Collecting annotations for induced musical emotion via online game with a purpose Emotify

A. Aljanaki, F. Wiering, R.C. Veltkamp

Technical Report UU-CS-2014-015
April 2014

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

Abstract

One of the major reasons why music is so enjoyable is its emotional impact. Indexing and searching by emotion would greatly increase the usability of online music collections. However, there is no consensus on the question which model of emotion would fit this task best. Such a model should be easy for listeners to use both to tag and to retrieve emotion, and should lead to unambiguous results. The latter is complicated not only due to linguistic issues, but also because musical emotion is a subjective phenomenon that depends on many extra-musical factors, such as mood of the listener, musical preferences, age, personality. We investigate this problem by creating a game with a purpose Emotify to collect emotional labels for a set of 400 musical excerpts in different genres. We use the Geneva Emotional Music Scales (GEMS) to annotate this corpus. In this technical report we analyze the data produced by the game. We find that the factors that influence induced musical emotion (in the order of decreasing importance) are musical preferences, mood and gender. We measure the agreement of listeners using Cronbach's alpha and find that it differs hugely per emotional category (amazement, sadness and solemnity are most inconsistent, and tenderness, power and joyful activation - most consistent categories) and does not differ significantly among the four tested musical genres (rock, pop, classical and electronic music).

1 Introduction

The emotional content of a musical piece is an essential, yet ambiguous part of it, even more so when it concerns induced (felt) emotion. With the current sizes of musical databases there is a growing need for automatic methods of music classification and similarity assessment, and emotion-based methods are potentially among the most useful access mechanisms for music collections. Training and evaluation of automatic emotion classification methods relies on a ground truth, and collecting such a ground truth involves human annotations. There is no objective judgement in case of induced emotion (that is, felt by the listener, as opposed to perceived emotion, which is a characteristic ascribed by the listener to the music). Due to subjectivity of felt emotion, the question about the degree of agreement between annotators arises. In order to answer this question, it is necessary to involve a big and varied set of participants. Recently, online games have become a popular way of large-scale data collection in music information retrieval. In this technical report we will present results from Emotify (www.emotify.org) - a game with a purpose for collecting labels on induced musical emotion, and analyse the data collected via this game.

Obtaining a ground-truth dataset of music labeled with emotions remains a challenging task in music emotion recognition research. Outside the laboratory, there are two possible ways of assembling a dataset labeled with emotion annotations: through social tag mining (relying on websites such as last.fm or allmusic.com) and in a more systematic way through user surveys or data collection games. Social tag mining makes it possible to collect a huge dataset, but lacks the homogeneity and control that a preselected emotional model and a controlled experimental setting provides. It is in most cases unfeasible in tag mining to measure the level of agreement between multiple users on certain tags (or it would be necessary to apply an additional cross-verification procedure as it was done in the case of data for the MIREX audio mood recognition task [2]). A controlled user experiment would be an ideal way of data collection. In this case, in addition to self-report, researchers can collect physiological measurements and exclude external factors that might influence the outcome. However, tasks involving music are very time-consuming. In the end, researchers seem to be left with a difficult choice between a small-scale or a very expensive survey. By creating a so-called 'game with a purpose' (GWAP) we are trying to avoid both pitfalls.

2 Related work

Organizing, categorizing and searching music by emotion is natural and convenient for music listeners. With the growing amounts of easily accessible music, it becomes very important for a musical database to provide such a functionality. Recently, automatic music emotion recognition received significant attention from the research community. A lot of work remains to be done on computational models of emotion and data collection.

2.1 Models of musical emotion

Several areas of science, such as psychology, musicology and neuroscience, have come up with general or domain-specific models of emotion. These can be divided into two groups: categorical and dimensional models. Categorical

models present emotions as consisting of several basic clusters or categories. The earliest attempt to create a specifically musical categorical model of emotion was undertaken by K. Hevner [1]. She created an ontology of eight emotional clusters, such as humorous, pathetic and dreamy, where each cluster contains from six to eleven adjectives.

Dimensional models arrange emotions in a continuous space along several (usually two or three) principal dimensions. The most popular dimensional model, frequently used in Music Information Retrieval, is the one proposed by J. Russell [9]. It consists of two dimensions: valence and arousal. The valence-arousal model is often criticized for its lack of granularity. For instance, anger and fear are placed very close to each other in the upper left quadrant of the valence-arousal plane. Moreover, music is capable of expressing much more subtle and even contradictory emotions (e.g. bitter-sweetness). It is impossible to present these on the valence-arousal plane. To address the pitfalls of existing models, a new domain-specific categorical emotional model called GEMS (Geneva Emotional Music Scale) was developed by Zentner et al. in 2008 [7]. The full GEMS scale consists of 45 terms, with shorter versions of 25 and 9 terms. Originally, the terms were collected in French, and later translated to English. These nine terms can in turn be grouped into 3 superfactors: vitality, sublimity and unease. We will use the shortest nine term version of GEMS in our online game (see table 1). GEMS is unique in that it addresses induced emotion, was created specifically for describing musical emotion, and has a level of emotional granularity that other models do not provide.

2.2 Datasets involving the GEMS model

GEMS has already been used as an underlying model for data collection, but none of the datasets was big, and the data is not publicly available. The biggest experiment, involving nearly 4000 participants, took place in 2010 [4] in Dublin. Participants listened to music and reported their emotional state, using several self-assessment questionnaires, GEMS among them. Physiological measurements were also recorded. The dataset contained 53 songs from different genres (rock, classical, pop, jazz, world etc.), specially selected for their emotional content. The analysis of the collected data is presented in the PhD thesis of Javier Jaimovich ([3]). Unfortunately, due to software error GEMS questionnaire had to be excluded from the final analysis. In 2010, Vuoskoski et al. performed a comparison of three emotional models (valence-arousal, 5 basic emotions and GEMS), using 16 excerpts from movie soundtracks [11]. The most consistent ratings were produced in the case of the two-dimensional valence-arousal model, while basic emotions and GEMS were less consistent. GEMS's consistency varied between categories. In 2011, Torres-Eliard et al. used GEMS for continuous emotion measurements [10]. Every rater controlled one GEMS dimension. Data on 36 musical excerpts were collected. The inter-rater agreement (based on the extent to which a single emotion was present in the music at a given moment of time) was satisfactory.

2.3 Musical games with a purpose

Recently, collaborative online games have become a popular way of collecting musical metadata. Some of these games were proposed for the collection of descriptive labels (tags) on short musical fragments, such as MajorMiner [8] and TagATune [6]. Some of the collected labels were also mood-related. A specifically emotion-targeted game called MoodSwings, for continuous emotional annotation of music, was created by Kim et al. [5]. In this game, players are paired up with a partner and both of them mark the perceived musical emotion on a per second basis on the valence-arousal plane. They earn points by guessing their opponent's position on the valence-arousal plane for the same fragment of music. The game we present, Emotify, is different from MoodSwings in several respects: it uses a categorical emotional model, it collects data on induced (not perceived) emotion, and the measurements are discrete rather than continuous.

3 Experiment description

Existing research on music-induced emotion is either dealing with data obtained through web-mining (which makes it very difficult to estimate inter-rater agreement), or with music selected for its strong and obvious emotional content. In the latter case, it is questionable how results are comparable to non-selected music, which may not have such explicit emotional content. Our experiment is intended to cover this gap and to collect data on induced emotion for a randomly selected large collection of music in different genres.

3.1 Music used in the game

We have assembled a set of 400 musical pieces from the Magnatune recording company (magnatune.com), 100 pieces from each of the four selected genres (**classical**, **rock**, **pop** and **electronic**). Genres were assigned by the recording company. The resulting dataset contains music from 241 different albums by 140 performers. The selection of pieces was random, as we were aiming at an ecologically valid musical dataset. There were several reasons to choose music from Magnatune: it is of good quality and it is generally little known (familiar music might precondition induced emotion). The music was reviewed manually and some recordings (around 2%) were removed because of their poor recording quality.

