Chapter 5
Hierarchical Emotion Detection in MIDI Files

5.1 Introduction

In our research, we concentrated on emotion detection in MIDI files [87] containing symbolic representation of music (key, structure, chords, instrument). The means of representation of music content in MIDI files is much closer to the description that is used by musicians, composers, and musicologists. To describe music, they use key, tempo, scale, sounds, etc. This way, we avoid the difficult stage of extracting separate notes, tracks, and instruments from audio files; and we can concentrate on the deciding element, which is the music content.

Listening to music is a particularly emotional activity. People need a variety of emotions, and music is perfectly suited to provide it to them. However, it turns out that musical compositions do not contain one type of emotion, e.g. only positive or only negative. During the course of one composition, these emotions can take on a variety of shades and change several times with varying intensity.

Apart from emotion detection, this chapter presents a strategy for the analysis of emotions contained within musical compositions. We present a method for tracking changing emotions during the course of a musical piece. The collected data allowed determining the dominant emotion in the musical compositions, presenting emotion histograms, and constructing maps visualizing the distribution of emotions over time.

5.2 Related Work

There are few papers dedicated to emotion detection in MIDI files; most focus on emotion detection in audio files [49, 118]. In addition to studies on emotion detection, there are papers on modifying MIDI file parameters with the aim of obtaining a specified emotion.

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_5

Wang et al. in [71] applied a hierarchical model for emotion detection. Emotion groups were created on the basis of Thayer's model and contained 2 emotions at the first level and 6 emotions at the second. The features used to build classifiers referred to pitch, intervals, tempo, instrument type, meter, and tonality.

Emotion detection in MIDI files can also be found in the work of DiPaola and Arya [17], who combined the emotional content of a piece with visualization elements. They used the detected emotion for animating a 3-D face. The features used referred to rhythm, volume, timbre, articulation, melody, and tonality.

Lin et al. [60] examined music emotion regression performance using audio, lyric, and MIDI features. Two sets of MIDI files were used: the first set was converted from audio files, and the second set was obtained from the Internet and musical score conversion. They found that the MIDI features performed better than the audio features.

A connection between MIDI files and emotion was presented in [9], where a computer program was used to produce performances with different emotional expressions. The program used a set of rules characteristic for each emotion (fear, anger, happiness, sadness, solemnity, tenderness), which were used to modify such parameters of MIDI files as tempo, sound level, articulation, tone onsets and delays.

Livingstone and Brown [61] proposed a dynamic music environment, where MIDI music tracks adjusted in real-time to the emotion in the computer games. Music emotion rules, which connect 8 emotion categories to musical elements such as mode, tempo, loudness, harmonic complexity and articulation, were collected and implemented.

Moriguchi et al. [69] proposed a system for controlling the degrees of emotions in MIDI files. Parameters such as timbre, tempo, number of performance tracks, and loudness of a given excerpt were used to modify the expressed emotion in the music when played back to the listener.

5.3 MIDI Music Data

The data set that was used in the conducted experiments consisted of 350 six-second MIDI excerpts and was described in detail in Chap. 3 Sect. 3.4. The hierarchical emotion model we used (Fig. 5.1) was based on Russell's circumplex model, and consisted of emotion categories on two levels, L1 and L2. The first level (L1) contains 4 categories, while the second level (L2) is related to the first, and is made up of 12 sub-emotions, 3 emotions for each emotion contained in the first level.

Data annotation was done by five music experts with a university music education. The amount of obtained examples labeled by emotions on the first level are presented in Table 5.1, and those labeled by emotions on the second level are presented in Table 5.2.

