Audio Features Dedicated to the Detection of Four Basic Emotions

Jacek Grekow

Audio Features Dedicated to the Detection of Four Basic Emotions

Jacek Grekow

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

Abstract. In this paper, we decided to study the effect of extracted audio features, using the analysis tool Essentia, on the quality of constructed music emotion detection classifiers. The research process included constructing training data, feature extraction, feature selection, and building classifiers. We selected features and found sets of features that were the most useful for detecting individual emotions. We examined the effect of low-level, rhythm and tonal features on the accuracy of the constructed classifiers. We built classifiers for different combinations of feature sets, which enabled distinguishing the most useful feature sets for individual emotions.

Keywords: Music emotion recognition, Audio feature extraction, Music information retrieval.

1 Introduction

One of the most important elements when listening to music is the expressed emotions. The emotions contained in music can alter or deepen the emotional state of the listener. For example, the Funeral March listened to during a funeral deepens the emotional state of the departed's loved ones; while light and relaxing music listened to at home after a hard day's work can restore the listener's good mood. The elements of music that affect the emotions are timbre, dynamics, rhythm, and harmony. Changes in the types of instruments used, the dynamics, rhythm, and harmony change the emotions found in the music.

In the era of the Internet, searching music databases for emotions has become increasingly important. Automatic emotion detection enables indexing files in terms of emotions [1]. Automatic emotion detection also enables creating visual emotion maps of musical compositions [2].

In this paper, we decided to study the effect of extracted audio features, using the analysis tool Essentia [3], on the quality of constructed music emotion detection classifiers. We selected features and found sets of features that were the most useful for detecting individual emotions. We examined the effect of low-level, rhythm and tonal features on the accuracy of the constructed classifiers.

Studies on emotion detection in music are mainly based on two popular approaches: categorical or dimensional. The categorical approach [4][5][6][7] describes emotions with a discrete number of classes - affective adjectives. In the

dimensional approach [8][9][10], emotions are described as numerical values of valence and arousal. In this way, the emotion of a song is represented as a point on an emotion space. In this work, we used the categorical approach.

An important phase in emotion detection is feature extraction. There are several other studies on the issue of emotion detection using different audio tools for musical feature extraction. Studies [6][11][12] used a collection of tools that use the Matlab environment called MIR toolbox [13]. Feature extraction library jAudio [14] was used in studies [7][9]. Feature sets extracted from PsySound [15] were used in paper [12], while studies [8][16] used the Marsyas framework [17]. The Essentia [3] library for audio analysis was used in studies [18][19].

There are also papers devoted to the evaluation of audio features for emotion detection within one program. Song et al. [6] 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.

An important paper in the area of music emotion recognition was written by Yang et al. [20], who did a comprehensive review of the methods that have been proposed for music emotion recognition. Kim et al. [21] presented another paper surveying the state of the art in automatic emotion recognition.

2 Music Data

In this research, we use four emotion classes: energetic-positive, energetic-negative, calm-negative, calm-positive. They are presented with their abbreviations in Table 1 and cover the four quadrants of the two-dimensional Thayer model of emotion [22]. They correspond to four basic emotion classes: happy, angry, sad, and relaxed.

Table 1. Description of mood labels

Abbreviation	Description
e1	energetic-positive
e2	energetic-negative
e3	calm-negative
e4	calm-positive

To conduct the study of emotion detection, we prepared two sets of data. One set was used for building one common classifier for detecting the four emotions, and the other data set for building four binary classifiers of emotion in music. Both data sets consisted of six-second fragments of different genres of music: classical, jazz, blues, country, disco, hip-hop, metal, pop, reggae, and rock. The tracks were all 22050Hz Mono 16-bit audio files in .wav format.

The author of this paper, a music expert with a university musical education, labeled the music samples. The music expert listened to six-second music samples

and then labeled them with one of the emotions (e1, e2, e3, e4). In the case when the music expert was not certain which emotion to assign, such a sample was rejected. In this way, each file was associated with only one emotion/label.

The first training data set for emotion detection consisted of 324 files, 81 files labeled as e1, 81 files labeled as e2, 81 files labeled as e3, and 81 files labeled as e4.

We obtained the second training data from the first set. It consisted of four sets of binary data. For example, data set for binary classifier e1 consisted of 81 files labeled e1 and 81 files labeled not e1 (27 files each from e2, e3, e4). In this way, we obtained four binary data sets (consisting of examples of "e" and "not e") for four binary classifiers e1, e2, e3, e4.

3 Feature Extraction

For feature extraction, we used Essentia [3], a tool for audio analysis and audio-based music information retrieval. Essentia is an open-source C++ library, which was created at Music Technology Group, Universitat Pompeu Fabra, Barcelona.

