An evaluation of feature types for mood estimation on a newly presented dataset of emotion and color responses to music

Primož Godec[1], Matevž Pesek[1], Mojca Poredoš[1], Gregor Strle[2], Jože Guna[3], Emilija Stojmenova[3], Matevž Pogačnik[3] and Matija Marolt[1]

[1]*University of Ljubljana, Faculty of computer and information science*
[2]*Institute of Ethnomusicology, Scientfic Research Centre of the Slovenian Academy of Sciences and Arts*
[3]*University of Ljubljana, Faculty of Electrotechnics*
E-mail: {matevz.pesek, matija.marolt}@fri.uni-lj.si, {primoz.godec, mojca.poredos}@lgm.fri.uni-lj.si
gregor.strle@zrc-sazu.si, {joze.guna, emilija.stojmenova, matevz.pogacnik}@fe.uni-lj.si

Abstract

This paper presents a new dataset gathered containing perceived and induced emotions for 200 audio clips. The gathered dataset also provides users' association of color for each clip, along with users' demographic and personal data, such as users' emotion state, preferred genres, music experience and daily inference, and others. With an online survey we collected more than 7000 responses for a dataset of 200 audio excerpts, thus providing about 37 user responses per clip. We introduced a new methodology for gathering user perception of emotions in a form of two new interfaces - the MoodGraph and MoodStripe. We present a preliminary evaluation of classifying the present emotions with a regression algorithm using MFCC and Chroma on the gathered dataset, and perform a comparison towards other datasets and algorithms.

1 Introduction

This paper tackles the problem of mood estimation in the field of music information retrieval (MIR). There are several applications of the mood estimation, for example in order to boost the efficiency of music recommendation systems. In order to develop such music recommendation system, an annotated dataset is needed for training. Several datasets were previously gathered, yet, to our knowledge, no dataset contained demographic and background data of the users who provided the annotations. Since the mood itself is highly subjective, we believe a dataset annotated by a large group of people can provide a solid base for further research in this field.

In order to obtain a statistically significant amount of user responses per song, we performed user response gathering over a dataset for 200 audio clips. In addition to responses on music clips, we collected some other participants' demographic data and perception of mood. This data might help us to understand difference in responses to audio clips and possibly find correlations between users with similar background. In order to evaluate the usefulness of the collected data and possibly highlight the importance of the relations between the modalities and the user's personal data, we performed a preliminary evaluation of the mood estimation algorithm using the regression for valence-arousal prediction, as described in [11]. This algorithm was tested on our dataset and Mood Swing Turk dataset [12].

Several datasets for mood estimation were previously presented. We intend to overcome some drawbacks of these datasets, such as a small amount of annotators, lack of background information (e.g. genre preference might bias one annotators perception regarding the perceived emotions). Eerola et al [5] performed a gathering of the film music dataset. Each sound track provides a single mean rating with label and values in three-dimensional valence-arousal-tension model. The dataset contains values for 361 film music clips. The Mood Swings Turk Dataset contains on average 17 valence-arousal ratings for each of the 240 audio clips [12]. Clips in this dataset are mostly excerpts from popular music. The Cal500 provides mood labels for 500 western popular songs [14] encompassing 3 annotations per song. The MTV Music Dataset contains 5 bipolar valence-arousal ratings for 192 popular songs [13]. These songs were obtained from the MTV channel playlists.

Mood estimation and perception has been explored on some of the mentioned datasets. However, there are several undiscovered relations that brought the problem to our attention. Music mood estimation has become a recognised task in the past year; the music information retrieval evaluation exchange (MIREX) organises mood classification task since 2007. Several machine-learning-driven approaches have already been presented for the mood estimation task. Schmidt et al. [11] use regression for mood classification. Panda et al. [8] use support vector machines, k-nearest neighbours, C4.5 and naive bayes. Support vector machine was also used by [6]. Barthet et al. [2] use support vector regression for classification.

The paper is structured as follows: section 2 describes the survey and it design, section 3 provides analyses of the gathered data and survey evaluation and section 4 concludes the paper and describes our future work.

2 Online survey

For the user responses gathering procedure, an online survey was used. In order to fully anticipate possible drawbacks of previously collected annotations, we carefully picked the set of labels for emotions by performing a preliminary study over a large label set and identifying the key labels most fitting the majority of participants.

