Chapter 7
Detection of Four Basic Emotions

7.1 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 this chapter, we study the quality of the constructed music emotion detection classifiers using audio features extracted by two different analysis tools: Essentia [8] and Marsyas [106]. We also decided to study the effect of extracted audio features on the quality of the 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.

7.2 Related Work

Studies on emotion detection in music are mainly based on two popular approaches: categorical or dimensional. In the dimensional approach, emotions are described as numerical values of valence and arousal. The categorical approach describes emotions with a discrete number of classes – affective adjectives. In this chapter, we used the categorical approach.

J. Grekow, *From Content-Based Music Emotion Recognition to Emotion Maps of Musical Pieces*, Studies in Computational Intelligence 747,
https://doi.org/10.1007/978-3-319-70609-2_7

One of the first papers on categorical emotion detection was a study by Li and Ogihara [57], who trained support vector machines (SVM) to classify music into one of 13 mood categories using a multi-label classification method. A labeled collection consisted of 499 sound files (30-seconds each) from the ambient, classical, fusion, and jazz genres. They used Marsyas to extract the timbral, rhythmic, and pitch features. The achieved accuracy was low, at a level of 45%.

Lu et al. [63] examined emotion detection and emotion tracking using intensity, timbre, and rhythm acoustic features. Emotion categories corresponded to the four quadrants on Thayer's two-dimensional (Energy-Stress) model [103]. To train, Gaussian Mixture Models were used on a set of 800 classical music clips (20 s each). The system of emotion detection achieved an average accuracy of 86%. In addition to emotion detection, emotion tracking through a music piece was presented, which divided the music into several segments.

The problem of multi-label classification of emotions in musical recordings was also presented by Wieczorkowska et al. [113]. The data set contained 875 samples with a length of 30 s each. For classification, the k-nearest neighbors (k-nn) algorithm was used.

In the community of Music Information Retrieval Evaluation eXchange (MIREX) for automatic music mood classification, five mood clusters were used for song categorization [43]. The Audio Mood Classification evaluation task was started for the first time in 2007. The ground truth set consisted of 600 clips (30 second each), with 120 in each mood cluster. The five emotion clusters, which were used by MIREX Audio Mood Classification, have not been frequently used in other music emotion detection works. Hu et al. in [44] indicates that the clusters might not be optimal and noticed some semantic overlap.

A popular emotion set used to categorize emotions in music turned out to be a collection consisting of 4 classes: happy, angry, sad, and relaxed. It corresponds to the four quadrants of the two-dimensional valence-arousal plane, which was used by Laurier in [54], where binary classifiers were constructed for each category. A data set of 1000 songs (30 s each) was divided between 4 categories. Classification accuracy was from 84% to 98%, and was obtained for the SVM algorithm with polynomial and linear kernel.

Four emotion classes (happy, angry, sad, relaxed) were also used in the categorical approach by Song et al. in [100]. The collected ground truth data set consisted of 2904 songs that were labeled with one of the four emotions. The highest accuracy, 53%, was achieved for SVM with polynomial kernel. Song et al. explored the relationship between musical features extracted by MIRtoolbox [53] and emotions. They compared the emotion prediction results for four sets of features: dynamic, rhythm, harmony, and spectral.

7.3 Music Data

In this research, we use four emotion classes: happy, angry, sad, and relaxed. They corresponds to the four quarters of Russell's model [88], which were formed by dividing a plane by two perpendicular axes: arousal and valence. The basic classes of emotions are assigned to the quarters as follows:

- happy – arousal high, valence high – Q1;
- angry – arousal high, valence low – Q2;
- sad – arousal low, valence low – Q3;
- relaxed – arousal low, valence high – Q4.

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 6-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 data set that was used in this experiment consisted of 324 six-second fragments and was described in detail in Chap. 3 Sect. 3.3.

Data annotation was done by five music experts with a university music education. The annotation process of music files with emotion classes was described in Chap. 3 Sect. 3.3.3. The amount of examples in the quarters on the A-V emotion plane is presented in Table 7.1.