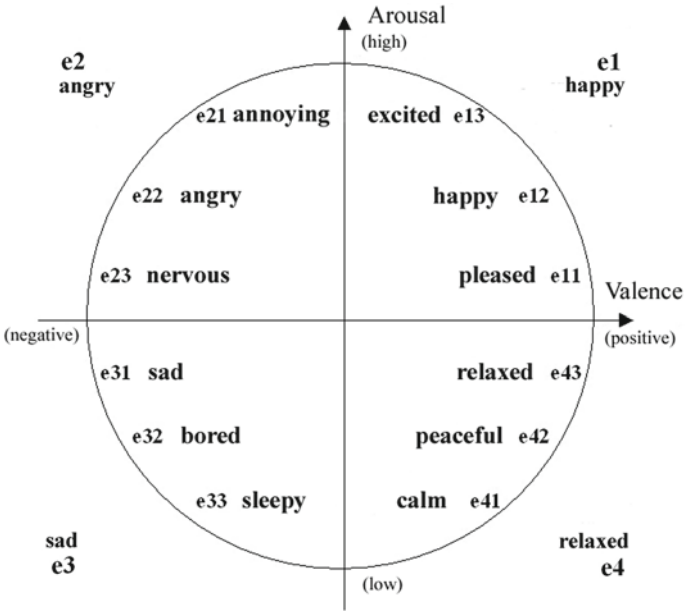

Fig. 5.1 Hierarchical model of emotions based on Russell's circumplex model

Table 5.1 Amount of MIDI examples labeled by emotions from the first level (L1)

Emotion abbreviation	Emotion	Amount of examples
e1	happy	89
e2	angry	105
e3	sad	79
e4	relaxed	77

5.4 Feature Extraction

We used our own software (written by the author) for feature extraction. 63 MIDI features were obtained for each 6-second labeled MIDI excerpt, and the extracted MIDI features were divided into four groups: rhythm, harmony, harmony-rhythm, and dynamic, and were described in Chap. 4 Sect. 4.3.

Table 5.2 Amount of MIDI examples labeled by emotions from the second level (L2)

Emotion abbreviation	Emotion	Amount of examples
e11	pleased	39
e12	happy	36
e13	excited	14
e21	annoying	35
e22	angry	36
e23	nervous	34
e31	sad	37
e32	bored	23
e33	sleepy	19
e41	calm	19
e42	peaceful	15
e43	relaxed	43

5.5 Construction of Classifiers

5.5.1 First Level Classifiers

5.5.1.1 One Classifier Recognizing Four Emotions

The classifier on the first level should be able to recognize the 4 categories of emotions that correspond to the four quarters of Russell's model: happy, angry, sad, and relaxed. From the music samples labeled by music experts and the extracted MIDI features, we created ARFF files that were the input data for the algorithms building classifiers.

We built classifiers for emotion detection using the following algorithms: J48, BayesNet, K-nn, SMO. J48 implements the C4.5 decision tree [82], BayesNet is an implementation of the Bayesian network classifier [41], K-nn represents K-nearest neighbours classifier [1], and SMO implements sequential minimal optimization algorithm for training a support vector classifier [78].

The classification results were calculated using a cross validation evaluation CV-10. The highest accuracy among all tested algorithms was obtained for the J48 algorithm (Table 5.3).

Table 5.3 Accuracy of classifiers obtained for first level (L1) classifiers

Classifier(%)	J48	BayesNet	K-nn	SMO
Accuracy after attribute selection	**76.00**	70.28	74.28	66.85
Accuracy after attribute selection	**82.00**	70.00	81.42	73.42

Table 5.4 Confusion matrix obtained for the J48 algorithm

		Predicted class			
		e1	e2	e3	e4
Actual class	e1	**66**	17	1	5
	e2	18	**83**	3	1
	e3	0	0	**71**	8
	e4	2	2	6	**67**

Classifier accuracy for algorithm J48 improved to **82.00%** after applying attribute selection (attribute evaluator: Wrapper Subset Evaluator [50], search method: Best First [117]).

From the confusion matrix (Table 5.4) obtained during classifier evaluation, 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, errors in differentiating emotions on the valence axis, between emotions e1 and e2, and between e3 and e4 are significantly more frequent.

The most important features for the detection of emotions on the first level (L1), selected from a rich set of features, presented in Chap. 4 Sect. 4.3, were:

- Size of AkD_S at every eighth (H),
- Median AkD_S at every eighth (H),
- Size of AkD_S at every new chord in a segment (H),
- Second Max AkD_B at every eighth (HR),
- Numerical Integration of AkM_B (HR),
- Median AkI_B (HR),
- First Strongest Rhythmic Pulse—SRP1 (R),

where H, HR, and R represent feature group abbreviations: harmony (H), harmony-rhythm (HR), and rhythm (R).

