency) is then estimated by analyzing the frequency deviation of the located spectral peaks. Next the Harmonic Pitch Class Profile (HPCP) feature is computed by mapping frequency and pitch class values (musical notes) using a logarithmic function [11]. The global HPCP vector is the average of the instantaneous values per frame, normalized to [0,1] to make it independent of dynamic changes. The resulting feature vector represents the average distribution of energy among the different musical notes. Finally, this vector is compared to minor and major reference key profiles based on music theory [17]. The profile with the highest correlation with the HPCP vector defines the mode.

In Figure 14, we represent the percentages of estimated major and minor music in the "happy" and "not happy" categories. We note that there is more major music in the "happy" than in the "not happy" pieces. In music theory and psychological research, the link between valence (positivity) and the musical mode has already been demonstrated [15]. Still, having empirical data from an audio feature automatically extracted showing the same tendency is an interesting result. We note also that the proportion of major music is also high in the "not happy" category, which is related to the fact that the majority, 64%, of the whole dataset is estimated as major.

Figure 14. Bar plot of the estimated mode proportions (in percentage) for the "happy" and "not happy" categories.

We have mentioned here some of the most relevant features showing their potential to individually discriminate between categories, however, we keep all the descriptors in our "bag-of-features"; those that are not obviously useful could be significant when combined with others in a linear or non-linear way. To capture these relationships and build the model, we tried several kinds of classification algorithms.

3.3 Classification Algorithms

Once the ground truth was created and the features extracted, we performed a series of tests with 8 different classifiers. We evaluated the classifiers using their implementations in Weka [46] with 10 runs of 10-fold cross-validation and parameter optimizations (See Table 3 for the mean accuracies). Next, we list the different classifiers we employed.

3.3.1 Support Vector Machines (SVMs)

Support Vector Machines [3], is a widely used supervised learning classification algorithm. It is known to be efficient, robust and to give relatively good performance. Indeed, this classifier is widely used in MIR research. In the context of a two-class problem in *n* dimensions, the idea is to find the "best" hyperplane separating the points of the two classes. This hyperplane can be of *n-1* dimensions and found in the original feature space, in the case that it is a linear classifier. Otherwise, it can be found in a transformed space of higher dimensionality using kernel methods (non-linear). The position of new observations compared to the hyperplane tells us in which class belongs the new input. For our evaluations, we tried different kernel methods: linear, polynomial, radial basis function (RBF) and sigmoid respectively called SVM linear, SVM poly, SVM RBF and SVM sigmoid, as shown in Table 3. To find the best parameters in each case we used the cross-validation method on the training data. For the linear SVM we looked for the best value for the cost *C* (penalty parameter), and for the others we applied a grid search to find the best values for the pair *(C, gamma)* [3]. For *C,* we used the range $[2^{-15},2^{15}]$ in 31 steps. For *gamma,* we used the range $[2^{15},2^{3}]$ in 19 steps. In the other cases than the linear SVM, once we have the best pair of values (*C,gamma*), we conduct a finer grid search on the neighborhood of these values. We note that from our data, the best parameter values highly depends on the category. Moreover, even if a RBF kernel is not always recommended for large feature sets compared to the size of the dataset [3], we achieved the best accuracy using this kernel for almost all categories. We used an implementation of the Support Vector Machines called libsvm [6].

3.3.2 Trees and Random Forest

The decision tree algorithm splits the training dataset into subsets based on a test attribute value. This process is repeated on each subset in a recursive manner (recursive partitioning). The random forest classifier uses several decision trees in order to improve the classification rate. We used an implementation of the C4.5 decision tree [33] (called J48 in Weka and in Table 3). To optimize the parameters of the decision tree, we performed a grid search on the two main parameters: *C* (the confidence factor used for pruning) from 0.1 to 0.5 in 10 steps and *M* (the minimum number of instances per leaf) from 2 to 20.

