Audio Engineering Society
Convention Paper

Presented at the 130th Convention
2011 May 13–16 London, UK

Using Support Vector Machines for Automatic Mood Tracking in Audio Music

Renato Panda[1], Rui Pedro Paiva[2]

[1] Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
panda@student.dei.uc.pt

[2] Department of Informatics Engineering, University of Coimbra, 3030-290 Coimbra, Portugal
ruipedro@dei.uc.pt

ABSTRACT

In this paper we propose a solution for automatic mood tracking in audio music, based on supervised learning and classification. To this end, various music clips with a duration of 25 seconds, previously annotated with arousal and valence (AV) values, were used to train several models. These models were used to predict quadrants of the Thayer's taxonomy and AV values, of small segments from full songs, revealing the mood changes over time. The system accuracy was measured by calculating the matching ratio between predicted results and full song annotations performed by volunteers. Different combinations of audio features, frameworks and other parameters were tested, resulting in an accuracy of 56.3% and showing there is still much room for improvement.

1. INTRODUCTION

Throughout our history, music has always been present, linked to us along all life stages and in various forms, both as individuals and in society. Used in areas as diverse as entertainment, sports, health or religion, music transmits different emotions and is perceived differently across cultures and civilizations.

During the past decades we have witnessed a gradual change in music distribution process and an increase both on supply and demand and in the diversity of content available. This coincided with the advent of the digital era, where we assisted a change from more expensive, low capacity distribution means like vinyls and tapes, to new methods such as the Compact Disk (CD). Nowadays, Internet and smaller digital audio formats allow users to have large collections and to easily listen to different songs instantaneously. A similar growth happened in distribution, with today's online music databases containing millions of songs from the most diverse artists and genres. This huge amount of information gives users an overload of information, hindering the simple process of browsing and discovering new songs. This issue highlighted the weaknesses existing in current search methods for multimedia databases. To find a desired type of music, the user has to search based on parameters like genre,

which may comprise a wide range of different music. In alternative, it's possible to browse by artist, album, year and similar information. The drawbacks are evident, namely, information needs to be added manually and users also need to know what they are looking for, inhibiting the process of discovery of new songs or artists. Thus, it would be valuable to have new search methods, able to query music based on its emotion or similarity and extract different kinds of information from audio.

Due to these limitations, we have seen an increasing interest in research areas related with music information retrieval (MIR) over the last decade. Nonetheless, areas like music emotion recognition (MER) are still largely unexplored in part due to its complexity. Different individuals can perceive different emotions from the same musical piece. The same also happens across different cultures and civilizations. Moreover, emotions are highly subjective and even when listeners agree in the perceived emotion, there's still ambiguity regarding its description. Additionally, it is not yet well-understood how and why music elements create specific emotional responses in listeners [1].

Most available studies on emotion recognition in music do not tackle the problem of emotion variations throughout a song. In fact, the aim is usually to find the single emotion that best describes an entire song. Of these, a few address the fact that emotions change during a song and propose ideas to track these changes over time [1], [2].

In this paper we aim to track mood in audio music, specifically its changes over time in terms of Thayer's quadrants [3]. To this end, we use a learning and classification approach with support vector machines (SVM), predicting mood models from extracted audio features. The method is evaluated on a mood-tracking-oriented dataset based on the one provided by Yang. These dataset annotations were used both directly as AV annotations and converted from arousal and valence (AV) values to quadrants of the Thayer plane. The trained models are then used to predict quadrant variations for full songs. To achieve this, two audio frameworks – Marsyas and MIR Toolbox – were used, to extract features in small intervals of 0.75 to 1.5 seconds, predicting quadrants at each specific moment. Several tests were run, using different windows sizes and feature sets, measuring performance of each framework features and parameters. Additionally, the method was evaluated with the subset of the best features, obtained in a previous study using feature selection algorithms such as forward feature selection (FFS) [4]. The predicted results were later compared with full song annotations gathered from volunteers in order to measure prediction accuracy, showing results hovering around 56%.

