Audio Features Dedicated to the Detection of Arousal and Valence in Music Recordings

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

***Abstract*—The aim of this paper was to discover what combination of audio features gives the best performance with music emotion detection. In our approach, emotion recognition was treated as a regression problem and a two-dimensional valence-arousal model was used to measure emotions in music. We used features extracted by Essentia and Marsyas, tools for audio analysis and audio-based music information retrieval. We examined the influence of different feature sets – 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. We found and presented features particularly dedicated to the detection of arousal and valence separately, as well as features useful in both cases.**

***Keywords*—music emotion detection; audio features; features selection**

I. Introduction

One of the most important elements when listening to music is the expressed emotions. The elements of music that affect the emotions are timbre, dynamics, rhythm, 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, genre, etc. 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.

The aim of this paper was to discover what combination of audio features gives the best performance with music emotion detection. In our approach, emotion recognition was treated as a regression problem and a two-dimensional valence-arousal model was used to measure emotions in music. We used features extracted by Essentia [1] and Marsyas [2], tools for audio analysis and audio-based music information retrieval.

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 [3], [4]. In the dimensional approach, emotion is described using dimensional space - 2D or 3D. Russell [5] proposed a 2D model, where the dimensions are represented by arousal and valence; used in [6], [7], [8]. The 3D model of Pleasure-Arousal-Dominance (PAD) was used in [9], [10].

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 [11], [12], [13], [14]. Music emotion recognition combining standard and melodic features extracted from audio was presented by Panda et al. in [12]. Song et al. [14] 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 at al. [15] 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 by Shyamala et al. in [16].

The rest of this paper is organized as follows. Section II describes the music annotated data set and the emotion model used. Section III presents tools used for feature extractions. Section IV describes regressor training and their evaluation. Section V is devoted to evaluating different combinations of feature sets. Section VI presents dedicated features to the detection of arousal and valence. Finally, Section VII summarizes the main findings.

II. 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. The training data were taken from the generally accessible data collection project MARSYAS[1]. The author selected samples and shortened them to the first 6 seconds. 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.

[1] http://marsyas.info/downloads/datasets.html

978-1-5090-5795-5/17/$31.00 © 2017 IEEE

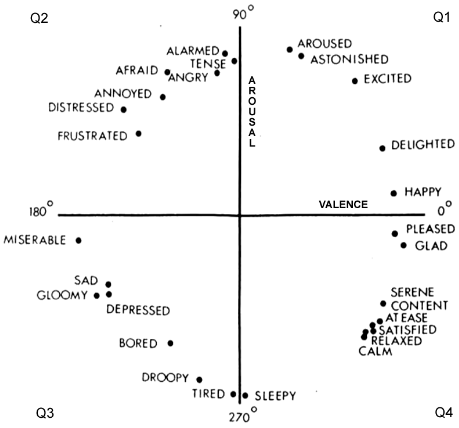

Fig. 1. Russell's circumplex model [5]

TABLE I
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

Data annotation was done by five music experts with a university musical education. Each annotator annotated all records in the dataset, which has a positive effect on the quality of the received data [17]. During annotation of music samples, we used the two-dimensional valence-arousal (V-A) model to measure emotions in music [5]. The model (Fig. 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 I.

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 [17].

III. FEATURES EXTRACTION

For feature extraction, we used Essentia [1] and Marsyas [2], which are tools for audio analysis and audio-based music information retrieval.

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. 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 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[2]. 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.

IV. REGRESSOR TRAINING

We built regressors for predicting arousal and valence. For training and testing, the following regression algorithms were used: SMOreg, REPTree, M5P. SMOreg algorithm [18] implements the support vector machine for regression. REPTree algorithm [19] builds a regression tree using variance and prunes it using reduced-error pruning. M5P implements base routines for generating M5 Model trees and rules [20], [21]. Before constructing regressors arousal and valence annotations were scaled between $[-0.5, 0.5]$.

We evaluated the performance of regression using the ten-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.

The highest values for determination coefficient (R^2) were obtained using SMOreg. After applying attribute selection (attribute evaluator: WrapperSubsetEval [22], search method: BestFirst [23]), we obtained $R^2 = 0.79$, for arousal and $R^2 = 0.58$ for valence. Mean absolute error reached values $MEA = 0.09$ for arousal and $MEA = 0.10$ for valence (Table II).

[2] http://essentia.upf.edu/documentation/algorithms_reference.html

TABLE II
R^2 AND MEA OBTAINED FOR SMOREG

	Essentia				Marsyas			
	Arousal		Valence		Arousal		Valence	
	R^2	MEA	R^2	MEA	R^2	MEA	R^2	MEA
Before attribute selection	0.48	0.18	0.27	0.17	0.63	0.13	0.15	0.16
After attribute selection	**0.79**	**0.09**	**0.58**	**0.10**	0.73	0.11	0.25	0.14

Predicting arousal is a much easier task for regressors than valence in both cases of extracted features (Essentia, Marsays) 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 $MEA = 0.13$ by Marsyas are better results than $R^2 = 0.48$ and $MEA = 0.18$ by Essentia. However, after selecting the most important attribute, Essentia turns out to be the winner with with $R^2 = 0.79$ and $MEA = 0.09$.

V. Evaluation of Different Combinations of Feature Sets

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

TABLE III
R^2 AND MEA FOR AROUSAL AND VALENCE OBTAINED FOR COMBINATIONS OF FEATURE SETS

	Arousal		Valence	
Features set	R^2	MEA	R^2	MEA
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	**0.79**	**0.09**	0.40	0.12
L+T	0.74	0.10	**0.56**	**0.10**
R+T	0.74	0.11	0.52	0.11
All (L+R+T)	**0.79**	**0.09**	**0.58**	**0.10**

The obtained results, presented in Table III, 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.