3.3.3 k-Nearest Neighbor (*k*-NN)

For a new observation, the *k*-NN algorithm looks for a number *k* of the closest training samples to decide on the class to predict. The result relies mostly on the choice of distance function, which might not be trivial in our case, and also in the choice of *k*. We tested different values of *k* (between 1 and 20) with the Euclidean distance function.

3.3.4 Logistic Regression

Logistic regression can predict the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression. To optimize it, we varied the ridge value [21].

3.3.5 Gaussian Mixture Models (GMMs)

GMM is a linear combination of Gaussian probability distributions. This approach assumes that the likelihood of a feature vector can be expressed with a mixture of Gaussian distributions. GMMs are universal approximations of density, meaning that with enough Gaussians, any distribution can be estimated. In the training phase, the parameters of the Gaussian mixtures for each class are learnt using the Expectation-Maximization algorithm, which iteratively computes maximum likelihood estimates [7]. The initial Gaussian parameters (means, covariance, and prior probabilities) used by the EM algorithm are generated via the k-means method [9].

3.4 Evaluation results

After independent parameter optimization for each classifier, the evaluation was made with 10 runs of 10 fold cross-validation. For comparison purposes, we show the mean accuracies obtained for each mood category and algorithm configuration separately. Each value in a cell represents the mean value of correctly classified data in the test set of each fold. Considering that each category is binary (for example, "angry" vs. "not angry"), the random classification accuracy is 50%.
The SVM algorithm with different kernels and parameters, depending on the category, achieved the best results. Consequently, we will choose the best configuration (SVM with polynomial kernel except for happy where we will use a linear SVM) for the integration in the final application.

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") peaked above 90%. Even though these results might seem surprisingly high, this is coherent with similar studies [37]. Also, the training examples were selected and validated only when they clearly belonged to the category or its complementary. This can bias the database and the model towards detecting very clear between-class distinctions.

	Angry	Happy	Relaxed	Sad	Mean Accuracy	Duration 10 folds
SVM linear	*95.79%*	***84.57%***	*90.68%*	*87.31%*	*89.58%*	14 s
SVM poly	***98.17%***	*84.48%*	***91.43%***	***87.66%***	***90.44%***	24 s
SVM RBF	*95.19%*	*84.47%*	*89.79%*	*87.52%*	*89.24%*	17 s
SVM sigmoid	95.08%	84.52%	88.63%	87.31%	*88.89%*	17 s
J48	*95.51%*	*80.02%*	*85.25%*	*85.87%*	*86.66%*	5 s
Random Forest	*96.31%*	*82.55%*	*89.47%*	*87.26%*	*88.90%*	13 s
k-NN	*96.38%*	*80.89%*	*90.08%*	*85.48%*	*88.21%*	4 s
Logistic Reg	*94.46%*	*73.60%*	*82.54%*	*76.38%*	*81.75%*	20 s
GMMs	*96.99%*	*79.91%*	*91.13%*	*86.54%*	*88.64%*	12 s

Table 3: Mean classification accuracy with 10 runs of 10-fold cross-validation, for each category against its complementary. In bold is the highest accuracy for each category. The last column is the duration, in seconds, for a 10-fold cross-validation (computed on a 1.86 Ghz Intel Core Duo).

	Angry	Happy	Relaxed	Sad
All features	*98.17%*	*84.57%*	*91.43%*	*87.66%*
MFCCs	*89.47%*	*57.59%*	*83.87%*	*81.74%*
Bark bands	*90.98%*	*59.82%*	*87.10%*	*83.48%*
Spectral complexity	95.86%	55.80%	88.71%	86.52%
Spectral centroid	*89.47%*	*50%*	*85.48%*	*83.04%*
Spectral skewness	*77.44%*	*52.23%*	*73.38%*	*73.48%*
Dissonance	*91.73%*	*62.05%*	*82.66%*	*79.57%*
Onset rate	*52.63%*	*60.27%*	*63.31%*	*72.17%*
Chords change rate	*74.81%*	*50%*	*69.35%*	*68.26%*
Mode	*71.43%*	*64.73%*	*52.82%*	*52.08%*

Table 4. Mean classification accuracy with 10 runs of 10-fold cross-validation, for each category against its complementary with feature sets made of one descriptor statistic.