Some basic emotion labels exist, e.g. [4]; however, there is no standard set in music and mood research to our knowledge. Others choose label sets intuitively, without any explanations [15]. In order to use an optimal set of labels, we prepared preliminary survey.

Participants were asked describe their emotional state on scale from 1 to 7 for each of 48 emotional labels. We also observed responses about color perception on a continuous color wheel used to describe connection between mood and colors. Depending on results of this questionnaire we selected 17 basic emotion labels which strongly correlate to three basic components that explain 64% of the variance to the dataset. Depending on participants' responses and results we also decided to restrict continuous color wheel; we developed a discrete-scale color wheel containing 49 colors.

2.1 The survey

By incorporating the conclusions of the preliminary survey, we performed a second online survey on a larger set of participants. We structured the survey into three sections. We captured participants' personal background and demographic information in first part. Users were asked to answer the demographic questions about the following aspects: age, area of living and native language along with information about users' music education, genre preference, and the amount of time listening to the music. We speculate about the importance of such data being gathered along with the mood perception responses for the annotated dataset.

Second part contains questions about participants' perception of mood and connection between color and mood. First, the participants were asked about their current mood state in a set of three separate tasks. A participant was asked to pin-point a location best matching his current mood in the valence-arousal space. The valence-arousal space is a 2D plane, describing the pleasantness on one and activeness on the second axis. We also inquired about the user's perception of current mood by selecting best-matching color in a color wheel.

Instead of gathering the participants' mood using the standardised Likert scale questionnaire, we introduced a new interface, described in [9], named the MoodStripe (see Fig. 1). The participant was asked to drag each label describing one emotion onto a one-dimensional canvas. The canvas possesses a continuous scale between *completely absent* and *significantly expressed*.

The second part contained also two tasks of capturing their perception of emotions. First, the user was asked to place 10 basic emotional labels onto valence-arousal space according to their view of the activeness and pleasantness of the emotion. For this task we developed a second interface named *the MoodGraph* (see Fig. 2), extending the moodStripe into a two dimensional canvas. The user was also asked to pick a best-matching color for each music excerpt.

In the third part of our survey, the participant was asked to respond to audio clips. The clips are 15 second long and was randomly selected from the dataset of 200 clips. To avoid overwhelming the participant, only 10

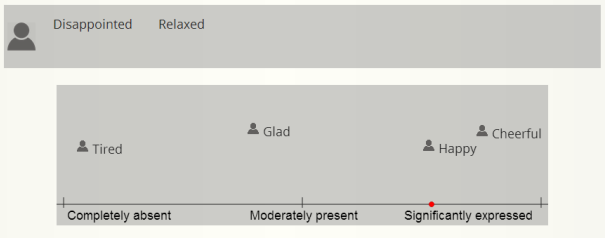

Figure 1: The new interface used in our survey, named *the MoodStripe*. With the drag-and-drop technique, a user can place emotion labels onto the plane depending on how the dragged emotion expressed. Placing the label towards the left side of the canvas reflects the absence of the emotions, where as placing it towards the right expresses the presence.

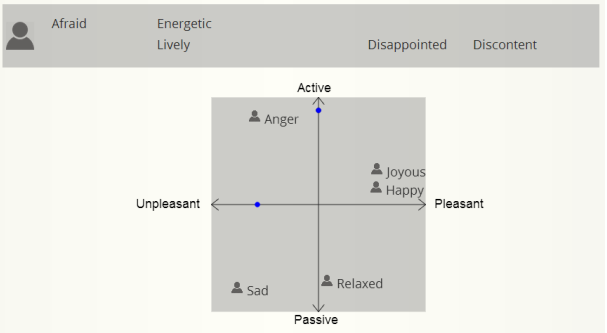

Figure 2: The MoodGraph - a two-dimensional extension of the moodStripe interface, with one category of nominal values. On the x axis *valence* (pleasantness) raises from the left to the right, on the y axis raises *arousal* (activeness) from the bottom to the top.

clips are presented during each participation. All selected audio clips are unknown for most of the participants, to avoid bias due to the familiarity of song. We gathered music from four sources. Eighty songs were chosen from the free online music service Jamendo. We selected songs from as many genre as possible. Next 80 clips was from film music dataset and is described here [5]. We also provide 20 clips from collection of slovenian folk music collection and 20 of the contemporary electro-acoustic music collection.