Before launching the data collection game, we decided to collect labels for a small subset of the data (which we will call **subset A**), to check how consistent the responses are. There were several reasons to conduct a pilot study before launching the game in order to control how well the participants will cope with the task. Firstly, we were using a sophisticated nine item scale. Secondly, the experiment was conducted online, not in a controlled environment. Thirdly, we were using a slightly different way of data collection than originally suggested by GEMS authors (instead of asking for a rating on a Likert scale for each of the emotions, we collected binary responses). The Likert scale is a psychometric scale commonly used in questionnaires.

When answering a question using a Likert scale, a participant has to choose one from several (usually 5 or 7) items typically ranging from "strongly disagree" to "strongly agree". However, this way of data collection is very slow, requires quite some mental effort, and is not suitable for a dynamic online game. Therefore we modified the user's task and asked to select several labels from a list instead. We also restricted users on how many labels should be selected, by explicitly demanding them to select no more than 3 labels.

Subset A consists of 60 songs, 15 songs in each of the four chosen genres, which constitutes 15% of the whole music set. For each of these songs we decided to collect (following a statistical rule of thumb) at least 10 measurements per variable (per category). That makes at least 90 annotations per song, since there are nine questions in the questionnaire (one for each of the nine GEMS categories). We count all labels given to a song independently (2 labels assigned by the same person are independently counted). The remaining 85% of the data (**subset B**) consists of 340 songs.

3.2 Questionnaire

We changed the wording of two GEMS categories: transcendence was replaced by solemnity and wonder by amazement (see Table 1). This was done because previous research showed that respondents might have problems with understanding these two categories [11].

Emotional category	Description	Superfactor
Amazement *	Feeling of wonder and happiness	Sublimity
Solemnity *	Feeling of transcendence, inspiration. Thrills.	
Tenderness	Sensuality, affect, feeling of love	
Nostalgia	Dreamy, melancholic, sentimental feelings	
Calmness *	Relaxation, serenity, meditativeness	
Power	Feeling strong, heroic, triumphant, energetic	Vitality
Joyful activation	Feels like dancing, bouncy feeling, animated, amused	
Tension	Nervous, impatient, irritated	Unease
Sadness	Depressed, sorrowful	

Table 1: GEMS categories with explanations as used in the game. The categories marked with asterisk were modified.

We collected the following personal data about participants: age, gender, first language, level of English (Beginner, Intermediate, Advanced), musical preferences and current mood (on a Likert scale from 1 - very bad to 5 - very good). The question about mood was included because the mood might influence participant's perception [12]. For every piece of music the participant listened to we collected:

- Emotional labels.

- Whether the participant is familiar with the piece.
- Whether the participant liked or disliked the piece.
- The order in which GEMS categories were presented to the participant (randomized between participants).
- Optionally, a new emotion definition or an explanation of given emotional labels.

3.3 Game design

Designing a GWAP is a matter of finding a reasonable compromise between user engagement (the fun factor) and research data collection. As compared to other GWAPs, Emotify is designed in an unusual way. A standard method of creating engagement in a game with a purpose is making the player compete with an opponent over guessing each other's answer (be it a tag, a category, genre or the cursor position on valence-arousal plane). Since Emotify is about collecting data on induced musical emotion, participants should not feel that their answer must depend on someone else's choice. Thus, the game cannot be competitive, or the focus of competition would have to lie in another dimension than the research data we are collecting. E.g., we could let players guess the composer, meanwhile asking to fill in the emotions that they felt. But such a design distracts the participants from the research goal and might result in poor data. Therefore, we decided to provide players with a reward by giving feedback on their own answers both during and after the game. Emotify thus resembles a psychological test, where the participant has to provide complete and serious answers in order to get meaningful feedback. Involving a social network, we also provide the possibility of inter-player comparison. The feedback that a player gets during the game consists of his score (similarity to other players) and the possibility to compare the emotional labels that he assigned to the averaged answers of other players or to the answers of his friends on Facebook. After completing 10 songs the player can review what kind of emotions he associated with the music that he liked or disliked.

There are two versions of the game – a Facebook application (http://apps.facebook.com/emotify/) and a stand-alone version (http://emotify.org/) (for those who do not possess or want to use or create a Facebook profile). Figure 1 shows a screenshot of the game interface.

Figure 1: Emotify interface. Calmness and tenderness are selected and highlighted. An explanation is shown for the hovered button (relaxation, serenity, meditativeness).

The game flow is as follows:

1. Players may authorize through Facebook or enter the game from the stand-alone website. They fill in their personal data.
2. The player selects the genre, and may switch between genres at any time.

3. In every genre the player is presented with a random sequence of musical excerpts. Each excerpt is one minute long. If a player is invited by someone through Facebook, he is presented with the same sequence as the player that sent him the invitation.
4. After listening to the one minute fragment (all fragments are incipits, the beginning of the song), the player selects up to three emotions from the list of nine.
5. The player also may indicate whether he liked or disliked the music and whether this fragment seems familiar. He may also provide a new emotion definition (for example, courage), if the nine categories seem not enough.
6. The player can skip any fragment.
7. After 10 fragments (not counting skipped fragments), the player gets feedback.

4 Collected data description

In this section we will describe the collected annotations and participant profiles. This description concerns both data subsets (the whole dataset, **subset A+B**).

4.1 Annotations

The annotations collected from the game are stored in the following format:

1. Song id (a value from 1 to 400).
2. Binary values (0 or 1) for each of the emotional categories.
3. User liked the song (binary).
4. User disliked the song (binary).
5. Commentary on why the emotional categories above were chosen (textual, optional).
6. Suggestion of a different emotional category (textual, optional).
7. The personal data of the participant who provided this annotation (age, mood, gender and all the other personal details discussed above).

The annotations and sound files are accessible online.

4.2 Participant profiles

A total number of1595 participants (651 females and 944 males) took part in the study and 15356 labels were collected for 400 songs during 7975 listening sessions. The average age of participants was 30.33 years (sd = 11.74). Figure 2 shows the distribution of participants over age and gender. The figure displays a beanplot, a data visualization technique alternative to boxplot. A thick line shows the average values of each sample, and the shape indicates density.

Participants listed different languages as their mother tongues: 37% English, 20% Russian, 20% Dutch, the remaining 23% of the participants indicated 40 other languages. The style preferences were as follows: 61% Rock, 55% Classical, 44% Pop and 43% Electronic (multiple genres were allowed). 10% of the participants reported that their English language proficiency was on the beginner level, 27% had intermediate level and 63% were advanced. On average, they listened to 8 songs, and spent 13 minutes and 40 seconds playing the game (sd = 12.62). The actual time spent in the game differed a lot per player. As we were advertising a game through online media, there were many players who merely examined the game and quit almost immediately, but there were also devoted players who spent a lot of time listening to music. In the experiment, participants had to select one, two or three main emotions they felt after listening to a one minute excerpt. For 37% of samples they selected only one emotion, 30% obtained two emotional labels and 33% three emotional labels. There were no complaints about not being able to select more than three labels, but 13 participants complained that they couldn't find an emotional category which would correspond exactly to what they felt.

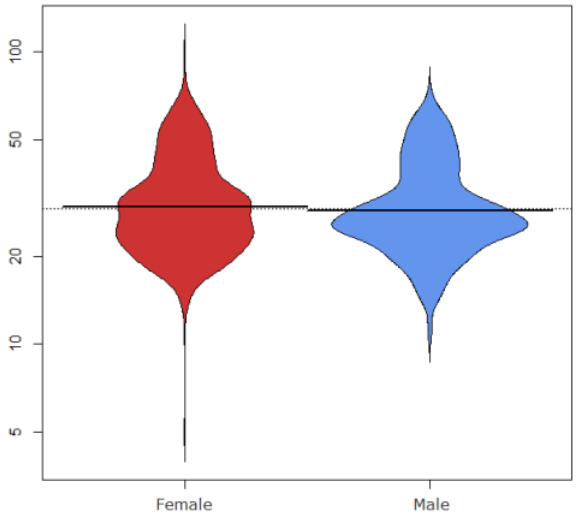

Figure 2: Beanplot: gender and age of participants.

