Information Processing and Management 52 (2016) 115–128

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Studying emotion induced by music through a crowdsourcing game

Anna Aljanaki*, Frans Wiering , Remco C. Veltkamp

Utrecht University, Princetonplein 5, De Uithof, 3584 CC Utrecht, Netherlands

ARTICLE INFO

Article history:
Received 14 May 2014
Revised 19 January 2015
Accepted 22 March 2015
Available online 21 April 2015

Keywords:
Music information retrieval
Game with a purpose
Music induced emotion
Crowdsourcing

ABSTRACT

One of the major reasons why people find music so enjoyable is its emotional impact. Creating emotion-based playlists is a natural way of organizing music. The usability of online music streaming services could be greatly improved by developing emotion-based access methods, and automatic music emotion recognition (MER) is the most quick and feasible way of achieving it. When resorting to music for emotional regulation purposes, users are interested in the MER method to predict their induced, or felt emotion. The progress of MER in this area is impeded by the absence of publicly accessible ground-truth data on musically induced emotion. Also, there is no consensus on the question which emotional model best fits the demands of the users and can provide an unambiguous linguistic framework to describe musical emotions. In this paper we address these problems by creating a sizeable publicly available dataset of 400 musical excerpts from four genres annotated with induced emotion. We collected the data using an online "game with a purpose" Emotify, which attracted a big and varied sample of participants. We employed a nine item domain-specific emotional model GEMS (Geneva Emotional Music Scale). In this paper we analyze the collected data and report agreement of participants on different categories of GEMS. We also analyze influence of extra-musical factors on induced emotion (gender, mood, music preferences). We suggest that modifications in GEMS model are necessary.

1. Introduction

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. Implementing such methods is not a straightforward task, not only due to MER (music emotion recognition) limitations, but also because the emotional content of a musical piece is an intrinsically ambiguous part of it. Within music-related emotions, an important distinction can be made between emotions that are expressed by music (while listener is not necessarily feeling them), and the emotions felt by listener as a response to music (which we refer to as induced emotions). There is no doubt that music can indeed arouse strong emotions in listeners (Krumhansl, 1997; Rickard, 2004). Many people use music for purposes of emotional self-regulation and music therapy (Gabrielsson, 2011), and it is important to develop methods that could automatically categorize and select music by these criteria. In this paper, we contribute to solving this problem.

* Corresponding author. Tel.: +31 302537886.
E-mail addresses: A.Aljanaki@uu.nl (A. Aljanaki), F.Wiering@uu.nl (F. Wiering), R.C.Veltkamp@uu.nl (R.C. Veltkamp).

http://dx.doi.org/10.1016/j.ipm.2015.03.004

The relationship between expressed and induced emotion is not direct. In Gabrielsson (2002), Gabrielsson argues that expressed and induced emotion can relate in four ways: *positive*, *negative*, *no systematic relation* or *no relation*, thus, positive relation should not always be assumed. Also, though a qualified listener can nearly always recognize emotion expressed in the music, emotion induction is less frequent. Recent studies suggest that listeners experience strong emotions only about 55% of the time they spend listening to music (Juslin & Laukka, 2004), or that in 65% of the musical episodes music affects how they feel (Juslin, Liljestrom, Vastfjall, Barradas, & Silva, 2008). Emotional responses can be measured from self-report, expressive behavior and physiological responses (heart rate, skin conductivity, blood pressure, as well as biochemical responses) (Krumhansl, 1997; Rickard, 2004). In case of music, pronounced expressive behavior is not the rule, and, arguably, self-report is the most widely used and the most informative measure, because it provides information on the otherwise inaccessible cognitive part of emotion (Zentner & Eerola, 2011). In this study, we will use self-report to measure induced emotional responses to music.

Musical emotions are not directly translatable into words. There is still no consensus between researchers on the most suitable model, despite numerous attempts to find one (Vuoskoski & Eerola, 2013). The choice of model is essential to the performance of MER algorithms. A model that fails to describe the phenomenon precisely will result in poor agreement between listeners, conflicting musical cues associated to different emotions, and impede accuracy of prediction. On the other hand, a model that oversimplifies the problem might result in better agreement, but would be less useful for listeners. Currently, a wide variety of emotion ontologies can be found not only in research, but in music industry as well, from the valence–arousal model used by Musicovery,[1] or ten categories ranging from *happy* and *fun* to *dramatic* and *stressful* by Aupeo,[2] to no ontology at all, but providing emotional playlists non-systematically created by users[3] and user-generated tags[4] instead. In 2008, a new domain-specific model was suggested, Geneva Emotional Music Scales (GEMS) (Zentner, Grandjean, & Scherer, 2008). It was developed specifically to describe emotion induced by music, and, as compared to other categorical models, GEMS describes refined positive responses to music in much more detail. Since 2008, GEMS has been used in some smaller scale studies with promising results (Baltes, Avram, Miclea, & Miu, 2011; Jaimovich, 2013; Torres-Eliard, Labbe, & Grandjean, 2011; Vuoskoski & Eerola, 2010). There have been no large scale studies conducted using this model, and no public data have been released.