3.5 Audio feature contribution

In this part, we evaluated the contribution of the audio features described in 3.2. In order to achieve this goal, we chose the best overall classifier for each category and we made 10 runs of 10-fold cross-validation with only one descriptor type statistic. We show in Table 4 the resulting mean accuracies for each configuration compared to the best accuracy obtained with all the features in the first row.

We observe that most of the descriptors give worst results for the "happy" category. This reflects also the results with all features, with a lower accuracy for "happy". Moreover, some descriptors like the spectral centroid and the chords change rate do not seem to contribute positively for this category. We also note that the mode helps to discriminate between "happy" and "not happy", like seen in Figure 14. It is also relevant for the "angry" category. However it does seem useful for "sad" against "not sad". It is also worth noticing that if some individual descriptors can give relatively high accuracies, the global system combining all the features is significantly more accurate.

3.6 Audio encoding robustness

The cross-validation evaluation previously described gives relatively satisfying results in general. It allows us to select the best classifier with the appropriate parameters. However, since the goal is to integrate this model into a working platform, we have to test the stability and robustness of the mood classification with low quality encodings. Indeed it should be able to process musical content of different quality (commercial or user generated). The original encodings of the training set were mp3 at 128 kbps (kilobits per second). We generated two modified versions of the dataset, lowering the bit rate to 64 kbps and 32kbps. In Figure 15, we represent the accuracy degradation of the classifier trained with the entire dataset and tested on the same one with the previously mentioned low-rate 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. Please note that the accuracies are different from Table 3 because in this case we are not performing cross-validation.

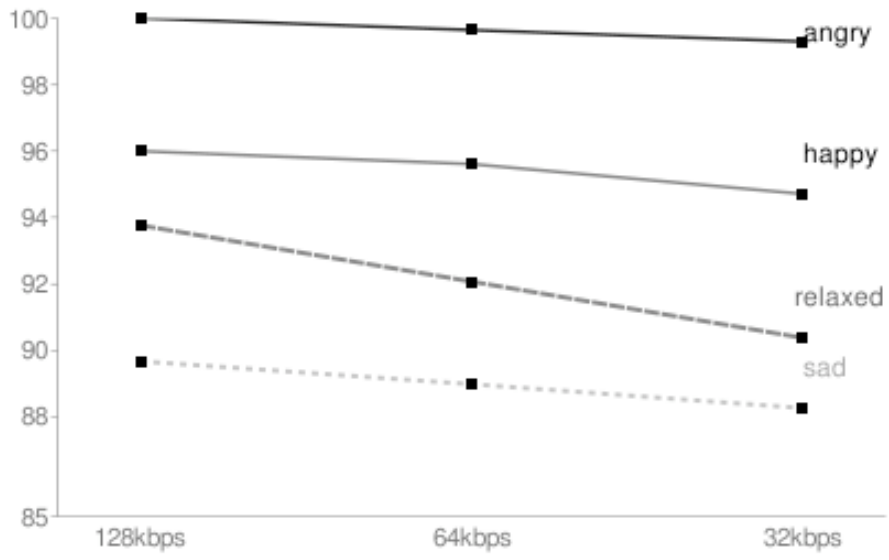

Figure 15: Effect of the audio bit rate reduction on the accuracy (in percentage) for the entire dataset.

We observe degradation due to encoding at a lower bit rate. However, in all cases, this does not seem to have a strong impact. The degradation, in percentage, compared to the original version at 128 kbps is acceptable. 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 also notice that the highest percentage of degradation is 3.6% obtained for the "relaxed" category (with 32 kbps). Even though there is a slight drop in the accuracy, the classification still gives satisfying results.