The selected features are mainly harmony and harmony-rhythm features. They present the statistics collected from a sequence of values as Chord Degree of Dissonance (AkD). Harmony-rhythm features pertain to statistics collected from sequences of parameters AkD, AkM, AkI, which describe the frequency ratios in chords, collected from musical segments at moments of the Strongest Pulses. They confirm the usefulness of the created features. In the selected features, we have representatives of all groups except for dynamic features, which in most likelihood were covered by other features. The First Strongest Rhythmic Pulse (the beat with the highest magnitude in the beat histogram) is also an important feature, which is a logical explanation of the effect of rhythm on the detected emotion.

Table 5.5 Accuracy obtained for High-Low Arousal and High-Low Valence classifiers using J48

Classifier	High-Low	High-Low
Accuracy (%)	**97.42**	82.57

5.5.1.2 Two Classifiers Recognizing Emotions on the Arousal and Valence Axes

To find which of the MIDI features are better suited for emotion differentiation on the arousal and valence axes from Russell's emotion model, we build two additional classifiers. The first one (High-Low Arousal) had the task of differentiating emotions from the top part of the semicircle of the model (e1, e2) from emotions from the bottom part of the semicircle (e3, e4). The second one (High-Low Valence) had to differentiate emotions from the left part of the semicircle (e2, e3) from emotions on the right part of the semicircle (e1, e4).

The classifiers were built using the J48 algorithm, which was the winner during building one classifier for the four emotions of L1. The classification results were calculated using a cross validation evaluation CV-10. For attribute selection we used attribute Wrapper Subset Evaluator and search method Best First. The accuracy obtained for the J48 algorithm after using attribute selection is presented in Table 5.5.

The high accuracy for the High-Low Arousal classifier (97.42%) indicates that the classifier, using the collected MIDI features, differentiates emotions in the top half from the bottom half of the model well. A slightly worse result (82.57%) was obtained for the High-Low Valence classifier, which confirms that recognizing emotions on the valence axis is more difficult than on the arousal axis.

Table 5.6 presents the most important features for detecting emotions on the arousal and valence axes. Both sets have selected rhythm features (R) describing the number of strong pulses (Relatively Strong Pulses) obtained from the beat histogram. We can also see the usefulness of features pertaining to the duration of notes (Note Duration) or number of notes per second (Note Density). It is interesting that in the feature set for the detection of High-Low Valence we only have harmony-rhythm features (HR) and not harmony features (H). We found that statistics from harmony features collected from musical segments at moments of the Strongest Pulses are more useful for detecting High-Low Valence than harmony features collected at every eighth (H). During High-Low Arousal detection, most harmony features (H) collected at every eighth note were enough.

Table 5.6 Selected features used for building High-Low Arousal and High-Low Valence classifiers

Classifier	Selected features
High-Low Arousal	Size of AkD_S at every eighth (H)
	First AkD_S at every eighth (H)
	Second AkD_S Percentage at every eighth (H)
	Third AkD_S at every eighth (H)
	Second AkD_B Percentage (HR)
	Relatively Strong Pulses 10 (RSP10) (R)
	Average Note Duration (R)
	Standard Deviation of Note Duration (R)
High-Low Valence	Numerical Integration of AkD_B (HR)
	First AkD_B Percentage (HR)
	Average of First 3 Max AkD_B (HR)
	Median AkM_B (HR)
	Average AkI_B (HR)
	Relatively Strong Pulses 30 (RSP30) (R)
	Note Density (R)

5.5.2 Second Level Classifiers

5.5.2.1 One Classifier Recognizing Twelve Emotions

While building the second level classifiers, we first decided to build a classifier that would differentiate 12 sub-emotions. We built classifiers for emotion detection using the following algorithms: J48, BayesNet, K-nn, SMO. The classification results were calculated using a cross validation evaluation CV-10.

Once again, we obtained the best accuracy for the studied algorithms with J48, with a value of 65.14% (Table 5.7); although the accuracy was lower by 17 % points than the accuracy of the classifier detecting 4 emotions on the first level (L1). In the case of detecting 12 sub-emotions, the classifier is clearly less accurate, which is connected with the greater number of classes with the same number of features. Also, the quality of data labeling by the experts at level L2 is generally lower than for level L1, which may also lower the results.

