Application of Free Choice Profiling for the Evaluation of Emotions Elicited by Music

Judith Liebetrau[1,2], Sebastian Schneider[1] and Roman Jezierski[1]

[1] Ilmenau University of Technology, Ilmenau, Germany
[2] Fraunhofer IDMT, Ilmenau, Germany
Judith.Liebetrau@tu-ilmenau.de

Abstract Music evokes and carries emotions. Despite many studies having investigated the relation between music and emotion, current research lacks a systematic and empirically derived taxonomy of musically induced emotions [1]. This work contributes to the question which musical features in particular are able to induce emotions while listening. Problems of defining and measuring emotions are explained. A method to measure affective states induced by music with the help of Free Choice Profiling (FCP) is outlined. Two FCP experiments, assessing the usefulness of the method for emotional research and the selection of test stimuli are described. The shown results are in line with psychological theories of emotions, i.e., the valence/arousal model.

Keywords: Free Choice Profiling, Measuring Emotions, Self-report.

1 Introduction

A diverse range of studies was carried out in the past to investigate how and in which way music influences the emotions of the listener, but still two main questions remain: What exactly is an emotion and how can it be measured? This paper contributes to the question which musical features in particular are able to induce emotions while listening; the research was conducted within a project funded by the German Research Foundation.

A broad overview of common measurement methods can be found in [2]. Although the term emotion is frequently used in literature, authors disagree on its definition, and a simple definition cannot be given. Scherer [3], for example, defines an emotion as an affective phenomenon, distinguishable from feelings, moods, or attitudes. Emotions, resulting from cognitive processes, are necessary for comprehension and appraisal of stimuli on the basis of knowledge. Seeing emotion as a phenomenon consisting of five components (cognitive, neurophysiological, motivational, motor expression, and subjective feeling), Scherer concludes that a universal measure would only become possible by taking into account changes of all components.

Due to the lack of an all-embracing measurement method, each component is measured on its own. The subjective experience of emotions can be assessed in different ways. One possibility is the measurement of changes in psychophysiological parameters like heart rate, heart rate variability and skin conductance during music perception. Numerous studies about the measurement of such physiological correlates

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

of emotions were carried out over the years (a brief review of these methods can be found in [2,4,17]), but this paper is focused on a second possibility: The assessment of the subjective experience of emotions during music perception based on self-reports [8]. In self-report methods, the subjects are stimulated to verbalize and express their emotions towards stimuli. Different techniques, called answering formats, exist to assess the participants' emotions, such as affective scales, free descriptions, or the use of "emotional space" [4]. Every answering format has different advantages and drawbacks when measuring emotions. The next section will explain in more detail advantages and disadvantages of common measuring methods.

2 Challenges in Measuring Subjective Experience of Emotions via Self-reports

In [2, p. 210] Zentner states that "there are four important limitations to self-report methodology […]: a) demand characteristics, b) self-presentation biases, c) limited awareness of one's emotions, and d) difficulties in the verbalization of emotion perception […]".

While the assessment of subjective experience with closed-response self-report methods such as adjective scales [5] or emotional spaces [6] is ensuring efficiency and, to some degree, a standardization of data collection, "the predetermined choices [of descriptors] might influence the participant to respond along the provided categories [and] the interpretation of the terms provided by the researcher might vary considerably across people […]" [2, p. 193]. One attempt to overcome the problems of closed-responses is the usage of a free response measurement: Subjects are allowed to explain the nature of the state they experience, i.e., an emotion while listening music, in their own words, for example in written form or an interview. A content analysis of the narrative establishes the link between music and the induced emotions. Unfortunately, the data treatment and interpretation of such a content analysis is not an easy task and cannot be automated. A second disadvantage lies in the different linguistic abilities of the subjects – some might lack an appropriate vocabulary to describe the emotional experience during listening to music. This might lead to a loss of information. A possible way to promote the advantages of both measurement approaches is the combination of open and closed-response format.

An approach using an open response format in combination with a closed-response format was presented in [8], the so-called Free-Choice Profiling (FCP). By applying FCP, subjects first define and identify individual attributes (emotional terms, also called descriptors) by themselves. The rating of intensity of the emotional experience during music perception is then done with the help of adjective scales, where the individual attributes are used as labels. Due to the design of the test method, it is taken into consideration that different subjects might use terms in different ways, or different terms with the same underlying meaning. The study mentioned in [8] was able to obtain clear and interpretable results consistent with music theory and emotional psychology. However, the study investigated only a small set of major/minor chord items, and as it was the first application of FCP in the field of emotions in music, questions of reliability and general feasibility of the method remained.

3 Experimental Design and Parameters

The promising results of [8] led to a research project funded by the German Research Foundation to verify FCP as a useful test methodology for the assessment of emotions and to enable a better classification of musical parts based of their emotional impact on music perception.

This paper presents two preliminary FCP studies of this ongoing project that were conducted with different scopes: The first experiment aimed at assessing the selection of suitable test stimuli, in terms of their degree of emotional impact. The target of the second study was to verify the usage of FCP as a suitable test methodology by using different test material. The test method, items, and participants for both experiments are explained in this section.

