

## Audio features dedicated to the detection and tracking of arousal and valence in musical compositions

Jacek Grekow

To cite this article: Jacek Grekow (2018) Audio features dedicated to the detection and tracking of arousal and valence in musical compositions, Journal of Information and Telecommunication, 2:3, 322-333, DOI: [10.1080/24751839.2018.1463749](https://doi.org/10.1080/24751839.2018.1463749)

To link to this article: <https://doi.org/10.1080/24751839.2018.1463749>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 27 Apr 2018.



Submit your article to this journal [↗](#)



Article views: 3032



View related articles [↗](#)



View Crossmark data [↗](#)



# Audio features dedicated to the detection and tracking of arousal and valence in musical compositions

Jacek Grekow

Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland

## ABSTRACT

The aim of this paper was to discover what combination of audio features gives the best performance with music emotion detection. Emotion recognition was treated as a regression problem, and a two-dimensional valence–arousal model was used to measure emotions in music. Features extracted by Essentia and Marsyas, tools for audio analysis and audio-based music information retrieval, were used. The influence of different feature sets was examined – low level, rhythm, tonal, and their combination – on arousal and valence prediction. The use of a combination of different types of features significantly improves the results compared with using just one group of features. Features particularly dedicated to the detection of arousal and valence separately, as well as features useful in both cases, were found and presented. This paper presents also the process of building emotion maps of musical compositions. The obtained emotion maps provide new knowledge about the distribution of emotions in an examined audio recording. They reveal new knowledge that had only been available to music experts until this point.

## ARTICLE HISTORY

Received 23 October 2017

Accepted 9 April 2018



## KEYWORDS

Music emotion detection;  
audio features; feature  
selection; emotion tracking

## 1. Introduction

One of the most important elements when listening to music is the expressed emotion. The elements of music that affect the emotions are timbre, dynamics, rhythm, tempo, and harmony. 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, and genre. One of the most important steps during building a system for automatic emotion detection is feature extraction from audio files. The quality of these features and connecting them with elements of music such as rhythm, harmony, melody and dynamics, shaping a listener's emotional perception of music, have a significant effect on the effectiveness of the built prediction models.

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 labelling music excerpts. It was presented, among others, in the

**CONTACT** Jacek Grekow  [j.grekow@pb.edu.pl](mailto:j.grekow@pb.edu.pl)  Faculty of Computer Science, Bialystok University of Technology, Wiejska 45A, Bialystok 15-351, Poland

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

following papers (Grekow, 2012; Li & Ogihara, 2003; Lu, Liu, & Zhang, 2006; Wieczorkowska, Synak, Lewis, & Raś, 2005). In the dimensional approach, emotion is described using dimensional space – 2D or 3D. Russell (1980) proposed a 2D model, where the dimensions are represented by arousal and valence; used in Grekow (2016, 2017b), Schmidt, Turnbull, & Kim (2010), and Yang, Lin, Su, & Chen (2008). The 3D model of Pleasure–Arousal–Dominance (PAD) was used in Deng & Leung (2015) and Lin, Chen, & Yang (2013).

Division into categorical approach and dimensional approach can be found in papers on examining features for music emotion recognition. Most papers, however, focus on studying features using a classification model (Grekow, 2015; Panda, Rocha, & Paiva, 2015; Saari, Eerola, & Lartillot, 2011; Song, Dixon, & Pearce, 2012). Music emotion recognition combining standard and melodic features extracted from audio was presented in Panda et al. (2015). Song et al. (2012) explored the relationship between musical features extracted by MIR toolbox and emotions. They compared the emotion prediction results for four sets of features: dynamic, rhythm, harmony, and spectral features. Baume, Fazekas, Barthet, Marston, & Sandler (2014) evaluated different types of audio features using a five-dimensional support vector regressor, in order to find the combination that produces the best performance. Searching for useful features does not only pertain to emotion detection. The issue of features selection improving classification accuracies for genre classification was presented in Doraisamy, Golzari, Norowi, Sulaiman, & Udzir (2008).

An important paper in the area of music emotion recognition was written by Yang & Chen (2012), who did a comprehensive review of the methods that have been proposed for music emotion recognition. Kim et al. (2010) presented another paper surveying the state of the art in automatic emotion recognition.