4.3 Feedback questionnaire

556 participants filled out a feedback questionnaire after completing the game. They were asked to rate how difficult it was to use the proposed emotional scale on a scale from 1 to 5 (1 means "very easy") and on average they gave rating of 2.92 (sd=1.07, mode=3). On average they rated their liking of music on a scale from 1 to 5 (1 means "disliked completely") as 3.16 (sd=1.08, mode=4). Also, participants were asked to indicate which GEMS categories were most difficult to understand and to associate with the emotions they felt (see Table 2, column 2). The feedback confirmed the conclusion of Vuoskoski and Eerola [11] that solemnity and amazement (or wonder an transcendence, as they were previously named) might confuse participants.

Emotion	Considered unclear	Frequency of selection
Amazement	31%	13%
Solemnity	31%	20%
Tenderness	12%	18%
Nostalgia	10%	26%
Calmness	3%	30%
Power	11%	18%
Joyful activation	11%	25%
Tension	13%	23%
Sadness	4%	30%

Table 2: Emotions marked as problematic. Second column: considered unclear by percentage of respondents (n=556). Third column: category assignment by percentage of song listenings (n=7975).

In the game it was also possible to suggest a new emotional category or comment on existing ones. We received 425 such comments. Table 3 summarizes some of the most popular comments. As we can see from the table 3, by far the most frequent suggestion is boredom. Overall, there were a lot of comments about liking or disliking the music. The emotions that users lacked were impetus, anger, fear and humour. Some suggestions that only occurred once were 'silly', 'religious', 'awkward'.

Category	Comment examples	Occurrence frequency
Disliking the music	boring, boredom, bored, annoyance, annoyed, ennui	68
Neutral	neutral, no emotion, indifferent	10
Liking music	interesting, nice, good	10
Impetus	anticipation, call to action, determination, hopefulness, impatient	8
Anger	aggression, anger, wild	6
Humour	humour, humorous, sarcastic	6
Fear	scared, fear, tense scene in a movie	6
Contentment	content, contented, satisfied	5

Table 3: Emotions that were suggested by Emotify users

4.4 Amounts of annotations

The annotations produced by the game are spread unevenly among the songs. Firstly, subset A has many more annotations than subset B. The reason is that for subset A we had to collect at least 10 annotations per variable, and for subset B we decided to let at least 10 participants listen to each song. The second reason is that users had quite a lot of freedom in the game - they could choose the preferred genre and skip songs they didn't like. Therefore, certain genres and songs had more annotations than the rest in the end. Figure 3 illustrates the spread of annotations among 400 songs of subsets A and B.

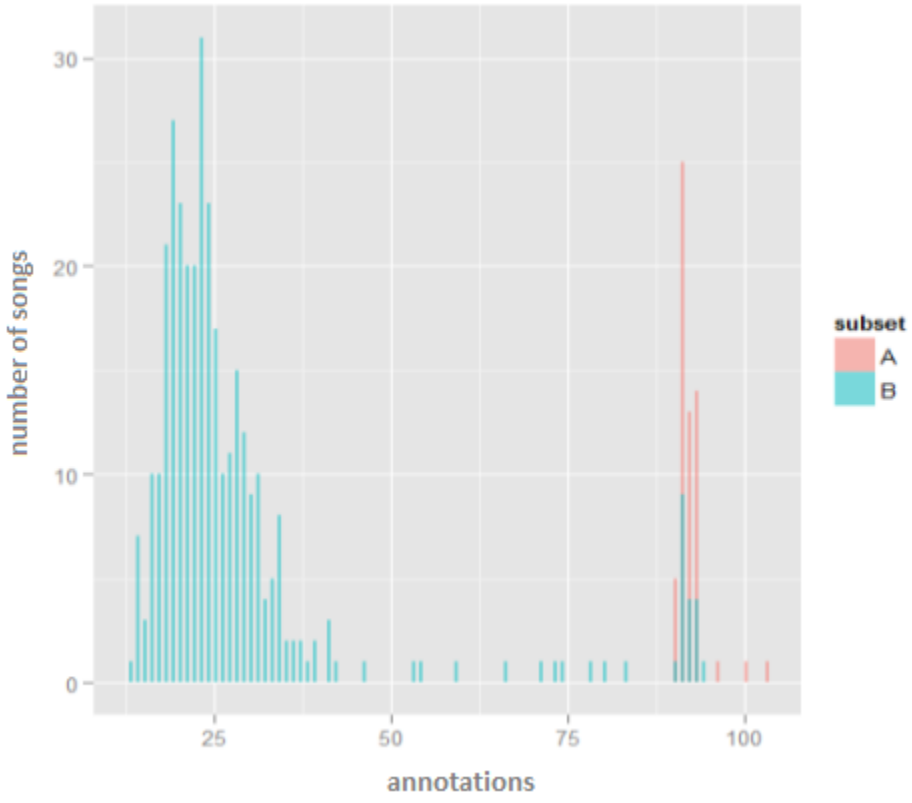

Figure 3: Histogram of the amount of annotations per song for subsets A and B.

The beanplots on Figure 4 show how the labels distributed between the four genres. From the plot we can see that classical music has most labels, and electronic music has least.

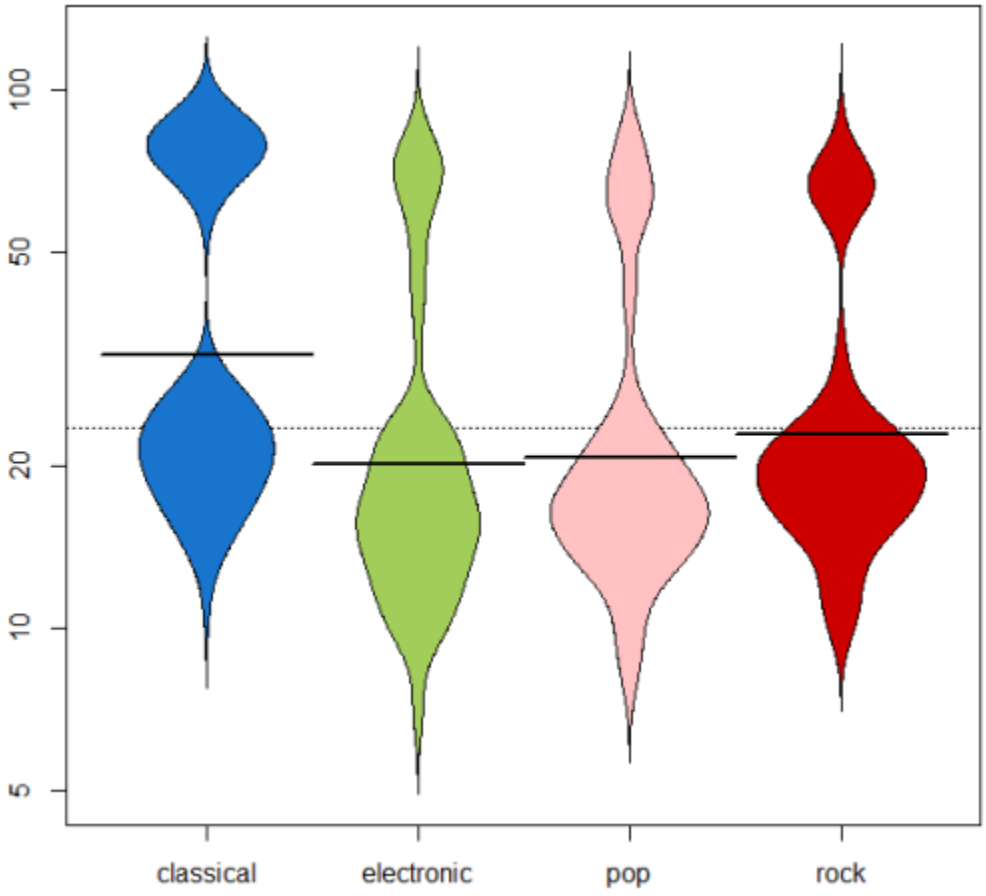

Figure 4: Number of labels for the four genres.

