Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009) Jyväskylä, Finland
Jukka Louhivuori, Tuomas Eerola, Suvi Saarikallio, Tommi Himberg, Päivi-Sisko Eerola (Editors)

Exploring Relationships between Audio Features and Emotion in Music

Cyril Laurier,[*1] Olivier Lartillot,[#2] Tuomas Eerola[#3], Petri Toiviainen[#4]

[*]*Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain*
[#]*Department of Music, University of Jyväskylä, Jyväskylä, Finland*
[1]cyril.laurier@upf.edu, [2]olivier.lartillot@campus.jyu.fi, [3]tuomas.eerola@campus.jyu.fi,
[4] petri.toiviainen@campus.jyu.fi

ABSTRACT

In this paper, we present an analysis of the associations between emotion categories and audio features automatically extracted from raw audio data. This work is based on 110 excerpts from film soundtracks evaluated by 116 listeners. This data is annotated with 5 basic emotions (fear, anger, happiness, sadness, tenderness) on a 7 points scale. Exploiting state-of-the-art Music Information Retrieval (MIR) techniques, we extract audio features of different kind: timbral, rhythmic and tonal. Among others we also compute estimations of dissonance, mode, onset rate and loudness. We study statistical relations between audio descriptors and emotion categories confirming results from psychological studies. We also use machine-learning techniques to model the emotion ratings. We create regression models based on the Support Vector Regression algorithm that can estimate the ratings with a correlation of 0.65 in average.

I. INTRODUCTION

Psychological studies have shown that emotions conveyed by music are not so subjective that they are invalid for mathematical modeling (Laurier & Herrera, 2009). Moreover, Vieillard et al. (2008) demonstrated that within the same culture, the emotional responses to music could be highly consistent. These results indicate that modeling emotion in music should be feasible.

Recently, in the Music Information Retrieval community, emotion classification of music has become a matter of interest, mainly because of the close relationship between music and emotions (Laurier & Herrera, 2009). In the present paper, we explore the relationships between emotions and audio features automatically extracted from the signal (raw digital data). There exist several studies dealing with automatic content-based mood or emotion classification, like for instance Lu (2006) or Yang (2008). Although this task is quite complex, satisfying results can be achieved if the problem is reduced to simple representations. However, almost every work differs in the way that it represents emotions. Similarly to psychological studies, there is no real agreement on a common model (Juslin & Västfjäll, 2008). Some consider a categorical representation based on mutually exclusive basic emotions such as “happiness”, “sadness”, “anger”, “fear” and “tenderness” (Lu, 2006), while others prefer a multi-labeling approach like Wieczorkowska (2005) (using a rich set of adjectives). The latter is more difficult to evaluate since they consider many categories. Other works use the dimensional representation (modeling emotions in a space), like Yang (2008). They model the problem with Thayer arousal-valence emotion plane (Thayer, 1996) 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. All these proposed approaches use a classification or regression problem and a "bag of features" (many features given to a classifier as a black box). However only a very few investigate the relationships between audio features and emotion dimensions or categories. In the rest of the paper, we show some relevant audio features to model emotions and we try to give some explanations based on musicology or psychological research. In Section 2, we will expose the method employed to build the ground truth and extract audio features. In Section 3, we will give some results about relevant audio features and how they relate to emotion ratings. In Section 4, we will detail our results in modeling the ratings using classification and regression techniques. Finally we will conclude in Section 5 and open the discussion about possible future works.

II. METHOD

A. Ground Truth

Our research is based on ground truth data created in a previous study by Eerola & Vuoskoski (submitted), where 110 15-second excerpts from film soundtracks were evaluated by 116 participants using 5 target emotions from the basic emotion model (fear, anger, happiness, sadness, tenderness) and 6 polar extremes from the three-dimensional model (valence, energy arousal and tension arousal). In the present study, we concentrate only on the basic emotion ratings. The participants were 116 university students aged 18-42 years (mean 24.7, SD 3.75, 68% females and 32% males). 48% of the participants were non-musicians, 41% had received some level of musical training, and 11% had music as a hobby for less than 3 years. It was decided to use film music because it is composed with the intention to convey powerful emotions, and could be considered as a relatively ‘neutral’ stimulus material in terms of music preferences and familiarity. Different film genres have been considered. The selection was made after a first experiment, detailed in Eerola & Vuoskoski (submitted), involving 360 rated excerpts. The aim was to keep an even distribution between the basic emotion and dimension rating values. The resulting dataset is made of 110 musical excerpts with mean ratings of basic emotions. Looking at these rating values, we note a high correlation between anger and fear ($r = .69$, $p < .001$), which shows an overlap between these two categories.