The aim of this paper was to discover what combination of audio features gives the best performance with music emotion detection. Emotion recognition was treated as a regression problem and a two-dimensional valence–arousal (V–A) model was used to measure emotions in music. Features extracted by Essentia (Bogdanov et al., 2013) and Marsyas (Tzanetakis & Cook, 2000), tools for audio analysis and audio-based music information retrieval, were used. This article is an extension of a conference paper (Grekow, 2017a) where the problem was preliminarily presented.

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 tools used for feature extraction. Section 4 describes regressor training and their evaluation. Section 5 is devoted to evaluating different combinations of feature sets. Section 6 presents dedicated features to the detection of arousal and valence. Section 7 describes results obtained by emotion tracking of two compositions. Finally, Section 8 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 22,050 Hz mono 16-bit audio files in .wav format. The training data were taken from the generally accessible data collection project Marsyas.<sup>1</sup> The author selected samples and shortened them to the first 6 s. This is the shortest possible length at which experts could detect emotions for a given segment. On the other hand, a short segment

ensures that emotional homogeneity of a segment is much more probable. The data set consisted of 324 samples.

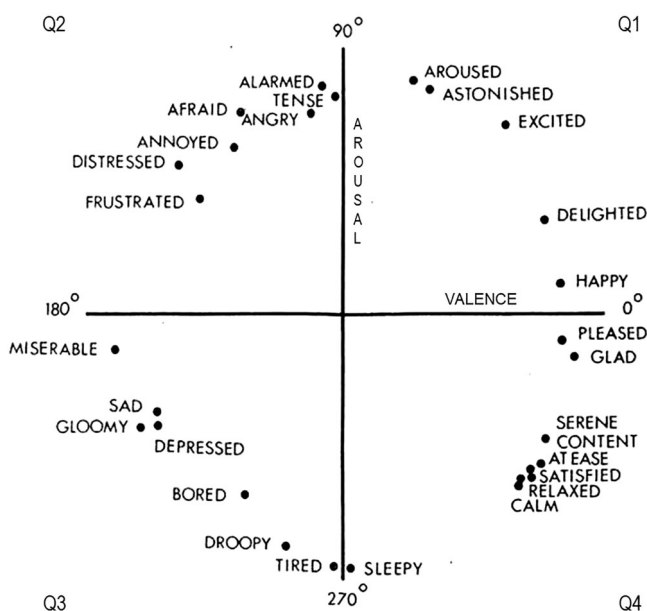
Data annotation was done by five music experts with a university musical education. Each annotator annotated all records in the data set, which has a positive effect on the quality of the received data (Aljanaki, Yang, & Soleymani, 2016). During annotation of music samples, we used the two-dimensional V–A model to measure emotions in music (Russell, 1980). The model (Figure 1) consists of two independent dimensions 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. The amount of examples in quarters on the A–V emotion plane is presented in Table 1.

Pearson correlation coefficient was calculated to check if valence and arousal dimensions are correlated in our music data. The obtained value  $r = -0.03$  indicates that arousal and valence values are not correlated, and the music data are a good spread in the quarters on the A–V emotion plane. This is an important element according to the conclusions formulated in Aljanaki et al. (2016).

### 3. Feature extraction

For feature extraction, tools for audio analysis and audio-based music information retrieval – Essentia (Bogdanov et al., 2013) and Marsyas (Tzanetakis & Cook, 2000) – were used.



**Figure 1.** A–V emotion plane – Russell's circumplex model (Russell, 1980).

**Table 1.** Amount of examples in quarters on the 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                 |

Marsyas software, 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. A full list of features extracted by Marsyas is available on the web site.<sup>2</sup> For each of these basic features, Marsyas calculates four statistic features:

- (1) *The mean of the mean* (calculate mean over the 20 frames, and then calculate the mean of this statistic over the entire segment)
- (2) *The mean of the standard deviation* (calculate the standard deviation of the feature over 20 frames, and then calculate the mean of these standard deviations over the entire segment)
- (3) *The standard deviation of the mean* (calculate the mean of the feature over 20 frames, and then calculate the standard deviation of these values over the entire segment)
- (4) *The standard deviation of the standard deviation* (calculate the standard deviation of the feature over 20 frames, and then calculate the standard deviation of these values over the entire segment).

