Chapter 3
Human Annotation

3.1 Introduction

Human annotation of music files is done in order to collect ground truth data, which will then be used to build an automatic emotion detection model. Annotation methods can be divided into two categories: expert-based and subject-based. Expert-based methods take advantage of music experts, musicians, people who work with music every day and most often play a musical instrument professionally. Music experts were used to determine music files in [39, 63, 104, 113]. Subject-based methods use people not connected professionally to music, non-musicians, and was employed in [2, 44, 51, 52, 64, 77, 105]. Subject-based methods also include a method of emotion tag collection directly from music websites such as Last.fm. This method was used in [54, 89, 116]. Due to the fact that survey respondents' answers can be somewhat subjective, the responses are often averaged.

In this paper, we used the expert-based method; we surveyed the opinions of five music experts, who are professionally involved with music every day, on the collected music samples. Each annotator annotated all music excerpts in the data set.

3.2 Length of a Musical Segment

An important element before carrying out annotation of music samples is to decide on the appropriate length of a musical fragment. The selected segment of a specific length will then undergo human annotation and audio features will be extracted. So what length should the indexed samples be or what is the shortest segment length necessary for an expert to be able to identify the emotion of a given fragment?

Many papers on automatic emotion detection have assumed a segment length of 20–30 s [44, 57, 63, 80, 104, 113]. These papers pertain to one static music emotion recognition, which assumes the emotion in a given segment does not change. Papers focusing on analyzing changes in emotions over time often use segments that are 1 s long [4, 51, 96, 119].

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_3

Selecting the appropriate segment length for emotion detection is quite important. On the one hand, a segment of a selected length is used during annotation by listeners—experts. On the other hand, it is used to extract audio features by a computer and to build an automatic emotion detection system. If to build training samples, we used a given length (i.e. a length of 6 s) of a musical segment, and on their basis we built emotion prediction models, prediction of new samples should also be of the same length, in our case 6 s.

The shorter a segment, the more detailed analysis of emotions is possible. Also, the shorter a segment, the more homogeneous the emotional content of the segment. On the other hand, a musical segment should not be too short during annotation since this will prevent a listener—expert from precisely identifying the emotion. Humans need time to determine the perceived emotion in music.

Bachorik et al. [5] investigated the length of time required for participants to initiate emotional responses to musical samples from a variety of genres by monitoring their real-time continuous ratings of emotional content and the arousal level of the music excerpts. On average, participants required 8 s of music before initiating emotional judgments.

Use of 6 s by Pampalk et al. [72] proved to be enough to build a system for content-based organization and visualization of music archives. From selected pieces of music in raw audio format, a geographic map was created where islands represented musical genres or styles.

Use of a segment shorter than 6 s hinders emotion detection by a listener [76, 111]. It enables differentiating two basic emotions, if a given fragment is happy or sad, but it prevents recognizing shades of emotions.

A segment with a duration of 6 s was used by Macdorman et al. [64] for automatic emotion prediction of song excerpts. At first 30-second segments were used, but due to the fact that pleasure and arousal typically change with musical progression, the segments were shortened to 6 s. The authors analyzed how pleasure and arousal ratings relate to the pitch, rhythm, and loudness of the song excerpts.

Xiao et al. [115] investigated the best segment duration for music mood analysis. Four versions of music data sets with a duration of clips of 4, 8, 16 and 32 s were tested. The results indicate that analyses of emotions in music should be based on shorter segments, no longer than 16 s, and the best performance was achieved by a segment length of 8 and 16 s.

In the report of the Emotion in Music task organized within the MediaEval benchmarking campaign, Aljanaki et al. [4] noticed problems with too short excerpts. The very short length of the annotated segments (0.5–1 s) allowed only capturing changes in dynamics and timbre. Simultaneously, it caused difficulty with capturing features pertaining to harmony and melody, which occur on a larger time scale.

In our experiment, the samples undergoing annotation were 6 s, which is the shortest possible length, determined experimentally, at which experts with a university music education could detect emotions for a given segment. A short segment

Table 3.1 Amount of examples of different genres of music

Genre	Amount of examples
Classical	67
Jazz	42
Blues	26
Country	50
Disco	27
Hip-hop	15
Metal	18
Pop	21
Reggae	22
Rock	36
All	324

ensures that the emotional homogeneity of a segment is much more probable. During the annotation, the experts sometimes provided their replies before 6 s were up, which suggests that a trained expert is able to identify an emotion before the end of 6 s.