For our work we decided to employ the GEMS model. Our motivation here was twofold. First: our need for a big data set about induced emotion. Secondly, we also felt that additional studies of GEMS were needed. The reasons are that in the original study (Zentner et al., 2008) that GEMS is based upon, mostly classical music was used, and, moreover, the study was conducted in French, and the terms were translated to English.

Obtaining ground truth remains a challenging task for MER research, where both music copyright and costs of annotation (with music annotation being a particularly time-consuming task) pose problems. 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. In most cases it is 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 case of the MIREX audio mood recognition task (Hu, Downie, Laurier, Bay, & Ehmann, 2007)). 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, firstly, such a setup lacks ecological validity, and secondly, 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.

In this paper we approach the problems described above by collecting our ground-truth data using a game with a purpose (GWAP). We advertised our game, Emotify, through social networks, and it attracted a big and varied set of participants.

1.1. Contribution

In this paper, we describe an experiment designed to study emotions induced by music and to collect ground-truth data which could be used in training machine listening algorithms. We created a game with a purpose and collected annotations for 400 musical excerpts using a domain-specific emotional model GEMS (Zentner et al., 2008). The annotations are publicly available.[5] We examine the model's usability in online context and analyze comments and suggestions of game players. We report degree of agreement between listeners on different emotional categories and genres. We also study the extra-musical factors that influence induced musical emotion.

The paper is organized as follows. In Section 2, related research, concerning music-related emotional models, datasets using GEMS, and musical GWAPs, is reviewed. Section 3 presents methods and procedure (the GWAP) of the experiment. In Section 4, we describe the dataset that we collected and released as an outcome of this study. In Section 5, we analyze

[1] www.musicovery.com.
[2] www.aupeo.com.
[3] www.stereomood.com.
[4] www.last.fm.
[5] http://www.projects.science.uu.nl/memotion/emotifydata/.

the consistency of responses made using GEMS model, and the feedback and suggestions from game players. In Section 6, we analyze the extra-musical factors that influence emotion. In Section 7, we discuss the main findings. Section 8 concludes the paper and suggests future work.

2. Related work

The research in this paper closely concerns two fields: music psychology and music information retrieval. In the last decade, both fields showed a lot of interest towards affective studies. We will not review all the work done in this domain. For review of recent studies on affective music psychology, we refer the reader to Juslin and Sloboda (2011). For review of automatic music emotion recognition, please consult (Yang & Chen, 2012). Below, we will discuss only the papers that raise issues closely related to our research, such as existing emotional models, experiments which involved GEMS, and music-related GWAPs.

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 models can be divided in two groups: categorical and dimensional models. Categorical models present emotions as consisting of several basic clusters. Dimensional models arrange emotions in a continuous space along several (usually two or three) principal dimensions. The most widely used dimensional model, frequently employed in Music Information Retrieval, was proposed by Russell (1980). It consists of two dimensions: valence and arousal. The valence–arousal (V–A) 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. Many researchers have concluded that V–A model fails to capture all the variance reflected by music (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Collier, 2007; Ilie & Thompson, 2006). Moreover, the V–A model is not specific to music and was not created to reflect induced emotion, which we are interested in for the purposes of our study. Being domain-specific might be crucial in case of music. In Scherer (2004), Scherer argues that everyday *utilitarian* emotions should be distinguished from *aesthetic* emotions, induced by works of art. Aesthetic emotions are usually much more subtle, and do not coincide with everyday emotions (for instance, shame or guilt are almost never felt in response to music (Zentner et al., 2008)). Musical emotions can also be contradictory (e.g. bitter-sweetness) (Hunter, Schellenberg, & Schimmack, 2008). It is impossible to present these on the valence–arousal plane.

