Chapter 9
Music Emotion Maps in the Arousal-Valence Space

9.1 Introduction

Emotions are a dominant element in music, and they are the reason people listen to music so often [81]. 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. [40, 85].

The emotional content of music is not always constant, and even in classical music or jazz changes often. Analysis of emotions contained in music over time is a very interesting aspect of studying the content of music. It can provide new knowledge on how the composer emotionally shaped the music or why we like some compositions more than others.

9.2 Related Work

Music emotion recognition concentrates on static or dynamic changes over time. Static music emotion recognition uses excerpts from 15 to 30 seconds and omits changes in emotions over time. It assumes the emotion in a given segment does not change. A regression approach and static emotion recognition was presented in [60, 109, 119].

Dynamic music emotion recognition analyzes changes in emotions over time. Methods for detecting emotions using a sliding window are presented in [32, 34, 51, 63, 96, 119]. Deng and Leung [16] proposed multiple dynamic textures to model emotion dynamics over time. To find similar sequence patterns of musical emotions, they used subsequence dynamic time warping for matching emotion dynamics. Aljanaki et al. [3] investigated how well structural segmentation explains emotion segmentation. They evaluated different unsupervised segmentation methods on the

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_9

task of emotion segmentation. Imbrasaite et al. [46] and Schmidt et al. [95] used Continuous Conditional Random Fields for dimensional emotion tracking.

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

The elements of music that affect the emotions are timbre, dynamics, rhythm, and harmony. 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.

Most papers, however, focus on studying features using a classification model [35, 36, 73, 90, 100]. Music emotion recognition combining standard and melodic features extracted from audio was presented by Panda et al. in [73]. Song et al. [100] 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. Baume et al. [6] evaluated different types of audio features using a five-dimensional support vector regressor in order to find the combination that produces the best performance.

9.3 Music Data

The data set that was used in this experiment 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 data set has been described in detail in Chap. 3 Sect. 3.3.

During the annotation of music samples, we used Russell's two-dimensional valence-arousal (V-A) model to measure emotions in music [88]. The model consists of two independent dimensions of valence (horizontal axis) and arousal (vertical axis). The annotation process of music files has been 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 9.1.

Table 9.1 Amount of examples in quarters on A-V emotion plane

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

9.4 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 has been described in Chap.6 Sect. 6.2.2. Essentia extractors have been described in Chap. 6 Sect. 6.2.1.

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.

9.5 Regressor Training

We built regressors for predicting arousal and valence using the WEKA package [114]. For training and testing, the following regression algorithms were used: SMOreg, REPTree, M5P. SMOreg algorithm [99] implements the support vector machine for regression. REPTree algorithm [41] builds a regression tree using variance and prunes it using reduced-error pruning. M5P implements base routines for generating M5 Model trees and rules [83, 110].

Before constructing regressors, arousal and valence annotations were scaled between $[-0.5, 0.5]$. We evaluated the performance of regression using the tenfold cross validation technique (CV-10).

The highest values for determination coefficient (R^2) were obtained using SMOreg (implementation of the support vector machine for regression) [99]. After applying attribute selection (attribute evaluator: Wrapper Subset Evaluator [50], search method: Best First [117]), we obtained $R^2 = 0.79$, for arousal and $R^2 = 0.58$ for valence. Mean absolute error reached values $MAE = 0.09$ for arousal and $MAE = 0.10$ for valence (Table 9.2).

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,

Table 9.2 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	**0.79**	**0.09**	**0.58**	**0.10**	0.73	0.11	0.25	0.14

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 $MAE = 0.09$.

9.6 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 Wrapper Subset Evaluator and search method Best First.

The obtained results, presented in Table 9.3, 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.

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

Table 9.3 R^2 and MAE for arousal and valence obtained for combinations of feature sets

Feature 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	**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**

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.