B. Audio Feature Extraction

In order to analyze the music from audio content, we automatically extracted a rich set of audio features based on temporal and spectral representations of the audio signal. For each excerpt of the dataset, we merged the stereo channels

into a mono mixture and its 200 ms frame-based extracted features were summarized with their component-wise statistics.

We extracted audio features such as:

- **Timbral**: Barkbands, 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
- **Rhythmic**: bpm, bpm confidence, zero-crossing rate, silence rate, onset rate, danceability

We obtained 200 feature statistics (minimum, maximum, mean, variance and derivatives). From this data we analyzed the distributions compared them to emotional ratings. We computed the correlations and tried machine-learning techniques to model emotion in music.

III. AUDIO FEATURES AND EMOTIONS

Exploring the data generated by the extraction of audio features is an arduous task. In this section, we will detail some interesting correlations found between single audio descriptor and emotion ratings. The ground truth is made of 110 excerpts, each one rated with the 5 basic emotions on a 7 point scale. We used means of these ratings to compare with the descriptor values. Moreover we evenly split the ground truth in 5 parts according to the emotion categories.

A. Dissonance

Consonance and dissonance are known to be relevant in emotion perception (Koelsch, 2006). The dissonance audio feature, also known as "roughness" (Sethares, 1998), 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. We computed the mean dissonance values over the frames of each excerpts and compare it to the ratings. In Table 1, we expose the correlation coefficients of the dissonance values with the emotion ratings.

Table 1. Correlation coefficients of the dissonance means with the emotion ratings

Emotion ratings	Correlation with dissonance
Happiness	0.09437118
Sadness	-0.581796517141
Tenderness	-0.502293010937
Fear	0.39795086
Anger	0.586563518825

Although there seems to be no correlation between the automatically extracted dissonance and the happy category, we note a positive correlation with fear and anger. This also relates to psychological studies stating that dissonant harmony may be associated with anger and unpleasantness (Wedin, 1927), (Hevner, 1936), (Zwicker & Fastl, 1999). We also observe a negative correlation with the sad and tender category.

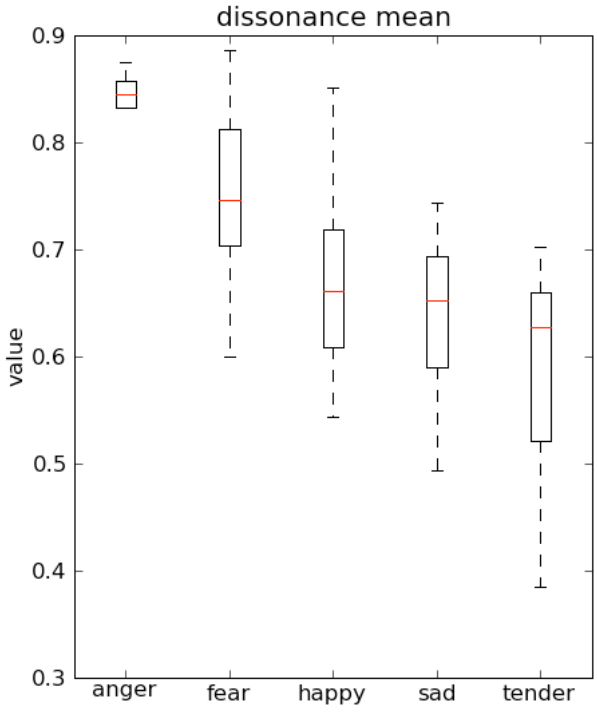

Figure 1. Box-and-whisker plot of the dissonance for each basic emotion categories

Looking at the distributions plotted in Figure 1, there is a clear link between dissonance and anger. Moreover we note that the happy, sad and tender excerpts have relatively lower values, which could indicate a correlation between consonant music and more pleasant emotions (if we consider that anger and fear are unpleasant emotions).

B. 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 (2006) explains how to compute an estimation of the mode from raw audio data. In Table 2, we represent the percentages of estimated major and minor music in the different emotion categories. We note a high percentage of the major mode in "happy" pieces (95% against 5%) and a high percentage of the minor mode in "sad" excerpts (85% against 15%). Moreover tender music appears to be mainly in major mode (95%). The fear category is mostly minor (80%). The only ambiguous case is for the anger category with a more even distribution. In music theory and psychological research, the link between valence (positivity of the emotion) and the musical mode has already been demonstrated. Music in a major mode tends to be more positive than music in a minor mode. The results we achieve

extracting the mode directly from the audio data confirms this statement, also showing the potential of this descriptor in detecting the emotion in music.