3.3 Audio Music Data

3.3.1 Data Set

The data set that was annotated by the music experts 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 training data were taken from the generally accessible data collection project MARSYAS.[1] The author selected samples and shortened them to the first 6 s, and as a result the data set consisted of 324 samples.

The amount of examples of different genres of music is presented in Table 3.1; the list of samples used in our experiments can be found on the web.[2]

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

[2] http://aragorn.pb.bialystok.pl/~grekowj/HomePage/EmoDataSet.

3.3.2 Music Experts

Data annotation was done by five music experts with a university music education[3] (Expert 1–5), which included the author of this book. The experts, aged 28–48 years, are active musicians, playing on one of the following instruments: piano, clarinet, percussion, accordion, double bass. With their many years of experience, they are tied with various styles of music, such as: classical, jazz, blues, pop, rock, disco, punk.

3.3.3 Annotation Process

Before the annotation process, the music experts were introduced to Russell's model, with quarters corresponding to four basic emotions on the arousal and valence axes. Next, a 15-minute training was conducted during which the annotators listened to composition fragments and marked values on the arousal and valence axes. The meanings of arousal and valence were explained and differences between perceived emotion and felt emotion were discussed. In our experiment, the music experts' task was to identify perceived emotions only.

After this training and after teaching the experts about the applied terminology, they began annotating the musical segments. The annotation was carried out using a web application with a database specifically created for this purpose by the author of this book (Fig. 3.1). The application was built using Java Enterprise Edition, Java Server Faces, Server Glassfish, and the MySQL database. The web application enables access to the formulas indexing music files through a web browser, and the collected annotations were saved in the database.

Each annotator annotated all records in the data set, which had a positive effect on the quality of the received data [4]. The process was synchronized, as a sample was played for the whole group of experts simultaneously. After each sample was played, the experts had several seconds to make a decision, i.e. select values on the arousal and valence axes, and then the next composition was played.

The annotation process was repeated after the first round of annotation of all compositions was completed. The second annotation round enabled the experts to check and correct their responses. During the first round, it can be assumed that the music experts weren't completely trained in identifying valence and arousal values, while during the second round, they had the possibility of correcting their initial responses. The corrections weren't major, but nevertheless they occurred. Repeating the annotation process was beneficial because the experts were able to verify their responses; they had a second chance to make a decision and the possibility to change their response. During the second round of annotation, corrections mainly involved the first several samples from the first round. The closer to the end of the second

[3]We would like to thank the following music experts for indexing music files: Wojciech Bronakowski, Mateusz Bielski, Wojciech Mickiewicz, Jan Mlejnek.

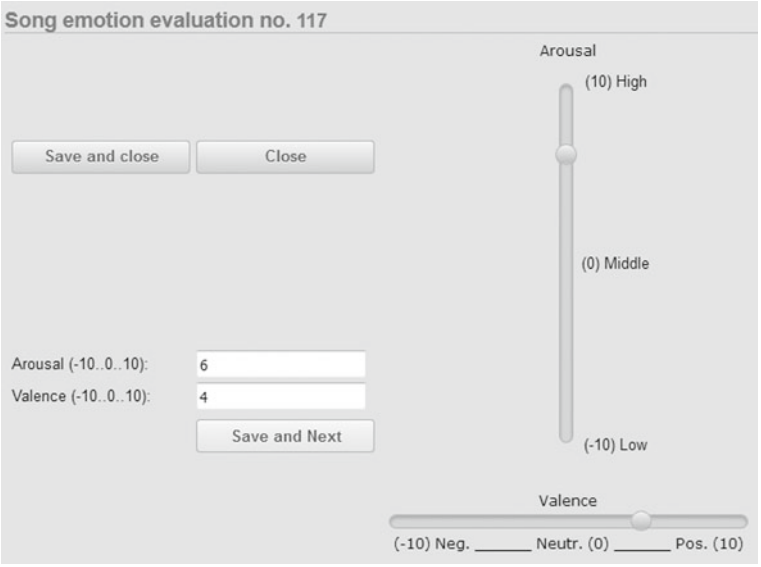

Fig. 3.1 View of the web application for labeling—survey for emotion labeling

round, the less corrections were being made. This indicates that the initial responses were not always correct, but the further into the indexing process the more correct the responses, which then did not need correction. There was a 30-minute break between the first and the second rounds of annotation. Despite the inconvenience, all annotators agreed with the validity of repeating the indexing process, since they were able to clarify their responses. Aljanaki et al. [4] also noted lower indexing quality at the beginning of their experiment; therefore, repeating the annotation process has a positive effect on the quality of the obtained data.