4.5 Aggregating data

The annotations provided by the game contain binary answers (emotion is either present or not, according to certain player). For each pair of participant and a song (we will call this a **listening**) there is a vector of emotions which were induced by this song in particular participant. For instance, song 1 was listened to by n participants, and they selected the following emotions: $(\text{amazement, power})_1$, $(\text{joyful activation, amazement})_2$ .. $(\text{power,tension})_n$. For computational purposes we need to aggregate these data, so that for each song we have just one estimation for each of the emotional categories. There are several ways to do this. The main decision that needs to be made is whether an emotion is given a fixed weight regardless of how many other emotions are selected, or whether each individual answer is weighted based on the number of selected emotions. Both decisions are plausible. On the one hand, selecting several emotions might indicate indecision and therefore it would be better to weight the answers. On the other hand, it might indicate emotional intensity and richness of the piece, in which case all the emotions should get an equal weight.

We are going to compare these two scores. The first score, which does not weight emotions per listening sessions, is calculated using Formula 1.

$$score^1_{ij} = \frac{1}{n}\sum_{k=1}^{n} a_k \tag{1}$$

where $score^1_{ij}$ is an estimated value of emotion i for song j, a_k is the answer of the k^{th} participant on a question whether emotion i is present in song j or not (the answer is either 0 or 1), and n is the total number of participants who listened to song j.

Formula 2 describes the second score, which takes into account how many other emotions a participant selected for this song:

$$score^2_{ij} = \frac{1}{n}\sum_{k=1}^{n} \frac{a^k}{\sum_{z=1}^{9} a^k_z} \tag{2}$$

where a^k_z are the answers to all emotions for this song by participant k, and the rest of the variables are the same as in the first measure.

Figure 5 illustrates the difference between the two scores. Each dot corresponds to a song. The horizontal axis shows the first score, the vertical axis shows the second score. From the plots we can see that the value of second score is always equal or smaller than the first. The difference between two scores is only big if an emotion is selected a lot for a certain song. An emotion that was often selected alongside other emotions would receive a smaller value than an emotion that was mostly selected independently.

4.5.1 Comparing graded answers and binary answers

In order to understand which of the scores is better we compared them to the answers we would receive if we asked participants to give answers on a Likert scale. We conducted a small scale experiment. We gave 15 rock songs to listen and rate to 6 participants. Each of them gave answers on a scale from 1 to 10 for each of the emotional categories. Table 4 shows the correlations between $score^1$ and $score^2$ grouped by emotion. The column titled coefficient shows the Pearson's correlation coefficient, or Pearson's r, for the first and second score.

For all the categories with a correlation significant on a 5% level (p-value < 0.05), the correlation was larger for $score^1$ (except for tension). Also, we can see that for all the emotions (except amazement and joyful activation), the correlation between Likert scale answers and the answers produced by the game are very high. The low correlation obtained in case of joyful activation can be explained by the fact that none of the songs among the 15 selected scored high on joyful activation, and this uniformity of the data deteriorated the performance metric. We conclude that the first score is closer to the original way of data collection. All further calculations in this report will be done using $score^1$.

4.6 Influence of button order on frequency of selection

For each of the participants the position of the buttons (the nine buttons with emotional labels on them) in the game interface was randomized. The buttons were placed in three rows of three, as shown on figure 1. We need to verify

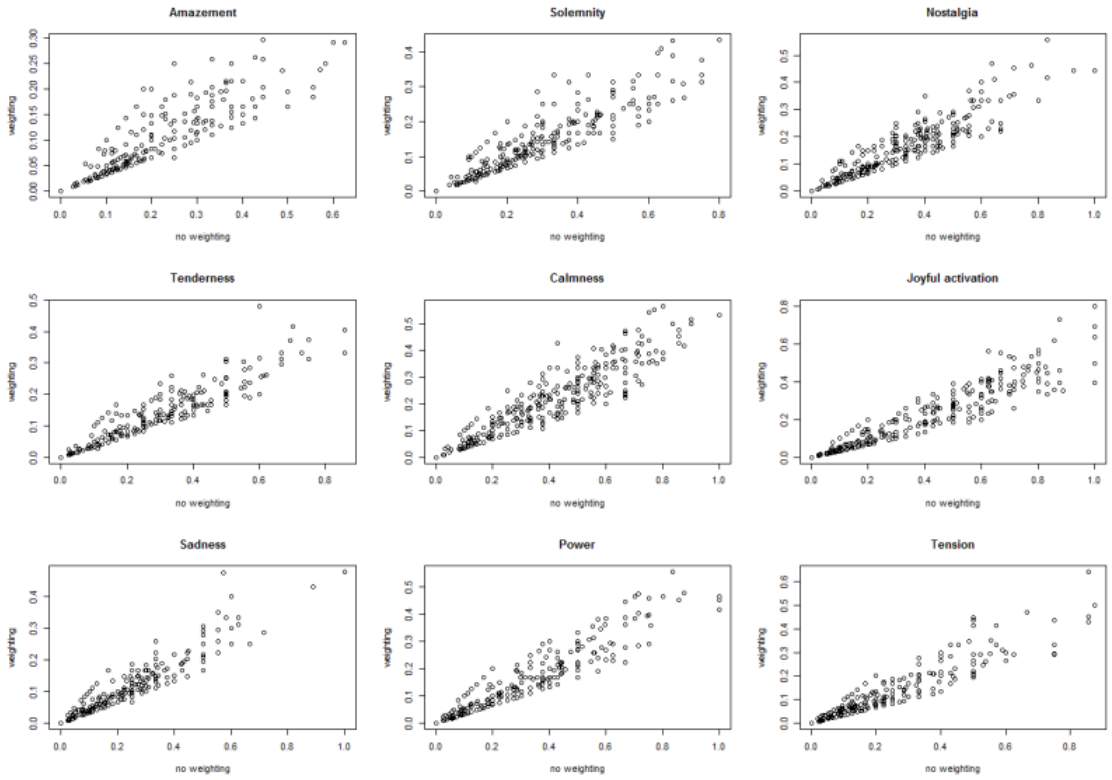

Figure 5: Each plot shows score[1] (unweighted, horizontal) against score[2](weighted, vertical) values for all songs, for a fixed emotion.

Emotion	*score*[1]	*score*[2]
	R	R
Amazement	0.37	0.42
Solemnity	0.72 *	0.62*
Tenderness	0.9 *	0.87*
Nostalgia	0.76 *	0.71*
Calmness	0.71 *	0.62*
Power	0.74 *	0.67*
Joyful activation	0.34	0.34
Tension	0.78 *	0.82*
Sadness	0.58 *	0.57*

Table 4: Correlation of *score*[1] and *score*[2] with Likert scale answers, the coefficients marked with asterisk are significant with p-value <0.05.

whether the buttons in certain positions were selected more often than buttons in other positions (regardless of the text on the button). Table 5 shows the frequencies with which buttons were selected. The position in the table corresponds to the button position on the screen. Since several buttons could be selected during one listening session, the percentages do not sum up to 100.

2021 | 19%	1840 | 17%	1969 | 19%
1916 | 18%	2027 | 19%	1946 | 19%
1804 | 17%	1862 | 18%	1816 | 17%

Table 5: Frequency of button selection (the absolute number of clicks and a percentage from all the listenings).

From examination of the table we can notice that the buttons in the lowest row were selected less frequently than the buttons in the first and second row. Table 6 shows the results of the pairwise Student's t-test. We can see that the difference between the second and third row is significant with p-value<0.05. The buttons in the lowest row were selected about 7% less often than the buttons above them.

1st and 3rd row		2nd and 3rd row	
1st row mean	1943	2nd row mean	1963
3rd row mean	1827	3rd row mean	1827
p-value	0.15	p-value	0.03

Table 6: T-test for button positions.