The earliest attempt to create a specifically musical categorical model of emotion was undertaken by Hevner (1936). She created an ontology of eight emotional clusters, such as humorous, pathetic and dreamy, where each cluster contained from six to eleven adjectives. In early seventies, L. Wedin proposed a three-dimensional model to describe musical emotion ("gaiety" vs. "gloom", "tension" vs. "relaxation" and "solemnity" vs. "triviality") (Wedin, 1972). None of these models were specifically developed for induced emotion. For a more comprehensive review of models of musical emotion we refer to Juslin and Sloboda (2011).

In 2008, a new domain-specific categorical emotional model called GEMS (Geneva Emotional Music Scale) was proposed (Zentner et al., 2008). GEMS is unique in that it addresses induced emotion, was created specifically for describing musical emotion, and has a level of granularity that other models do not provide. Zentner et al. conducted four consecutive studies to derive the model. First, a list of music-related terms was compiled both for induced and perceived emotion. It showed that these two types of emotion differ from each other, the major difference being the bias for positive emotions in case of induced emotions. In the following studies, a structure of music-induced emotions was examined through factor analysis of questionnaires. As a result, the GEMS scale was created. Through further factor analysis, shorter versions of the scale were added. The full GEMS scale consists of 45 terms, with shorter versions of 25 and 9 terms. These nine terms can in turn be grouped into 3 superfactors: vitality, sublimity and unease. Originally, the terms were collected in French, and later translated to English. In 2012, an additional research was conducted to improve the short GEMS scale (Coutinho & Scherer, 2012). In this research, the problem of classical music overrepresentation in the original work behind GEMS was addressed. The experiment confirmed the nine-factor structure of GEMS. It was suggested to add new terms related to feelings of harmony, interest and boredom. The final results from the study are still unpublished, so we used the original short nine term version of GEMS for our online game (see Table 1).

2.2. Experiments involving the GEMS model

In this section we will describe research papers that used GEMS as an underlying model for data collection. The biggest one, involving nearly 4000 participants, took place in 2010 (Jaimovich, Coghlan, & Knapp, 2012) in Dublin. Participants listened to music and reported their emotional state, using several self-assessment methods, 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 (2013). Unfortunately, due to a software error, the answers to GEMS questionnaire had to be discarded. 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 (Vuoskoski & Eerola, 2010). The most consistent ratings were produced

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

Emotional category	Explanation	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	

in the case of the two-dimensional valence–arousal model, while basic emotions and GEMS were less consistent, with GEMS's possessing both the most consistent (joyful activation, tension) and inconsistent (wonder, transcendence) categories. In 2011, K. Torres-Eliard et al. used GEMS for continuous emotion measurements (Torres-Eliard et al., 2011). Every rater controlled one GEMS dimension. Data on emotion *expressed* in 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 found to be in the range of good agreement (Cronbach's alphas ranged from 0.84 to 0.98). In Baltes et al. (2011), GEMS was used in a study of operatic performance, and this self-report measure showed significant correlation with physiological parameters (such as systolic blood pressure, respiratory sinus arrhythmia, etc.).

In the original study that introduced GEMS (Zentner et al., 2008), a small-scale experiment with 16 classical pieces showed that GEMS equips listeners with a more adequate instrument to measure musical emotion and results in better agreement than V–A or basic emotions. As this experiment was very small, based on one genre only, and the questions asked for V–A model were unconventional, this finding needs further investigation. In all further studies that we described above, none of the datasets was big, and the data is not publicly available. This is why we conclude that additional experimentation is needed.

2.3. Musical GWAPs

Collaborative online games are a popular way of collecting musical metadata, since it is easy to entertain people with music. Some of these games were proposed for the collection of descriptive labels (tags) on short musical fragments, such as MajorMiner (Mandel & Ellis, 2008) and TagATune (Law, von Ahn, Dannenberg, & Crawford, 2007), where the collected labels could also be mood-related. A specifically emotion-targeted GWAP called MoodSwings, for continuous emotional annotation of music, was created by Kim, Schmidt, and Emelle (2008). 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 GWAP 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. Methods

In this section, we will explain the design of our experiment: the structure of our musical dataset, the modifications to the GEMS questionnaire that we made in order to adapt it to an online game, and the design decisions behind the GWAP. The last element is discussed in more detail in Aljanaki et al. (2014a).

3.1. Music

In existing research on musical emotion, music is often selected for its strong and obvious emotional content (Jaimovich, 2013; Vuoskoski & Eerola, 2010; Zentner et al., 2008). In such a case, it is questionable how obtained results are comparable to non-preselected music. In our experiment, to provide ecological validity, we intentionally chose music randomly from a larger collection. We assembled a set of 400 musical pieces from the Magnatune recording company (magnatune.com), 100 pieces from each of 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. 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 (Schubert, 2007)). The music was reviewed manually and some recordings (around 2%) were removed because of insufficient quality.