To build binary classifiers, we prepared the second training data from the first set, which consisted of four sets of binary data. For example, the 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). Thus, we obtained four binary data sets (consisting of 81 examples of 'e' and 81 examples of 'not e') for four binary classifiers e1, e2, e3, e4. To make the number of examples uniform in the binary data sets for the four classes, the number of examples labeled e1 was reduced to 81 and the number of those labeled e2 and e3 was reduced to 81.

Table 7.1 Amount of examples in quarters on A-V emotion plane

Basic emotion	Emotion abbreviation	Quarter	Arousal-Valence	Amount of examples
Happy	e1	Q1	High-High	93
Angry	e2	Q2	High-Low	70
Sad	e3	Q3	Low-Low	80
Relaxed	e4	Q4	Low-High	81

7.4 Feature Extraction

For feature extraction, we used Essentia [8] and Marsyas [106], which are tools for audio analysis and audio-based music information retrieval. Marsyas framework was described in Chap. 6 Sect. 6.2.2, and Essentia extractors were described in Chap. 6 Sect. 6.2.1.

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

7.5 Results

7.5.1 Construction of One Classifier Recognizing Four Emotions

We built classifiers for emotion detection using the following algorithms: J48, RandomForest, BayesNet, K-nn, SMO (SVM). The classification results were calculated using a cross validation evaluation CV-10.

The first important result was that during the construction of the classifier for 2 data sets obtained from Marsyas and Essentia, the highest accuracy among all tested algorithms was obtained for SMO algorithm [79]. SMO was trained using polynomial kernel. The second best algorithm was RandomForest.

The best results we obtained using the feature set from Essentia. The results obtained for SMO algorithm are presented in Table 7.2. The classifier accuracy improved to 64.51% after applying attribute selection (attribute evaluator: Wrapper Subset Evaluator [50], search method: Best First [117]). In Essentia, tonal and rhythm features greatly improve classifier accuracy. These features are not available in Marsyas and thus Essentia obtains better results.

The confusion matrix (Table 7.3), obtained during classifier evaluation, shows that the most recognized emotions were e2 and e4 (F-measure = 0.68), and the next emotion was e1 (F-measure = 0.64). The hardest emotion to recognize was e3 (F-measure = 0.59).

Table 7.2 Accuracy obtained for SMO algorithm

	Essentia (%)	Marsyas (%)
Before attribute selection	62.04	54.01
After attribute selection	**64.51**	58.02

Table 7.3 Confusion matrix for the best result

		Predicted class			
		e1	e2	e3	e4
Actual class	e1	**66**	10	4	13
	e2	21	**42**	5	2
	e3	14	2	**42**	22
	e4	11	0	11	**59**

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

The most important features, for the detection of four basic emotions, after applying attribute selection were:

- Dissonance (L),
- Melbands Crest (L),
- Melbands Kurtosis (L),
- Melbands Spread (L),
- Spectral Complexity (L),
- Spectral Energy (L),
- Spectral Kurtosis (L),
- Spectral Spread (L),
- Spectral RMS (L),
- Harmonic Pitch Class Profile Entropy (T),
- Tuning Diatonic Strength (T),

where L and T represent feature group abbreviations: low-level (L), tonal (T).

The results were not satisfactory; classifier accuracy was too low (64.51%). It is difficult to build a good classifier that differentiates four emotions equally well.

7.5.2 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. In this experiment, we used features extracted from Essentia, which was selected as the winner in the previous experiment.