4.7 Influence of English language proficiency

For almost 62% of the game participants, the first language was another language than English. From these participants, 10% indicated that their level of English fluency is "Beginner", 27% indicated 'Intermediate' and 63% 'Advanced'.

The group of beginner-level participants was too small for their answers to be separately compared with other groups. This is why we studied the effect of removing those participants and computed intra-class correlation coefficients for each of the songs with and without beginner-level participants. Removing 'beginners' didn't affect intra-class correlation coefficients significantly.

5 Data analysis

In this section we describe the annotations in more detail and present the statistical analysis that was carried out on them. We are answering the following questions:

1. Do the personal parameters, such as gender, mood and liking the music, influence induced emotion, and how?
2. How consistent are the annotations, i.e. how similar is induced emotion between participants?
3. Are there significant differences between genres?
4. Are there redundant emotional categories in the chosen emotional model?

5.1 Influence of individual parameters on induced emotion

5.1.1 Influence of mood on induced emotion

The first hypothesis we test is whether people perceive music differently when their mood is better or worse. The participants' mood in the questionnaire ranged from 1 (very bad) to 5 (very good). We have conducted a Chi-square test on category selection frequencies sorted by participants' mood and found significant differences for the categories

sadness, tenderness and calmness. Table 7 shows how often a category was selected by a participant when he or she was in a certain mood. The clearest tendency is observed for sadness. The lower the participant's mood, the more often he or she selects sadness as an emotion he or she feels when listening to music. Participants who indicated that their mood is 'very bad' selected sadness almost twice as often as participants whose mood was indicated as 'very good'. A similar trend is observed for calmness - the lower the mood, the more calmness the music induces. A slight opposite trend is observed for amazement - the better person feels, the more amazement is induced by music.

Emotion	Participant's mood					Chi sq	P Value
	1	2	3	4	5		
Amazement	14%	12%	15%	16%	18%	9.1	0.05
Solemnity	17%	21%	22%	22%	24%	5.4	0.24
Tenderness	**23%**	**19%**	**20%**	**23%**	**18%**	**13.7**	**0.007**
Nostalgia	25%	27%	26%	28%	26%	3.4	0.48
Calmness	**56%**	**43%**	**40%**	**44%**	**44%**	**12.4**	**0.0297**
Power	20%	17%	19%	20%	21%	5.07	0.28
Joyful activation	23%	25%	29%	27%	29%	8.154	0.08
Tension	21%	15%	14%	15%	16%	6.424	0.16
Sadness	**28%**	**17%**	**14%**	**15%**	**15%**	**34.46**	**6.003e-07**

Table 7: Mood and frequency of selection of emotional category.

5.1.2 Influence of gender

In this section we test, whether females tend to report that they feel different emotions than males. We conduct a Chi-square test for each emotional category (by gender and genre) to find out whether the means are significantly different.

Table 8 shows how frequent a certain emotion was felt for a certain genre by a group of female and a group of male participants. On the right side of the table we see the P-values from a Chi-square test. The values significant on a 5% level are shown in red. The most significant difference was observed in pop music category, where only 8% of the male participants felt amazed by the music, as opposed to 18% of female participants. The same tendency is observed in other genres and emotional categories: whenever there is a difference, females select a category more often.

5.1.3 Influence of musical preferences on liking

Liking and disliking the music appears to be very important for induced emotions, and is even sometimes regarded as a musically induced emotion per se. In the previous section we mentioned that a lot of comments from the users explained whether they liked or disliked the music. There was also an option in the game to mark the song as liked or disliked. In this section we test how dependent liking and disliking the music in the game was on a preference for this musical genre.

From table 9 we see that in all cases people who report frequently listening to genre X, tend to like songs in genre X more and dislike those less than those who do not prefer this musical genre. Though the difference between these groups of listeners exist, it is not as big as might be expected, and for pop and electronic music the differences on disliking of the music were not even statistically significant (all the values in the table, except the values marked with asterisk, are significant on a 5% level). Therefore, we can't reliably predict whether a person will like the music just on the basis of his genre preferences.

Table 10 shows, how often were emotions selected in two conditions - when the music was liked or disliked. We conduct a Chi-square test and find, that these differences are significant (p-value < 0.0001). We can see that tension and sadness are felt five times more often, and amazement is felt four times less often when the person dislikes the music.

Genre	Frequency of emotion (males)				Frequency of emotion (females)				P-value			
	Classical	Rock	Pop	Electro	Classical	Rock	Pop	Electro	Classical	Rock	Pop	Electro
Amazement	**14%**	15%	**7%**	13%	**17%**	13%	**15%**	10%	**0.01**	0.49	**2.34e-07**	0.08
Solemnity	26%	14%	15%	22%	24%	14%	13%	23%	0.16	0.54	0.2	0.13
Tenderness	20%	19%	27%	7%	21%	18%	24%	12%	0.14	0.74	0.12	0.19
Nostalgia	26%	29%	32%	14%	24%	33%	36%	11%	0.54	0.09	0.06	0.07
Calmness	34%	24%	35%	29%	33%	27%	32%	25%	0.54	0.17	0.2	0.09
Power	13%	**20%**	10%	26%	15%	**24%**	13%	25%	0.4	**0.04**	0.06	0.72
Joyful activation	27%	23%	26%	28%	28%	24%	20%	29%	0.74	0.72	0.88	0.72
Tension	17%	20%	**15%**	36%	10%	18%	**20%**	40%	0.3	0.26	**0.009**	0.07
Sadness	17%	21%	22%	**9%**	18%	19%	24%	**14%**	0.75	0.45	0.2	**0.007**

Table 8: Frequency of emotion selection for different genre, for female and male participants

Genre	Regular Listeners		Non-Listeners	
	Liked songs	Disliked songs	Liked songs	Disliked songs
Classical	60%	4%	48%	12%
Rock	40%	24%	30%	35%
Pop	39%	26% *	30%	29% *
Electronic	37%	25% *	27%	30% *

Table 9: Liking and disliking music from different genres by regular and irregular listeners

Emotion	Liked music	Disliked music
Amazement	8%	2%
Solemnity	11%	6%
Tenderness	11%	5%
Nostalgia	12%	11%
Calmness	17%	12%
Power	10%	8%
Joyful activation	15%	8%
Tension	5%	27%
Sadness	6%	27%

Table 10: Frequency of emotion selection for liked and disliked music (as a percentage from all the liked/disliked listenings).

5.2 Listener agreement on emotional categories

From **subset A** only 25 songs possess at least one emotional category that is selected by the majority (more than a half) of the respondents. The highest percentage of respondents to select a category unanimously was 77%. The most frequent highly selected categories were calmness and joyful activation, both for 8 songs, tension, for 7 songs, and the least frequent were power, nostalgia and tenderness. The rest of the categories (amazement, solemnity and sadness) in most cases weren't selected by more than one third of participants unanimously. We assessed listener agreement on the proposed emotional terms. In order to do so we calculated Cronbach's alpha values for each of the emotional categories for all the 60 pieces in the dataset (see Figure 6).

Genre	Amazement	Solemnity	Tenderness	Nostalgia	Calmness	Power	Joyful Activation	Tension	Sadness
Classical	0.7	0.48	0.75	0.81	0.92	0.9	0.96	0.55	0.78
Rock	0.36	0.7	0.86	0.81	0.72	0.87	0.92	0.75	0.81
Pop	0.31	0.72	0.85	0.64	0.9	0.82	0.91	0.83	0.46
Electronic	0.48	0.72	0.85	0.6	0.78	0.82	0.87	0.75	0.7
Average	0.46	0.65	0.82	0.71	0.83	0.85	0.91	0.72	0.69

Table 11: Cronbach's alpha values for categories calculated on subset A.

In psychological research, Cronbach's alpha above 0.7 is viewed as an acceptable reliability indicator, and three categories do not pass that threshold: amazement, solemnity and sadness. The latter has rather high values in all genres except electronic music.