4. Integration in the PHAROS Project

After explaining the method we used to build the ground truth, extract the features, select the best classification model and evaluate the results and robustness, we discuss here the integration of this technology in the PHAROS search engine framework.

4.1 The PHAROS project

PHAROS[7] (Platform for searcHing of Audiovisual Resources across Online Spaces) is an Integrated Project funded 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 that 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 and integrated the automatic music mood annotation model previously described.

4.2 Integration of the mood annotator

As a search engine, PHAROS uses automatic content annotation to index audiovisual 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 library for Support Vector Machines [6]. The SVMs were trained with full ground truth datasets and optimal parameters. Using a standard XML representation format defined in the project, we wrapped this implementation into a webservice, which could be accessed by other modules of the PHAROS platform. Furthermore, exploiting the probability output of the SVM algorithm, we provided a confidence value for each mood classifier. This added a floating point value that is used for ranking the results of a query by the annotation probability (for instance from the less to the most happy).

The resulting annotator extracts audio features and predicts the music's mood at a high speed (more than twice real-time), with the same performance level than what was presented in the previous section (using the same database). This annotator contributes to the overall system by allowing for a flexible and distributed usage. In our tests, using a cluster of 8 quad-core machines, we can annotate

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

1 million songs (using 30-seconds of each) in 10 days. The mood annotation is used to filter automatically the content according to the needs of users 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 that are 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 of allowing people to find the music they want to hear.

5. User evaluation

In the context of the PHAROS project, a user evaluation has been conducted. The main goal of these evaluations was to assess the usability of the PHAROS platform and in particular, the utility of several annotations.

Protocol

26 subjects participated in the evaluation. They were from the general public, between 18 and 40 years old (27 in average), all of them self-declared eager music listeners and last.fm users. The content processed and annotated for this user evaluation was made of 2092 30-second music videos. After a presentation of the functionalities on site, the users were then directly using an online installation of the system from their home. During 4 weeks, they could test it with some tasks they were asked to do every two days. The task related to our algorithm was to search for some music and to refine the query using a mood annotation. One query example could be to search for "music" and then refine with the mood annotation "relaxed". They had to answer a questionnaire at the end of the study:

-*"Do you find it interesting to use the mood annotation to refine a query for music?"*
-*"Do you find the "mood" annotation innovative?"*
-*"Does the use of the mood annotation correspond to your way of searching for audiovisual information?"*