Table 7.4 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	75.92	80.24	74.07	72.84
After attribute selection	87.04	87.65	82.71	87.04

Table 7.5 Selected features used for building binary classifiers

Classifier	Selected features	Classifier	Selected features
e1	Barkbands Kurtosis (L) Dissonance (L) High Frequency Content (L) Spectral Centroid (L) Spectral Complexity (L) Spectral Entropy (L) Spectral Strong Peak (L) Beats Loudness (R) Chords Strength (T) Key Strength (T) Chords Histogram (T)	e3	Melbands Crest (L) Melbands Kurtosis (L) Pitch Salience (L) Spectral Energy (L) Spectral Entropy (L) Key Strength (T)
e2	Barkbands Flatness (L) Melbands Flatness (L) Silence Rate (L) Spectral Entropy (L) Onset Rate (R) Chords Strength (T)	e4	Barkbands Kurtosis (L) Barkbands Skewness (L) Barkbands Spread (L) Melbands Crest (L) Spectral Complexity (L) Beats Loudness Band Ratio (R) Harmonic Pitch Class Profile (T)

Once again, we obtained the best results for SMO algorithm, which are presented in Table 7.4. Accuracy improved (7–15 % points) for all four classifiers after applying attribute selection (attribute evaluator: Wrapper Subset Evaluator, search method: Best First).

The best classifier accuracy was obtained for emotion e2 (87.65%); the results were also high for e1 and e4 (87.04%). Summarizing, accuracy is higher than 80% for all emotions, which is a big improvement of accuracy in comparison with the previous experiment, where we used one classifier recognizing four emotions (64.51%).

Table 7.5 presents the most important features obtained after feature selection (attribute evaluator: Wrapper Subset Evaluator, search method: Best First) for each emotion. Each classifier dedicated to recognizing only one emotion has its own set of features, different from the rest, consisting of a combination of low-level, rhythm, and tonal features. We can notice a domination of low-level features. Features describing

Table 7.6 Classifier accuracy for emotions e1, e2, e3, and e4 obtained for combinations of feature sets

Feature set	Classifiers for e1 (%)	Classifiers for e2 (%)	Classifiers for e3 (%)	Classifiers for e4 (%)
L	79.01	86.42	77.77	85.80
R	76.54	83.33	78.93	77.16
T	75.93	79.63	82.10	77.77
L + R	77.16	89.50	80.25	85.80
L + T	**87.04**	**92.59**	**82.71**	86.42
R + T	**87.04**	83.95	**82.71**	75.30
All (L + R + T)	**87.04**	87.65	**82.71**	**87.04**

spectrum occur in all four sets. Features describing energy in the Barkbands of a spectrum occur in three sets (e1, e2, e4). Features describing energy in the Melbands of a spectrum occur in three sets (e2, e3, e4). Tonal features, Chord Strength, and Key Strength are also important since they are included in two sets each.

7.5.3 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 (Table 7.6). The best results for each classifier have been marked in bold.

The obtained results indicate that the use of all groups (low-level, rhythm, tonal) of features resulted in the best accuracy or equal with use of 2 groups of features, in most cases (e1, e3, e4). The only exception was classifier e2, where using the set L + T (low-level, tonal) had better results (92.59%) than using all features – accuracy 87.65%.

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

7.5.4 Emotion Maps

The result of emotion tracking of musical compositions are emotion maps. We used the best obtained classifier for predicting four emotions to analyze musical compositions. The compositions were divided into 6-second segments with a 3/4 overlap.

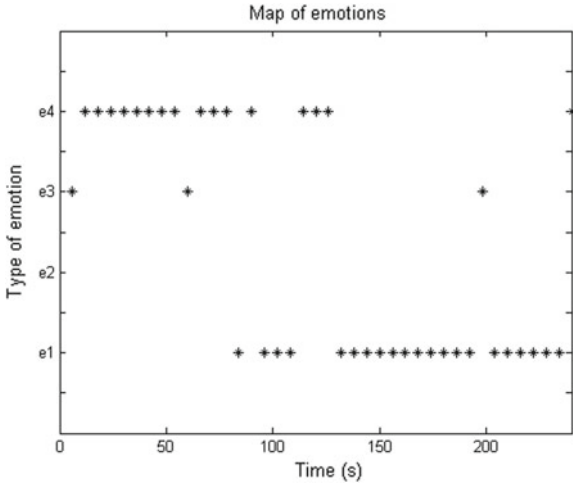

Fig. 7.1 Emotion map for the song Let It Be by Paul McCartney (The Beatles)