3.1 Test Method in General

FCP, a method common in food research, was used to identify individual attributes (emotional terms) and to rate the liking and/or intensity of those attributes. The procedure, which is outlined in detail in [8, 9], helps to identify significant attributes, discrimination, and panelist performance. It takes into consideration that different subjects might use terms in different ways, or different terms with the same underlying meaning. In recent years, FCP was also successfully applied and refined in the field of user experience to assess multimodal quality perception [18].

As mentioned in [18] the FCP is structured into four different parts, referred to as *introduction*, *attribute elicitation*, *refinement of attributes,* and *sensory evaluation*. In the *introduction* the nature of descriptive evaluation, in particular the use of the participant's own attributes to describe the perceived emotionality of test items, is explained in detail. This first step of the method is the most crucial part, because here the cornerstone for the assessment is laid. Subjects have to understand the method correctly, but special care must be taken not to influence them in a certain direction. Therefore, the participants are shown how to find attributes that define emotions with an easy task of a different perceptual domain[1]. The *attribute elicitation* aims at finding individual emotion attributes that characterize each participant's emotional perception of the different test stimuli. In this study, participants listened to a small representative subset of test items (see Section 4.2) and wrote down the perceived emotions using their own words, without any limitation concerning the number of attributes. No additional technique like repertory grid method [10] or natural grouping [10] was used as support for the elicitation of attributes. In the third step a *refinement of attributes* was done. Here, strong attributes were chosen out of all developed attributes according to two rules: First, attributes must be unique and each attribute must describe only one aspect of emotion. Second, the participants must be able to define the attribute in their own words. Hence, the participants had to write down a definition of each of the attributes left over for the final evaluation. For the *sensory evaluation* all generated attributes were printed out on paper together with 10 cm long

[1] For example, as it was the case in this study, the participants are asked to describe the emotional impact of different movies or photos.

scales, labeled with “min” and “max”[2]. These individual score cards, one for every test item, were used for the evaluation of all test items, which were presented randomly one by one. The subjects were advised to mark the perceived strength of each attribute for each test item.

3.2 Test Items

Experiment 1: The test items consisted of eight specifically designed major/major and minor/minor chord combinations, derived from the circle of fifths (see Figure 1). Each item consisted of two chords played one after the other: C/F, C/G, C/B, C/D♭ (major), as well as c/f, c/g, c/b, c/d♭ (minor). These are the two chords located next to C (F and G), and the ones furthest away (B and D♭). To assess the selection of suitable test stimuli, in terms of their degree of emotional impact, these musical phrases were also varied in instrument choice and tempo. Their length varied, depending on instrument choice and tempo, from approx. 2.5s to 7.5s. The decision to use these basic musical structures was made in order to exclude as many other variables as possible, including familiarity with well-known musical pieces.

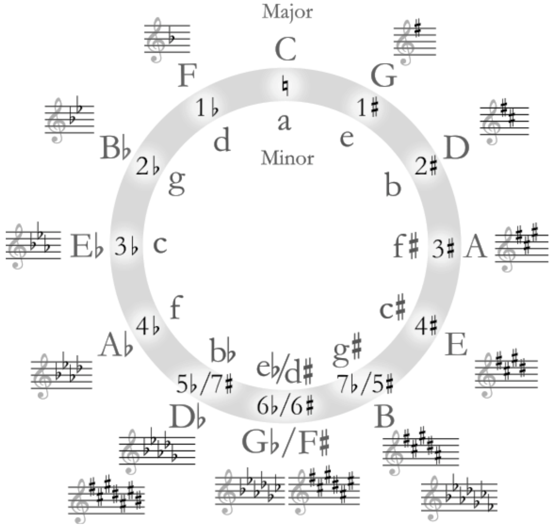

Fig. 1. Circle of fifths. (from: en.wikipedia.org, licensed under CC BY-SA 3.0)

Three different instruments were used: Violin, piano, and synthesizer. Violin was chosen because of the possibility to induce sad emotions [12]. Former studies indicate that artificial instruments, e.g., a synthesizer, can lead to a decreased recognition of sad emotions [12]. Piano was chosen because of its broad usage and popularity in other studies, i.e., to allow comparability [13]. Furthermore, three different tempi (30,

[2] Where “min” means that the attribute is not perceived at all, while “max” refers to its maximum pronounced form.

70, and 120 bpm) were used as conditions. The previous FCP study [8] used 30 bpm only and indicated that this tempo already induces a slightly sad emotional offset, so this tempo was used in this study for comparison. 70 bpm was chosen as it is close to resting heart rate and can be seen as a "normal" state of activation for the listener. 120 bpm is a common tempo for modern dance music as well as for modern marches ("march tempo") and can be regarded as more activating. This results in a total set of 72 musical phrases.