Essentia is an open-source C++ library, which was created at Music Technology Group, Universitat Pompeu Fabra, Barcelona. Essentia 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. A full list of features is available on the web site.<sup>3</sup> 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, labelled 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

Regressors for predicting arousal and valence were built. For training and testing, the following regression algorithms were used: SMOreg, REPTree, M5P. SMOreg algorithm (Smola & Schölkopf, 2004) implements the support vector machine for regression. REPTree algorithm (Hall et al., 2009) builds a regression tree using variance and prunes it using reduced-error pruning. M5P implements base routines for generating M5 Model trees and rules

(Quinlan, 1992; Wang & Witten, 1997). Before constructing regressors arousal and valence annotations were scaled between  $[-0.5, 0.5]$ .

The performance of regression was evaluated using the 10-fold cross-validation (CV-10) technique. The whole data set was randomly divided into 10 parts, 9 of them for training and the remaining 1 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.

The highest values for determination coefficient ( $R^2$ ) were obtained using SMOreg. After applying attribute selection [attribute evaluator: WrapperSubsetEval (Kohavi & John, 1997), search method: BestFirst (Xu, Yan, & Chang, 1988)],  $R^2 = 0.79$  for arousal and  $R^2 = 0.58$  for valence were obtained. Mean absolute error reached values  $MAE=0.09$  for arousal and  $MAE=0.10$  for valence (Table 2).

Predicting arousal is a much easier task for regressors than valence in both cases of extracted features (Essentia, Marsyas) and values predicted for arousal are more precise.  $R^2$  for arousal were comparable (0.79 and 0.73), but features which describe valence were much better using Essentia for audio analysis. The obtained  $R^2 = 0.58$  for valence are much higher than  $R^2 = 0.25$  using Marsyas features. In Essentia, tonal and rhythm features greatly improve prediction of valence. These features are not available in Marsyas and thus Essentia obtains better results.

One can notice the significant role of the attribute selection phase, which generally improves prediction results. Marsyas features before attribute selection outperform Essentia features for arousal detection.  $R^2 = 0.63$  and  $MAE=0.13$  by Marsyas are better results than  $R^2 = 0.48$  and  $MAE=0.18$  by Essentia. However, after selecting the most important attribute, Essentia turns out to be the winner with  $R^2 = 0.79$  and  $MAE=0.09$ .

## 5. Evaluation of different combinations of feature sets

The effect of various combinations of Essentia feature sets – low-level (L), rhythm (R), tonal (T) – on the performance obtained for SMOreg algorithm was evaluated. The performance of regression was evaluated using the CV-10 technique. Attribute selection, with attribute evaluator WrapperSubsetEval and search method BestFirst, was also used.

The obtained results, presented in Table 3, indicate that the use of all groups (low-level, rhythm, tonal) of features resulted in the best performance or equal to best performance by combining feature sets. The best results have been marked in bold. Detection of arousal using the set L+R (low-level, rhythm features) has equal results as using all groups. Detection of valence using the set L+T (low-level, tonal features) has only little worse results than using all groups.

**Table 2.**  $R^2$  and  $MAE$  obtained for SMOreg.

|                            | Essentia    |             |             |             | Marsyas |       |         |       |
|----------------------------|-------------|-------------|-------------|-------------|---------|-------|---------|-------|
|                            | Arousal     |             | Valence     |             | Arousal |       | Valence |       |
|                            | $R^2$       | $MAE$       | $R^2$       | $MAE$       | $R^2$   | $MAE$ | $R^2$   | $MAE$ |
| Before attribute selection | 0.48        | 0.18        | 0.27        | 0.17        | 0.63    | 0.13  | 0.15    | 0.16  |
| After attribute selection  | <b>0.79</b> | <b>0.09</b> | <b>0.58</b> | <b>0.10</b> | 0.73    | 0.11  | 0.25    | 0.14  |