9.7 Selected Features Dedicated to the Detection of Arousal and Valence

Table 9.4 presents 2 sets of selected features, which using the SMOreg algorithm obtained the best performance by detecting arousal (Sect. 9.6). 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 9.5 presents 2 sets of selected features, which using the SMOreg algorithm obtained the best performance by detecting valence (Sect. 9.6). 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 Entropy, Key Strength.

Comparing the sets of features dedicated to arousal (Table 9.4) and valence (Table 9.5), 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).

Table 9.4 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)
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)
Beats Per Minute (BPM) Histogram (R)	Beats Loudness (R)
BPM of the Most Salient Tempo (R)	**Danceability (R)**
Danceability (R)	**Onset Rate (R)**
Onset Rate (R)	**Beats Loudness Band Ratio (R)**
Beats Loudness Band Ratio (R)	
Chords Strength (T)	
Harmonic Pitch Class Profile Entropy (T)	
Key Strength (T)	
Chords Histogram (T)	

Table 9.5 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 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)

9.8 Emotion Maps

The result of emotion tracking are emotion maps. We used the best obtained models for predicting arousal and valence to analyze musical compositions. The compositions were divided into 6-second 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 4 different ways:

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

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

9.8.1 Emotion Maps of Two Compositions

Figures 9.1 and 9.2 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.

Emotion maps present two different emotional aspects of these compositions. The first significant difference is distribution on the quarters of the Arousal-Valence map. In Let It Be (Fig. 9.1b), the emotions of quadrants Q4 and Q1 (high valence and low-high arousal) dominate. In Sonata Pathetique (Fig. 9.2b), the emotions of quarter Q4 (low arousal and low valence) dominate with an incidental emergence of emotions of quarter Q3 (low arousal and low valence).

Another noticeable difference is the distribution of arousal over time. Arousal in Let It Be (Fig. 9.1c) has a rising tendency over time of the entire song, and varies from low to high. In Sonata Pathetique (Fig. 9.2c), in the first half (s. 0–160) arousal has very low values, and in the second half (s. 160–310) arousal increases incidentally but remains in the low value range.

The third noticeable difference is the distribution of valence over time. Valence in Let It Be (Fig. 9.1d) remains in the high (positive) range with small fluctuations, but it is always positive. In Sonata Pathetique (Fig. 9.2d), 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 Pathetique is mostly calm and predominantly positive.

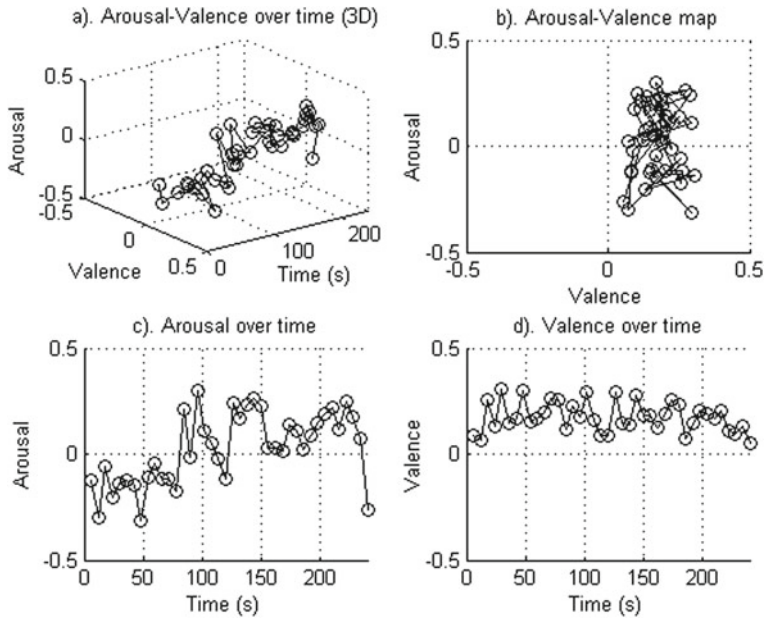

Fig. 9.1 A-V maps for the song Let It Be by Paul McCartney (The Beatles)

