Music Emotion Maps in Arousal-Valence Space

Jacek Grekow(✉)

Faculty of Computer Science, Bialystok University of Technology,
Wiejska 45A, 15-351 Bialystok, Poland
j.grekow@pb.edu.pl

Abstract. In this article we present the approach in which the detection of emotion is modeled by the pertinent regression problem. Conducting experiments required building a database, annotation of samples by music experts, construction of regressors, attribute selection, and analysis of selected musical compositions. We obtained a satisfactory correlation coefficient value for SVM for regression algorithm at 0.88 for arousal and 0.74 for valence. The result applying regressors are emotion maps of the musical compositions. They provide new knowledge about the distribution of emotions in musical compositions. They reveal new knowledge that had only been available to music experts until this point.

Keywords: Emotion detection · Emotion tracking · Audio feature extraction · Music information retrieval

1 Introduction

Emotions are a dominant element in music, and they are the reason people listen to music so often [12]. Systems searching musical compositions on Internet databases more and more often add an option of selecting emotions to the basic search parameters, such as title, composer, genre, etc. The emotional content of music is not always constant, and even in classical music or jazz changes often. Analysis of emotions contained in music over time is a very interesting aspect of studying the content of music. It can provide new knowledge on how the composer emotionally shaped the music or why we like some compositions more than others.

Music emotion recognition, taking into account the emotion model, can be divided into categorical or dimensional. In the categorical approach, a number of emotional categories (adjectives) are used for labeling music excerpts. It was presented in the following papers [5,6,11]. In the dimensional approach, emotion is described using dimensional space - 2D or 3D. Russell [13] proposed a 2D model, where the dimensions are represented by arousal and valence; used in [15,18]. The 3D model of Pleasure-Arousal-Dominance (PAD) was used in [3,10].

Music emotion recognition concentrates on static or dynamic changes over time. Static music emotion recognition uses excerpts from 15 to 30 s and omits changes in emotions over time. It assumes the emotion in a given segment does not change. Dynamic music emotion recognition analyzes changes in emotions

K. Saeed and W. Homenda (Eds.): CISIM 2016, LNCS 9842, pp. 697–706, 2016.
DOI: 10.1007/978-3-319-45378-1_60

over time. Methods for detecting emotion using a sliding window are presented in [9,11,15,18]. Deng and Leung [3] proposed multiple dynamic textures to model emotion dynamics over time. To find similar sequence pasterns of musical emotions, they used subsequence dynamic time warping for matching emotion dynamics. Aljanaki et al. [1] investigated how well structural segmentation explains emotion segmentation. They evaluated different unsupervised segmentation methods on the task of emotion segmentation. Imbrasaite et al. [7] and Schmidt et al. [14] used Continuous Conditional Random Fields for dimensional emotion tracking.

In our study, we used dynamic music emotion recognition with a sliding window. We experimentally selected a segment length of 6 s as the shortest period of time after which a music expert can recognize an emotion.

The rest of this paper is organized as follows. Section 2 describes the music annotated data set and the emotion model used. Section 3 presents features extracted by using tools for audio analysis. Section 4 describes regressor training and their evaluation. Section 5 presents the results of emotion tracking. Finally, Sect. 6 summarizes the main findings.

2 Music Data

The data set that was annotated consisted of 324 six-second fragments of different genres of music: classical, jazz, blues, country, disco, hip-hop, metal, pop, reggae, and rock. The tracks were all 22050 Hz mono 16-bit audio files in .wav format.

Data annotation was done by five music experts with a university musical education. During annotation of music samples, we used the two-dimensional valence-arousal (V-A) model to measure emotions in music [13]. The model consists of two independent dimensions (Fig. 1) of valence (horizontal axis) and arousal (vertical axis). Each person making annotations after listening to a music sample had to specify values on the arousal and valence axes in a range from −10 to 10 with step 1. On the arousal axis, a value of −10 meant low while 10 high arousal. On the valence axis, −10 meant negative while 10 positive valence.