Table 5.7 Accuracy of classifiers obtained for second level (L2) classifiers

Classifier(%)	J48	BayesNet	K-nn	SMO
Accuracy before attribute selection	**60.00**	56.00	54.28	52.28
Accuracy after attribute selection	**65.14**	58.00	63.55	60.00

Table 5.8 Four classifiers for the second level (L2)

Name of classifier	Detected emotions of second level
CL21	e11, e12, e13
CL22	e21, e22, e23
CL23	e31, e32, e33
CL24	e41, e42, e43

Table 5.9 Accuracy of classifiers obtained for 4 second level (L2) classifiers

Name of classifier(%)	CL21	CL22	CL23	CL24
Before attribute selection	76.40	74.42	70.88	80.51
After attribute selection	84.76	84.76	84.81	**93.50**

5.5.2.2 Four Classifiers Recognizing Sub-emotions

In order to improve emotion detection accuracy on the second level (L2), we decided to build 4 classifiers, one for each quarter of Russell's emotion model (Table 5.8). Each of the classifiers specializes in detecting 3 sub-emotions for the respective quarter, CL21—detects emotions in the first quarter, CL22 in the second, CL23 in the third, and CL24 in the fourth.

To create the specific classifiers for level L2, we used samples from a given category. In other words, to build classifier CL21 for emotions e11, e12, and e12 we only used sampled labeled as e1 at level L1. We did the same for the remaining classifiers, CL22, CL23, and CL24.

Thus, we obtained 4 classifiers specializing in detecting sub-emotions at level L2. To build the classifiers, we used algorithm J48, which was the winner when we built classifiers at levels L1 and L2. The obtained accuracy for each classifier before and after attribute selection is presented in Table 5.9.

Notice the clear improvement in accuracy (84.74–93.50%) compared with the accuracy obtained for the classifier detecting 12 emotions at level L2 (65.14%). These results confirm the usefulness of the classifiers and that 4 classifiers specializing in detecting 3 sub-emotions for each quarter of the model detect emotions better than one classifier detecting all 12 emotions.

5.6 Hierarchical Classification

Level L1 and L2 classifiers were used for hierarchical emotion detection in music files (Fig. 5.2). Emotion detection in music files was done analogous to the used hierarchical model of emotions with categories on two levels L1 and L2.

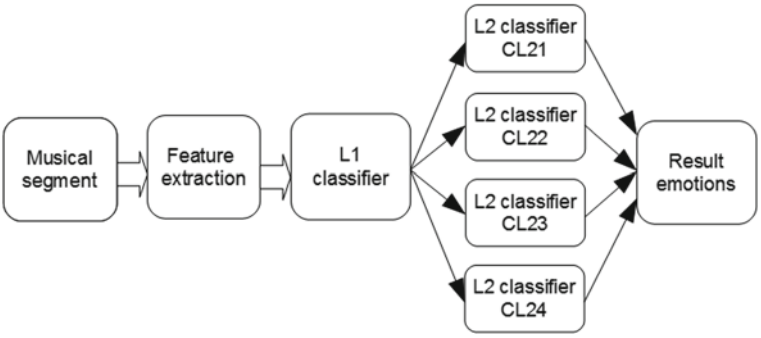

Fig. 5.2 Hierarchical emotion detection in a musical segment

First, a musical segment underwent feature extraction; then, the obtained features vector representing the musical segment was classified at level L1. One classifier (Sect. 5.5.1.1) detecting 4 emotions—e1—happy, e2—angry, e3—sad, e3—relaxed—was used. Next, depending on the results of the first classifier, the appropriate level L2 classifier was selected (Sect. 5.5.2.2); its task was to detect 3 sub-emotions. For example, if at the first level the detected emotion was e1, then at the second level a classifier detecting sub-emotions for e1, i.e. classifier CL21, was used. If the result of classification at level L1 was emotion e2, then at level L2 we used classifier CL22. And so on for the remaining cases. The result of hierarchical classification of musical segments was the detection of emotions on two levels, L1 and L2.

5.7 Emotion Tracking in MIDI Files

5.7.1 System Construction

The proposed system for tracking emotions in a musical composition is shown in Fig. 5.3. It consists of a database of musical compositions, composition segmentation, hierarchical emotion detection, and the result presentation module. The resulting emotion labels were used to designate the consecutive segments of a musical composition. The collected data allowed for the analysis of a musical composition in terms of the emotions contained therein.

When using the system, the user first selects a musical composition from the database, then cross-indexes it for emotion. Finally, an analysis and visualization of the obtained results are conducted.