We used Essentia version 2.0.1 (published in 02/2014), which 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.

The use of Essentia software entailed getting through installation procedures and the documentation. Launching the program required compiling the source code (C ++) and installing additional libraries.

Extracted features by Essentia are divided into three groups: low-level, rhythm and tonal features (Table 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 emotion, music data sets served as input data for the Essentia tool used for feature extraction. For each 6-second file from the music data set, we obtained a representative single feature vector. The obtained lengths of feature vectors had 471 features.

4 Results

4.1 The Construction of One Classifier Recognizing Four Emotions

We built classifiers for emotion detection using the WEKA package [23]. During the construction of the classifier, we tested the following algorithms: J48, RandomForest, BayesNet, IBk (K-nn), SMO (SVM). The classification results were calculated using a cross validation evaluation CV-10.

Table 2. The feature set obtained from Essentia

Group	Group Abbreviation	Features
Low-level features	L	Average Loudness
		Energy of the Barkbands
		Energy of the Erbbands
		Energy of the Melbands
		Dissonance
		Dynamic Complexity
		HFC (High Frequency Content)
		Pitch Salience
		Silence Rate
		Spectral Centroid
		Spectral Complexity
		Spectral Energy
		Spectral Energy Band High
		Spectral Energy Band Low
		Spectral Energy Band Middle High
		Spectral Energy Band Middle Low
		Zero Crossing Rate
		GFCC (Gammatone Feature Cepstral Coefficients)
		MFCC (Mel-Frequency Cepstral Coefficients)
Rhythm features	R	Beats Loudness
		Beats Loudness Band Ratio
		BPM (The mean of the most salient tempo)
		BPM Histogram
		Danceability
		Onset Rate
Tonal features	T	Chords Changes Rate
		Chords Number Rate
		Chords Strength
		Key Strength
		Chords Histogram
		HPCP (Harmonic Pitch Class Profile)

The first important result was that during the construction of the classifier we obtained the highest accuracy among all tested algorithms for SMO algorithm. SMO was trained using polynominal kernel.

Table 3. Accuracy obtained for SMO algorithm

	Accuracy
Before attribute selection	59.26%
After attribute selection	**64.50%**

The results obtained for SMO algorithm are presented in Table 3. The result (classifier accuracy) improved to **64.50%** after applying attribute selection (attribute evaluator: WrapperSubsetEval [24], search method BestFirst).

The confusion matrix (Table 4), obtained during classifier evaluation, shows that the most recognized emotion was e2 (F-measure 0.727), and the next emotions were e1 and e3 (F-measure 0.653 and 0.65). The hardest emotion to recognize was e4 (F-measure 0.544).

From the confusion matrix, we can conclude that usually fewer mistakes are made between the top (e1, e2) and bottom (e3, e4) quadrants of the Thayer model. At the same time, recognition of emotions on the valence axis (positive-negative) is more difficult.

Table 4. Confusion matrix for the best result

classified as –	a	b	c	d
a = e1	**62**	9	4	6
b = e2	19	**56**	3	3
c = e3	9	4	**51**	17
d = e4	19	4	18	**40**

The most important features (with group abbreviation) after applying attribute selection were:

- Energy of the Erbbands (L),
- MFCC (L),
- Onset Rate (R),
- Beats Loudness Band Ratio (R),
- Key Strength (T),
- Chords Histogram (T).

In the selected features, we have a representative of low-level (L), rhythm (R) and tonal (T) features. This means that features of each of the three groups are important/useful during emotion detection.

The results were not satisfactory; classifier accuracy was too low (**64.50%**). It is difficult to build a good classifier that differentiates four emotions equally well. Some emotions have better recognition (e2) and others worse (e1, e3, e4), which lowers total classifier accuracy.

4.2 The Construction of Binary Classifiers

To improve emotion detection accuracy, we decided to build specialized binary classifiers for each emotion. A binary classifier algorithm can better analyze data sets for the presence of a given emotion.

During the construction of the binary classifiers, we tested the following algorithms: J48, RandomForest, BayesNet, IBk (K-nn), and SMO (SVM) on the prepared binary data. We calculated the classification results using a cross validation evaluation CV-10.

Table 5. Classifier accuracy for emotions e1, e2, e3, and e4 obtained for SMO

	Classifiers for e1	Classifiers for e2	Classifiers for e3	Classifiers for e4
Before attribute selection	66.05%	87.04%	77.16%	65.43%
After attribute selection	80.86%	**90.74%**	87.03%	77.16%

Once again, we obtained the best results for SMO algorithm. The results are presented in Table 5. Accuracy improved (3-14 percentage points) for all four classifiers after applying attribute selection (attribute evaluator: WrapperSubsetEval, search method BestFirst).