Value determination on the A-V axes was unambiguous with a designation of a point on the A-V plane corresponding to the musical fragment. The data collected from the five music experts was averaged. Figure 2 presents the annotation results of a data set with A-V values. Each point on the plane, defined by values of arousal and valence, represents one of 324 musical fragments. As can be seen in the figure, the musical compositions fill the four quadrants formed by arousal and valence almost uniformly. The amount of examples in quarters on the A-V emotion plane is presented in Table 1.

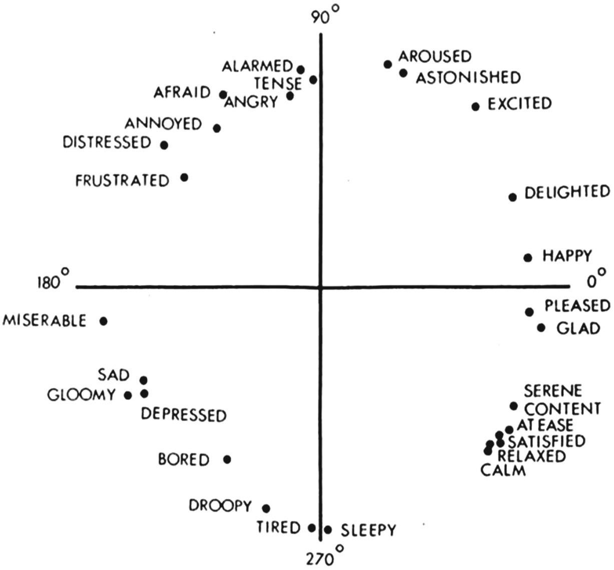

Fig. 1. Russell's circumplex model [13]

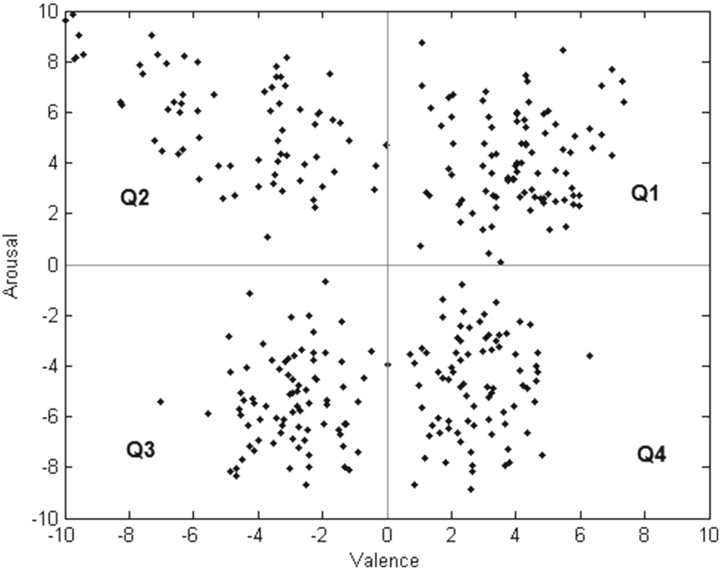

Fig. 2. Data set on A-V emotion plane

Table 1. Amount of examples in quarters on A-V emotion plane

Quarter abbreviation	Arousal-Valence	Amount of examples
Q1	High-high	93
Q2	High-low	70
Q3	Low-low	80
Q4	Low-high	81

3 Features Extraction

For feature extraction, we used Essentia [2] and Marsyas [16], which are tools for audio analysis and audio-based music information retrieval.

Marsyas software (version 0.5.0), written by George Tzanetakis, is implemented in C++ and retains the ability to output feature extraction data to ARFF format. With this tool, the following features can be extracted: Zero Crossings, Spectral Centroid, Spectral Flux, Spectral Rolloff, Mel-Frequency Cepstral Coefficients (MFCC), and chroma features - 31 features in total. For each of these basic features, Marsyas calculates four statistic features (the mean of the mean, the mean of the standard deviation, the standard deviation of the mean, and the standard deviation of the standard deviation).