Experiment 2: Several participants of experiment 1 mentioned that the test items appeared to be too short for eliciting distinctive emotions. The second experiment therefore aimed to investigate whether longer test items were perceived differently. New items were created according to Table 1. As the first experiment took the participants nearly 60 min. on average, it was decided to use fewer items to compensate for the longer item length. Only two different tempi (70 and 120 bpm) and only one instrument (synthesizer) were used. The item length ranged from 6s to 10s. The reduced set of 16 test items led to a much shorter rating time of only 15 min. on average.

Table 1. Chord combinations used in the second experiment.

Item number	Chord combination
1	C-G-C-F-C-G-C-F-C
2	a-e-a-d-a-e-a-d-a
3	C-A-C-D-C-A-C-E-C
4	a-f#-a-b-a- f#-a-c#-a
5	C-F#-C-B-C-F#-C-D♭-C
6	a-d#-a-g#-a-d#-a-b♭-a
7	C-E♭-C-B♭-C-E♭-C-A♭-C
8	a-c-a-g-a-c-a-f-a

3.3 Test Panel

In the first test 24 subjects, 9 female and 15 male, participated. The average age was 24.8 years.

The number of subjects in the second experiment was 10, with an average age of 24.7 years. Half of the participants were male and the other half female.

Any participant took part in only one of the experiments. Although some subjects reported slight hearing damages, none of the subjects were rejected from test participation and analysis.

4 Experiment 1

The first experiment aimed at assessing the selection of suitable test stimuli and a general re-evaluation of FCP for the assessment of emotional impact while listening to musical phrases.

4.1 Test Facilities

The tests were conducted in the Audio Lab at Ilmenau University of Technology, a room compliant to ITU-R BS.1116-1 [14], to EBU 3276, and to DIN 15996. Its exact dimensions are 8.4m x 7.6m x 2.8m. Two identical Genelec 1030A loudspeakers were used in the test, placed on stands at ear height of the seated subjects. Participants were seated in the sweet spot position in front of a desk with a flat screen monitor, keyboard, and mouse. The arrangement of the speakers and the listening positions are in accordance with ITU-R BS.1116-1.

4.2 Test Procedure in Detail

Introduction: Each participant received a short introduction about the test in general and the test method FCP. They were handed out a privacy policy and had to fill out a short questionnaire regarding demographics, musical knowledge, and their current mood. For a better understanding of the attribute elicitation and listening task, each subject was asked to imagine two different (known) movies and to verbalize the differences in the emotions they associated with them. The supervisor took care to avoid giving predetermined attributes that might influence them in a certain direction.

Attribute Elicitation: During this stage, each participant assessed a representative selection of 16 of the final test items, that is one item for each instrument, tempo, and key, and wrote down the verbal descriptors with which they would have to rate these items in the fourth part (attribute rating). A graphical user interface (GUI) was used, allowing the subjects to listen to each item as often as they wanted. During this part, the supervisor left the room for the control room, in order not to disturb the participant. The participants were seated in a 90° position to the control room window, thus the supervisor remained available, either via eye contact or a microphone connection.

Attribute Refinement: After the participant signaled that he/she was done, the supervisor and the participant reviewed the attribute list together. The participant decided if some words could be summarized to one single term or should be renamed. After this, the participant was asked to give a brief explanation for each term, if possible. The attribute list was reviewed once again afterwards.

Attribute Rating: Starting with a short rating test of 3 items and all of his or her attributes, each participant carried out a training task. In case the participant felt the need to apply changes, they were allowed to modify their descriptors one last time.

After this, the actual test started, where each subject rated all 72 items with the complete set of their descriptors. The test allowed the participants to listen to each test item as often as they wanted, but it was not allowed to revise ratings of prior items. It was planned to have a rating software right from the beginning, but due to a computer failure the first 4 subjects did a rating on paper with a list of their attributes on the left and for each attribute a 10 cm long rating scale on the right side of the sheet. The scales were labeled with min and max. Later subjects carried out the rating with software. The design of the graphical user interface was similar to the rating sheets. If participants took longer than 60 minutes, they were asked to take a break of approximately 10 minutes before they went on.

4.3 Results

The data was analyzed with a Multiple Factor Analysis (MFA) [15], a widely used method in sensory profiling. As each participant uses his/her own vocabulary, a multi-dimensional perceptual space – the verbal descriptors representing the dimensions – is created. MFA is very similar to Principal Component Analysis (PCA)[3]: it compares the individuals' perceptual spaces and combines them into a single global one. An MFA provides mainly two outputs: a) The mean location of the test items on the global space and b) the location of the verbal descriptors on these dimensions.