We randomly divided our musical corpus into two subsets, maintaining the genre ratio (15 songs from each of the four genres). The smaller subset of the data (which will be called **subset A**) consists of 60 songs. The remaining 85% of the corpus

(**subset B**) consists of 340 songs. We collected different amounts of annotations for **subset A** and **subset B**. The smaller subset was intended to be used to investigate listener agreement on GEMS categories, the bigger subset was intended to be distributed as a public dataset of annotated music. In **subset A**, each song is annotated with at least 10 measurements per variable, which makes it at least 90 annotations per song, since there are nine questions in the questionnaire. We count all labels given to a song independently, thus if a person assigns 2 labels to a piece of music, we count each of those. For **subset B**, at least 10 people listened to and annotated each song.

3.2. Questionnaire

3.2.1. GEMS questionnaire adaptations

In order to adapt the GEMS questionnaire to an online game, we made several modifications. Originally, GEMS is designed to be answered using a Likert scale ranging from 1 to 5. 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 of several items typically ranging from "Not at all" to "Very much". This way of data collection is, however, very slow, requires quite some mental effort, and is not suitable for a dynamic online game. Therefore we modified the task and asked to select several labels from a list instead. This means that for each emotion we obtain one value, which is either 1 or 0 (emotion is present or not), which results, for each song, in a vector of 9 binary values.

We also restricted users on how many labels they could select, by explicitly demanding them to select no more than 3 labels. We did this because we wanted the players to select only the strongest emotions. As we abandoned the Likert scale, limiting the number of responses was the only way to measure the strength of emotion.

Following the findings from Torres-Eliard et al. (2011) and Vuoskoski and Eerola (2010), where it was discovered that participants have trouble with understanding certain categories of GEMS, we changed the wording of three GEMS categories by replacing them with one of the emotions from the list of explanatory synonyms that accompany each GEMS category. *Transcendence* was changed to *solemnity*, *wonder* to *amazement*, and *peacefulness* to *calmness* (see Table 1).

3.2.2. Personal questions

We also collected the following personal data about participants: age, gender, first language, level of English (Beginner, Intermediate, Advanced), musical preferences (we specifically asked the participants to report their preferences on the four selected genres, and added an open question where other preferred genres could be indicated), and current mood (on a Likert scale from 1 (*very bad*) to 5 (*very good*)).

3.2.3. Other information

For every piece of music the participant listened to we collected, apart from the data described above:

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

3.3. Game design

We launched a game with a purpose called *Emotify* in March 2013. As a platform, we used both a social network (a Facebook application (apps.facebook.com/emotify)) and a stand-alone website (www.emotify.org). Using a social network as the platform for a GWAP simplifies dissemination, but for those who do not possess or want to use or create a Facebook profile, we provided a stand-alone version. Fig. 1 shows a screenshot of the game interface. Involving a social network gave us the possibility to provide users with inter-player comparison in a non-competitive manner. The feedback that a player received during the game consisted of his score (similarity to other players) and the possibility to compare the emotional labels that he assigned to the average answers of other players or to the answers of his friends on Facebook. This comparison was only available after the player provided his own answers. After completing 10 songs, the players received feedback on which kinds of emotions they associated with the liked or disliked music. The gameflow is as follows.

1. The player authenticates through Facebook (or alternatively, enters the game from the stand-alone website) and provides personal details: age, gender, musical preferences, first language, level of English, and current mood. At this stage, the player is also provided instructions and is asked to report his or her personal emotions in response to music.
2. The player is randomly assigned to one of four musical genres (rock, pop, classical and electronic music) and can switch to any other if he or she so wishes. The player may also switch at any later time.
3. In every genre, the player is presented with a random sequence of musical excerpts, each one minute in length. If a player is invited by a friend through Facebook, he or she is presented with the same (whenever possible) sequence as the player who sent the invitation. This constraint is necessary in order to enable comparison between them.
4. After listening to the one-minute fragment, the player selects up to three emotions from a list of nine. This limitation should encourage players to think more carefully about the choices and name only the strongest emotions.

Fig. 1. Emotify interface. Calmness and tenderness are selected and highlighted. An explanation is shown for the hovered button. (Sensuality, affect, feeling of love.)