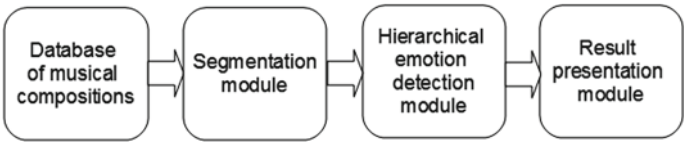

Fig. 5.3 Construction of the emotion tracking system

5.7.2 Musical Composition Segmentation

Emotions in musical compositions are not constant. In a fragment lasting several seconds, there may be just one emotion or it may change many times. The emotion can change in very different ways over the course of a musical composition lasting only a few minutes depending on the musical content of the piece. Emotion reflects what is happening in the musical composition, for example, if the pace of the composition increases, the emotion changes in the direction of the upper quadrants (e1, e2) of Russell's model. If the sounds of the piece begin to be less consonant, the expressed emotions come from the left lateral quadrants (e2, e3) of Russell's model.

Some pieces may have many emotional changes (e.g. musical compositions of varying moods or affecting the listener with a whole range of musical means such as different pace, variable rhythm, dynamics, etc.), while others may be based on one unchanging emotion (e.g. musical compositions of uniform structure with a steady pace, dynamics, and rhythm).

Detection of emotion was conducted in our research on 6-second segments, with each consecutive segment shifted by 2 s; thus, successive segments overlapped at a 2/3 ratio (Fig. 5.4). This allowed exactly tracking and detecting even the slightest

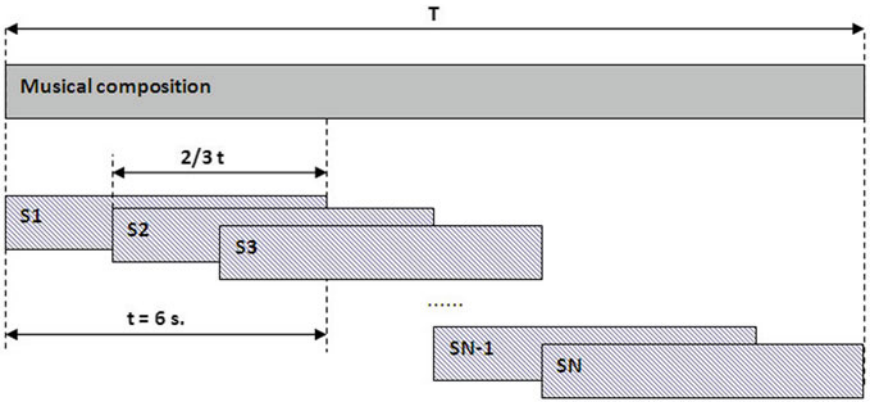

Fig. 5.4 Division of a musical piece into segments

change of emotion in the examined musical composition. For a musical composition lasting $T = 120\,\text{s}$, $N = 60$ segments ($S1$, $S2$, ..., $S59$, $S60$) were analyzed, and for each L1 and L2 level of emotion detection was performed.

5.8 Results of Emotion Tracking in MIDI Files

The result of tracking emotions of a musical composition are segments with emotions described on two levels: the higher, more general, L1; and the lower, more detailed, L2. Analysis of the obtained emotions confirms the assumption that emotions are not uniform in a musical piece.

5.8.1 Emotion Histograms of Musical Compositions

Emotions can change throughout a musical composition. Some emotions are more common than others and their type is not always the same. The first method used for presenting the distribution of emotions in a musical composition is emotion histograms. Figures 5.5 and 5.6 present the emotion histograms of two compositions: Ludwig van Beethoven's Sonata No. 23 F minor, Opus 57, part 1 (Appassionata), and Frédéric Chopin's Prelude in C minor Op.28, No. 20. On the presented graphs, the horizontal axis corresponds to the type of emotion, and the height of the bar indicates how often a specific emotion occurred.