Table 2. Distribution of the Mode feature automatically extracted from the audio content over the different categories

Category	Major	Minor
Happiness	**95%**	5%
Sadness	15%	**85%**
Tenderness	**95%**	5%
Fear	20%	**80%**
Anger	40%	60%

C. Onset Rate

Rhythm is an important musical feature to express different emotional aspects (Juslin & Laukka, 2004). From psychological studies we note that, roughly, the faster the more arousing is a musical piece. One possibility to look at rhythmic information is to compute the onset rate (number of onset in one second). An onset is an event in the music (any note, drum etc…). The onset times are estimated looking for the peaks in the amplitude envelope. The onset rate is an estimation of the number of events in one second, which is related to a perception of the speed. In Figure 2, we compare the onset rate values for the different emotion categories. It shows that "happy" songs have higher values, which confirms some psychological results (Juslin & Laukka, 2004) that "happy" music tends to be fast. It seems also quite coherent to have lower values for sad and tender excerpts. Moreover, observing high values for the fear category is coherent. However, surprisingly, the anger category has a wide range of onset rate values, which means that this descriptor might not be that useful for this particular category.

D. Loudness

The loudness of a musical piece is seen as a relevant musical feature to express emotions (Juslin & Laukka, 2004). To get an estimation of the loudness, we compute the energy of the signal within a 2 seconds window; we normalize it and take the mean over the excerpt. In Figure 3, we can observe the distributions of the loudness values for each emotion category. We note a low range of high values for the happy and anger categories. This seems quite related with arousal. For the tender excerpts, the values are relatively higher that one could expect. The importance of loudness to manipulate emotion in music (Nagel, 2008) is confirmed by these different distributions over the emotion categories.

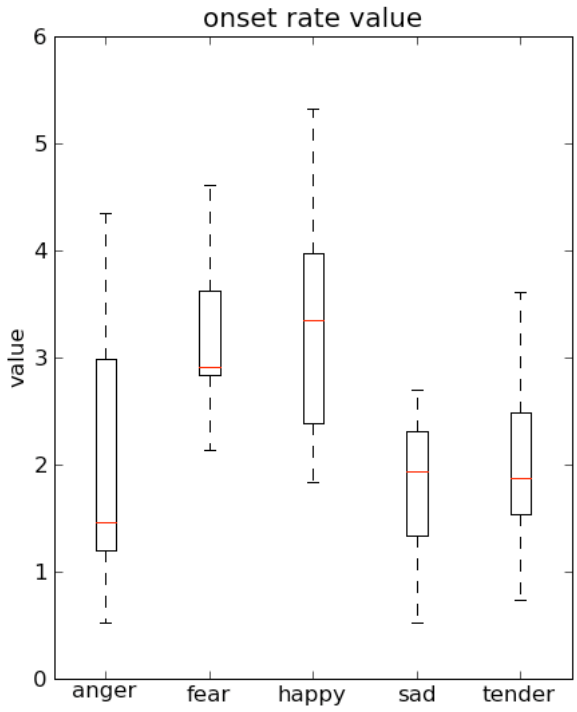

Figure 2. Box-and-whisker plot of the onset rate for the each basic emotion categories

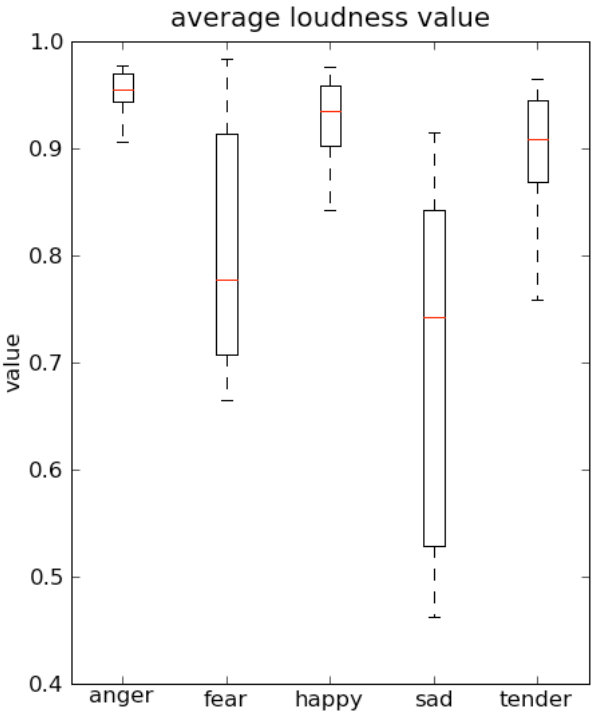

Figure 3. Box-and-whisker plot of the loudness

IV. MACHINE LEARNING APPROACHES

In order to know how well these audio features can model emotions, we employed machine-learning techniques using all the 200 features mentioned in Section 2. To model category ratings, we used Support Vector Machines (SVM) as a classifier (Boser, 1992) with the libsvm implementation[1], Radial Basis Function kernel, optimizing the cost and gamma parameters using a grid search. To model numerical ratings, we used the SVMReg regression algorithm implemented in the WEKA classification software (Witten & Frank, 1999).