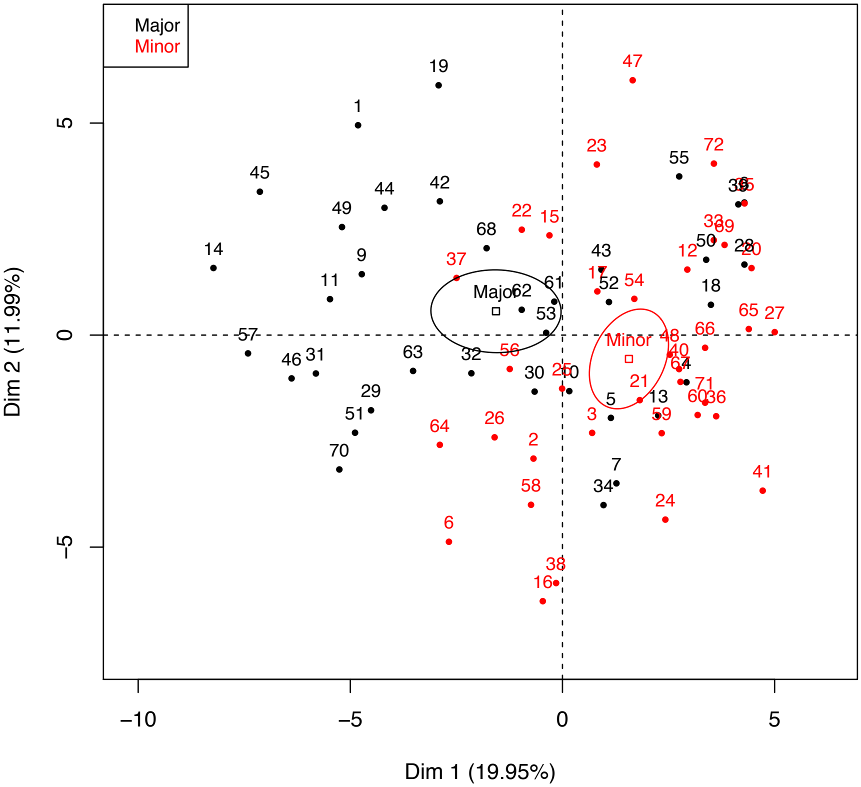

Fig. 2. Graph of the first two dimensions of experiment 1 with an explained total variance of 32%. Shown are the major and minor chord-combination groups and their respective confidence ellipses for the mean of each group.

[3] In fact an MFA computes nested PCAs.

Figure 2 shows a graph of the first two dimensions with the mean location of the test items and confidence ellipses for the means of major and minor categories. The non-overlapping ellipses clearly indicate that these categories were rated significantly different. In total, the first two dimensions declare only 32% of the total variance of the original data. One reason for this could be that there was little agreement among participants.

Still the arrangement of the test items on the first two dimensions is sensibly interpretable in several other ways beyond key[4]: All chord-combinations of each key, except the ones featuring B and D♭, were rated significantly distinguishable and were ordered from left to right on dimension 1 according to their distance to C on the circle of fifths (see Figure 1).

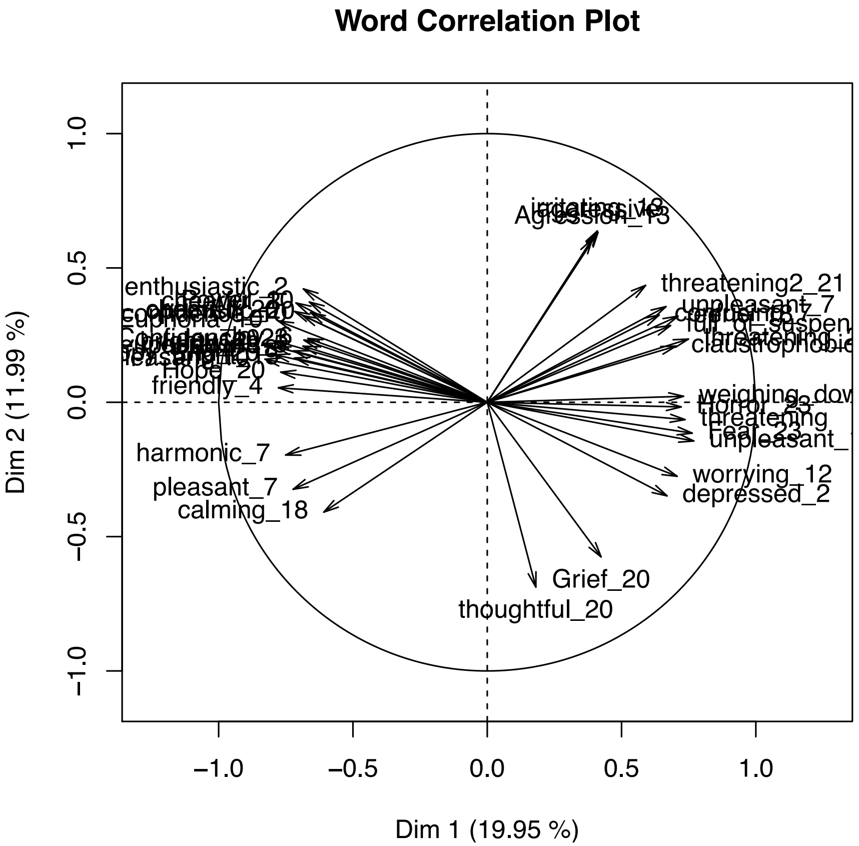

Fig. 3. Word chart of the significant listener descriptors for the first two dimensions of experiment 1. The numbers behind the descriptors refer to the listener.