Figure 5.5a presents the histogram of L1 level emotions in Beethoven's Appassionata sonata, in which emotion e2 (angry) occurs in 76% of the segments and is dominant. The second, most significant, emotion is e1 (happy), which occurs in 20%

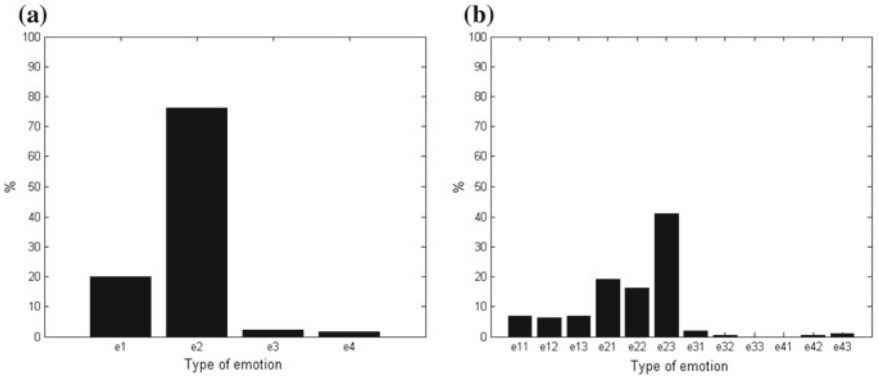

Fig. 5.5 Histogram of L1 (**a**) and L2 (**b**) level emotions in L.v. Beethoven's Sonata No. 23 F minor, Opus 57 (Appassionata), part 1

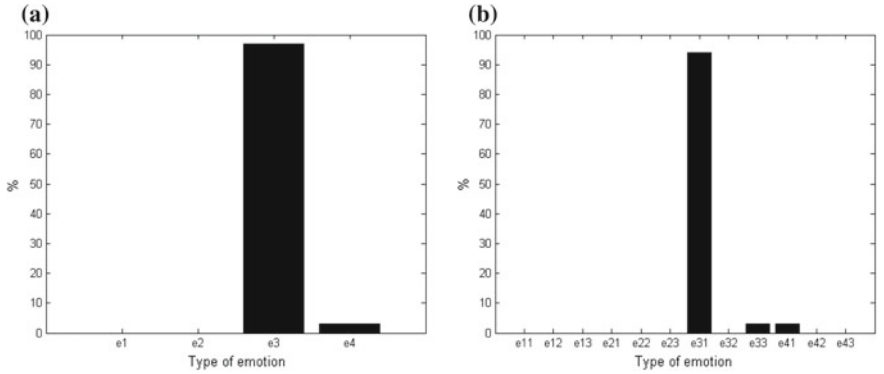

Fig. 5.6 Histogram of L1 (**a**) and L2 (**b**) level emotions in F. Chopin's Prelude in C minor Op.28, No. 20

of the segments. We notice that the occurrence of emotions e4 (relaxed) and e3 (sad) is very rare in the given composition and occur in about 2% of the segments each.

Figure 5.5b presents the histogram of L2 level emotions in Beethoven's Appassionata sonata. Analyzing it and comparing it with the L1 level histogram (Fig. 5.5a), we can see in detail how an emotion from level L1 (e2) breaks down into emotions from level L2: e21 (annoying), e22 (angry), and e23 (nervous). Notice the domination of emotion e23 (40%). The sub-emotions of the first quarter of Russell's model (e1) are e11 (pleased), e12 (happy), and e13 (excited), and they occur in about 7% of segments each.

The contrast to the presented histograms of Beethoven's Appassionata (Fig. 5.5) is the histograms of F. Chopin's Prelude in C minor Op.28, No. 20 (Fig. 5.6). Not only is there a different main emotion, its domination is much greater. We can notice a great domination of the main emotion from level L1, e3 (97%), and the domination of emotion e31 (94%) from level L2. The occurrence of other emotions is marginal.

Emotional diversity is much richer in Appassionata; there are two main emotions, e1 and e2, from level L1, with various shades of emotions at level L2. In Prelude, we have one dominating emotion, e3, from level L1 and one, e31, from level L2. In other words, there is a lack of diversity in shades of emotions.

5.8.2 Emotion Maps

Another method used to analyze emotions in a musical composition is detailed maps showing the distribution of emotions for the duration of a piece (Figs. 5.7, 5.8, 5.9 and 5.10). The horizontal axis shows the time in seconds and the vertical axis the emotions occurring at a given moment.