A. Learning categories

Our ground truth is divided into 5 categories: fear, anger, happiness, sadness and tenderness. The mean accuracy of a SVM classifier on 10 runs of 10-fold cross-validation is 66%. In Table 3 we show the confusion matrix of this SVM classifier. It shows some confusion of the classifier between anger and fear (high valence, high arousal), and anger and happiness (high arousal). Considering that the baseline for a 5 categories classification task is 20%, it shows that audio features can model these categories in a quite satisfying way. Moreover the perceptual overlap between the anger and fear categories noticed within the listener ratings might lower the accuracy of the classifier. These two categories might not be so clearly mutually exclusive.

Table 3. Confusion Matrix of the SVM classifier

Classified as ->	Anger	Fear	Happy	Sad	Tender
Anger	15	3	4	0	0
Fear	5	13	1	3	0
Happy	5	0	13	0	4
Sad	1	1	1	16	3
Tender	0	0	3	3	16

B. Learning ratings

Using SVM Regression and a Radial Basis Function kernel (RBF), we modeled the ratings of basic emotions. We tried to model the mean rating values (averaged over all listeners). In Table 4, we show the mean correlations of the models with the ratings, based on 10 runs of 10-fold cross validation for each category. For a comparison purpose, we also computed the results using the Linear Regression algorithm. It shows a relatively high positive correlation between the model based on audio features and the ratings. The most difficult category to model seems to be the "happiness" category. Nevertheless, we obtain a mean correlation of 0.65. We observe that the SVM Regression models are more accurate than the Linear Regression ones. For all emotion categories except fear, this difference is statistically significant ($p<0.05$). These results are quite encouraging and demonstrate that we can model the emotion ratings of the listeners to a certain extent. We should also keep in mind that we are considering means ratings and that there is a "glass-ceiling effect" imposed by the inconsistency between listeners and also by the self-inconsistency of each person between different ratings.

Table 4. Correlation between the SVM and Linear regression models based on audio features for the different emotion categories. '*' means that the result of one model is significantly higher than the result of the other ($p<0.05$).

Category	SVM (corr)	Linear Regression (corr)
Anger	**0.65***	0.54
Fear	**0.67**	0.65
Happiness	**0.59***	0.48
Sadness	**0.69***	0.62
Tenderness	**0.67***	0.50

V. CONCLUSION

Based on music made to create emotions (film soundtracks), we observed from the audio analysis a confirmation of some psychological results regarding relevant musical features to express emotions. We reported on important relations between the dissonance, onset rate and loudness audio descriptors and the annotated categories. We showed the correlation of the mode estimation with valence (except for instances classified as "anger"). We also modeled the emotion ratings and created emotion classifiers using Support Vector Machines. This works helps to demonstrate that the information we can automatically extract from the audio signal is relevant and can be used to classify music by emotion. However we should also notice the limitation of these techniques due to the subjectivity of this problem. Future works will consist in comparing the relevant features between different genres, trying to find other relationships between audio descriptors and musical emotions and designing new audio features especially useful for this task.

ACKNOWLEDGMENT

We are very grateful to all the human annotators that helped to create the ground truth. This research has been partially funded by the EU Project PHAROS IST-2006-045035.

REFERENCES

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.

Eerola & Vuoskoski (submitted), A comparison of the discrete and dimensional models of emotion in music.

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

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

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).

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

Koelsch, S., Fritz, T., Cramon, D. Y. V., Müller, K., & Friederici, A. D. (2006). Investigating emotion with music: an fmri study. *Human Brain Mapping*, 27 (3), 239-250.

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

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

Laurier C, Herrera, P (2009). Automatic Detection of Emotion in Music: Interaction with Emotionally Sensitive Machines. *Handbook of Research on Synthetic Emotions and Sociable Robotics: New Applications in Affective Computing and Artificial Intelligence*. chap. 2, (pp. 9-32). IGI Global.

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.

Nagel, F., Kopiez, R., Grewe, O. & Altenmüller, E. (2008) Psychoacoustical correlates of musically induced chills. *Musicae Scientiae*, 12(1), 101-113

Sethares WA (1998). *Tuning Timbre Spectrum Scale*. Springer-Verlag

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

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.

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

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

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

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.

Zwicker, E. & Fastl, H. (1999). Psychoacoustics: Facts and models (2nded.). Berlin: Springer.