The participant performed two tasks on each provided audio clip: selecting the best-matching color reflecting the audio, and to describe the perceived and induced emotions of the audio. The latter was performed by using a two-category moodgraph. We provided two categories of labels, one with perceived and the other with induced emotion labels. The user was asked to select at least one label from each category and place them into the valence-arousal space (Fig. 2).

3 Results

We collected 7187 responses from 1357 participants in our survey. The dataset contains 200 audio clips, resulting in collecting 37 responses per audio clip on average. To our knowledge, there is no mood-music dataset with such high ratio per clip. Each response contains at least

Table 1: Results for our and Mood Swing dataset in prediction valence arousal using regression using MFCC and Chroma. Three rates were used to measure accuracy: average distance to mean valence-arousal value, average distance to nearest valence-arousal value, average distance to mean valence-arousal value measured in number of standard deviations.

Feature	Our dataset			Mood Swing dataset		
	Avg. distance	Near. distance	Avg. dist in sdt	Avg. distance	Near. distance	Avg. dist in sdt
MFCC	0.2243	0.0611	0.5357	0.2448	0.0641	0.6514
Chroma	0.2215	0.0614	0.4993	0.3316	0.1026	0.8940

two positioned labels - one describing a perceived and one an induced emotion, and a picked color tone for the audio clip.

The following section presents a preliminary demographic analysis, along with a performed experiment using the regression for mood estimation on the collected dataset.

3.1 Demographic analysis

The basic demographic characteristics of the 952 participants are as follows. The average age of participants was 26.5 years, the youngest had 15, the oldest 64 years. 65% of participants are women, 66% are from urban areas. 50% have no music education, 47% do not play instruments or sing. The amount of music listening per day is evenly spread from less than 1 hour to over 4 hours. 3% claimed they were under the influence of drugs when taking the survey.

3.2 Predicting valence-arousal values using regression

We performed a simple mood estimation using the regression on the collected dataset. We implemented the regression algorithm described by [11]. The least squares method on mel-frequency cepstral coefficients (MFCC) [7] and Chroma [3] features were used. The MFCCs was calculated with the cepstral coefficients 20, Chroma was calculated using 12 bins. Features was calculated using LibROSA python library on a 15 seconds audio clips from dataset. The dataset was divided into 2 training (70%) and testing (30%) sets for the least squares method [1]. Training performed by using the mean values of valence and arousal components of the responses. The algorithm separately evaluated the valence and arousal components using following equation:

$$y = Xb \tag{1}$$

where X is features matrix. Each row in X presents feature vector for single audio clip. b is the transformation vector from feature matrix to prediction vector y.

Results using regression algorithm are shown in table 1. For regression with MFCC and regression with Chroma average distance between estimated and mean valence-arousal value was calculated. We also provide the average distance to nearest value in dataset and the average distance between the mean valence-arousal point and the prediction, measured in the size of the standard deviation of data. Same algorithm was also used on the features Schmidt et al. provides for Mood Swing dataset [12] to compare results.

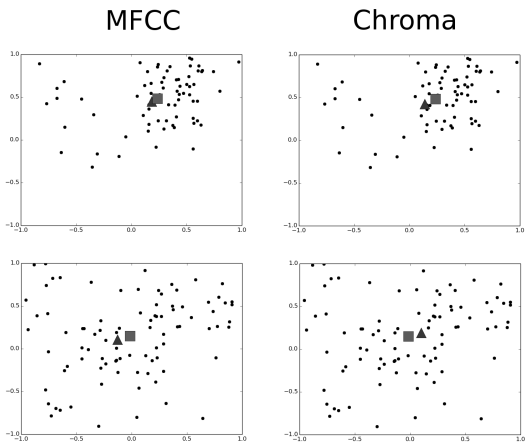

Figure 3: Image shows 4 valence-arousal spaces, where pleasantness raises from left to right on abscissa and activeness for bottom to top on ordinate axis. The first row shows prediction on songs with IDs 13 (first row) and 71 (second row) using MFCC and Chroma. The triangle presents regression algorithm prediction, where as the square marks mean value of all valence-arousal values gathered with survey. Dots indicate all gathered valence-arousal responses.

The results show a better correlation between MFCC and valence-arousal than this between Chroma and valence arousal. The table also shows significantly better results obtained on our dataset, compared to the results on the Mood Swing dataset.

[novo] We also performed a regression algorithm on chroma features calculated with the Compositional hierarchical model described in [10], what gave us the best results on a regression algorithm. The average distance is 0.1862, the distance to nearest value is 0.0719 and the distance measured in standard deviation is 0.4459.