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 $MEA = 0.14$) and only rhythm features for valence ($R^2 = 0.15$ and $MEA = 0.15$).

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, we can conclude that 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.

VI. Selected Features Dedicated to the Detection of Arousal and Valence

Table IV presents 2 sets of selected features, which using the SMOreg algorithm obtained the best performance by detecting arousal (Section V). 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 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 V presents 2 sets of selected features, which using the SMOreg algorithm obtained the best performance by detecting valence (Section V). 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.

TABLE IV
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)
Melbands Crest (L)	Erbbands Flatness (L)
Melbands Flatness (L)	Erbbands Skewness (L)
Melbands Kurtosis (L)	**Melbands Crest (L)**
Melbands Skewness (L)	**Melbands Kurtosis (L)**
Melbands Spread (L)	Silence Rate (L)
Spectral Energy (L)	**Spectral Energy (L)**
Spectral Entropy (L)	**Spectral Entropy (L)**
Spectral Flux (L)	**Spectral Flux (L)**
Spectral Kurtosis (L)	**Spectral Rolloff (L)**
Spectral Rolloff (L)	**Spectral Skewness (L)**
Spectral Skewness (L)	Beats Count (R)
BPM Histogram (R)	Beats Loudness (R)
Danceability (R)	**Danceability (R)**
Onset Rate (R)	**Onset Rate (R)**
Beats Loudness Band Ratio (R)	**Beats Loudness Band Ratio (R)**
Chords Strength (T)	
Beats Loudness Band Ratio (R)	
HPCP Entropy (T)	
Key Strength (T)	
Chords Histogram (T)	

TABLE V
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 Salience (L)
Spectral Energy (L)	Silence Rate (L)
Zero Crossing Rate (L)	Spectral Centroid (L)
GFCC (L)	Spectral Energy (L)
MFCC (L)	Spectral Spread (L)
Beats Loudness (R)	**Zero Crossing Rate (L)**
Onset Rate (R)	**GFCC (L)**
Beats Loudness Band Ratio (R)	**MFCC (L)**
Chords Strength (T)	**Chords Strength (T)**
HPCP Entropy (T)	**HPCP Entropy (T)**
Key Strength (T)	**Key Strength (T)**
Chords Histogram (T)	Key Scale (T)

Comparing the sets of features dedicated to arousal (Table IV) and valence (Table V), 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).

VII. CONCLUSIONS

In this article, we studied the usefulness of audio features during emotion detection in music files. 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.

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

We can conclude that 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 affect on improving the effectiveness of automatic emotion detection in music files.

ACKNOWLEDGMENT

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

REFERENCES

[1] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. Zapata, and X. Serra, "ESSENTIA: an audio analysis library for music information retrieval," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013, pp. 493–498.

[2] G. Tzanetakis and P. Cook, "Marsyas: A framework for audio analysis," *Org. Sound*, vol. 4, no. 3, pp. 169–175, 2000.

[3] J. Grekow, "Mood tracking of musical compositions," in *Proceedings of the 20th International Conference on Foundations of Intelligent Systems*, ser. ISMIS'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 228–233.

[4] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 1, pp. 5–18, 2006.

[5] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[6] J. Grekow, "Music emotion maps in arousal-valence space," in *Computer Information Systems and Industrial Management: 15th IFIP TC8 International Conference, CISIM 2016, Vilnius, Lithuania, Proceedings*. Springer International Publishing, 2016, pp. 697–706.

[7] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proceedings of the International Conference on Multimedia Information Retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 267–274.

[8] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 2, pp. 448–457, 2008.

[9] J. J. Deng and C. H. Leung, "Dynamic time warping for music retrieval using time series modeling of musical emotions," *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 137–151, 2015.

[10] Y. Lin, X. Chen, and D. Yang, "Exploration of music emotion recognition based on midi," in *Proceedings of the 14th International Society for Music Information Retrieval Conference*, 2013.

[11] J. Grekow, "Audio features dedicated to the detection of four basic emotions," in *Computer Information Systems and Industrial Management: 14th IFIP TC 8 International Conference, CISIM 2015, Warsaw, Poland, September 24-26, 2015, Proceedings*. Springer International Publishing, 2015, pp. 583–591.

[12] R. Panda, B. Rocha, and R. P. Paiva, "Music emotion recognition with standard and melodic audio features," *Applied Artificial Intelligence*, vol. 29, no. 4, pp. 313–334, 2015.

[13] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, pp. 1802–1812, 2011.

[14] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vitória, Porto, Portugal*, 2012, pp. 523–528.

[15] C. Baume, G. Fazekas, M. Barthet, D. Marston, and M. Sandler, "Selection of audio features for music emotion recognition using production music," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*, 2014.

[16] S. Doraisamy, S. Golzari, N. M. Norowi, M. N. Sulaiman, and N. I. Udzir, "A study on feature selection and classification techniques for automatic genre classification of traditional malay music," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA*, 2008, pp. 331–336.

[17] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Emotion in music task: lessons learned," in *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands*, 2016.

[18] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[20] R. J. Quinlan, "Learning with continuous classes," in *5th Australian Joint Conference on Artificial Intelligence*. Singapore: World Scientific, 1992, pp. 343–348.

[21] Y. Wang and I. H. Witten, "Induction of model trees for predicting continuous classes," in *Poster papers of the 9th European Conference on Machine Learning*. Springer, 1997.

[22] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[23] L. Xu, P. Yan, and T. Chang, "Best first strategy for feature selection," in *Proceedings of the 9th International Conference on Pattern Recognition*, 1988, pp. 706–708 vol.2.