5. The player also may indicate whether he or she liked or disliked the music and whether he or she knows the song. The player may also provide a new emotion definition if none of the nine corresponds to what he or she is feeling.
6. At any time, it is possible to skip listening and go to another song or another genre.
7. There is a countdown from 10 to 1, saying that after 10 fragments the player will receive final feedback on his or her emotional perception of music. The countdown should encourage players to listen to at least 10 fragments to earn a "reward". Players may continue after listening to 10 fragments, but we prefer them not to do so, because feeling emotional content of music requires concentration and sensitivity, which is difficult to maintain for a long period of time.

Before starting with the game, the players were explained that they will be asked to describe what they feel in response to music, and also they were encouraged to skip the song if it fails to elicit any emotions in them. We tried to encourage more personal induced emotion responses by providing feedback in a style of psychological questionnaire.

For more details about the game we refer the reader to Aljanaki et al. (2014a).

4. Annotations

Creating a publicly accessible dataset was one of the main motivations of this study. We have made the data available online.[6] Below, we list statistics on game players and describe the size and contents of our dataset.

4.1. Participants

1778 participants (747 females, 1031 males) took part in the study and 16191 labels were collected for 400 songs during 8358 listening sessions. The average age of participants was 30.32 years (sd = 11.74). Participants listed different languages as their first language: 38% English, 19% Dutch, 19% Russian, the remaining 24% of the participants indicated 41 other languages (mostly European, with some Chinese, Hindi, etc.). The style preferences were as follows: 61% Rock, 55% Classical, 44% Pop and 43% Electronic (multiple genres were allowed). 11% of the participants reported that their English language proficiency was on the beginner level, 26% were on intermediate level and 63% were advanced. On average, they listened to 8 songs, and spent 13 min and 40 s playing the game (sd = 12.62). The actual time spent in the game differed a lot over all players. 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

[6] http://www.projects.science.uu.nl/memotion/emotifydata/.

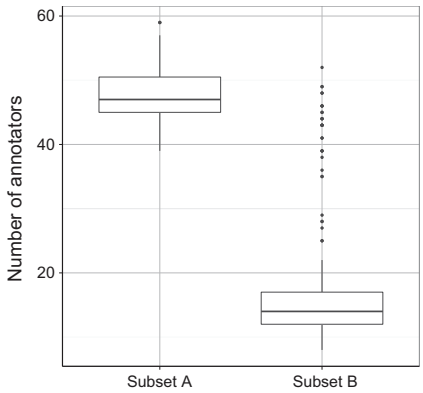

Fig. 2. Histogram of amount of annotators per song for subsets A and B.

Table 2
Frequency of button selection (the absolute number of clicks and a percentage from all the listenings).

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

complaints about not being able to select more than three labels, but a small amount of participants complained that they could not find an emotional category which would correspond exactly to what they felt, and about 7% of participants reported not being able to find exact emotion they felt in GEMS model and suggested new emotion definitions.

4.2. Amounts of annotations

The annotations produced by the game are spread unevenly among the songs, which is caused both by design of the experiment and design of the game. Participants could skip songs and switch between genres, and they were encouraged to do so, because induced emotional response does not automatically occur on every music listening occasion. Therefore, less popular (among our particular sample of participants) genres received less annotations, and the same happened to less popular songs. Boxplots in Fig. 2 illustrate the spread of annotations among 400 songs. On average, each song from **subset A** was annotated by 48 participants (sd = 4.46) and each song from **subset B** by 16 participants (sd = 8.7).

4.3. Confounding factors

4.3.1. Influence of button order on frequency of selection

For each of the participants, positions of the buttons (the nine buttons with emotional labels on them) in the game interface was randomized. The buttons were placed as shown in Fig. 1. We needed to verify, whether the buttons in certain positions were selected more often than buttons in other positions (regardless of the text on the button). Table 2 shows the frequencies of button selections in a listening session. 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.

By examining the table we notice that the buttons in the lowest row were selected less frequently than the buttons in the first and second row. Table 3 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. As far as buttons were randomized per every session, this effect should not influence the quality of annotations.

Table 3
T-test for button positions.

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

4.3.2. Influence of English language proficiency

For almost 62% of the game participants, the first language was another language than English. From these participants, 15% indicated that their level of English fluency is "Beginner", 36.5% indicated "Intermediate" and 48.5% "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" did not affect intra-class correlation coefficients significantly, which led us to believe that the level of their understanding was satisfactory enough for their answers to not degrade the quality of our dataset.

5. GEMS model comprehensibility and consistency of participant's responses

One of the objectives of our experiment was to test whether the GEMS model is suitable for large-scale music categorization and retrieval. In our game, we involve a varied sample of participants from different age groups and linguistic backgrounds, which resembles an actual composition of users of online music services. We explicitly ask them to provide feedback on using GEMS, and we also use implicit consistency measures.