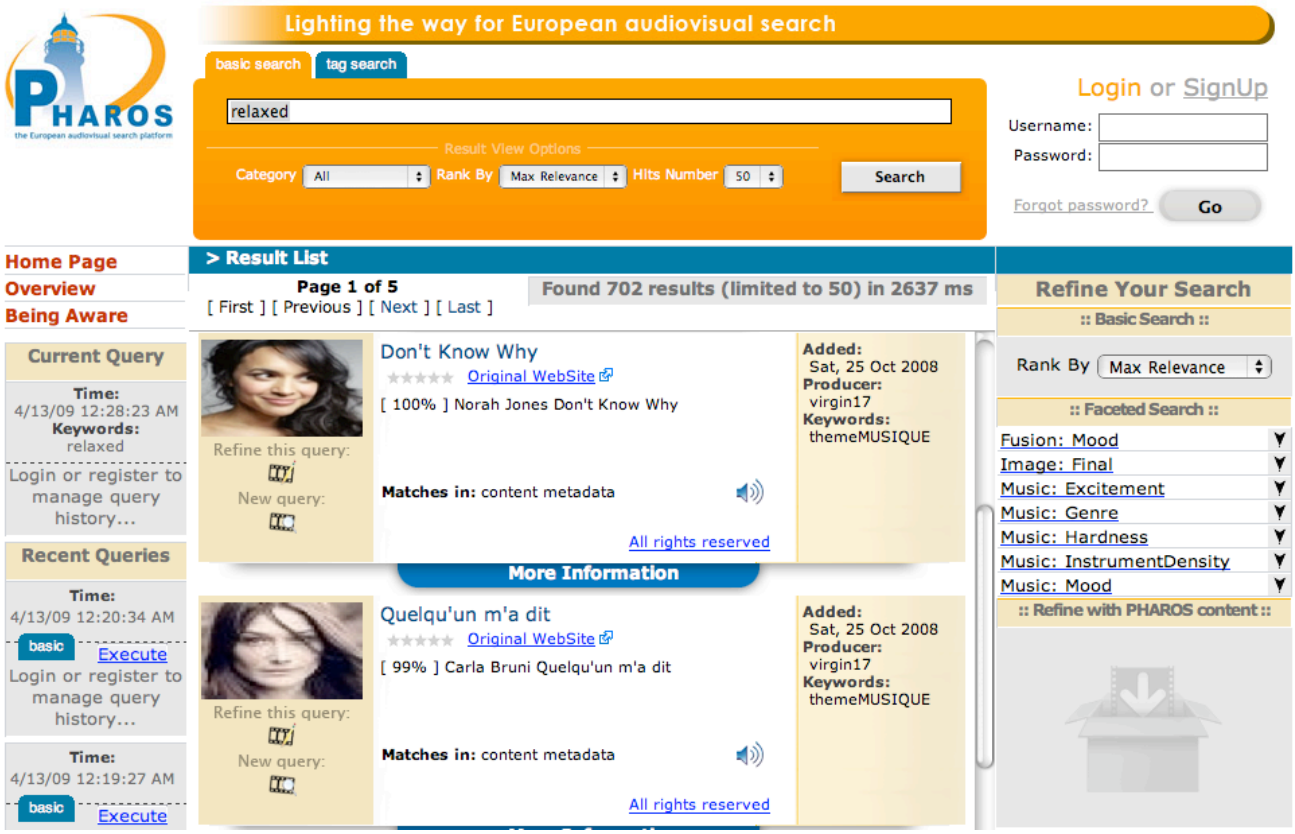

Figure 16. Screenshot of the PHAROS interface used for the user evaluation.

Results

As a general comment, there is difficulty for users to understand directly a content-based annotation. Some effort and thinking has to be done to make it intuitive and transparent. For instance what is "sad=0.58" (music annotated sad with a confidence of 0.58), is it really sad? Is it very sad? The confidence, or probability, value of one annotation is quite relative to other instances and most of all to the training set. This can be used for ranking results but might not be shown to the end-user directly. We should prefer nominal values like "very sad" or "not sad" for instance. Another important point seen when analyzing the comments from the users is the need to focus on precision. Especially in the context of a search engine, people will only concentrate on the first results and may not go to the second page. Instead, they are more likely to change their query. Several types of musical annotations were proposed to the user (genre, excitement, instrument, color, mode and key). From this list, mood was ranked as the second best in utility, just after musical genre (which is often given as metadata). Users had to rate on a scale from 0 to 10 their answer to several questions (0 would be "I strongly disagree" and 10 "I strongly agree"). We summarize here the answers to the questions related to the mood annotation:

-*"Do you find it interesting to use the mood annotation to refine a query for music?"* Users answered positively with a mean of 8.66, standard deviation of 1.85, showing a great interest to use this annotation.

-*"Do you find the "mood" annotation innovative?"* The mean of answers was also positive with 6.18 in average (standard deviation 3.81).

-*"Does the use of the mood annotation correspond to your way of searching for audiovisual information?"* Here users agreed with an average of 6.49 (standard deviation 3.47).