The best results were obtained for emotion e2 (90.74%) and the worst results for emotion e4 (77.16%). We can conclude that in our case, emotions with a negative valence (e2, e3) are recognized better by approx. 10 percentage points than emotions with a positive valence (e1, e4). The obtained binary classifier accuracy results were higher (12-26 percentage points) than the accuracy of one classifier recognizing four emotions.

Table 6 presents the most important features obtained after feature selection (attribute evaluator: WrapperSubsetEval, search method BestFirst) for each emotion. In each feature set, we had a representative of low-level, rhythm features, even though we had different sets for each emotion. Only in the case of classifier e4, tonal features were not used. The energy of the bands was important for e1, e2, and e4 classifiers, but they differed as to which bands they pertain: e1 - Barkbands, e2 - Erbbands, and Melbands, e4 - Barkbands and Erbbands. High Frequency Content, which is characterized by the amount of high-frequency content in the signal is important for e3 and e4 classifiers. Beats Loudness Band Ratio (the beat's energy ratio on each band) was very important for emotion

detection because it was used in all sets. Another important feature was the tonal feature: Chords Histogram, which was used by e2 and e3 classifiers.

Table 6. Selected features used for building binary classifiers

Classifier	Selected features
e1	Energy of the Barkbands (L)
	Onset Rate (R)
	Beats Loudness Band Ratio (R)
	Key Strength (T)
e2	Average Loudness (L)
	Dissonance (L)
	Energy of the Erbbands (L)
	Energy of the Melbands (L)
	MFCC (L)
	Beats Loudness Band Ratio (R)
	Chords Changes Rate (T)
	Chords Histogram (T)
e3	High Frequency Content (L)
	Silence Rate (L)
	Spectral Energy Band Middle Low (L)
	Beats Loudness Band Ratio (R)
	Key Strength (T)
	Chords Histogram (T)
e4	Energy of the Barkbands (L)
	Energy of the Erbbands (L)
	High Frequency Content (L)
	Pitch Salience (L)
	Beats Loudness Band Ratio (R)

Feature sets seem to logically describe the nature of each emotion. More energetic emotions are described by features pertaining to energy and rhythm, and more calm emotions by parameters such as rhythm and the amount of high frequency.

4.3 Evaluation of Different Combinations of Feature Sets

During this experiment, we evaluated the effect of various combinations of feature sets - low-level (L), rhythm (R), tonal (T) - on classifier accuracy obtained for SMO algorithm. We calculated the classification results using a cross validation evaluation CV-10. To improve the classification results, we used attribute selection (attribute evaluator: WrapperSubsetEval, search method BestFirst). The obtained results are presented in Table 7.

The obtained results indicate that the use of all groups (low-level, rhythm, tonal) of features resulted in the best accuracy in most cases (e1, e2, e3). The only exception was classifier e4, where using the set L+T (low-level, tonal) had better results (80.24%) than using all features - accuracy 77.16%.

Table 7. Classifier accuracy for emotions e1, e2, e3, and e4 obtained for combinations of feature sets

Features set	Classifiers for e1	Classifiers for e2	Classifiers for e3	Classifiers for e4
L	72.22%	88.27%	79.62%	77.77%
R	73.45%	82.09%	81.48%	72.83%
T	72.22%	81.48%	76.54%	69.75%
L+R	77.16%	88.88%	77.77%	77.16%
L+T	79.01%	**90.74%**	**86.41%**	**80.24%**
R+T	**80.86%**	**90.12%**	79.01%	73.45%
All (L+R+T)	**80.86%**	**90.12%**	**87.03%**	77.16%

The use of individual feature sets L, R or T did not have better results than their combinations. Combining feature sets R+T (rhythm and tonal features) improved classifier results in the case of classifiers e1 and e2. Combining feature sets L+T (low-level and tonal features) improved classifier results in the case of classifiers e2, e3 and e4.

5 Conclusions

In this paper, we studied the effect of extracted audio features, using the analysis tool Essentia, on the quality of constructed music emotion detection classifiers. The research process included constructing training data, feature extraction, feature selection, and building classifiers.

We built a classifier recognizing four basic emotions, but its accuracy was not satisfactory (64.50%). We then built binary classifiers dedicated to each emotion with accuracy from 77% to 90%. We obtained information on which features are useful in the detection of particular emotions.

We examined the effect of low-level, rhythm and tonal feature sets on the accuracy of the constructed binary classifiers. We built classifiers for different combinations of feature sets, which enabled distinguishing the most useful feature sets for individual emotions. The obtained results present a new and interesting view of the usefulness of different feature sets for emotion detection.