**Table 3.**  $R^2$  and MAE for arousal and valence obtained for combinations of feature sets.

| Features set | Arousal     |             | Valence     |             |
|--------------|-------------|-------------|-------------|-------------|
|              | $R^2$       | MAE         | $R^2$       | MAE         |
| L            | 0.74        | 0.10        | 0.49        | 0.12        |
| R            | 0.68        | 0.11        | 0.15        | 0.15        |
| T            | 0.53        | 0.14        | 0.48        | 0.12        |
| L+R          | <b>0.79</b> | <b>0.09</b> | 0.40        | 0.12        |
| L+T          | 0.74        | 0.10        | <b>0.56</b> | <b>0.10</b> |
| R+T          | 0.74        | 0.11        | 0.52        | 0.11        |
| All (L+R+T)  | <b>0.79</b> | <b>0.09</b> | <b>0.58</b> | <b>0.10</b> |

The use of individual feature sets L, R or T did not achieve better results than their combinations. Worse results were obtained when using only tonal features for arousal ( $R^2 = 0.53$  and  $MAE=0.14$ ) and only rhythm features for valence ( $R^2 = 0.15$  and  $MAE=0.15$ ).

Combining feature sets L+R (low-level and rhythm features) improved regressors results in the case of arousal. Combining feature sets L+T (low-level and tonal features) improved regressors results in the case of valence.

In summary, low-level features are very important in the prediction of both arousal and valence. Additionally, rhythm features are important for arousal detection, and tonal features help a lot for detecting valence. The use of only individual feature sets L, R, or T does not give good results.

## 6. Selected features dedicated to the detection of arousal and valence

Table 4 presents two sets of selected features, which using the SMOreg algorithm obtained the best performance by detecting arousal (Section 5). Features marked in bold are in both groups. Notice that after adding tonal features T to group L+R, some of the features were replaced by others and some remained without changes. Features

**Table 4.** Selected features used for building the arousal regressor.

| Features from set L+R+T              | Features from set L+R                |
|--------------------------------------|--------------------------------------|
| Average Loudness (L)                 | Barkbands Kurtosis (L)               |
| Barkbands Spread (L)                 | Dissonance (L)                       |
| <b>Melbands Crest (L)</b>            | Erbbands Flatness (L)                |
| Melbands Flatness (L)                | Erbbands Skewness (L)                |
| <b>Melbands Kurtosis (L)</b>         | <b>Melbands Crest (L)</b>            |
| Melbands Skewness (L)                | <b>Melbands Kurtosis (L)</b>         |
| Melbands Spread (L)                  | Silence Rate (L)                     |
| <b>Spectral Energy (L)</b>           | <b>Spectral Energy (L)</b>           |
| <b>Spectral Entropy (L)</b>          | <b>Spectral Entropy (L)</b>          |
| <b>Spectral Flux (L)</b>             | <b>Spectral Flux (L)</b>             |
| Spectral Kurtosis (L)                | <b>Spectral Rolloff (L)</b>          |
| <b>Spectral Rolloff (L)</b>          | <b>Spectral Skewness (L)</b>         |
| <b>Spectral Skewness (L)</b>         | Beats Count (R)                      |
| BPM Histogram (R)                    | Beats Loudness (R)                   |
| <b>Danceability (R)</b>              | <b>Danceability (R)</b>              |
| <b>Onset Rate (R)</b>                | <b>Onset Rate (R)</b>                |
| <b>Beats Loudness Band Ratio (R)</b> | <b>Beats Loudness Band Ratio (R)</b> |
| Chords Strength (T)                  |                                      |
| Beats Loudness Band Ratio (R)        |                                      |
| HPCP Entropy (T)                     |                                      |
| Key Strength (T)                     |                                      |
| Chords Histogram (T)                 |                                      |