Essentia is an open-source C++ library, which was created at Music Technology Group, Universitat Pompeu Fabra, Barcelona. Essentia (version 2.1 beta) contains a number of executable extractors computing music descriptors for an audio track: spectral, time-domain, rhythmic, tonal descriptors, and returning the results in YAML and JSON data formats. Extracted features by Essentia are divided into three groups: low-level, rhythm, and tonal features. Essentia also calculates many statistic features: the mean, geometric mean, power mean, median of an array, and all its moments up to the 5th-order, its energy, and the root mean square (RMS). To characterize the spectrum, flatness, crest and decrease of an array are calculated. Variance, skewness, kurtosis of probability distribution, and a single Gaussian estimate were calculated for the given list of arrays.

The previously prepared, labeled by A-V values, music data set served as input data for tools used for feature extraction. The obtained lengths of feature vectors, dependent on the package used, were as follows: Marsyas - 124 features and Essentia - 530 features.

4 Regressor Training

In this paper emotion recognition was treated as a regression problem. We built regressors for predicting arousal and valence using the WEKA package [17]. For training and testing, the following regression algorithms were used: SMOreg, REPTree, M5P.

We evaluated the performance of regression using the 10-fold cross validation technique (CV-10). The whole data set was randomly divided into ten parts, nine of them for training and the remaining one for testing. The learning procedure was executed a total of 10 times on different training sets. Finally, the 10 error estimates were averaged to yield an overall error estimate.

For measuring the performance of regression, we used correlation coefficient (CF). Correlation coefficient measures the statistical correlation between the actual values and the predicted values. The correlation coefficient ranges from 1 to −1. Value 1 means perfectly correlated results, 0 there is no correlation, and −1 means that the results are perfectly correlated negatively.

Table 2. Correlation coefficient obtained for SMOreg

	Essentia		Marsyas	
	Arousal	Valence	Arousal	Valence
Before attribute selection	0.69	0.51	0.79	0.39
After attribute selection	**0.88**	**0.74**	0.85	0.54

The highest values for correlation coefficient (CF) were obtained using SMOreg (implementation of the support vector machine for regression) and are presented in Table 2. CF improved to 0.88 for arousal and 0.74 for valence after applying attribute selection (attribute evaluator: WrapperSubsetEval [8], search method: BestFirst). Predicting arousal is a much easier task for regressors than valence in both cases of extracted features (Essentia, Marsays). CF for arousal were comparable (0.88 and 0.85), but features which describe valence were much better using Essentia for audio analysis. The obtained CF 0.74 was much higher than 0.54 using Marsyas features.

5 Results of Emotion Tracking

We used the best obtained models for predicting arousal and valence to analyze musical compositions. The compositions were divided into 6-s segments with a 3/4 overlap. For each segment, features were extracted and models for arousal and valence were used.

The predicted values are presented in the figures. For each musical composition, the obtained data was presented in 4 different ways:

1. Arousal-Valence over time;
2. Arousal-Valence map;
3. Arousal over time;
4. Valence over time.

Simultaneous observation of the same data in 4 different projections enabled us to accurately track changes in valence and arousal over time, such as tracking the location of a prediction on the A-V emotion plane.

5.1 Emotion Maps

Figures 3 and 4 show emotion maps of two compositions, one for the song Let It Be by Paul McCartney (The Beatles) and the second, Sonata Pathetique (2nd movement) by Ludwig van Beethoven.

Emotion maps present two different emotional aspects of these compositions. The first significant difference is distribution on the quarters of the Arousal-Valence map. In Let It Be (Fig. 3b), the emotions of quadrants Q4 and Q1 (high valence and low-high arousal) dominate. In Sonata Pathetique (Fig. 4b), the emotions of quarter Q4 (low arousal and low valence) dominate with an incidental emergence of emotions of quarter Q3 (low arousal and low valence).