5.1. Feedback questionnaire

After completing 10 excerpts, game players were offered to view their scores ("reward") and were asked to fill in a feedback questionnaire. 556 participants did so. They were asked to rate how difficult it was to use GEMS on a scale from 1 to 5 (where 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 (where 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 4, column 2).

From the feedback we can see that we did not manage to improve the situation (Vuoskoski & Eerola, 2010) with categories *wonder* and *transcendence* by giving them synonymous names. A very big number of participants (one third) considered them unclear. The most comprehensible categories were calmness and sadness. Those were most often selected as well. The rest of the categories were considered unclear by approximately one tenth of players.

5.2. Listener agreement on emotional categories

We collected especially a big amount of data for one subset of songs (**subset A**), in order to examine the inter-rater agreement (see Section 3). Here we will analyze these songs.

From the 60 songs of **subset A**, only 25 songs possessed at least one emotional category that was selected by 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 (7 songs), and the least frequent were power, nostalgia and tenderness. The rest of the categories (amazement, solemnity and sadness) in most cases were not selected by more than one third of participants unanimously. Though most of the songs failed to reach majority vote on any of the emotional categories, all of the songs demonstrate agreement that is much better than random.

To assess agreement, we calculated Cronbach's alpha per category (see Table 5). Cronbach's alpha is a coefficient of internal consistency, commonly used in psychometric tests (Cronbach, 1951). In psychological research, Cronbach's alpha above 0.7 is viewed as acceptable agreement, and three categories do not pass that threshold: amazement, solemnity and sadness. The last one has rather high values in all genres except pop music, and the first in the list has a high value in the classical genre. For solemnity, all values are above 0.7 except for the classical genre. We conducted the Tukey HSD test on Cronbach's alpha values between genres and did not find any significant differences.

Table 4
Second column: considered unclear by percentage of respondents ($n = 556$). Third column: how often an emotion was selected in listening sessions ($n = 8358$).

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 5
Cronbach's alpha values per category per genre (subset A).

Genre	Amazement	Solemnity	Tenderness	Nostalgia	Calmness	Power	J. activation	Tension	Sadness
Classical	0.70	0.48	0.75	0.81	0.92	0.90	0.96	0.55	0.78
Rock	0.36	0.70	0.86	0.81	0.72	0.87	0.92	0.75	0.81
Pop	0.31	0.72	0.85	0.64	0.90	0.82	0.91	0.83	0.46
Electronic	0.48	0.72	0.85	0.60	0.78	0.82	0.87	0.75	0.70
Average	0.46	0.65	0.82	0.71	0.83	0.85	0.91	0.72	0.69

Table 6
Emotions that were suggested by players of Emotify.

Group	Category	Examples	Occurrence frequency
1	Disliking the music	Boring, boredom, bored, annoyance, annoyed, ennui	68
2	Neutral	Neutral, no emotion, indifferent	10
3	Liking music	Interesting, nice, good	10
4	Impetus	Anticipation, determined, hopeful, impatient, call to action	8
5	Humor	Humor, humorous, sarcastic, silly	7
6	Anger	Aggression, anger, wild	6
7	Fear	Scared, fear, tense scene in a movie	6
8	Contentment	Content, contented, satisfied	5

5.3. Suggestions to modify the model

Players were given the opportunity to suggest a new emotional term that was missing from the model, or comment on existing ones. We received 437 such comments. Of them, 125 comments suggested new emotional terms, and the rest explained the reasons behind choosing from a list of GEMS terms or contained other notions.

Table 6 lists the most frequent semantic groups of comments, ordered by popularity. As we can see from the table, by far the most frequent suggestion is not related to emotion induced by music but to disliking it—boredom. In groups 1 and 2 we placed all the comments which referred to the fact that music failed to induce any emotion in the respondent. Though we asked the participants to skip the fragments which did not induce any emotion in them, not all the participants did so. Group 3 contains comments on liking the music. Groups 1–3 confirm the findings of Coutinho and Scherer (2012). In Coutinho and Scherer (2012), it was discovered that feelings of interest, boredom ("bored", "indifferent", "weary") and feelings of harmony and clarity are lacking from the model. Indeed, when reporting their induced emotion, participants find it very important to be able to report their interest, engagement and enjoyment (or, on the other hand, indifference, boredom and irritation from disliked music). These emotions can also be regarded as music-induced emotions and should be included in the model.