In all cases the mood annotation and its integration into the PHAROS platform was greatly accepted and highly considered by users. They also rated it as the most innovative musical annotation overall. In Figure 16, we show a screenshot of the version of the PHAROS platform installed for the user evaluation. As an open framework, a PHAROS installation can be instantiated with different configurations, features and user interfaces. In this study we used an instance created by taking advantage of Web Ratio[8] (an automatic tool to generate web interface applications). In this screenshot, the user is searching for "relaxed" music. They enter "relaxed" as a keyword and are browsing the musical results. The ones shown here were rated as "relaxed" (respectively 100% and 99%) thanks to the automatic music mood annotator we describe in this article.

6. Discussion and Conclusion

We presented an approach for automatic music mood annotation introducing a procedure to exploit both the wisdom of crowds and the wisdom of the few. We detailed the list of audio features used and revealed some results using those most relevant. We reported the accuracies of optimized classifiers and tested the robustness of the system against low bit rate mp3 encodings. We explained how the technology was integrated in the PHAROS search engine and used it to query for, refine and rank music. We also mentioned the results from a user evaluation, showing a real value for the users in an information retrieval context. However, one may argue that this approach with 4 mood categories is simple when compared to the complexity of human perception. This is most likely true. Nevertheless, this is an important first step for this new type of annotation. So what could be done to improve it? 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 and add useful information (without increasing the noise for the user). Then, we can try to make better predictions by using a larger ground truth dataset or by designing new audio descriptors especially relevant for this task. Another option would be to generate analytical features [30], or to combine several classifiers to try to increase the accuracy of the system. We could also consider the use of other contextual information like metadata, tags, or text found on the Internet (from music blogs for instance). It has also been shown that lyrics can help to classify music by mood [19]. Indeed, multimodal techniques would allow us to capture more emotional data but also social and cultural information not contained in the raw audio signal. We should also focus on the user's needs to find the best way to use the technology. There is a clear need to make the annotation more understandable and transparent. Mood representations can be designed to be more usable than only textual labels. Finally, the mood annotation could be personalized, learning from the user's feedback and his/her perception of mood. This would add much value, although it might require more processing time per user, thus making the annotation less scalable. Nevertheless, it could dramatically enhance the user experience.

7. Acknowledgments

We are very grateful to all the human annotators that helped to create our ground truth dataset. We also want to thank all the people contributing to the Music Technology Group (Universitat Pompeu Fabra, Barcelona) technologies and, in particular, Nicolas Wack, Eduard Aylon and Robert Toscano. We are also grateful to the entire MIREX team, specifically Stephen Downie and Xiao.

[8] http://www.webratio.com

We finally want to thank Michel Plu and Pascal Bellec from Orange R&D for the user evaluation data and Piero Fraternali, Alessandro Bozzon and Marco Brambilla from WebModels for the user interface. This research has been partially funded by the EU Project PHAROS IST-2006-045035.

8. References

[1] Andric A, & Haus G (2006) Automatic playlist generation based on tracking user's listening habits. Multimedia Tools and Applications, 29(2):127-151

[2] Berenson ML, Goldstein M, Levine D (1983) Intermediate Statistical Methods and Applications: A Computer Package Approach. Prentice-Hall

[3] Boser BE, Guyon, IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In COLT '92: Proceedings of the fifth annual workshop on Computational learning theory, (pp. 144-152). New York, NY, USA: ACM

[4] Bigand E, Vieillard S, Madurell F, Marozeau J, Dacquet A (2005) 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

[5] Casey MA, Veltkamp R, Goto M, Leman M, Rhodes C, Slaney M (2008) Content-based music information retrieval: Current directions and future challenges. Proceedings of the IEEE, 96(4): 668-696

[6] Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1-38

[8] Downie, JS (2008) The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. Acoustical Science and Technology, 29(4): 247-255

[9] Duda, R. O. and Hart, P. E. (1973) Pattern Classification and Scene Analysis. John Wiley & Sons Inc, Somerset, New Jersey, U.S.A