[4] The respective graphs cannot be shown here due to space reasons, but are available on request.

The second dimension (y-axis) features the faster tempi and the synthesizer-sound on the upper part, while the lower part primarily contains the 30 bpm tempo and violin- and piano-sounds. The instrument synth was rated significantly higher than violin and piano, and 30bpm and 120bpm can be clearly separated in the second dimension.

To identify the perceived emotions the participants associate with these dimensions, Figure 3 shows the respective word chart of the first two dimensions. Only those descriptors contributing to both dimensions with an $R^2 \geq 0.5$ are plotted, hence not all descriptors of all participants are present. All verbal descriptors were originally given in German and translated by the authors, the English translations shown here may hence convey slightly different meanings. Word charts tend to be crowded, therefore Table 2 gives an overview of these attributes, ordered by participant.

The first dimension (x-axis) features positive descriptors on the left side (excited, happiness, confidence, euphoria, harmonic, pleasant, calming, etc.), and negative ones on the right side (fear, horror, menacing, aggressive, irritating, unpleasant, depressed, etc.). This conforms very well with the concept of "valence" in emotional psychology [2, 4]. The second dimension (y-axis) does not contain many descriptors (which results in its low explained variance), but they are clearly interpretable: the lower part shows descriptors of low activity, such as: calming, pleasant, harmonic, but also grief, thoughtful, depressed and unpleasant. The upper part contains descriptors that are clearly active, for example, aggressive and excited. This again conforms with another well-known concept: arousal [2, 4].

Table 2. Significant descriptors contributing to dimensions 1 and 2 of experiment 1. The numbers behind the descriptors refer to the listener.

Attribute	Attribute	Attribute
threatening	claustrophobic_7	calming_18
aggressive	Euphoria_10	light_18
depressed_2	Joy_10	cheerful_20
cheerful_2	Power_10	Hope_20
enthusiastic_2	self-confidence_10	thoughtful_20
friendly_4	pleasant_10	Happy_End_20
full_of_suspense_4	unpleasant_10	Grief_20
weighing_down_4	bright_12	optimistic_20
euphoric_5	worrying_12	threatening_21
Joy_6	Agression_13	threatening_21
pleasant_7	irritating_13	Confidence_23
harmonic_7	bright_15	Fear_23
confusing_7	euphoric_18	Horror_23
unpleasant_7	eerie_18	

5 Experiment 2

Several participants of experiment 1 mentioned that the test items appeared to be too short for eliciting distinctive emotions. To assess the effect of longer test items a second experiment was conducted. The two experiments are comparable in their procedure, but in this second experiment we made a slight change to the preparation task for the *attribute elicitation*. The attribute election procedure itself stayed the same. Minor changes were the use of a different test facility room and test items.

5.1 Test Facilities

The tests were conducted in the Audio Lab at Fraunhofer IDMT compliant to ITU-R BS.1116-1 [14], to EBU 3276 and to DIN 15996. Its exact dimensions are 6.90 x 4.60 x 2.70. Two identical K&H O-510 loudspeakers were used in the test, placed at ear height of the seated subjects. Participants were seated in the sweet spot position. The arrangement of the speakers and the listening positions are in compliance with ITU-R BS.1116-1. The changes in test facilities are considered not to bias the results. The room characteristic is in line with the room characteristics of the first experiment. Although the test equipment is not exactly the same like experiment 1, the same class of high quality loudspeaker was used for the tests.

5.2 Test Procedure in Detail

The general procedure of this experiment was very similar to the first experiment (see section 4): The introductory task of imagining two movies was replaced, because for some participants the task was too abstract and they had problems understanding the intention of the attribute elicitation task. Instead, participants were now handed out five different images[5], which were taken from the International Affective Picture System (IAPS)[6] database. They were asked to explain what emotions these images elicited and to verbalize the similarities and dissimilarities.

5.3 Results

Experiment 2 was analyzed in the same manner as experiment 1 (cf. Section 4.3). Figure 4 shows the graph of the mean location of the test items on the first two dimensions. It is apparent that the explained variance is higher (42.6%) than in experiment 1 (32%), which can be interpreted as a slightly higher agreement among the participants on what they perceive.

[5] The images portrayed: 1) several woodlice, 2) a woman and a child close together, 3) a wolf, 4) a rabbit, and 5) a lonely road through grass-covered plains.