Another noticeable difference is the distribution of arousal over time. Arousal in Let It Be (Fig. 3c) has a rising tendency over time of the entire song, and varies from low to high. In Sonata Pathetique (Fig. 4c), in the first half (s. 0–160) arousal has very low values, and in the second half (s. 160–310) arousal increases incidentally but remains in the low value range.

The third noticeable difference is the distribution of valence over time. Valence in Let It Be (Fig. 3d) remains in the high (positive) range with small fluctuations, but it is always positive. In Sonata Pathetique (Fig. 4d), valence, for the most part, remains in the high range but it also has several large declines in valence (s. 120, 210, 305), which makes valence more diverse.

Arousal and valence over time were dependent on the musical content. Even on a short fragment of music, these values varied significantly. From the course of arousal and valence, it appears that Let It Be is a song of a decisively positive nature with a clear increase in arousal over time. Sonata Pathetique (2nd movement) is mostly calm and predominantly positive.

5.2 Features Describing Arousal and Valence Values in Musical Compositions

To analyze and compare changes in arousal and valence over time (time series), we proposed the following parameters:

1. Mean value of Arousal;
2. Mean value of Valence;
3. Standard deviation of Arousal;
4. Standard deviation of Valence;
5. Mean of derivative of Arousal;
6. Mean of derivative of Valence;
7. Standard deviation of derivative of Arousal;
8. Standard deviation of derivative of Valence;
9. Quantity of changing sign of Arousal (QCA) – describes how often Arousal changes between top and bottom quarters of the A-V emotion model;
10. Quantity of changing sign of Valence (QCV) – describes how often Valence changes between left and right quarters of the A-V emotion model;
11. QCE – is the sum of QCA and QCV;
12. Percentage representation of emotion in 4 quarters of A-V emotion model (4 parameters).

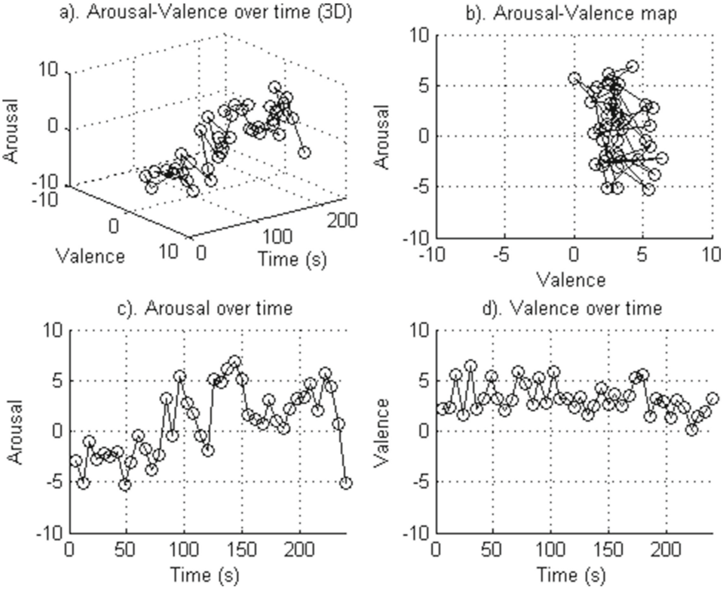

Fig. 3. A-V maps for the song let it be by Paul McCartney (The Beatles)