This paper is organized as follow: In section 2, we describe relevant work that has been done in the area. In section 3, we present the feature extraction process and the employed frameworks. In section 4 we approach the used strategy for classification and regression to create the mood models. In section 5, the quality of the used annotations is discussed. Experimental results for mood tracking are presented and discussed. Finally, conclusions from this study are drawn in section 6 as well as considerations for future improvements.

2. RELATED WORK

For long, mood and emotions has been a major subject in psychology studies, with psychologists aiming to create the best model to represent emotions. This task is complex and has been subject to discussion over the years, largely because of the subjectivity inherent to emotions. Different persons have different perceptions of the same stimulus and often use different words to describe similar experiences.

The existent theoretical models can be separated in two different approaches: categorical and dimensional models. As the name implies, in categorical models emotions are organized in different categories such as anger, fear, happiness or joy. Since this approach is categorical, there is no distinction between songs grouped in the same category, even if there are obvious differences in terms of how strong the evoked emotions are. On the other side, dimensional models map emotions to a plane, using several axes, with the most common approach being a two dimensional model using arousal and valence values. While the ambiguity of such models is reduced, it is still present, since for each quadrant there are several emotions. As an example, emotions such as happiness and excitation are both represented by high arousal and positive valence. To solve this, dimensional models have been further divided in discrete – described above, and continuous. Continuous models eliminate the existent ambiguity since each point on the emotion plan denotes a different emotional state [1].

One of the most effective and simple models was proposed by Thayer in 1989 [3]. It consists on a two dimensional model based on energy and stress, splitting the plane in four distinct quadrants: Contentment, representing calm and happy music; Depression, referring to calm and anxious music; Exuberance, referring to happy and energetic; and Anxiety, representing frantic and energetic music (Figure 1). In this model, emotions are placed far from the origin, since it is where arousal and valence values are higher and therefore emotions are clearer. This model can be considered discrete, as approached in this article, with the four quadrants used as classes or continuous, as used in previous classification works. [1]

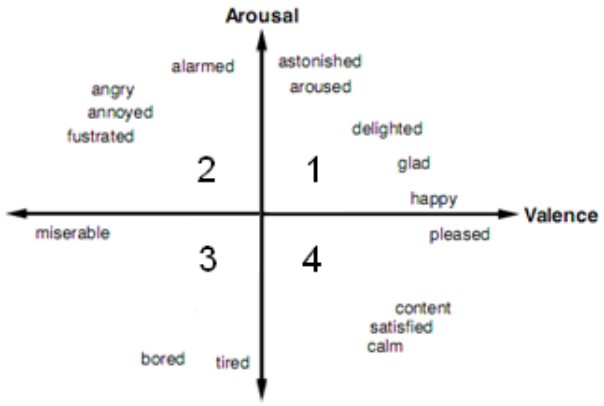

Figure 1 Thayer's model of mood (adapted from [5]).

The relation between music and emotions has been the subject of some studies in the past century, where researchers tried to identify the influence of certain musical attributes such as tempo, mode, rhythm and others on the human subjects. [6]

Few works have tried to identify emotions based on the audio signals. Of these, most of them focused on recognizing the dominant emotion in a given music piece, with few assuming that mood changes along the song and aiming to track it.

Two of the most relevant studies were conducted by Lu et al. [2] and Yang et al. [1]. Although these papers are not devoted to mood tracking, they approach the subject, shedding some light over it and presenting interesting points and ideas.

In [2], Lu et al. proposed a solution to mood detection and tracking using four different emotions represented by the quadrants of the Thayer's model. Two distinct approaches, using hierarchical and non-hierarchical frameworks based on Gaussian mixture models (GMM) are used, together with features of intensity, timbre and rhythm, de-correlated via the Karhunen-Loeve transform. Results from the hierarchical framework, although more complex, were generally better, making a better use of sparse training data, important when the available training data is limited. The detection process is later extended to mood tracking and used to detect changes across quadrants. The new approach tries to find potential mood change boundaries. This method consists in a two-step mood tracking scheme. First, the goal is to find potential boundaries, recurring to intensity (using the intensity outline to detect possible boundaries), timbre and rhythm (to check for possible mood changes in possible boundaries). Then, the musical clip is divided into several independent segments, each containing a constant mood. In mood tracking tests, the results showed that about 84.1% of the boundaries are recalled and the precision is about 81.5%. These values should be regarded with caution, since only nine classical music pieces were used. Some other limitations are the low number of mood categories and the absence of important mood-related features, such as mode and articulation. Other drawback is the need of two passes over each song. The first step is used to detect potential boundaries, while the second is used to divide the musical clip into segments of constant mood.