[6] http://csea.phhp.ufl.edu/Media.html

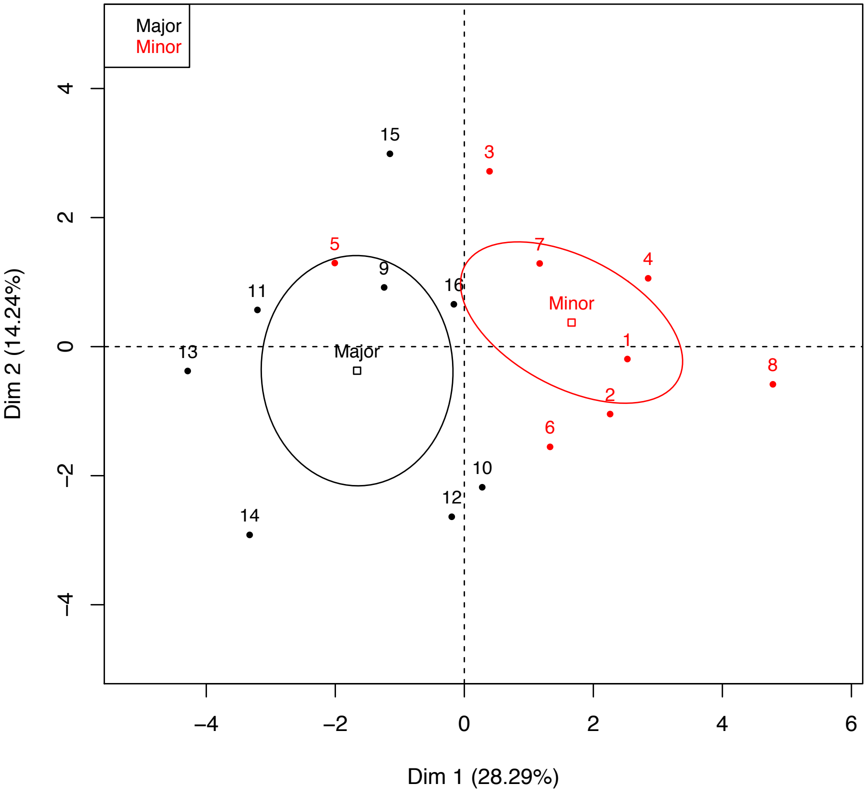

Fig 4. Graph of the first two dimensions of experiment 2 with an explained total variance of 42,6%. Shown are the major and minor chord-combination groups and their respective confidence ellipses for the mean of each group.

Considering the position of the test items on these dimensions, the picture is partially similar to the one of experiment 1: On the first dimension (the x-axis), the left hand side contains all major chords except one, while the right hand side contains all minor chords except one. Furthermore[7], the items are – as it was the case in experiment 1 – sorted according to the circle of fifths, with the neighboring chord-structures on the left hand side and the opposing chord-structures on the right hand side of each key group. The location of the items on the second dimension (y-axis) is not as obvious as in experiment 1, but it can be noted that the items rated most positive on this dimension are the faster ones (120 bpm), while the most negative ones are the slower ones (70 bpm), and that these groups are significantly different. The word chart (Fig. 5, see p. 13) of the first two dimensions matches the item location chart: On dimension 1 (x-axis), positive descriptors can be found on the left hand side: happy, cheerful, euphoria, impressive, heroic, festive, etc.; negative descriptors

[7] As before, the respective graphs cannot be shown here due to space reasons but are available on request.

are located on the right side of the axis: menacing, danger, loneliness, exhausted, thoughtful, etc. Compared to Figure 3 the second dimension (y-axis) is not that clearly marked as in experiment 1, but in general the more "active" descriptors (happy, cheerful, menacing) are located on the positive side of this dimension, while the negative side contains mostly "inactive" descriptors: heroic, festive[8], loneliness, exhausted, thoughtful, etc. Again, Table 3 shows all the descriptors significantly contributing to the dimensions 1 and 2.

Table 3. Significant descriptors contributing to dimensions 1 and 2 of experiment 2. The numbers behind the descriptors refer to the listener.

Attribute	Attribute	Attribute
boring_1	Suspense_4	cheerful_8
not_harmonic_1	monotone_5	Depression_9
heroic_2	menacing_6	happy_9
euphoric_2	ominous_6	exhausted_9
menacing_2	promising_6	Loneliness_10
depressing_2	carefree_6	Euphoria_10
Success_3	delighted_7	festive_10
Danger_3	thoughtful_7	Party_mood_10
annoying_4	sad_7	impressive_10

In summary, the location of the items on these dimensions and the respective descriptors concur with experiment 1 in that the first dimension can easily be interpreted as "valence". In the case of the second dimension, it seems that the participants knew what they wanted to rate, but then had problems to actually discern the items. This is not surprising as the difference in activation between 120 and 70 bpm is clearly much lower than the difference between 120 and 30 bpm, as it was the case in experiment 1. Nonetheless, the second dimension can easily be interpreted as "arousal".

[8] In the case of "heroic" and "festive" it might be argued that these are active descriptors, but the German connotation of the original descriptors is more that of a ceremonial atmosphere.