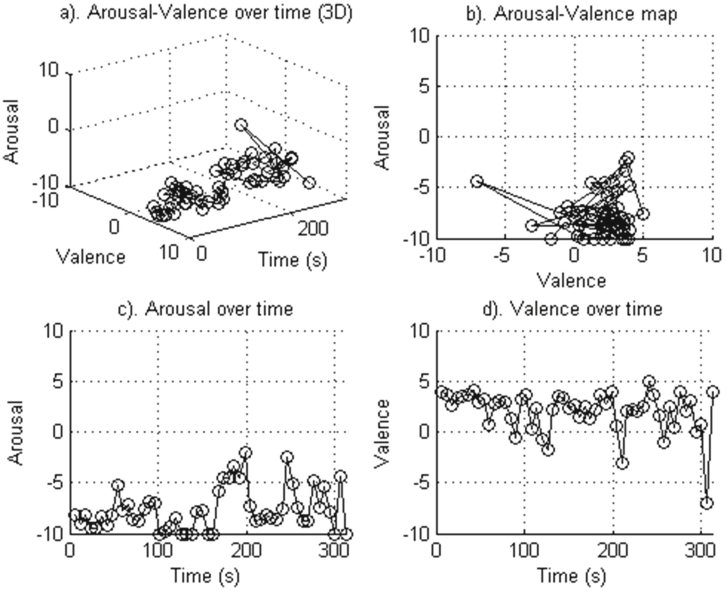

Fig. 4. A-V maps for Piano Sonata No. 8 in C minor, Op. 13 (Pathetique), part 2, by Ludwig van Beethoven

Analysis of the distribution of emotions over time gives a much more accurate view of the emotional structure of a musical composition. It provides not only information on which emotions are dominant in a composition, but also how often they change, and their tendency. The presented list of features is not closed. We will search for additional features in the future.

5.3 Comparison of Musical Compositions

Another experiment was to compare selected well-known Ludwig van Beethoven's Sonatas with several of the most famous songs by The Beatles. We used nine musical compositions from each group for the comparison (Table 3). This experiment did not aim to compare all the works of Beethoven and The Beatles, but only to find the rules and most important features distinguishing these 2 groups.

Table 3. List of musical compositions

L. v. Beethovens Sonatas	The Beatles
Sonata Appassionata, part 1	Hey Jude
Sonata Appassionata, part 2	P.S. I Love You
Sonata Appassionata, part 3	While My Guitar Gently Weeps
Sonata Waldstein, part 1	I'll Follow The Sun
Sonata Waldstein, part 2	It's Only Love
Sonata Waldstein, part 3	Yesterday
Sonata Pathetique, part 1	Michelle
Sonata Pathetique, part 2	Girl
Sonata Pathetique, part 3	Let It Be

Each sample was segmented and arousal and valence were detected. Then, 15 features, which were presented in the previous section, were calculated for each sample. We used the PART algorithm [4] from the WEKA package [17] to find the decision-making rules differentiating the two groups.

It turned out that the most distinguishing feature for these two groups of musical compositions was the *Standard deviation of Valence*. It was significantly smaller in The Beatles' songs than in Beethoven's compositions. *Standard deviation of Valence* reflects how big deviations were from the mean. The results show that in Beethoven's compositions valence values were much more varied than in the songs of The Beatles.

To find another significant feature in the next stage, we removed the characteristic that we found previously (Standard deviation of Valence) from the data set. Another significant feature was *Standard deviation of Arousal*. In Beethoven's compositions, the values of the *Standard deviation of Arousal* were much greater than in the Beatles' songs. This proves the compositions have a greater diversity of tempo and volume.

In the next analogous stage, the feature we found was *Standard deviation of derivative of Arousal*. It reflects the magnitude of changes in arousal between the studied segments. We found higher values of *Standard deviation of derivative of Arousal* in Beethoven's compositions.

The interesting thing is that in the group of the most important distinguishing features we did not find features describing emotion type (*Mean value of Arousal*, *Mean value of Valence* or *Percentage representation of emotion in 4 quarters*). This is confirmed by the fact that we cannot assign common emotions to the different sample groups (Beethoven, The Beatles); in all groups, we have emotions from the four quadrants of the emotion model. Features that better distinguish between the two groups of compositions were features pertaining to changes in emotions and their distribution in the musical compositions.