In Fig. 5.7, presenting a map of L1 level emotions for Beethoven's Appassionata, we notice that e2 is dominant throughout the entire piece. From the map, we can see

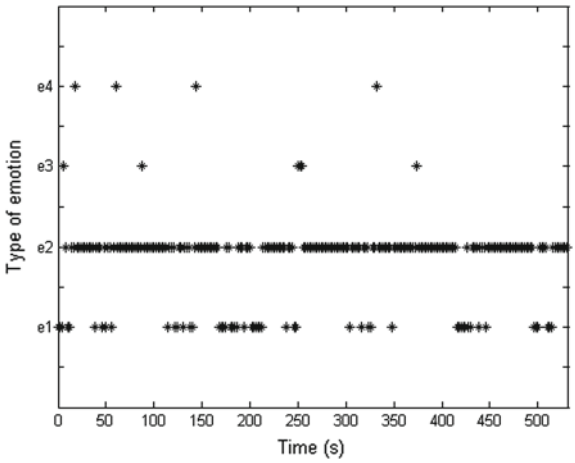

Fig. 5.7 Map of L1 level emotions in L.v. Beethoven's Sonata No. 23 F minor, Opus 57 (Appassionata), part 1

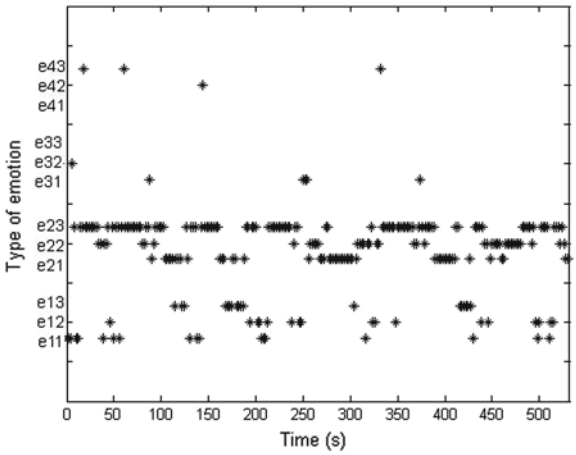

Fig. 5.8 Map of L2 level emotions in L.v. Beethoven's Sonata No. 23 F minor, Opus 57 (Appassionata), part 1

when the second most frequent emotion, e1, occurs. The occurrence of emotions e3 and e4 is incidental.

By analyzing the map of L2 level emotions for L.v.Beethoven's Appassionata (Fig. 5.8), we can notice the detailed distribution of emotions. The set of occurring emotions is quite rich. One could make an attempt to find patterns in the presented map, for example, we noticed the subsequent occurrence of emotions e23, e22, e21 forming falling 'stairs' in several moments of the piece: s. 70–100, s. 240–260, s. 360–390.

The contrast to the presented maps of Beethoven's Appassionata is the maps of F. Chopin's Prelude in C minor Op.28, No. 20 (Figs. 5.9 and 5.10). The dominating emotion, e3, throughout the entire composition from level L1 is presented in the form of a horizontal line. A short change in emotions to e4 occurs only in the 44th second (Fig. 5.9). A similar horizontal line for e31 occurs at level L2 (Fig. 5.10).

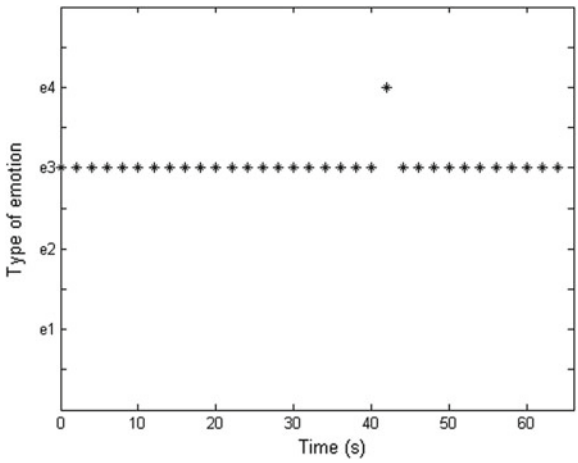

Fig. 5.9 Map of L1 level emotions in F. Chopin's Prelude in C minor Op.28, No. 20