Other semantic groups of comments, not related to liking of disliking the music, are introduced in groups 4–8. *Anger* (group 6) and *fear* (group 7), according to Zentner et al. (2008), are often expressed by music, but are unlikely to be induced by it. Most likely, respondents were confusing what they perceive in music and what it induces in them. Impetus (group 4) was the next most suggested semantic group after feelings of interest and boredom. Less frequently suggested semantic groups were humor (group 5) and contentment (group 8). These emotions, along with boredom and interest, also are not covered by GEMS.

Some suggestions that only occurred once were "religious" and "awkward".

5.4. Factor analysis and correlation analysis of the categories

In this section we analyze the relationships between GEMS categories and conduct factor analysis to compare our findings to the three-factor structure obtained in Zentner et al. (2008).

5.4.1. Averaging labels across participants

In the game, we collected responses as binary vectors. For purposes of analysis, we needed to average them. We experimented with two scores to average the responses. In one, an emotion is given a fixed weight regardless of how many other emotions are selected. In the other, each individual answer is weighted based on the number of selected emotions. In the end, the first score was chosen on basis of experiment that showed that it is closer to answers collected using Likert scales. The first score is calculated using Formula (1).

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

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

Table 7
Correlations between emotional categories.

	Solemnity	Tenderness	Nostalgia	Calmness	Power	Joyful activation	Tension	Sadness
Amazement	−0.08	−0.13	−0.23	−0.26	0.16	0.41	−0.11	−0.32
Solemnity		−0.17	−0.18	0.08	0.06	−0.26	−0.20	0.14
Tenderness			0.51	0.53	−0.59	−0.38	−0.50	0.26
Nostalgia				0.43	−0.49	−0.41	−0.45	0.42
Calmness					−0.64	−0.51	−0.41	0.25
Power						0.41	0.42	−0.28
J. activation							0.03	−0.64
Tension								−0.03

5.4.2. Correlation analysis

We used the score described above to average the annotations, and calculated correlations (Spearman's correlation coefficient, as the data is not distributed normally) between GEMS categories (see Table 7). Before doing correlation analysis, we excluded the annotations from those listening sessions where participants indicated that they disliked the music. The reasons for doing so will be explained in Section 6.

Strong positive correlations mean that the correlated categories were either often selected together (co-occurring emotions), or were often selected by different people for the same music (confused and potentially redundant categories). Prominent examples are: tenderness and nostalgia with $r = 0.55$ and $p < 0.001$ (compare to Zentner et al. (2008) $r = 0.5$), power and joyful activation with $r = 0.41$ and $p < 0.001$ (compare to Zentner et al. (2008) $r = 0.38$). The strongest correlations are negative (sadness and joyful activation with $r = -0.64$ and power and tenderness with $r = -0.64$).

5.4.3. Factor analysis

We performed maximum likelihood exploratory factor analysis to compare our findings with results described in Zentner et al. (2008). Based on a Scree plot (a test based on component eigenvalues) and parallel analysis (a test that uses a comparison with random data set of the same size), we retained three factors that explained 54% of variance. We extracted and rotated them using oblique promax rotation (we have no reason to believe that the factors must be orthogonal). Table 8 shows the factor loadings of GEMS categories. The first component correlates mostly with sadness, nostalgia and tension, and negatively correlates with joyful activation and amazement. We will name this factor **Sadness**. The second component correlates with calmness, nostalgia and tenderness, and negatively correlates with power. This factor will be called **Peacefulness**. The third component correlates with nostalgia and tenderness and negatively correlates with tension and solemnity. This factor will be referred to as **Melancholy**.

In Zentner et al. (2008), the GEMS categories are grouped into three superfactors (see Table 1). Our factors do not conform to those exactly. In Zentner et al. (2008), tension and sadness contribute to one factor **Unease** and are correlated with $r = 0.22$. In our case, sadness and tension are not correlated at all ($r = 0.03$). Factor **Sublimity** is similar to our factor **Peacefulness**, but categories *amazement* and *solemnity* are not contributing to this factor anymore. In fact, *solemnity* in our study is negatively correlated with *tenderness* and *nostalgia*, while it was positively correlated in Zentner et al. (2008) ($r = 0.42$ and $r = 0.33$, respectively). See Table 7 for comparison.

6. Influence of personal factors on induced emotion

Personal and situational factors can significantly affect the emotion induced by music in the listener (Thompson, Graham, & Russo, 2005; Dibben, 2004). In this section, we will examine the degree of this influence for various factors.