**Table 5.** Selected features used for building the valence regressor.

| Features from set L+R+T       | Features from set L+T         |
|-------------------------------|-------------------------------|
| High Frequency Content (L)    | Melbands Crest (L)            |
| Melbands Kurtosis (L)         | Melbands Spread (L)           |
| Melbands Skewness (L)         | Pitch Saliency (L)            |
| <b>Spectral Energy (L)</b>    | Silence Rate (L)              |
| <b>Zero Crossing Rate (L)</b> | Spectral Centroid (L)         |
| <b>GFCC (L)</b>               | Spectral Energy (L)           |
| <b>MFCC (L)</b>               | Spectral Spread (L)           |
| Beats Loudness (R)            | <b>Zero Crossing Rate (L)</b> |
| Onset Rate (R)                | <b>GFCC (L)</b>               |
| Beats Loudness Band Ratio (R) | <b>MFCC (L)</b>               |
| <b>Chords Strength (T)</b>    | <b>Chords Strength (T)</b>    |
| <b>HPCP Entropy (T)</b>       | <b>HPCP Entropy (T)</b>       |
| <b>Key Strength (T)</b>       | <b>Key Strength (T)</b>       |
| Chords Histogram (T)          | Key Scale (T)                 |

found in both groups seem to be particularly useful for detecting arousal. Different statistics from spectrum and mel bands turned out to be especially useful: Spectral Energy, Entropy, Flux, Rolloff, Skewness, and Melbands Crest, Kurtosis. Also, three rhythm features belong to the group of more important features because both sets contain: Danceability, Onset Rate, Beats Loudness Band Ratio.

Table 5 presents two sets of selected features, which using the SMOreg algorithm obtained the best performance by detecting valence (Section 5). Particularly important low-level features, found in both groups, were: Spectral Energy and Zero Crossing Rate, as well as Mel Frequency Cepstrum Coefficients (MFCC) and Gammatone Feature Cepstrum Coefficients (GFCC). Particularly important tonal features, which describe key, chords, and tonality of a musical excerpt, were: Chords Strength, Harmonic Pitch Class Profile (HPCP) Entropy, Key Strength.

Comparing the sets of features dedicated to arousal (Table 4) and valence (Table 5), we notice that there are much more statistics from spectrum and mel bands in the arousal set than in the valence set. MFCC and GFCC were useful for detecting valence and were not taken into account for arousal detection.

Features that turned out to be universal, useful for detecting both arousal and valence, by using all features (L+R+T), are:

- Melbands Kurtosis (L)
- Melbands Skewness (L)
- Spectral Energy (L)
- Beats Loudness Band Ratio (R)
- Chords Strength (T)
- Harmonic Pitch Class Profile (HPCP) Entropy (T)
- Key Strength (T)
- Chords Histogram (T).

### 7. Emotion maps

The result of emotion tracking is emotion maps. The best obtained models for predicting arousal and valence to analyse musical compositions were used. 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.

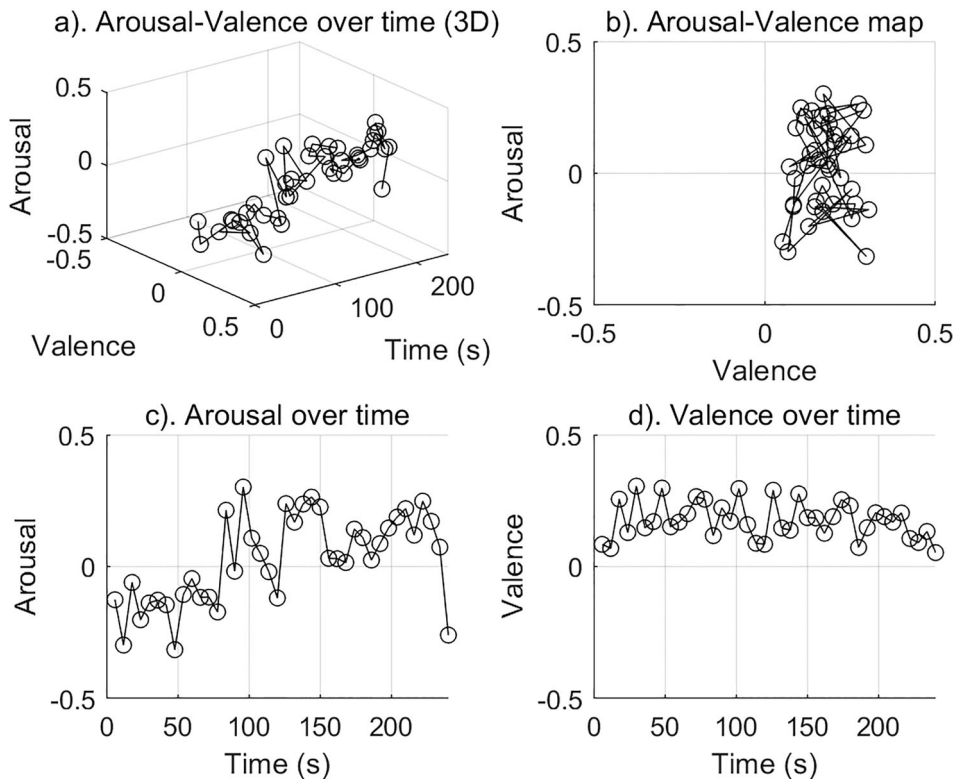
The predicted values are presented in the figures in the form of emotion maps. For each musical composition, the obtained data was presented in four different ways:

- (1) Arousal-Valence over time
- (2) Arousal-Valence map
- (3) Arousal over time
- (4) Valence over time.

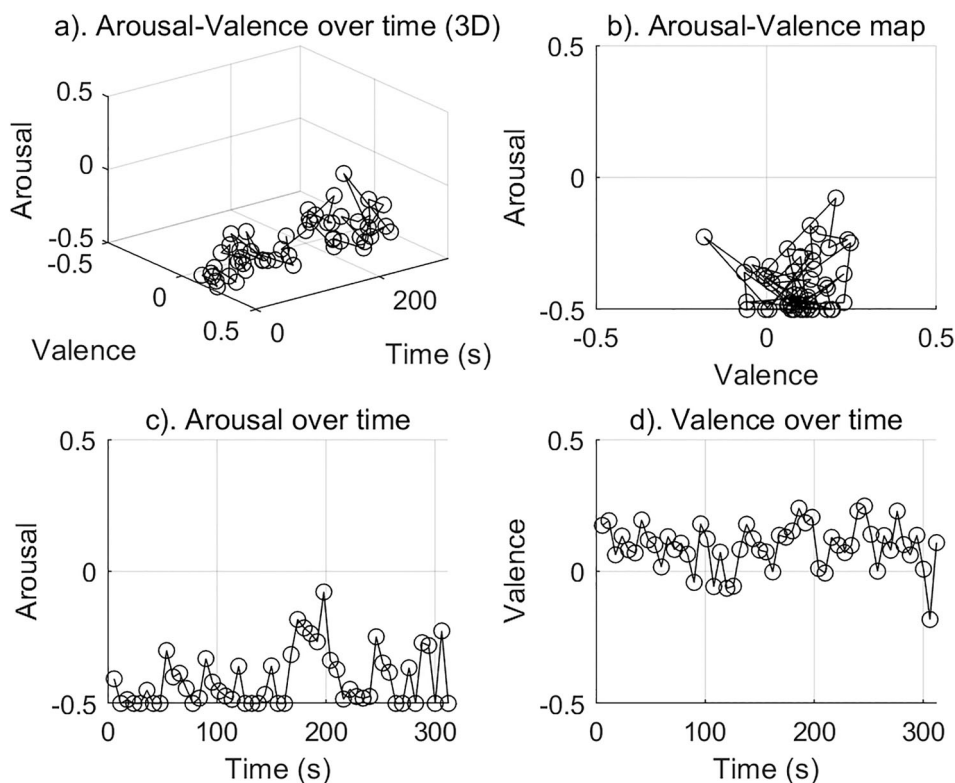
Simultaneous observation of the same data in four different projections enabled us to accurately track changes in valence and arousal over time, such as tracking the location of a prediction on the A-V emotion plane.

Figures 2 and 3 show emotion maps of two compositions, one for the song Let It Be by Paul McCartney (The Beatles) and the second, Piano Sonata No. 8 in C minor, Op. 13 (Pathétique), 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 A-V map. In Let It Be (Figure 2b), the emotions of quadrants Q4 and Q1 (high valence and low-high arousal) dominate. In Sonata Pathétique (Figure 3b), the emotions of quarter Q4 (low arousal and high valence) dominate with an incidental emergence of emotions of quarter Q3 (low arousal and low valence).