Yang et al [1] propose a different solution, based on a continuous model using the Thayer's plane. Here, the authors map each song to a point in the plane, using regression and audio features from various frameworks to predict the arousal and valence values. The approach is evaluated using R^2 statistics, reaching an accuracy of up to 58.3% for arousal and 28.1% for valence. Although its focus is classification for the entire 25 second clips, the author discusses the possibility of extending such a system to the mood tracking problem, extracting audio features for smaller segments and treating them as a normal regression problem. A somewhat similar solution was studied by Korhonen et al. [7], modelling the emotional content of music as a function of time and musical features. Using system identification techniques, the authors attain an average R^2 statistic of 21.9% for valence and 78.9% for arousal. The study was conducted recurring to a very small data set composed of 6 excerpts of classical songs.

In this paper we use four categories to track mood over time based on Thayer's plane. Using support vector

classification, the category of each segment is predicted, measuring the matching rating against real annotations of full songs in order to understand the importance of each framework features to this problem.

3. AUDIO FEATURES EXTRACTION

Several studies have examined the relation between musical attributes and emotions over the years. In a recent overview, Friberg [6] lists the following features as relevant for music mood analysis: timing, dynamics, articulation, timbre, pitch, interval, melody, harmony, tonality and rhythm. Other common features not included in that list are, for example, mode, loudness or musical form [8].

However, many of the listed features are usually hard to extract from audio signals or still require further study from a psychological perspective. As a result, many of the features currently in use for MER are typically applied in other contexts such as speech recognition and genre classification. These features usually describe audio attributes such as pitch, harmony, loudness and tempo, mostly calculated recurring to the short time spectra of the audio waveform.

From the available audio frameworks, we opted to use two – Marsyas and MIR Toolbox, based on the features extracted, performance and the fact they are able to extract features for smaller segments. Some interesting alternatives were only capable of extracting values for an entire piece or harder to obtain and use.

Marsyas[1] is a software framework, known for its speed and lightweight. Used in many academic and industry projects, it is able to extract a considerable amount of features and has proven its value by achieving top results in some MIR competitions[2]. Some of the drawbacks are the lack of documentation and lacking some features considered relevant to MER such as tonality and dissonance features. The following features were extracted with Marsyas: zero crossing rate, spectral centroid, rolloff, spectral flux, Mel frequency cepstral coefficients (MFCCs), chroma, spectral crest factor (SCF), spectral flatness measure (SFM), linear spectral pairs (LSP) and linear prediction cepstral coefficients (LPCC).

MIR Toolbox[3] is a free audio framework written for MATLAB. It is considerably heavier and slower than Marsyas but with several advantages, providing some high and low level features not commonly available, has great documentation and it is easier to use. In our work, MIR Toolbox was used to extract: root mean square energy, notes onset time, zero crossings rate, rolloff, high-frequency energy, MFCCs, sensory dissonance (roughness), spectral peaks variability (irregularity), spectral flux, spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral flatness, spectral entropy, chroma, key strength (major, minor), pitch, mode and tonal centroid. Memory constraints prevented us from extracting some extra features.

Several experiments were conducted, testing different combinations of features and window sizes for frame-level features. The values varied from initial tests with 2048 up to 32768 samples per window, nearly 1.5 seconds.

All songs used were converted to WAV PCM format, with 22050 Hz sampling rate, 16 bits quantization and mono. The entire feature sets were normalized with values from 0 to 1.