[10] Farnsworth, PR (1954) A study of the Hevner adjective list. The Journal of Aesthetics and Art Criticism, 13(1):97–103, 1954

[11] Gómez E (2006) Tonal description of music audio signals. PhD thesis, Universitat Pompeu Fabra

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

[13] Hevner K (1936) Experimental studies of the elements of expression in music. American Journal of Psychology, 58:246-268

[14] Hu X, Downie JS, Laurier C, Bay M, Ehmann AF (2008) The 2007 MIREX audio mood classification task: Lessons learned. In Proceedings of the 9th International Conference on Music Information Retrieval, pp 462–467, Philadelphia, PA, USA, 2008

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

[16] Juslin PN, Västfjäll D (2008) Emotional responses to music: The need to consider underlying mechanisms. Behavioral and Brain Sciences, 31 (5)

[17] Krumhansl CL (1997) An exploratory study of musical emotions and psychophysiology. Canadian journal of experimental psychology, 51(4):336–353

[18] Laurier C, Herrera P (2007) Audio music mood classification using support vector machine. Music Information Retrieval Evaluation eXchange (MIREX) extended abstract

[19] Laurier C, Grivolla J, Herrera P (2008) Multimodal music mood classification using audio and lyrics. In Proceedings of the International Conference on Machine Learning and Applications. San Diego, CA, USA

[20] Laurier C, Herrera P (2009) Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines. IGI Global. pp. 9-32

[21] Le Cessie S, Van Houwelingen JC (1992) Ridge estimators in logistic regression. Applied Statistics, 41 (1), 191-201

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

[23] Lidy T, Rauber A, Pertusa A, Iñesta JM (2007) MIREX 2007: Combining Audio And Symbolic Descriptors For Music Classification From Audio. MIREX 2007 - Music Information Retrieval Evaluation eXchange, Vienna, Austria, September 23-27, 2007

[24] Lindström E (1997) Impact of melodic structure on emotional expression. In Proceedings of the Third Triennial ESCOM Conference, (pp. 292-297)

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

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

[27] Mandel M, Poliner GE, Ellis DP (2006) Support vector machine active learning for music retrieval. Multimedia Systems, 12(1)

[28] Mandel M, Ellis, DP (2007) Labrosa's audio music similarity and classification submissions. MIREX 2007 - Music Information Retrieval Evaluation eXchange, Vienna, Austria, September 23-27, 2007

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

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

[31] Peeters G (2004) A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Tech. rep., IRCAM

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

[33] Quinlan, R. J. (1993) C4.5: programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc

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

[35] Sethares WA (1998) Tuning Timbre Spectrum Scale. Springer-Verlag

[36] Shi YY, Zhu X, Kim HG, Eom KW (2006) 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 Toronto, Canada, pp 1085–1088

[37] Skowronek J, McKinney MF, van de Par S (2007) A demonstrator for automatic music mood estimation. In Proceedings of the International Conference on Music Information Retrieval, Vienna, Austria

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

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

[40] Thayer RE (1996) The Origin of Everyday Moods: Managing Energy, Tension, and Stress. Oxford University Press, Oxford

[41] Tzanetakis G, Cook P (2002) Musical genre classification of audio signals. IEEE Transactions on Audio, Speech and Language Processing, 10(5):293-302

[42] Tzanetakis G (2007) Marsyas-0.2: a case study in implementing music information retrieval systems. In Intelligent Music Information Systems

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

[44] Wedin L (1972) A Multidimensional study of perceptual-emotional qualities in music. Scandinavian Journal of Psychology, 1972;13(4):241-57

[45] Wieczorkowska A, Synak P, Lewis R, and Ras Z (2005) Extracting emotions from music data. In Foundations of Intelligent Systems, Springer-Verlag, pp 456-465

[46] Witten IH, Frank E (1999). Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco

[47] Yang YH, Lin YC, Su YF, Chen HH (2008) A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech, and Language Processing, 16(2):448–457