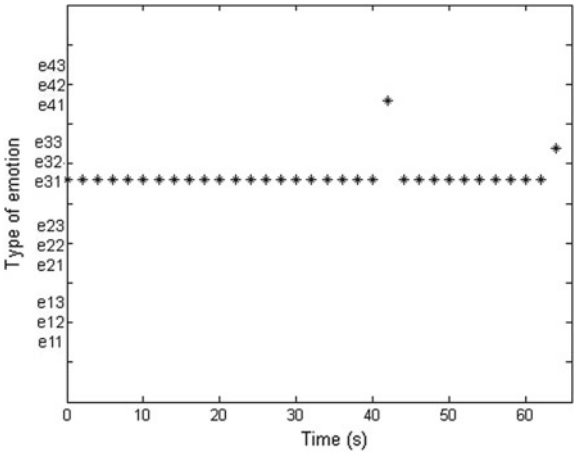

Fig. 5.10 Map of L2 level emotions in F. Chopin's Prelude in C minor Op.28, No. 20

Comparing the maps of the two compositions, we can see how much they vary in their distribution of emotions over time.

5.8.3 Quantity of Changes of Emotion

Because some pieces may have many emotional changes (e.g., songs with varying moods), while others may be based on a single dominant emotion (e.g. musical compositions with a steady pace, dynamics and rhythm, etc.), we introduced the Quantity of Changes of Emotion (QCE) in a musical composition, which is the sum of the number of changes of emotion in adjacent segments. To make the indicator

Table 5.10 The dominant emotion and the Quantity of Changes of Emotion (QCE) in a piece

Piece	QCE L1	QCE L2	Dominating emotion in L1 (percentage)	Dominating emotion in L2 (percentage)
Appassionata, part 1	24.34	39.33	e2 (76%)	e23 (40%)
Prelude Op.28, No. 20	5.88	8.82	e3 (97%)	e31 (94%)

values independent from the length of the piece, the obtained sum was divided by the number of N segments.

$$QCE = \frac{\sum_{i=1}^{N-1} f(i)}{N} * 100 \tag{5.1}$$

$$f(i) = \begin{cases} 1, & \text{if } Emotion(i) \neq Emotion(i+1) \\ 0, & \text{if } Emotion(i) = Emotion(i+1) \end{cases} \tag{5.2}$$

where i is the number of the segment in the piece, N the number of segments in the composition, and $Emotion(i)$ represents the emotion of the i segment. The function $f(i)$ indicates whether the adjacent segments have a different (value 1) or same (value 0) emotion. The more changes of emotion in a musical composition, the greater the QCE value.

Table 5.10 presents the obtained values for the quantity of changes of emotion and the dominating emotions for the presented two compositions. We can notice that the dominant emotion percentages are much higher in the Prelude than in the Appassionata. Also, the QCE in the Prelude has smaller values than the Appassionata at both levels L1 and L2. From the obtained results, we can conclude that the Prelude is more emotionally homogeneous with a greater dominance of individual emotions and Beethoven's Appassionata is more diverse emotionally.

The method of creating emotions in a musical piece depends on the composer. These emotions can be presented in the forms of histograms, maps, or using such parameters as QCE. A search for parameters describing the emotional distributions in compositions could be an interesting continuation of this work in the future.

5.9 Conclusions

In this chapter we presented emotion detection in pieces of classical music in the form of MIDI files. A hierarchical model of emotions consisting of two levels, L1 and L2, was used. A collection of harmony and rhythm MIDI features extracted from music files allowed for emotion detection with an average of 82% accuracy at level L1. The built classifiers detecting emotions at level L2, i.e. the sub-emotions of level L1, achieved accuracy between 84 and 93%. They were built so that they specialize

in detecting emotions from a selected group of sub-emotions, which improved their effectiveness.

During feature selection, we found the most useful MIDI features to build a classifier recognizing four emotions on the first level. We also found the most important features to distinguish emotions on the arousal and valence axes from Russell's emotion model; harmony-rhythm MIDI features proved to be particularly useful for emotion detection on the valence axis.

A strategy for the analysis of emotions contained within MIDI musical compositions was presented. We constructed the system for tracking changing emotions during the course of a musical piece, and the collected data allowed determining the dominant emotion in the musical compositions, presenting emotion histograms, and constructing maps visualizing the distribution of emotions in time.

The amount of changes of emotions during a piece may be different; therefore, we introduced a parameter evaluating the quantity of changes of emotions in a musical composition. The information obtained about an emotion in a piece made it possible to analyze the musical compositions, thus providing new knowledge about the compositions and the method of their emotional development.