4. QUADRANTS BASED CLASSIFICATION AND AV REGRESSION

Statistical classifiers are supervised machine learning procedures used to evaluate individual items based on their characteristics and correctly placing them into groups, using as basis a training set of previously labelled items. From the various classification methods available, we opted to use support vector machines (SVMs) as part of our solution since it is capable of both classification (SVC) and regression (SVR). SVMs are one of the most used classifiers nowadays, providing good results in previous studies [1].

The process is divided in two phases: training the model and predicting. During the first stage, the classifier is trained by analyzing a given set of examples, previously identified as belonging to a given category. In a simplified manner, the model is built by mapping all training examples to a plane, in a way that every category is divided by a clear gap as wide as possible. In phase two, the new, unclassified, examples are then

[1] http://marsyas.info/

[2] http://www.music-ir.org/mirex/wiki/2010:MIREX2010_Results

[3] https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

mapped into that same space and predicted to be of a determined category based on which region of the gap they fall on.

In the training process a dataset provided by Yang [1] was used. The dataset consists of 189 clips with 25 seconds each from various genres, mainly Pop/Rock from both western and oriental artists. The clips were selected by specialists, representing the 25 seconds that best represented each song. Each clip was later annotated with arousal and valence values ranging between -1.0 and 1.0, by at least 10 volunteers each. The AV annotations were transformed into quadrant annotations and used with audio features extracted from each clip to train a SVC model. Each framework was initially used separately, to understand how each would perform in the specific problem.

A second approach was also developed, based on regression instead of classification. To this end, AV annotations were used to train 2 different regressors (SVR), one for arousal and other for valence. The created models were then used to predict AV values for full song feature vectors, in a similar way as in the quadrants approach, with the particularity that two models were used simultaneously, one to predict arousal and the other to predict valence. This resulted in the prediction of AV moods across time. Those AV values were then converted to quadrants and compared against manual mood tracking annotations.

In both approaches, the train process was tested using two distinct methods: creating the model using all songs' feature vectors as well as averaging each songs' vectors, resulting in one feature vector per song.

Given the high number of features available, a feature selection algorithm was used to reduce the feature set, selecting an optimal subset for our problem and improving the experimental results, by using only features evaluated as relevant. To this end forward feature selection (FFS) was used. The algorithm is quite simple, starting with an empty feature set, and testing each one individually, ending with a ranking of the best features and its performance. A brief description of the algorithm's steps is presented below, starting with an empty "optimal set":

- Pick one of the remaining features.
- Train a classifier using the optimal set plus the chosen feature.
- Test the previously trained classifier and store the results.
- Repeat for each one of the remaining features.
- Add the best feature to the end of "optimal set" list.
- Repeat the entire procedure until no remaining features are left.

The testing process was conducted using a total of 29 full songs based on Yang's 25-sec song dataset. For each song, a large number of feature vectors were obtained, representing small, equal time segments. Each of these vectors was then used with the emotion models trained before, outputting a predicted category. After the process is concluded, the result is the mood tracking and its changes over time.

In order to measure how the system performs, the matching rate between predictions and real annotations is essential. Creating manual annotations for entire music pieces is difficult and very time consuming. Due to this, we opted to create quadrant annotations, made by two volunteers with music experience. Volunteers were asked to listen carefully to each song a few times and mark the quadrant changes over time, creating a mood track as exemplified in Figure 2.

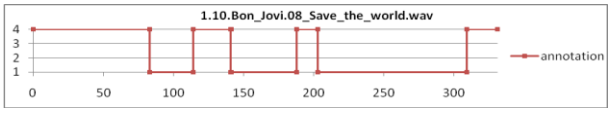

Figure 2 Example of a tracking annotation.

A total of 57 songs were analysed by volunteers, with the matching rate between both annotations being calculated. For our tests only the songs with a matching rate of 80% or higher, a total of 29 songs, were used.

The manual annotations were then compared with the predicted ones, giving us the percentage of matching between both, indicating the accuracy of our solution.