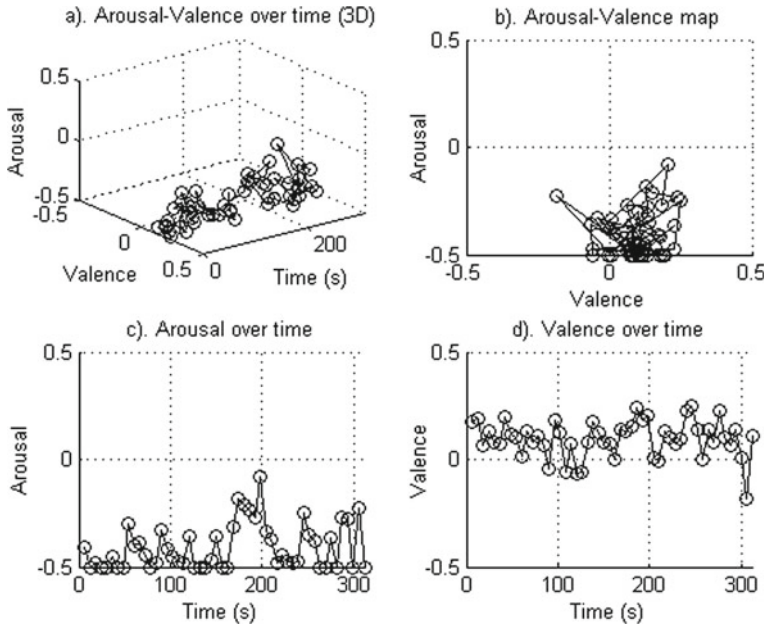

Fig. 9.2 A-V maps for Piano Sonata No. 8 in C minor, Op. 13 (Pathetique), 2nd movement, by Ludwig van Beethoven

9.8.2 Features Describing Emotion Maps

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

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

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

9.8.3 Comparison of Musical Compositions

Another experiment was to compare selected well-known Ludwig van Beethoven's Sonatas with several of the most famous songs by The Beatles. We used nine musical compositions from each group for the comparison (Table 9.6). This experiment did not aim to compare all the works of Beethoven and The Beatles, but only to find the rules and most important features distinguishing these 2 groups.

Each sample was segmented and arousal and valence were detected. Then, 15 features, which were presented in the previous section, were calculated for each sample. We used the PART algorithm [24] from the WEKA package [114] to find the decision-making rules differentiating the two groups.

It turned out that the most distinguishing feature for these two groups of musical compositions was the *Standard deviation of valence*. It was significantly smaller in The Beatles' songs than in Beethoven's compositions (Fig. 9.3). *Standard deviation of valence* reflects how big deviations were from the mean. The results show that in Beethoven's compositions valence values were much more varied than in the songs of The Beatles.

Table 9.6 List of musical compositions

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

Fig. 9.3 Box plot of *Standard deviation of valence* in The Beatles' and in Beethoven's compositions

To find another significant feature in the next stage, we removed the characteristic that we found previously (*Standard deviation of valence*) from the data set. Another significant feature was *Standard deviation of arousal*. In Beethoven's compositions, the values of the *Standard deviation of arousal* were much greater than in the Beatles' songs (Fig. 9.4). This proves the compositions have a greater diversity of tempo and volume.

In the next analogous stage, the feature we found was *Standard deviation of derivative of arousal*. It reflects the magnitude of changes in arousal between the studied segments. We found higher values of *Standard deviation of derivative of arousal* in Beethoven's compositions (Fig. 9.5).

An example of a feature that is unsuitable for differentiating between two examined groups of compositions is presented in Fig. 9.6. Overlapping values of the feature *Percentage representation of emotion e4*, obtained for compositions by The Beatles and Beethoven, cause that the usefulness of this feature to differentiate the way emotions are shaped in the studied groups is small.