Table 11 shows values of Cronbach's alpha for each category and also the averaged (across genres) values.

We also calculate the Cronbach's alpha value for each of the songs from the subset A. Table 12 shows those values averaged by genre. As the Tukey HSD test showed, the differences were too small to conclude that the consistency of responses was different between genres.

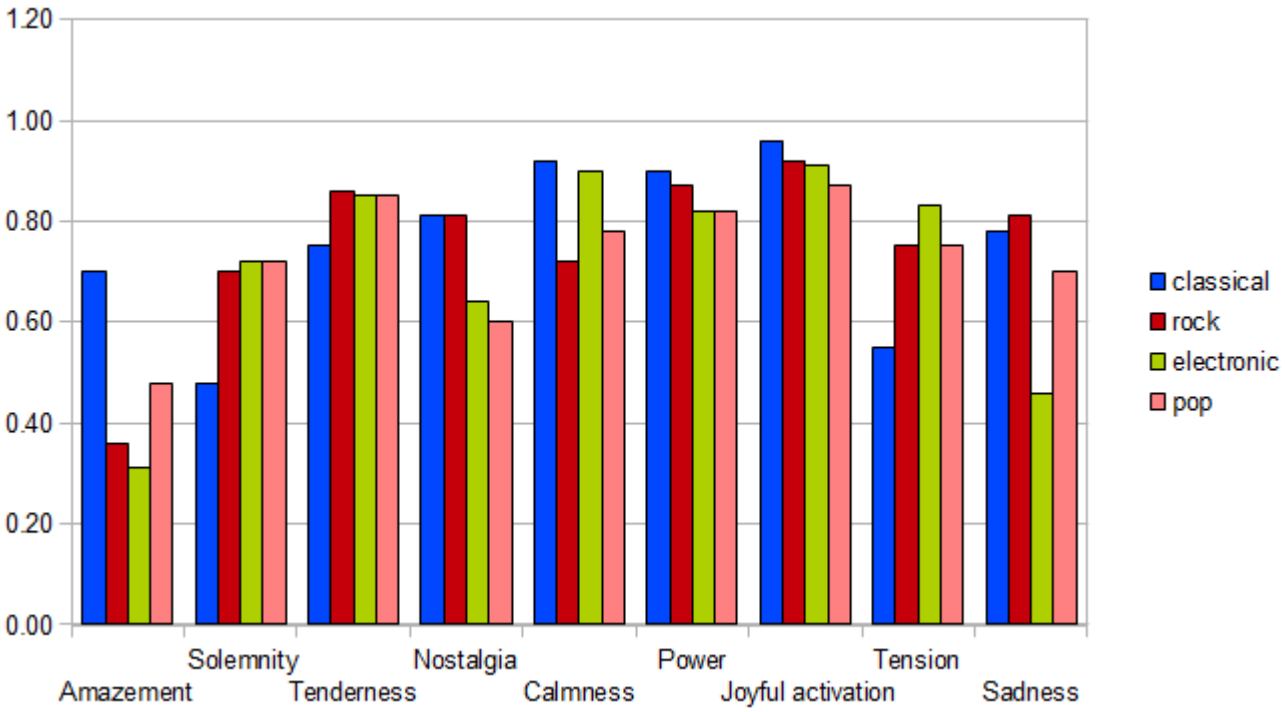

Figure 6: Cronbach's alpha for each category for each genre.

Genre	Cronbach's Alpha
Classical	0.84
Rock	0.72
Pop	0.8
Electronic	0.86

Table 12: Cronbach's alpha values for songs from subset A.

5.3 Effect of musical preference on response consistency

For each piece, participants could indicate whether they liked the music or not. Responding to this question was not obligatory, yet for more than a half of listening sessions we have this information. Figure 7 shows how preference is distributed across genres.

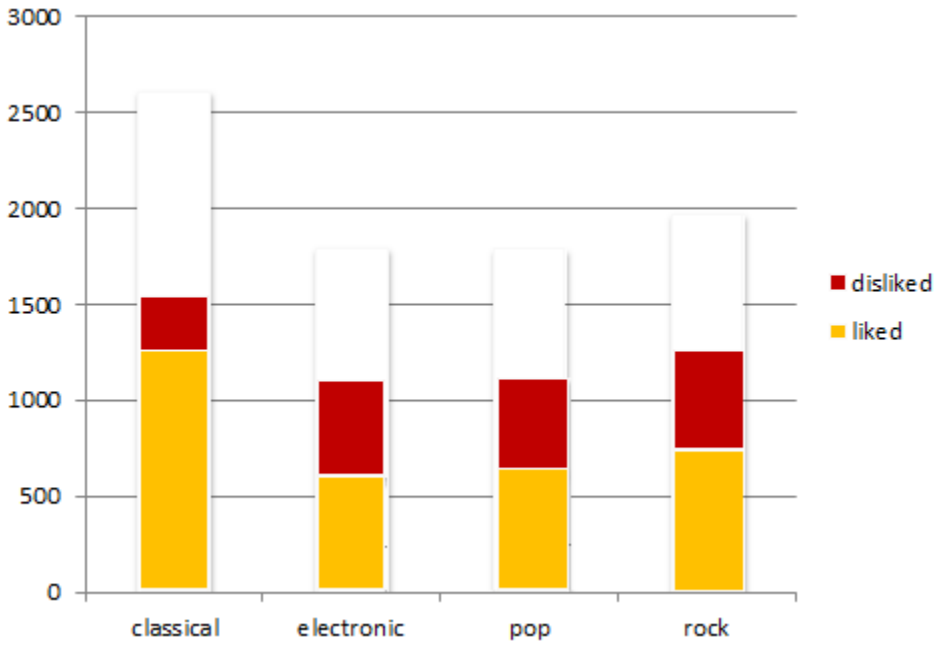

Figure 7: Liking and disliking indications.

Naturally, some of the songs were liked better by participants and some songs received more "dislikes" than likes. A positive dependency was discovered between the consistency of the ratings (as measured by intraclass correlation coefficients) and liking the music.

First, we selected the three least consistent and the three most consistent songs from subset A (table 13). The first three songs (inconsistent group) received comments such as "too formulaic", "horrible", "burned rubber". For the last three songs (the liked group) there was just one comment: "blues makes everyone enjoy it while experiencing divine sadness through good riffs ". It seems that the least consistent songs were also found least enjoyable. From table 13 we can clearly see that for the consistent group of songs, the ratio of likes and dislikes is much higher (the songs were enjoyed more often than not enjoyed).

Title	Genre	Cronbach's Alpha	Like/Dislike ratio
Young Again	Rock	0.1854219949	0.85
Persephone	Pop	0.3997159091	0.56
Self Possession	Rock	0.5261951164	1.36
The Spirit	Rock	0.9574060887	3
Trio Sonata in D minor - Allegro	Classical	0.9578780681	13.5
Suite in Fa Major - Adagio	Classical	0.9583095188	4.5

Table 13: 6 songs from subset A.

We computed the ICC (intra-class correlation) for all the responses, and excluded the responses from participants disliking the song. The remaining responses were more consistent (mean ICC = 0.18 as compared to ICC =0.16, with p-value = 0.01).

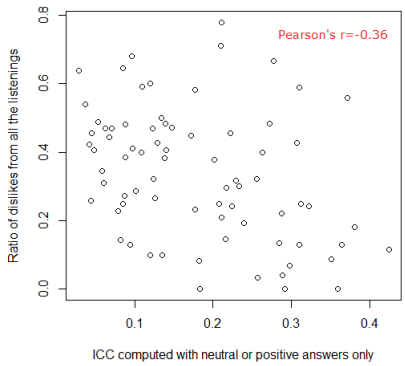

Figure 8: Scatterplot of song ICC and its dislike ratio (dislikes to likes) for **subset A**.