**Figure 2.** A-V maps for the song Let It Be by Paul McCartney (The Beatles).



**Figure 3.** A–V maps for Piano Sonata No. 8 in C minor, Op. 13 (Pathétique), 2nd movement, by Ludwig van Beethoven.

Another noticeable difference is the distribution of arousal over time. Arousal in *Let It Be* (Figure 2c) has a rising tendency over time of the entire song and varies from low to high. The last sample with low valence is the exception; the low valence is the result of the slow tempo at the end of the composition. In *Sonata Pathétique* (Figure 3c), 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* (Figure 2d) remains in the high (positive) range with small fluctuations, but it is always positive. In *Sonata Pathétique* (Figure 3d), valence, for the most part, remains in the high range but it also has several declines (s. 90, 110, 305), which makes valence more diverse.

Arousal and valence over time were dependent on the music content. Even in 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, while *Sonata Pathétique* is mostly calm and predominantly positive.

## 8. Conclusion

In this article, the usefulness of audio features during emotion detection in music files was studied. Different features sets were used to test the performance of built regression models intended to detect arousal and valence. Conducting experiments required

building a database, annotation of samples by music experts, construction of regressors, attribute selection, and evaluation of various group features. Features extracted by Essentia, due to their variety and quantity, are better suited for detecting emotions than features extracted by Marsyas.

The influence of different feature sets – low-level, rhythm, tonal, and their combination – on arousal and valence prediction was examined. The use of a combination of different types of features significantly improved the results compared with using just one group of features. Features particularly dedicated to the detection of arousal and valence separately, as well as features useful in both cases, were found and presented.

In conclusion, low-level features are very important in the prediction of both arousal and valence. Additionally, rhythm features are important for arousal detection, and tonal features help a lot for detecting valence.

The obtained results confirm the point of creating new features of middle and higher levels that describe elements of music such as rhythm, harmony, melody, and dynamics shaping a listener's emotional perception of music. They are the ones that can have an effect on improving the effectiveness of automatic emotion detection in music files.

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 obtained emotion maps are algorithm-based and may contain a certain inaccuracy. Their precision is dependent on the accuracy of the regressors used to recognize emotions. It deserves to be improved in future papers dedicated to the construction of emotion maps.

## Notes

1. <http://marsyas.info/downloads/datasets.html>
2. <http://marsyas.info/doc/manual/marsyas-user/bextract.html>
3. [http://essentia.upf.edu/documentation/algorithms\\_reference.html](http://essentia.upf.edu/documentation/algorithms_reference.html)

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This research was realized as part of study no. S/WI/3/2013.

## Notes on contributor

**Jacek Grekow** received an MSc degree in Computer Science from the Technical University in Sofia, Bulgaria in 1994, and a PhD degree from the Polish-Japanese Institute of Information Technology in Warsaw, Poland 2009. He also obtained a Master of Arts degree from The Fryderyk Chopin University of Music in Warsaw 2007. His primary research interests are music information retrieval, emotions in music, and music visualization.