Classifier accuracy could be better. The process of searching for and assessing new features describing audio files continues.

Acknowledgments. This paper is supported by the S/WI/3/2013.

References

1. Grekow, J. and Ras, Z.W.: Emotion Based MIDI Files Retrieval System. In: Ras, Z.W., Wieczorkowska, A.A. (eds.) Advances in Music Information Retrieval. SCI, vol. 274, pp. 261-284. Springer, Heidelberg (2010)

2. Grekow, J.: Mood tracking of musical compositions. Foundations of Intelligent Systems: ISMIS 2012, Lecture Notes in Artificial Intelligence, Vol. 7661, pp. 228-233 (2012)
3. Bogdanov, D., Wack N., Gomez 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 Conference on Music Information Retrieval, pp. 493-498 (2013)
4. Lu, L., Liu, D., Zhang, H. J.: Automatic mood detection and tracking of music audio signals. IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 1, pp. 5-18 (2006)
5. Grekow, J., Ras, Z.W.: Detecting emotions in classical music from MIDI files. In: Rauch, J., Ras, Z.W., Berka, P., Elomaa, T. (eds.) ISMIS 2009. LNCS (LNAI), vol. 5722, pp. 261-270. Springer, Heidelberg (2009)
6. Song, Y., Dixon, S., Pearce, M.: Evaluation of Musical Features for Emotion Classification. In Proceedings of the 13th International Society for Music Information Retrieval Conference (2012)
7. Xu, J., Li, X., Hao, Y., Yang, G.: Source Separation Improves Music Emotion Recognition. ACM International Conference on Multimedia Retrieval (2014)
8. Yang,Y.-H., Lin,Y.-C., Su,Y.-F., Chen, H.H.: A regression approach to music emotion recognition. IEEE Transactions on Audio, Speech, and Language Processing 16(2), pp. 448-457 (2008)
9. 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)
10. Schmidt, E. M., Turnbull, D. and Kim Y. E.: Feature Selection for Content-Based, Time-Varying Musical Emotion Regression. Proc. ACM SIGMM International Conference on Multimedia Information Retrieval, Philadelphia, PA (2010)
11. Saari, P., Eerola, T., Fazekas, G., Barthet, M., Lartillot, O. and Sandler, M.: The role of audio and tags in music mood prediction: a study using semantic layer projection. In: Proceedings of the 14th International Society for Music Information Retrieval Conference (2013)
12. Aljanaki, A., Wiering, F., Veltkamp, R.C.: Computational modeling of induced emotion using GEMS. In: Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR), pp. 373-378 (2014)
13. Lartillot, O., Toiviainen, P.: MIR in Matlab (II): A Toolbox for Musical Feature Extraction from Audio. In International Conference on Music Information Retrieval, pp. 237-244 (2007)
14. McKay C., Fujinaga I., Depalle P.: jAudio: a feature extraction library. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR05), pp. 600-603 (2005)
15. Cabrera, D.: PSYSOUND: A computer program for psychoacoustical analysis. In: Proceedings of the Australian Acoustical Society Conference, pp. 47-54 (1999)
16. Grekow, J.: Mood Tracking of Radio Station Broadcasts, Foundations of Intelligent systems: ISMIS 2014, Lecture Notes in Computer Science, Volume 8502, pp. 184-193 (2014)
17. Tzanetakis, G., Cook, P.: Marsyas: A framework for audio analysis. Organized Sound 10, 293-302 (2000)
18. Laurier, C.: Automatic Classification of Musical Mood by Content-Based Analysis. PhD thesis, UPF, Barcelona, Spain (2011)

19. Sarasua, A., Laurier C., Herrera P.: Support Vector Machine Active Learning for Music Mood Tagging. 9th International Symposium on Computer Music Modeling and Retrieval (CMMR), London (2012)
20. Yang Y.-H., Chen, H.H.: Machine Recognition of Music Emotion: A Review. ACM Transactions on Intelligent Systems and Technology, Volume 3 Issue 3, Article No. 40 (2012)
21. Kim, Y., Schmidt, E., Migneco, R., Morton, B., Richardson, P., Scott, J., Speck, J., Turnbull, D.: State of the Art Report: Music Emotion Recognition: A State of the Art Review. In Proceedings of the 11th International Society for Music Information Retrieval Conference, pp. 255-266 (2010)
22. Thayer, R.E.: The Biopsychology of Mood and Arousal. Oxford University Press (1989)
23. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)
24. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence, 97(1-2), pp. 273-324 (1997)