3.3.4 Results

As a result of the annotation by five music experts (Expert 1–5), we obtained data describing 324 fragments. Each fragment received five opinions, which is a total of 1620 of all annotations.

Figure 3.2 presents the 324 response, illustrated on the Arousal-Valence plane, from Expert 1. Each point represents one music excerpt, and its location on the plane is described by the arousal and valence values provided by Expert 1 for a given composition. As can be seen from the graph, responses can be found in all quarters of the A–V emotion model.

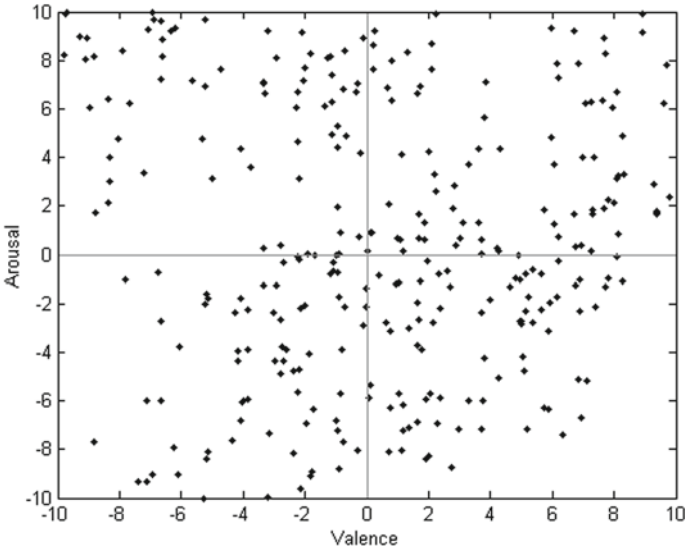

Fig. 3.2 Responses on the A–V emotion plane from Expert 1

Figure 3.3 presents the 324 responses from Experts 2–5. In all cases, there are responses in all quarters of the A–V emotion model, although the spread of point locations differ slightly.

Data collected from the five music experts were averaged; Fig. 3.4 presents the averaged responses for 324 compositions. Each point, with coordinates comprised of averaged values of arousal and valence, represents one music excerpt.

To check if in our music data valence and arousal dimensions are correlated, the Pearson correlation coefficient was calculated [13]. The obtained value $r = -0.03$ indicates that arousal and valence values are not correlated, and the music data are a good spread in the quarters on the A–V emotion plane. This is an important element according to the conclusions formulated by Aljanaki et al. in [4].

The values provided by the experts on the arousal and valence axes were in the range $[-10, 10]$. The mean of all collected values for arousal was: -0.16, and the mean for valence: 0.11. Both values are close to zero, which indicates a good distribution of samples on both sides of the valence and arousal axes. The mean standard deviation of the obtained responses from five experts on the arousal axis was 1.63, and on the valence axis 1.46. Both values are of a similar order and constitute about 8% of the entire range of values on the axes.

Considering the internal consistency of the collected data, Cronbachs α [15] for arousal and valence achieved high values of 0.94 and 0.86, respectively. From these values we can conclude that the agreement of the experts' opinions was greater for labeling arousal values than valence. Valence is usually more difficult to recognize, and here the experts' replies differed slightly more than in the case of arousal.