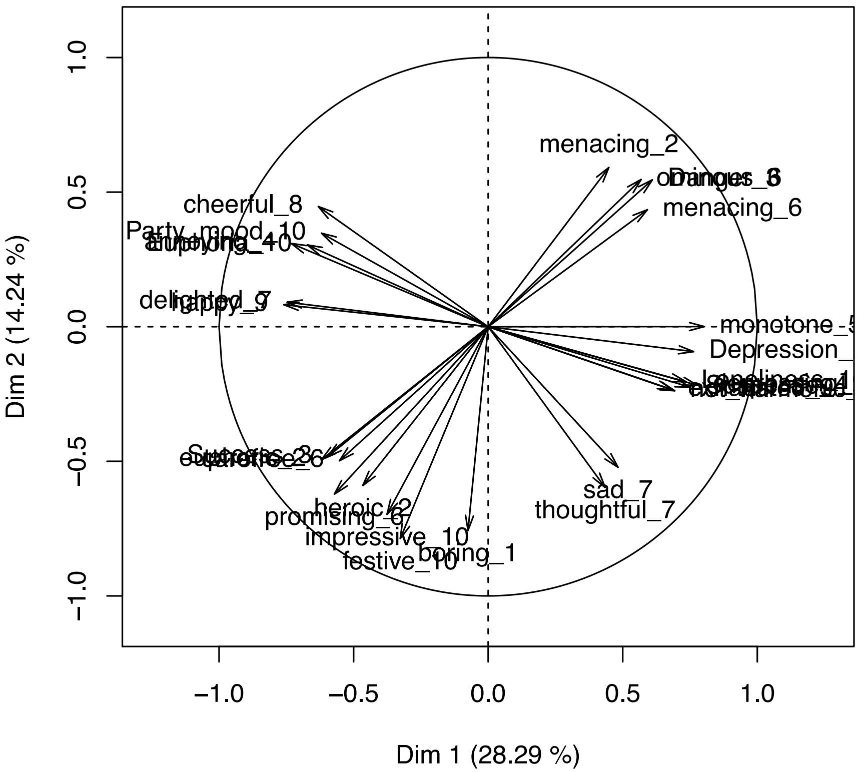

Fig. 5. Word chart of the significant listener descriptors for the first two dimensions of experiment 2. The numbers behind the descriptors refer to the listener.

6 Conclusion and Further Work

In this paper, we propose and investigate FCP as a test method to overcome drawbacks of common self-report methods, to assess the emotional state of a subject during music perception. By applying FCP, subjects define individual attributes (emotional terms) by themselves. The rating of the intensity of the emotional experience during music perception is done with the help of adjective scales, where for each subject their individual defined attributes are used as labels. To prove the feasibility of FCP for the evaluation of emotions elicited by music and to assess the selection of suitable test stimuli, two experiments were carried out.

The results of experiment 1 showed that the subjects rate the emotional impression according to dimensions of valence and arousal, which are commonly proposed by emotional psychology. Furthermore, simple major and minor chord combinations could directly be linked to the dimension of valence, with the participants being able to sort the chord-samples according to the circle of fifths. The second dimension features the faster tempi and the synthesizer-sound on one side, while the other side primarily shows the 30 bpm violin- and piano-sounds. This leads to the conclusion that the second dimension represents “arousal”. Although the detailed analysis shows clearly interpretable results, the results only declare 32% of the total variance of the system. This further leads to the conclusion that there is rather large disagreement between the participants on what they perceive, and the "least common denominator" is fairly small.

Verbal comments of the participants led to the assumption that the musical phrases were too short to elicit emotion. Experiments examining the lower bound of length in which emotions can be perceived have been conducted, e.g., [19, 20], finding that excerpts as short as 250-500ms are sufficient to elicit emotions. However, these studies examined emotions on a very basic level, reducing the spectrum to a binary happy/sad decision [20] or neutral/moving [19]. Thus, it remains unclear whether participants are able to precisely classify their perceived emotions in a multi-dimensional space with such short pieces. Furthermore, the studies used excerpts of classical and well-known musical pieces. This poses the question whether participants rated their actual perceived emotions or rather their remembered emotions based on familiarity.

To prove the hypothesis that longer stimuli are more easily classified, a second experiment was conducted where longer test stimuli were used. While the items of experiment 1 consisted of two chords with a maximum item length of 7s were used, experiment 2 had items with nine chords per item and a maximum length of 10s.

The explained variance of the first two dimensions in experiment 2 is slightly higher than in experiment 1 with 42.6%, which can be interpreted as a slightly higher agreement among the participants on what they perceive. In general, the results of the first experiment were confirmed. Unfortunately, the extension of the musical phrases did not lead to a significant higher explained variance.

Although the results are easily interpretable and sensible, the low explained variances of the results are puzzling. One explanation could be that emotion is a very subjective experience, which is not easy to describe or indicate.

Furthermore, the test method cannot solve the problem of awareness of an emotion as mentioned in [2, p. 210 et seq.]. When defining an emotion as consisting of several components, it is questionable which part of an emotion is accessible at all and which part is accessed in a self-report. This could be another reason for the low explained variance of the test results.

FCP is able to approach two of the four problems raised by Zentner ([2, p. 210 et seq.], also see Section 2): Because no fixed responses are given, participants do not feel the need to comply with certain emotional concepts and will not feel *demand characteristics*. Secondly, the *difficulties in verbalization of musical emotions* are partly compensated by FCP’s ability to directly compare and group correlating descriptors of all participants. Hence, it is not so important that the participant is able to express emotions with a complex vocabulary, but rather that he/she is able to discern and rate the perceived emotions.