Even when the listenings which were disliked were excluded from the dataset, there still remained a correlation between the ratio of dislikes and response consistency, as shown in Figure 8. The scatterplot only shows **subset A**, because removing songs from **subset B** has a side-effect. As this subset has less data, ICC is more negatively affected by removing even more data points, even though the data points removed are less consistent.

We conclude that either people can understand an emotion of the song better when they like it, or people like the song more when it's easier to understand its emotion.

5.4 Genre differences on emotional categories

In this section we test whether certain emotions are more frequently induced by certain genres. For each of the 400 songs from **subset A+B** we calculated the score[1] for each of the emotions. Then we used an unrelated one-way ANOVA to test for the differences in variance. We found significant (p <0.01, df (degrees of freedom) = 3) differences in solemnity, tenderness, nostalgia, power, tension and sadness, and no difference in amazement, calmness and joyful activation. Figure 9 shows boxplots of the distribution of emotions per genre. Each boxplot shows how often a certain emotion was selected for this genre.

We conducted a Tukey HSD test to find out which combinations of genre and emotion differ from the rest. Tukey's test is a multiple comparison test, used to pairwise compare means of samples (it is an equivalent of a t-test for multiple samples).

We found, that between rock and pop there is no significant difference. We list other differences for the following six emotional categories:

- **Solemnity**. The genres that are different are classical (M=0.27) and electronic (M=0.29), as opposed to pop (M=0.15) and rock (M=0.15). This result suggests that solemnity is occurring more often in classical and electronic music, than in pop and rock music.
- **Tenderness**. Only electronic music (M=0.08) is significantly different from all other genres (M=0.23), suggesting that electronic music rarely induces tenderness.
- **Nostalgia**. Rock (M=0.31) and pop (M=0.35) genres differ from all the other genres (M=0.19).
- **Power**. Pop (M=0.16) and classical (M=0.16) music differ from rock (M=0.24) and electronic (M=0.3) music, suggesting that rock and electronic music is more powerful than pop and classical.
- **Tension**. Electronic music (M=0.28) is different from all the other genres (M=0.13), suggesting that electronic music induced tension more often.
- **Sadness**. Electronic music (M=0.11) is only significantly lower than pop (M=0.18).

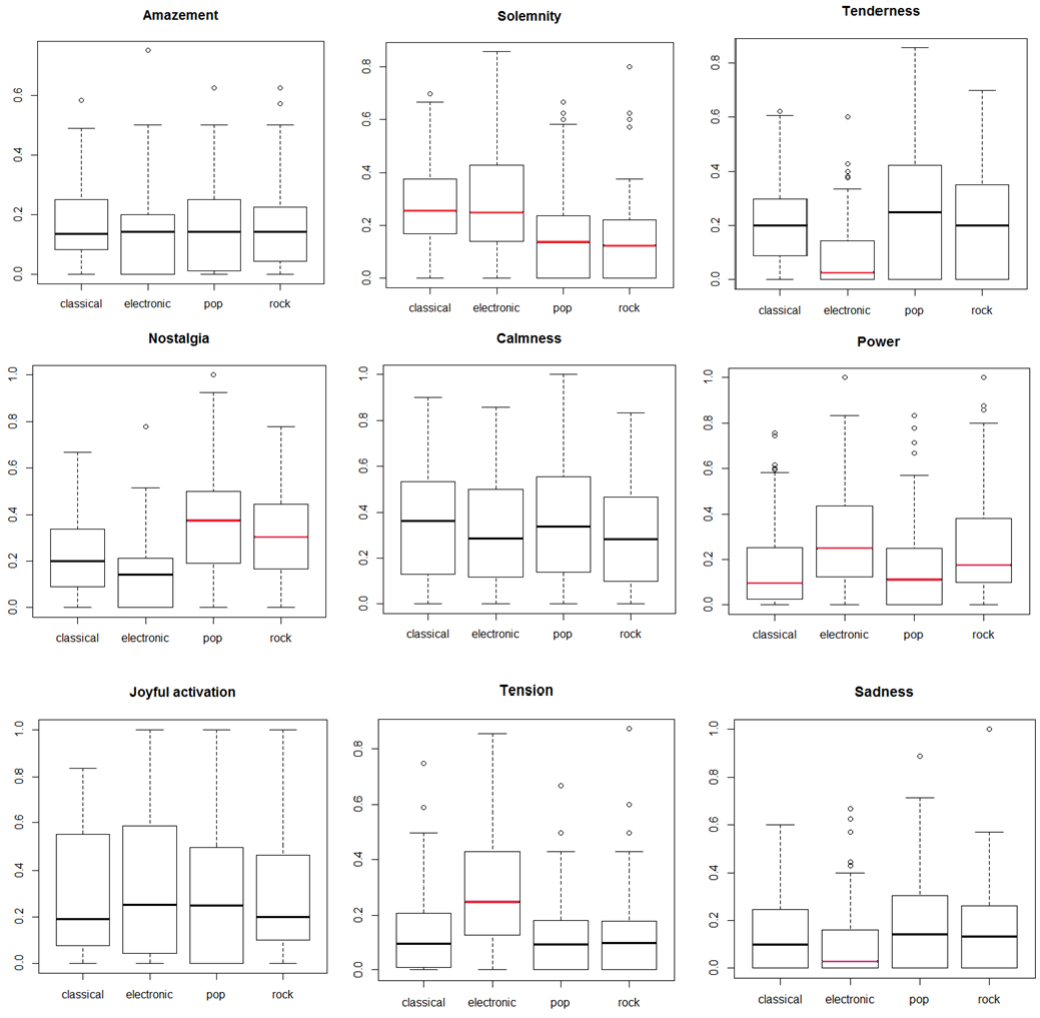

Figure 9: Boxplots showing how frequently emotion is induced in a certain genre.

5.5 Factor analysis

In order to find out whether the GEMS model contains redundant emotions, we do a factor analysis. Figure 10 shows a correlation matrix and a scatterplot matrix for 9 emotional categories. As far as data is non-normally distributed, we conducted a non-parametric test and calculated a Spearman's correlation coefficient. Each data point is a song, and both axes show score[1] for the corresponding emotional category. The number of asterisks indicates the p value of the correlation (*- <0.05, ** - <0.01, *** - <0.001).