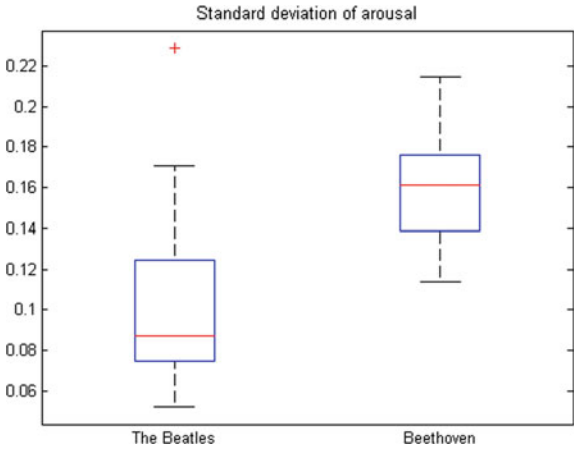

Fig. 9.4 Box plot of *Standard deviation of arousal* in The Beatles' and in Beethoven's compositions

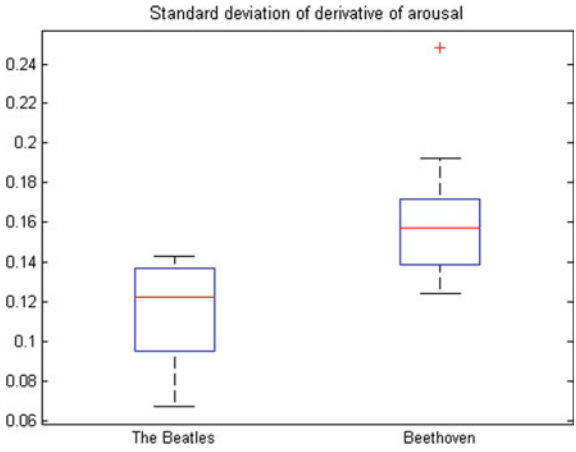

Fig. 9.5 Box plot of *Standard deviation of derivative of arousal* in The Beatles' and in Beethoven's compositions

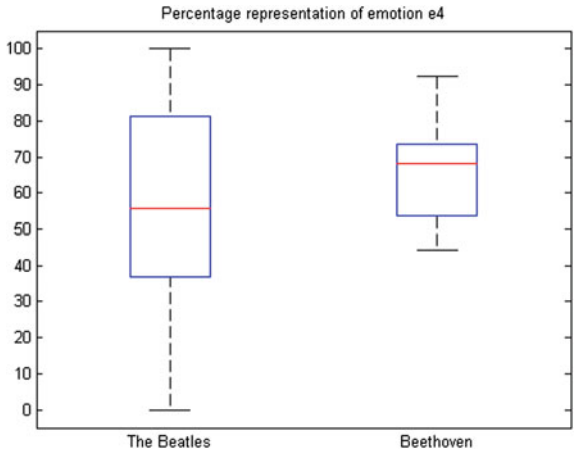

Fig. 9.6 Box plot of *Percentage representation of emotion e4 (relaxed)* in The Beatles' and in Beethoven's compositions

The interesting thing is that in the group of the most important distinguishing features we did not find features describing the emotion type (*Mean value of arousal*, *Mean value of valence*, or *Percentage representation of emotion in 4 quarters*). This is confirmed by the fact that we cannot assign common emotions to the different sample groups (Beethoven, The Beatles); in all groups, we have emotions from the four quadrants of the emotion model.

We can conclude that features that better distinguish between the two groups of compositions were features pertaining to changes in emotions and their distribution in the musical compositions.

9.9 Conclusions

In this chapter, we presented the detection of emotions as a problem of regression. The result of applying regressors are emotion maps of the musical compositions. Conducting experiments required the construction of regressors, attribute selection, and analysis of selected musical compositions.

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

In this chapter, we also studied the usefulness of audio features during emotion detection. Different feature sets were used to test the performance of built regression models intended to detect arousal and valence. 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. These features can have an affect on improving the effectiveness of automatic emotion detection in music files.