To further investigate the dependency of the linguistic abilities on the rating and see if FCP really solves the addressed problem, we plan to conduct new experiments, using the same test items as in experiment 1. The next experiment will use a pictorial rating system called SAM (Self-Assessment Manikin, Fig. 5) [16], a common and well-researched rating system in emotional research. This rating system assesses the three dimensions valence, arousal and dominance in a non-verbal way and is thus suited to be used by children and/or non-native speakers. The participants of this experiment will consist of "Amazon Mechanical Turk" (MTurk) workers[9]. This so-called "clickworker"-platform allows to offer easy tasks that can be solved with a few mouse-clicks, such as annotation tasks, to registered workers. Amazon MTurk is a cheap and efficient way to have many test items annotated by a lot of people in order to build a ground truth. The results will be compared to those of the experiments already conducted.

Acknowledgments This work was supported by the German Research Foundation (DFG)-funded project "Fundamental research on the development of algorithms for the analysis of emotion-psychological characteristics of music signals".

References

1. Zentner, Marcel; Grandjean, Didier; Scherer, Klaus R. (2008): "Emotions evoked by the sound of music: Characterization, classification, and measurement. " In: Emotion 8 (4), 494–521
2. Sloboda, John A.; Juslin, Patrik N.; Frijda, Nico H. (Hg.) (2010): Handbook of music and emotion. Oxford: Oxford University Press
3. Scherer, K. R. (2005). "What are emotions? And how can they be measured?" Social Science Information, 44(4), 695–729
4. Nagel, Frederik (2007): Psychoacoustical and psychophysiological correlates of the emotional impact and the perception of music. 1. Aufl. Göttingen: Sierke
5. Hevner, K.: "Experimental studies of the elements of expression in music." American Journal of Psychology, 48, 246–286, 1936
6. Russell, J. Russell: "A circumplex model of affect." Journal of Personality and Social Psychology 39 (1980), pp. 1161–1178, 1980
7. Zentner, Marcel; Eerola, Tuomas (2010): Self-report measures and Models. In: Patrik N. Juslin, John A. Sloboda und Nico H. Frijda (Hg.): Handbook of music and emotion. Oxford: Oxford University Press, 187–221
8. Schneider, S.; Raschke, F.; Gatzsche, G.; Strohmeier, D.: "Free Choice Profiling and Natural Grouping as Methods for the Assessment of Emotions in Musical Audio Signals", 126th AES Convention, 2009 Munich
9. Lawless, H.T., and Heymann, H. "Sensory evaluation of food: principles and practices". Chapman & Hall, New York. 1999
10. Jack, F.R. and Piggott, J.: "Free choice profiling in consumer research," Food quality and Preference, vol. 3, no. 3, pp. 129–134, 1992
11. Williams, A.A., Langron, S.P. "The use of Free-choice Profiling for the Evaluation of Commercial Ports." In: Journal of the Science of Food and Agriculture 35, pp. 558-568, 1984
12. Behrens, Green: "The Ability to Identify Emotional content of Solo Improvisations Performed Vocally and on Three Different Instruments"; Psychology of Music ,Volume 21(1):20-33 (1993)
13. Hailstone, Julia; Omar, Rohani; Henley, Susie; Frost, Chris; Kenward, Michael; Warren, Jason: "It's not what you play, it's how you play it: Timbre affects perception of emotion in

[9] https://www.mturk.com/mturk/welcome

music." In: The Quart. J. of Expt. Psych (The Quarterly Journal of Experimental Psychology), 1962, No. 11, 2141–2155, 2009

14.Recommendation ITU-R BS.1116-1 (10/1997) Methods for the subjective assessment of small impairments in audio systems including Multichannel Sound Systems. International Telecommunication Union, Radiocommunication Assembly

15.Abdi H. and Valentin, D.: "Multiple factor analysis (mfa)," Encyclopedia of measurement and statistics, pp. 657–663, 2007.16

16.Bradley, Margaret M.; Lang, Peter J. (1994): Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential. In: Journal of Behavior Therapy and Experimental Psychiatry (Vol. 25, No. I.), S. 49–59

17.Stemmler, G. (2003): ''Methodological Considerations in the Psychophysiological Study of Emotion'', In: R.J. Davidson, K.R. Scherer and H. Goldsmith (eds) Handbook of the Affective Sciences, pp. 225–55. New York and Oxford: Oxford University Press.

18.D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, "Open profiling of quality: a mixed method approach to understanding multimodal quality perception," Advances in Multimedia, vol. 2010, Article ID 658980, 28 pages, 2010.

19.S. Filipic, B. Tillmann, and E. Bigand, "Judging familiarity and emotion from very brief musical excerpts," *Psychonomic Bulletin & Review*, vol. 17, no. 3, pp. 335–341, Jun. 2010.

20.I. Peretz, L. Gagnon, and B. Bouchard, "Music and emotion: perceptual determinants, immediacy, and isolation after brain damage," *Cognition*, vol. 68, no. 2, pp. 111–141, 1998.