5. MOOD TRACKING EVALUATION

5.1. Annotations evaluation

During the training process, AV annotations provided by Yang [1] were used. The set was created from 195 clips with 25 seconds each, annotated by a total of 253

different volunteers of different backgrounds, in a subjective test, resulting in at least 10 annotations per song. The author tried to have a balanced set of songs between quadrants. Nonetheless, after the subjective test, that was not accomplished, with a clear deficit in quadrant 2. Another problem is the fact that more than 70% of the songs are placed near the origin of the Thayer's model. This situation could have been caused by a significant difference in annotations for the same songs, which could be a consequence of high subjectivity in the emotions conveyed by those songs.

Yang's annotations were then transformed into quadrants annotations and used to create the classifier models. The unbalanced distribution of songs per quadrant is shown below (Figure 3). The situation may cause negative consequences during the training process, affecting the predicted results.

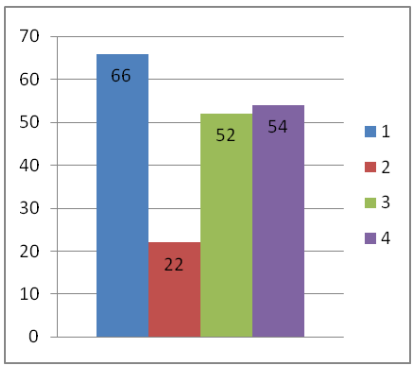

Figure 3 Yang's dataset distribution across quadrants.

To measure accuracy of the mood tracking approach described in this paper, we used manual annotations from full songs created by two volunteers. In addition to being difficult and very time consuming, this process proved to be very complex since users were not only expected to annotate the evoked emotion but also to mark the cue points where they considered the emotion changes. Depending on the volunteers and songs, there may be changes generally detected by many volunteers, however the inverse is also bound to happen, with great differences between volunteers. To reduce the effect of this problem, a high number of annotations per song is needed. This is not the case so far, since we only have annotations from two users in this preliminary study. However, we are presently conducting an annotation study with a large population.

Besides the number of annotations, another limitation is the low number of songs employed and the fact that it is limited to western music.

One last problem with the provided annotations is the fact that, in some cases, they are inconsistent when compared with Yang's annotations. For instance, there are cases where our volunteers annotated only changes between quadrants one and four (e.g.: happy and peaceful), while Yang's subjects used quadrant 3 (e.g.: angry). There are several possibilities for this, from subjectivity to cultural differences. However, this cannot be further investigated due to the small population and number of annotations.

Clearly, the current annotations are not sufficient. To overcome the problems, new, more complete, sets for training and testing should be gathered. Especially in terms of full song annotations, the number of songs of the dataset must be increased and more volunteers be used to manually annotate mood changes.

5.2. Mood tracking results

The results were measured by calculating the matching ratio between predictions and the respective annotations for full songs.

Without feature selection and using the quadrants approach, the results between both frameworks' feature sets were generally similar. Using Marsyas features, a maximum accuracy of 53.45% was obtained with 1.5 seconds windows. In the case of MIR Toolbox features, results were close, with 52.7% matching and really small changes with both windows sizes. This was the only time Marsyas scored higher than MIR Toolbox, coming second in all other tests. Using both frameworks the accuracy increased marginally. As for feature selection, using a subset of 38 features from MIR Toolbox, results rose nearly 2%, while Marsyas dropped 1%.

By analyzing the predictions for each song it was possible to observe a huge difference in matching rate in several cases. Results were good when songs were mainly on quadrant 1 and also 4. On the other hand, results collapsed when predicting quadrants 3 and especially 2.

To further investigate this problem we applied our second approach, using the AV annotations to train a new regression model. This model was then used to

predict AV values for the full songs. To measure accuracy the predicted mood tracking was only then converted to quadrants and matched against our manual annotations. The results showed little to no change in case of Marsyas, with the best results actually dropping to 48.9%. However, MIR Toolbox results went up to 55.95% for 1.5 seconds windows. This is consistent with another study we are conducting presently, where MIR Toolbox performed better for valence prediction, while Marsyas features proved to be quite ineffective. Applying feature selection improved all results with MIR Toolbox reaching 56.30%.