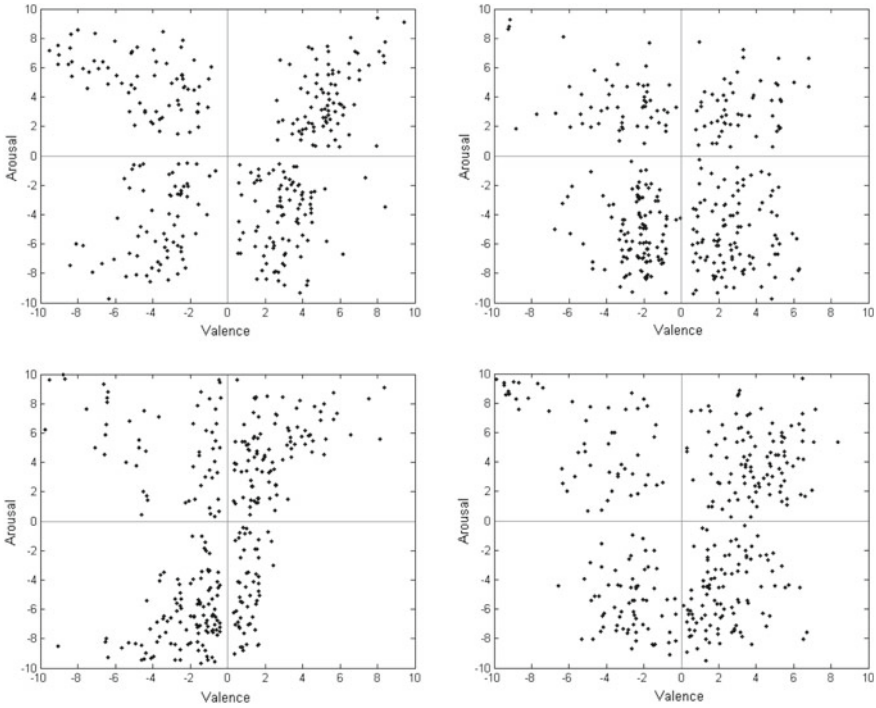

Fig. 3.3 Responses on the A–V emotion plane from Experts 2–5

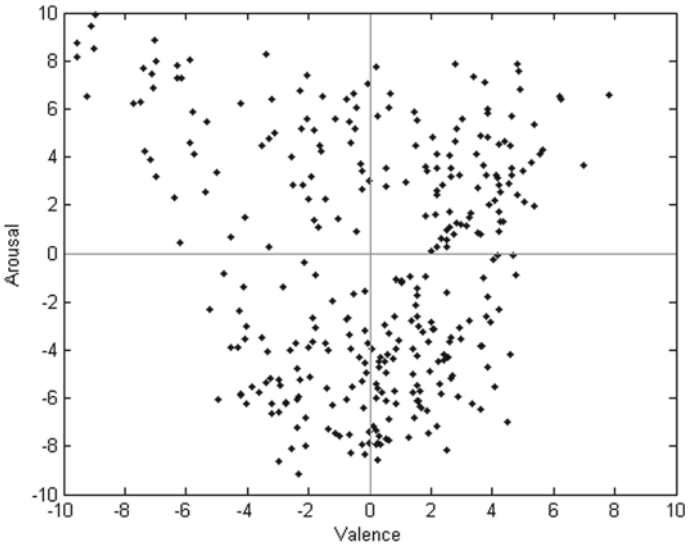

Fig. 3.4 Averaged values of responses from five music experts on the A–V emotion plane

Table 3.2 Amount of examples in quarters on the A–V emotion plane

Quarter abbreviation	Arousal-Valence	Basic emotion	Emotion label	Amount of examples
Q1	High-High	happy	e1	93
Q2	High-Low	angry	e2	70
Q3	Low-Low	sad	e3	80
Q4	Low-High	relaxed	e4	81

The amount of examples in the quarters on the A–V emotion plane is presented in Table 3.2. The arousal and valence values identify belonging to a given quarter of the model and emotion class simultaneously.

3.4 MIDI Music Data

3.4.1 Data Set

In this work, emotion detection experiments were conducted on audio files as well as MIDI files. For emotion detection experiments in MIDI files, we prepared a separate database with 83 compositions of classical music, which contains compositions by such eminent composers as:

- Franz Schubert (1797–1828),
- Ludwig van Beethoven (1770–1827),
- Felix Mendelssohn Bartholdy (1809–1847),
- Frédéric Chopin (1810–1849),
- Robert Schumann (1810–1856),
- Edvard Grieg (1843–1907),
- Isaac Albniz (1860–1909).

All compositions were piano-based; this way we rejected the aspect of studying the effect of various instruments on the perceived emotions. From the collected compositions, we extracted 350 six-second segments, an average of four fragments from each composition, which differed in tempo, volume, complexity, harmony, and dynamics.

The 350 music excerpts were annotated by five music experts with a university music education, people who have professional experience in playing and listening to music. To label emotions in MIDI files, we used a hierarchical model of emotions.

3.4.2 Hierarchical Emotion Model

The model we chose is based on Russell's circumplex model (Fig. 3.5). Following the example of this model, we created a hierarchical model of emotions consisting of two levels, L1 and L2.

The first level contains four categories of emotions that correspond to the four quarters of Russell's model (Table 3.3). In the first group (e1), pieces of music can be found that convey positive emotions and have a quite rapid tempo, are happy and arousing (excited, happy, pleased). In the second group (e2), the tempo of the pieces is fast, but the emotions are more negative, expressing annoying, angry, and nervous. In the third group (e3) are pieces that have a negative valence and low arousal, expressing sad, bored, and sleepy. In the last group (e4) are pieces that have low arousal and positive valence and express calm, peaceful, and relaxed.