Table 2: Average distances between the predictions and the mean valence-arousal value shown separately for the valence and arousal arousal component. Distance for arousal is significantly smaller.

	MFCC	Chroma	MP Chroma
Valence	0.1734	0.1826	0.1494
Arousal	0.0871	0.0940	0.0898

The table 2 shows the accuracy in predictions separately for valence and arousal values measured in average distance between prediction and mean valence-arousal value from dataset. Results indicate a more accurate prediction

of the regression method for arousal component.

4 Conclusion and future work

We gathered well annotated dataset with a lot of responses with online survey, which will be publicly available presently. It contains participants' demographical data, users mood and color perception and most importantly, a significant amount of mood and color responses per audio clip in the dataset. Each of these responses contains induced emotions and perceived emotions with valence-arousal values. It also contains color perception for audio. Unlike several others datasets ours will also provides audio clips used in survey, due to the copyright of the original audio.

This dataset opens new possibilities for research mood evaluation from audio. Mood evaluation is an important for music recommendation systems based on mood. Personal background data of the participants in our dataset provides new possibilities to research usage of such data for the music recommendation.

We will shortly begin with the second run of survey. This survey will be presented in English language and will contains an additional set of audio clips. With the second run we intend to raise the number of responses per clip and enlarge number of audio clips in dataset. We also plan on comparing the collected data for English-speaking versus Slovene-speaking participants. We will continue testing the mood evaluation algorithms on dataset and intend to develop a new method which considers demographic and background data of the annotators.

We will further explore correlation between mood and colors. We intend to develop a music visualisation, based on the results of our study. The visualisation will be used in music recommendation interfaces, substituting the text and other data on the user's screen with a visual representation of the audio.

References

[1] Hervé Abdi. The method of least squares. *Encyclopedia of Measurement and Statistics. CA, USA: Thousand Oaks*, 2007.

[2] Mathieu Barthet, David Marston, Chris Baume, György Fazekas, and Mark Sandler. Design and evaluation of semantic mood models for music recommendation. In *Proc. International Society for Music Information Retrieval Conference*, 2013.

[3] Juan Pablo Bello and Jeremy Pickens. A robust mid-level representation for harmonic content in music signals. In *ISMIR*, volume 5, pages 304–311, 2005.

[4] Tim Dalgleish and Michael J Power. Basic emotions. In *Handbook of cognition and emotion*. Wiley Online Library, 1999.

[5] Tuomas Eerola and Jonna K Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 2010.

[6] Cyril Laurier, Perfecto Herrera, M Mandel, and D Ellis. Audio music mood classification using support vector machine. *Music Information Retrieval Evaluation eXchange (MIREX) extended abstract*, 2007.

[7] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *ISMIR*, 2000.

[8] R Panda, R Malheiro, B Rocha, A Oliveira, and RP Paiva. Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. *Proc. CMMR*, 2013.

[9] Matevz Pesek, Primoz Godec, Mojca Poredoš, Gregor Strle, Joze Guna, Emilija Stojmenova, Matevz Pogacnik, and Matija Marolt. Gathering a dataset of multi-modal mood-dependent perceptual responses to music. In *Proceedings of the 2nd Workshop Emotions and Personality in Personalized Services (EMPIRE 2014), Aalborg, Denmark*, pages 1613–0073, 2014.

[10] Matevz Pesek and Matija Marolt. Chord estimation using compositional hierarchical model. In *6th International Workshop on Machine Learning and Music, held in conjunction with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML/PKDD*, volume 2013, 2013.

[11] Erik M Schmidt and Youngmoo E Kim. Projection of acoustic features to continuous valence-arousal mood labels via regression. In *10th International Society for Music Information Retrieval Conference. ISMIR*, 2009.

[12] Erik M Schmidt and Youngmoo E Kim. Modeling musical emotion dynamics with conditional random fields. In *ISMIR*, pages 777–782, 2011.

[13] Björn Schuller, Clemens Hage, Dagmar Schuller, and Gerhard Rigoll. 'mister dj, cheer me up!': Musical and textual features for automatic mood classification. *Journal of New Music Research*, 39(1):13–34, 2010.

[14] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(2):467–476, 2008.

[15] Bin Wu, Simon Wun, Chung Lee, and Andrew Horner. Spectral correlates in emotion labeling of sustained musical instrument tones. In *ISMIR*, pages 415–420, 2013.