A summary with the most relevant results, obtained with 1.5 seconds windows is shown below (Table 1).

	All features		Feature selection	
	Quadrants	AV	Quadrants	AV
Marsyas	53.45%	48.90%	52.55%	50.89%
MIR Toolbox	52.70%	55.95%	54.51%	56.30%
Marsyas + MIR T.	53.66%	55.95%	54.72%	54.96%

Table 1 Mood tracking results

6. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed an approach to automatic mood tracking in the Thayer's plane, using classification and regression. We also studied the influence of both frameworks features in the given problem and studied the creation of full song annotations to validate mood tracking problems.

We tested two distinct ideas regarding the annotations used. In the first case, we undertook the training process with quadrant annotations, by converting the AV annotations to quadrants of the Thayer's plane. They were then used to train a support vector classifier (SVC) and predict the quadrant of each segment. The second case was based on regression, by using the AV values directly to train two different support vector regressors (SVR). This approach resulted in the prediction of AV values, which were only then converted to quadrants in order to measure accuracy.

To measure the accuracy of our solution, the matching ratio of the predicted results against the full song annotations was calculated. So far the attained accuracy reaches 56.3% for 1.5 seconds segments using a subset of MIR Toolbox features and the AV approach. It is also noticeable that, for the quadrants approach, with or without feature selection, the obtained results are very similar between feature sets of both. The opposite is verified when using the AV approach, with MIR Toolbox performing better by more than 6%.

Analyzing the resultant mood tracks, a higher accuracy predicting segments in the first and fourth quadrants is observable, as opposed to the remaining two quadrants, which depend on negative valence. Previous classification studies demonstrated that valence depends on a small set of specific features to perform acceptably. Another possible cause is the fact that our set of songs is small and unbalanced, in part due to the subjectivity and complexity of emotions and the fact that subjects tend to more frequently agree when identifying happier songs and their arousal variation, than songs with a negative valence.

Several aspects are likely to improve our results and we aim to approach them in a near future. Some of the most relevant are:

- Use a smaller and more relevant set of audio features, obtained by using better feature selection algorithms.
- A balanced data set, that will improve SVC and SVR training, as well as feature ranking.
- Improved tracking annotations, carried out by a higher number of subjects and over a bigger data set.
- Extra audio features, that showed to be good in mood classification and tracking problems, according to the literature such as [1].

7. ACKNOWLEDGEMENTS

This work was supported by the MOODetector project (PTDC/EIA-EIA/102185/2008), financed by the Fundação para Ciência e Tecnologia - Portugal.

8. REFERENCES

[1] 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, vol. 16, Feb. 2008, pp. 448-457.

[2] 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, Jan. 2006, pp. 5-18.

[3] R. E. Thayer, "The Biopsychology of Mood and Arousal," Oxford University Press, New York, 1989, pp. 234.

[4] S.L. Chiu, "Selecting input variables for fuzzy models," Journal of Intelligent and Fuzzy Systems, vol. 4, 1996, pp. 243-256.

[5] Y.E. Kim, E.M. Schmidt, R. Migneco, B.G. Morton, P. Richardson, J. Scott, J.A. Speck, and D. Turnbull, "Music Emotion Recognition: A State of the Art Review," International Society of Music. Information Retrieval (ISMIR), 2010, pp. 255-266.

[6] A. Friberg, "Digital Audio Emotions - An Overview of Computer Analysis and Synthesis of Emotional Expression in Music," Proc. of the 11th Int. Conference on Digital Audio Effects (DAFx-08), 2008, pp. 1-6.

[7] M.D. Korhonen, D. a Clausi, and M.E. Jernigan, "Modeling Emotional Content of Music Using System Identification," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 36, Jun. 2006, pp. 588-599.

[8] O. C. Meyers, "A mood-based music classification and exploration system", MSc thesis, Massachusetts Institute of Technology, 2007.