6 Conclusions

In this article we presented the approach in which the detection of emotion was modeled by the regression problem. Conducting experiments required building a database, annotation of samples by music experts, construction of regressors, attribute selection, and analysis of selected musical compositions. We obtained a satisfactory correlation coefficient value for SVM for regression algorithm at 0.88 for arousal and 0.74 for valence.

The result applying regressors are emotion maps of the musical compositions. They provide new knowledge about the distribution of emotions in musical compositions. They reveal new knowledge that had only been available to music experts until this point. The proposed parameters describing emotions can be used in the construction of a system that can search for songs with similar emotions. They describe in more detail the distribution of emotions, their evolution, frequency of changes, etc.

Acknowledgments. This research was realized as part of study no. S/WI/3/2013 and financed from Ministry of Science and Higher Education funds.

References

1. Aljanaki, A., Wiering, F., Veltkamp, R.C.: Emotion based segmentation of musical audio. In: Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, pp. 770–776 (2015)
2. Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., Serra, X.: ESSENTIA: an audio analysis library for music information retrieval. In: Proceedings of the 14th International Society for Music Information Retrieval Conference, Curitiba, Brazil, pp. 493–498 (2013)
3. Deng, J.J., Leung, C.H.: Dynamic time warping for music retrieval using time series modeling of musical emotions. IEEE Trans. Affect. Comput. **6**(2), 137–151 (2015)

4. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization. In: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 144–151. Morgan Kaufmann Publishers Inc., San Francisco (1998)
5. Grekow, J.: Mood tracking of musical compositions. In: Chen, L., Felfernig, A., Liu, J., Raś, Z.W. (eds.) ISMIS 2012. LNCS, vol. 7661, pp. 228–233. Springer, Heidelberg (2012)
6. Grekow, J.: Mood tracking of radio station broadcasts. In: Andreasen, T., Christiansen, H., Cubero, J.-C., Raś, Z.W. (eds.) ISMIS 2014. LNCS, vol. 8502, pp. 184–193. Springer, Heidelberg (2014)
7. Imbrasaite, V., Baltrusaitis, T., Robinson, P.: Emotion tracking in music using continuous conditional random fields and relative feature representation. In: 2013 IEEE International Conference on Multimedia and Expo Workshops, San Jose, CA, USA, pp. 1–6 (2013)
8. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97** (1–2), 273–324 (1997)
9. Korhonen, M.D., Clausi, D.A., Jernigan, M.E.: Modeling emotional content of music using system identification. Trans. Sys. Man Cyber. Part B **36**(3), 588–599 (2005)
10. Lin, Y., Chen, X., Yang, D.: Exploration of music emotion recognition based on MIDI. In: Proceedings of the 14th International Society for Music Information Retrieval Conference (2013)
11. Lu, L., Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signals. Trans. Audio Speech Lang. Proc. **14**(1), 5–18 (2006)
12. Pratt, C.C.: Music as the Language of Emotion. The Library of Congress. U.S. Govt. Print. Off., Washington (1950)
13. Russell, J.A.: A circumplex model of affect. J. Pers. Soc. Psychol. **39**(6), 1161–1178 (1980)
14. Schmidt, E.M., Kim, Y.E.: Modeling musical emotion dynamics with conditional random fields. In: Proceedings of the 2011 International Society for Music Information Retrieval Conference, pp. 777–782 (2011)
15. Schmidt, E.M., Turnbull, D., Kim, Y.E.: Feature selection for content-based, time-varying musical emotion regression. In: Proceedings of the International Conference on Multimedia Information Retrieval, MIR 2010, pp. 267–274. ACM, New York (2010)
16. Tzanetakis, G., Cook, P.: Marsyas: a framework for audio analysis. Org. Sound **4**(3), 169–175 (2000)
17. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco (2005)
18. Yang, Y.H., Lin, Y.C., Su, Y.F., Chen, H.H.: A regression approach to music emotion recognition. Trans. Audio Speech Lang. Proc. **16**(2), 448–457 (2008)