## References

- Aljanaki, A., Yang, Y.-H., & Soleymani, M. (2016). *Emotion in music task: Lessons learned. Working Notes Proceedings of the MediaEval 2016 Workshop*. Hilversum, Netherlands.
- Baume, C., Fazekas, G., Barthet, M., Marston, D., & Sandler, M. (2014). *Selection of audio features for music emotion recognition using production music*. Audio Engineering Society Conference: 53rd International Conference: Semantic Audio.
- Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., & Serra, X. (2013). *ESSENTIA: An audio analysis library for music information retrieval. Proceedings of the 14th International Society for Music Information Retrieval Conference* (pp. 493–498). Curitiba, Brazil.
- Deng, J. J., & Leung, C. H. C. (2015). Dynamic time warping for music retrieval using time series modeling of musical emotions. *IEEE Transactions on Affective Computing*, 6(2), 137–151.
- Doraisamy, S., Golzari, S., Norowi, N. M., Sulaiman, M. N., & Udzir, N. I. (2008). *A study on feature selection and classification techniques for automatic genre classification of traditional Malay music. ISMIR'08, 9th International Conference on Music Information Retrieval* (pp. 331–336). Philadelphia, PA: Drexel University.
- Grekow, J. (2012). Mood tracking of musical compositions. In L. Chen, A. Felfernig, J. Liu, & Z. W. Raś (Eds.), *Proceedings of the 20th International Conference on Foundations of Intelligent Systems* (pp. 228–233). Berlin/Heidelberg: Springer.
- Grekow, J. (2015). Audio features dedicated to the detection of four basic emotions. In K. Saeed & W. Homenda (Eds.), *Computer Information Systems and Industrial Management: 14th IFIP TC 8 International Conference, CISIM 2015, Proceedings* (pp. 583–591). Springer International Publishing, Cham.
- Grekow, J. (2016). Music emotion maps in arousal–valence space. In K. Saeed & W. Homenda (Eds.), *Computer Information Systems and Industrial Management: 15th IFIP TC8 International Conference, CISIM 2016, Proceedings* (pp. 697–706). Springer International Publishing, Cham.
- Grekow, J. (2017a). Audio features dedicated to the detection of arousal and valence in music recordings. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (pp. 40–44). IEEE, Gdynia, Poland.
- Grekow, J. (2017b). Comparative analysis of musical performances by using emotion tracking. In M. Kryszkiewicz, A. Appice, D. Slezak, H. Rybinski, A. Skowron, & Z. W. Raś (Eds.), *Foundations of Intelligent Systems: 23rd International Symposium, ISMIS 2017, Proceedings* (pp. 175–184). Springer International Publishing, Cham.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.
- Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J. J., & Turnbull, D. (2010). *State of the art report: Music emotion recognition: A state of the art review. Proceedings of the 11th International Society for Music Information Retrieval Conference, ISMIR'10, Utrecht, Netherlands* (pp. 255–266).
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1–2), 273–324.
- Li, T., & Ogihara, M. (2003). *Detecting emotion in music. ISMIR'03, 4th International Conference on Music Information Retrieval*, Baltimore, MD, October 27–30, 2003.
- Lin, Y., Chen, X., & Yang, D. (2013). *Exploration of music emotion recognition based on midi. Proceedings of the 14th International Society for Music Information Retrieval Conference* (pp. 221–226). Curitiba, Brazil.
- Lu, L., Liu, D., & Zhang, H.-J. (2006). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1), 5–18.
- Panda, R., Rocha, B., & Paiva, R. P. (2015). Music emotion recognition with standard and melodic audio features. *Applied Artificial Intelligence*, 29(4), 313–334.
- Quinlan, R. J. (1992). *Learning with continuous classes. 5th Australian Joint Conference on Artificial Intelligence* (pp. 343–348). Singapore: World Scientific.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.

- Saari, P., Eerola, T., & Lartillot, O. (2011). Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(6), 1802–1812.
- Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). *Feature selection for content-based, time-varying musical emotion regression. Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 267–274). New York, NY: ACM.
- Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, 14(3), 199–222.
- Song, Y., Dixon, S., & Pearce, M. (2012). *Evaluation of musical features for emotion classification. Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR'12* (pp. 523–528). Porto, Portugal: Mosteiro S. Bento Da Vitória.
- Tzanetakis, G., & Cook, P. (2000). Marsyas: A framework for audio analysis. *Organised Sound*, 4(3), 169–175.
- Wang, Y., & Witten, I. H. (1997). *Induction of model trees for predicting continuous classes. Poster papers of the 9th European Conference on Machine Learning*. Springer.
- Wieczorkowska, A., Synak, P., Lewis, R., & Raś, Z. W. (2005). Extracting emotions from music data. In M.-S. Hacid, N. V. Murray, Z. W. Raś, & S. Tsumoto (Eds.), *Foundations of Intelligent Systems: 15th International Symposium, ISMIS'05, Proceedings, Saratoga Springs, NY, May 25–28, 2005* (pp. 456–465). Berlin/Heidelberg: Springer.
- Xu, L., Yan, P., & Chang, T. (1988). *Best first strategy for feature selection. Proceedings of the 9th International Conference on Pattern Recognition* (Vol. 2, pp. 706–708), Rome, Italy.
- Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, 3(3), 40:1–40:30.
- Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2), 448–457.