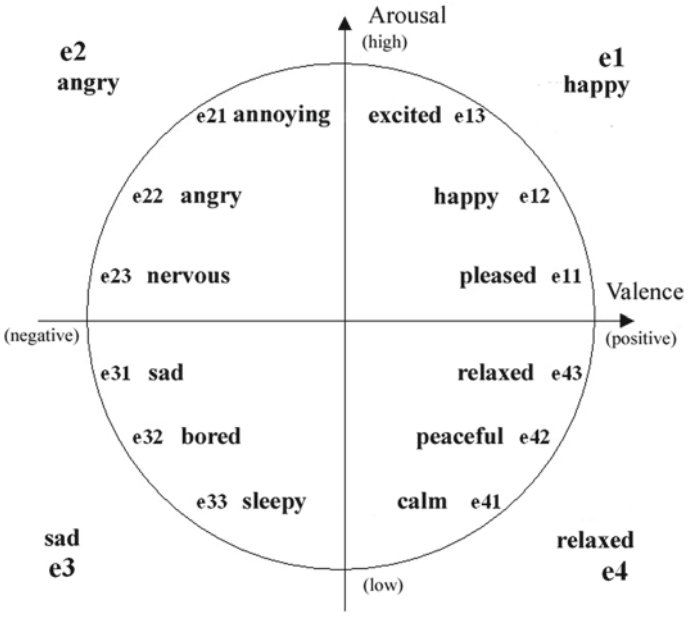

Fig. 3.5 Hierarchical model of emotions based on Russell's circumplex model

Table 3.3 Description of emotion categories in L1, the first level

Abbreviation	Description	Arousal-Valence
e1	happy	High-High
e2	angry	High-Low
e3	sad	Low-Low
e4	relaxed	Low-High

Table 3.4 Description of emotion categories in L2, the second level

Abbreviation	Description	Abbreviation	Description
e11	pleased	e31	sad
e12	happy	e32	bored
e13	excited	e33	sleepy
e21	annoying	e41	calm
e22	angry	e42	peaceful
e23	nervous	e43	relaxed

The second level is related to the first, and is made up of twelve sub-emotions, three emotions for each emotion contained in the first level (Table 3.4). In our hierarchical model of emotions, we have four categories in level L1 and twelve categories in level L2. The emotion categories in L1 are a generalization of the more detailed emotions in L2. Category names in L1 are also found in L2, for example, the entire group e1 in level L1 has been described by the adjective happy, and it includes the emotions excited, happy, and pleased in level L2.

3.4.3 Annotation Process

Six-second music samples were listened to and then labeled with one of the emotions of the second level (L2). Labeling with an emotion from the second level (L2) automatically indicated the parent emotion from the first level (L1). For example, if an expert selected emotion e13 (excited) from level L2, this meant they also selected emotion e1 (happy) from level L1. Thus, the samples were labeled with emotions from two levels of the hierarchical model. The short 6-second length of each segment ensured that the studied music fragments were relatively homogeneous emotionally, which allowed labeling a segment with one emotion. Each annotator annotated all records in the data set.

3.4.4 Results

The data collected from the five music experts were averaged by selecting an emotion that occurred the most often among the experts' responses. The amount of obtained examples labeled by emotions on the first level is presented in Table 3.5, and those labeled by emotions on the second level are presented in Table 3.6.

Considering the internal consistency of the collected data, Cronbachs α [15] obtained a value of 0.90 for data in level L1. Cronbachs α for collected data in level L2 obtained a value of 0.88, which means it was lower than for level L1. This

Table 3.5 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

Table 3.6 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

is logical since we have more categories at level L2 as well as more differences in experts' opinions.

3.5 Summary

In this chapter, we presented two music data sets, audio and MIDI, that underwent annotation by music experts. We used specifically written web applications to collect data, which facilitated indexing musical compositions by many experts simultaneously. Each annotator annotated all records in the data set, and data collected from the music experts were averaged. The obtained music data are a good spread in the quarters of Russell's emotion plane.

The collected labeled audio music excerpts will serve as ground truth data during automatic emotion detection using the categorical (Chaps. 7 and 8) and dimensional (Chaps. 9 and 10) approaches. In the dimensional approach, we will use the collected arousal and valence values. In the categorical approach, we will use the four emotion classes corresponding to the four quarters of Russell's model: happy, angry, sad, and relaxed. The audio music excerpts labeled by four emotion classes will also be used

for initial assessment of the usefulness of the designed audio features in Chap. 6 Sect. 6.3. The collected labeled music MIDI excerpts will serve as ground truth data during categorical hierarchical emotion detection in MIDI files in Chap. 5 as well as for initial assessment of the usefulness of the designed MIDI features in Chap. 4 Sect. 4.3.