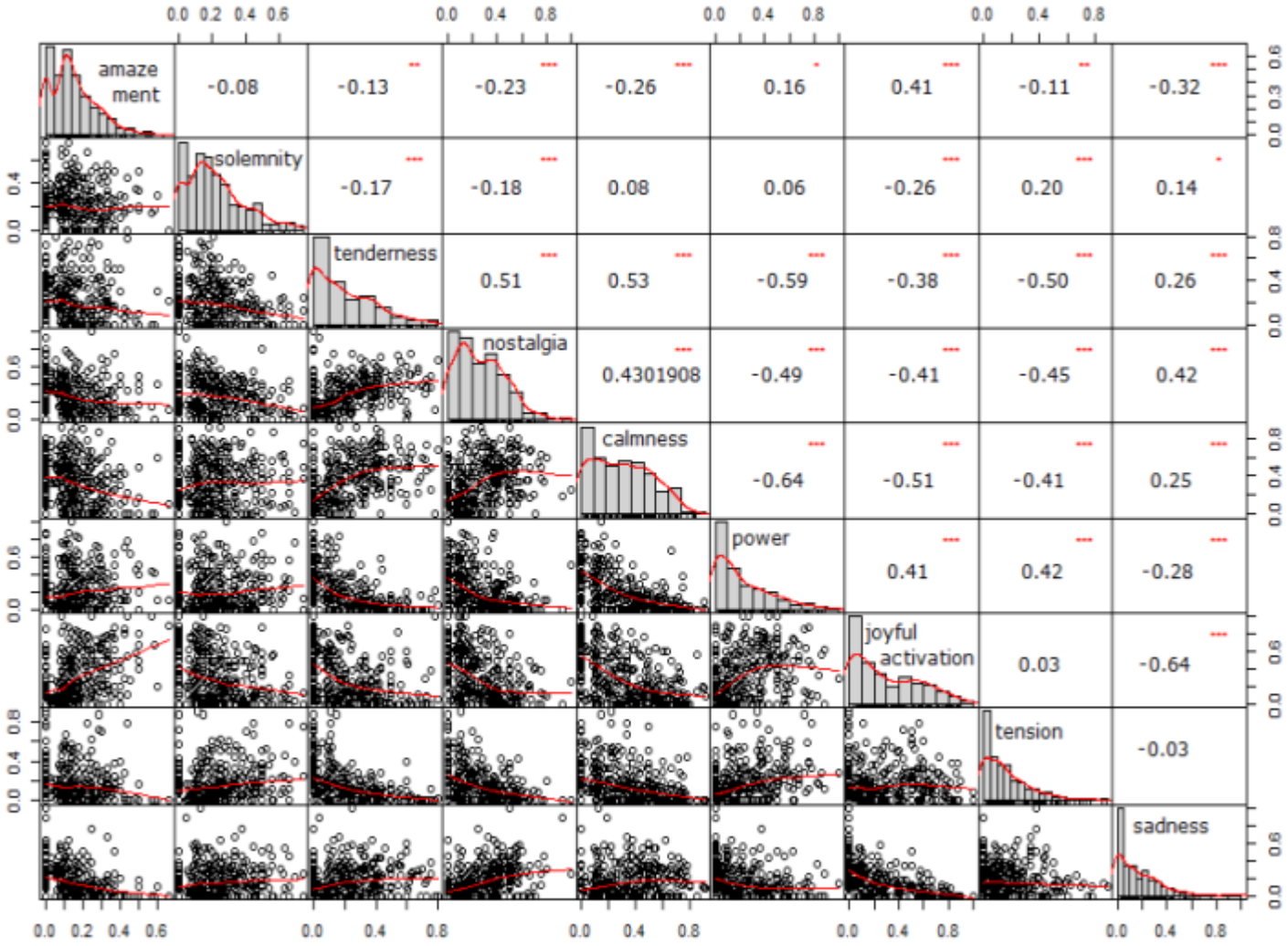

Figure 10: Correlation coefficients and scatterplots for emotional categories.

We see that some of the categories are strongly correlated. It means that they either were often selected together, or were often selected by different people for the same music (participants could not distinguish beween them, therefore they might be redundant). Prominent examples are: negatively correlated joyful activation and power (r=0.41, p<2.2e-16) tenderness and nostalgia (r=0.51, p<2.2e-16).

For a factor analysis we exclude one category (amazement) from our dataset, because it is least correlated with the rest of the emotions, and we have reasons to suspect that this is because (as it was explained above) this category is much more inconsistent and mostly dependent on participants' musical taste and mood.

According to the Scree test (a test based on component eigenvalues) there are four factors in the data. We extract and rotate them using varimax rotation. Table 11 shows the factor loadings on these four factors. The first component

correlates mostly with tenderness and calmness. The second component correlates with nostalgia and sadness. The third component correlates with tension and to a lesser degree with power. The fourth component correlates with solemnity.

	1	2	3	4
solemnity	0.00	0.08	0.11	0.95
tenderness	0.74	0.16	-0.16	-0.32
nostalgia	0.29	0.62	-0.48	-0.29
calmness	0.85	0.10	-0.17	0.17
power	-0.78	-0.15	0.22	0.01
joyful_activation	-0.51	-0.69	-0.20	-0.25
tension	-0.29	0.03	0.91	0.10
sadness	0.04	0.91	-0.01	0.06

Figure 11: Factor loadings of emotional categories on the four factors.

Let us compare these factors to the three superfactors from [7] . In [7], the first superfactor is sublimity, correlating with wonder, transcendence, tenderness, nostalgia and peacefulness. In our case all these categories are indeed correlated, except for solemnity, which is now a separate fourth factor, and amazement, which we excluded from analysis. The second superfactor (vitality) is correlated with power and joyful activation. In our model, power and joyful activation are indeed correlated (r=0.41), and joyful activation together with power negatively correlates with the first factor. The third superfactor is called unease and correlates with sadness and tension. This relationship does not hold for our data: sadness and tension are not correlated at all. Instead, sadness together with nostalgia contributes to the second factor.

We can see that some of the categories have similar loadings on factors. Nostalgia and tenderness are only distinguished by nostalgia correlating with sadness, and tenderness less so. Power and joyful activation can be distinguished by power positively correlating with tension, and joyful activation negatively correlating with solemnity.

6 Conclusion

This technical report describes an online game with a purpose and a dataset collected using this game. The data was collected using a modified GEMS questionnaire. We modified two categories, wonder and transcendence, by replacing with a subcategory, amazement and solemnity. Nevertheless, these two categories were still considered unclear by one third of participants, which means that this might not be a linguistic issue.

We measured consistency of responses of the game participants, and found that though people can agree on some emotions (joyful activation, power, tenderness, tension, calmness), other emotions are more subjective (sadness, amazement, solemnity). There were no significant differences on user agreement between musical genres.

Factor analysis indicated that GEMS could be best explained by four factors, contrary to original results[7]. Three categories - solemnity, sadness and tension - behave differently than they did in [7]. Sadness and tension are not correlated, and solemnity is not correlated with nostalgia or tenderness, but with tension.

We found that emotions, reported by a participant, are strongly affected by his mood, which might be an indication that the emotions reported were indeed induced, not perceived. Apart from mood, reported emotion were also strongly dependant on liking (and, in particular, disliking) the music, and, to a lesser degree, by gender. No influence of age or mother tongue was found. We found, that in order to predict induced emotion, several extra-musical factors must be taken into account, but it is not enough to rely on knowing the preferred genres to predict liking of the music.

As a result of this research, we created a dataset of non-preselected music in different genres, labelled with emotion, that can be used as a ground truth for computational modeling based on GEMS. Unfortunately, not all of the categories participants were sufficiently able to understand and agree upon. At least, amazement labels demonstrate very poor consistency and can't be used. In this research we also demonstrated, that for a successful computation model of induced emotion, a lot of external factors should be taken into account, such as current mood of participant, his musical

preferences on a more fine-grained scale than genre, gender. Integrating all this knowledge into a computational model is a direction for future research.

References

[1] K. Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48:246–268, 1936.

[2] X. Hu, S. J. Downie, C. Laurier, M. Bay, and A. F. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. *Proceedings of the 9th International Conference on Music Information Retrieval*, pages 462–268, 2008.

[3] J. Jaimovich. *Emotion Recognition from Physiological Indicators for Musical Applications*. PhD thesis, Queen's University Belfast, Belfast, United Kingdom, 2013.

[4] J. Jaimovich, N. Coghlan, and R.B. Knapp. Emotion in motion: A study of music and affective response. *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval*, pages 29–44, 2012.

[5] Y. Kim, E. Schmidt, and L. Emelle. Moodswings: A collaborative game for music mood label collection. *Proceedings of the International Conference on Music Information Retrieval*, pages 231–236, 2008.

[6] E. L. M. Law, L. von Ahn, R. B. Dannenberg, and M. Crawford. Tagatune: a game for music and sound annotation. *Proceedings of the 8th International Conference on Music Information Retreival*, pages 361–364, 2007.

[7] K. R. Scherer M. Zentner, D. Grandjean. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8:494–521, 2008.

[8] M. Mandel and D. Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[9] J.A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[10] K. Torres-Eliard, C. Labbe, and D. Grandjean. Towards a dynamic approach to the study of emotions expressed by music. *Proceedings of the 4th International ICST Conference on Intelligent Technologies for Interactive Entertainment*, pages 252–259, 2011.

[11] J. K. Vuoskoski and T. Eerola. Domain-specific or not? the applicability of different emotion models in the assessment of music-induced emotions. *Proceedings of the 10th International Conference on Music Perception and Cognition*, pages 196–199, 2010.

[12] T. Wildschut, C. Sedikides, J. Arndt, and C. Routledge. Nostalgia. content, triggers, functions. *Journal of Personality and Social Psychology*, 91:975–993, 2006.