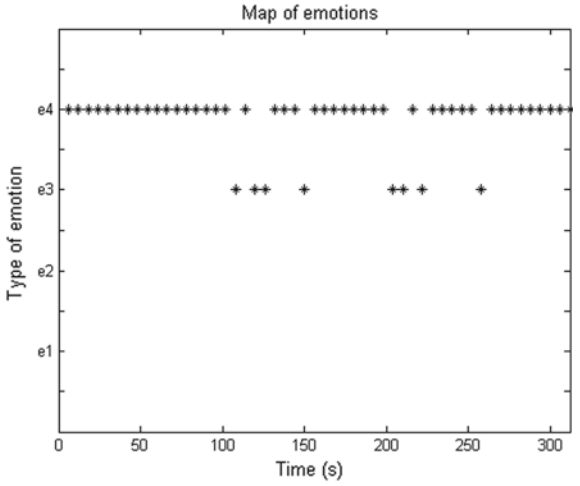

Fig. 7.2 Emotion map for Piano Sonata No. 8 in C minor, Op. 13 (Pathetique), 2nd movement, by Ludwig van Beethoven

For each segment, features were extracted and classifiers for emotion detection were used.

From the created emotion maps, we can find out:

- which emotion or emotions are dominant throughout the entire composition,
- how often changes in emotions occur and in which directions,
- which emotions are shaped and developed during the duration of the composition: at the beginning, in the middle, and at the end of the piece.

Figures 7.1 and 7.2 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 (Pathetique), 2nd movement, by Ludwig van Beethoven. The

horizontal axis shows the time in seconds and the vertical axis the emotion occurring at a given moment.

From the emotion maps of the presented compositions, we can notice their diametrically different emotional character. In the case of Sonata Pathetique, presented in Fig. 7.2, the dominating emotion is e4 (relaxed); and in Let It Be, presented in Fig. 7.1, the dominating emotions are e1 (happy) and e4 (relaxed). Analyzing the development of emotions over time, we notice that in Let It Be there is a different emotion at the beginning and a different one at the end of the composition; in the beginning part of the piece (up until about 90 s.) e4 dominates, and in the end (from 130 s.) e1 dominates, while emotion e3 (sad) occurs sporadically and for a short time (s. 10, 60, 200). In Sonata Pathetique, emotion e4 dominates throughout the entire composition, with short changes in the direction of emotion e3 (s. 120, 200).

The presented emotion maps can have various applications. They can be used to search the database for compositions with a similar or specified distribution of emotions. After extracting parameters describing emotions on a map, we could compare groups of compositions or even compositions by various composers [32].

The emotion maps of compositions using four basic emotions (happy, angry, sad and relaxed) are, however, an oversimplification of the many shades of emotions occurring in music. The detailed distribution of emotions over time of the aforementioned compositions is presented in Chap. 9 Sect. 9.8.1, where we used the Arousal-Valance plane to build emotion maps.

7.6 Conclusions

In this chapter, we presented the detection of four basic emotions in music files. We built a classifier recognizing four basic emotions, but its accuracy was not satisfactory (64%). We then built 4 binary classifiers dedicated to each emotion, with much higher accuracy, from 82% to 87%.

We studied the effect of the extracted audio features on the quality of the constructed music emotion detection classifiers. We obtained information about which features are useful in the detection of particular emotions. The use of all three groups (low-level, rhythm, tonal) of features resulted in the best accuracy or equal with the use of two groups of features, in most cases of binary classifiers.

As a result of emotion tracking of musical compositions, we constructed emotion maps visualizing the distribution of emotions over time. Emotion maps provide new knowledge about the distribution of four emotions in musical compositions and can be used to search for compositions with a specified distribution of emotions, among others.