6.1. Influence of mood

Recent findings suggest that people perceive music differently depending on their mood. In Dibben (2004), participants' arousal was manipulated with physical exercise prior to listening to music. Their self-reported induced emotion changed, while perceived emotion did not differ between groups that did or did not exercise. Likewise, in Zagrodski (2013) no effect of previous mood was observed for recognizing perceived emotion from music.

We were expecting to find an effect of mood in our induced musical emotion study. We conducted a Chi-square test on category selection frequencies grouped by participants' mood and found significant differences for the categories sadness, tenderness and calmness (Table 9). The clearest tendency is observed for sadness. The lower the participant's mood, the more often he or she selects sadness as an emotion induced by music. Participants who indicated that their mood was "very bad" selected sadness almost twice as often as participants whose mood was "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 the person feels, the more amazement is induced by music.

Table 8
Factor loadings of GEMS categories.

	Sadness	Peacefulness	Melancholy
Amazement	−0.42	−0.09	0.13
Solemnity	0.26	0.18	−0.56
Tenderness	0.16	0.31	0.41
Nostalgia	0.39	0.46	0.65
Calmness	0.05	0.93	−0.08
Power	−0.08	−0.53	−0.24
Joyful activation	−0.86	−0.22	0.19
Tension	0.24	−0.28	−0.60
Sadness	0.75	−0.24	0.16

Table 9
Mood and frequency of selection of emotional category. Bold if p-value < 0.05.

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	5.1	0.28
Joyful activation	23	25	29	27	29	8.2	0.08
Tension	21	15	14	15	16	6.4	0.16
Sadness	**28**	**17**	**14**	**15**	**15**	**34.5**	**6.003e−07**

Table 10
Frequency of emotion selection per different genre, per gender. Bold if p-value < 0.05.

Emotion chosen per session	Genre (males) (%)				Genre (females) (%)				p-value			
	C	R	P	E	C	R	P	E	C	R	P	E
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
J. 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**

6.2. Influence of gender

We conducted a Chi-square test on frequencies of each emotional category grouped by gender and genre (Table 10).

We did not find many significant differences. For example, in the pop music category, only 8% of the male participants felt *amazed*, as opposed to 18% of female participants. Females also more often felt *amazed* when listening to classical music, and more often indicated they felt *power* when listening to rock.

6.3. Influence of musical preference

Liking and disliking the music appears to be very important for induced emotions, and is even sometimes considered to be a musical emotion per se. From Table 11 we can see that selection of emotional category is strongly dependent on whether participant liked or disliked the music, especially for such categories as amazement, joyful activation, tension and sadness. It is important to understand whether we can rely on knowledge about preferred genres to predict whether someone would like the music.

From Table 12 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 this difference exists, it is not as big as might be expected, and for pop and electronic music the differences between liking and disliking the music were not even statistically significant.

Table 11
Category assignment by percentage of song listenings ($n = 8358$) depending on liking or disliking the music.

Emotion	When music is liked (%)	When music is disliked (%)
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 12
Liking and disliking music by genre preference. Significant on a 5% level except if marked with asterisk.

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*

6.4. Effect of liking the music on response consistency

For more than half of the listening sessions, participants reported whether they liked or disliked the music (they could skip this question if they did not have an opinion). There was a positive dependency between the consistency of the ratings (as measured by intraclass correlation coefficients) and liking the music. When the disliked listening sessions were excluded, the data showed more consistency (mean ICC = 0.18 as compared to ICC = 0.16, significant on t-test with p-value < 0.01).

Even when the disliked listenings were excluded from the dataset, there still remained a correlation between the ratio of likes and response consistency, shown in Fig. 3. The scatterplot only shows subset A, because subset B does not have enough songs to analyze it by removing part of them (without disliked listening sessions).

This means 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.

7. Discussion

In this paper we presented a GWAP for music induced emotion annotation and analyzed the data collected using this GWAP. We were aiming at improving automatic music emotion recognition methods by creating a new sizeable and public dataset, providing further testing to the GEMS model, and studying the extra-musical factors that contribute to emotion induction.

We modified the GEMS model because it was found that participants find some of the categories confusing (Torres-Eliard et al., 2011; Vuoskoski & Eerola, 2010). Two of the modified categories (*amazement* and *solemnity*, previously *wonder* and *transcendence*) still resulted in low agreement between participants (0.46 and 0.65 in terms of Cronbach's alpha, interpreted as Unacceptable and Acceptable, respectively), which might be caused by two issues. Firstly, low agreement might imply that these categories are inherently more subjective and depend on situational, cultural and other factors. The feedback questionnaire also showed that these two categories are less understood (*amazement* and *solemnity* were considered unclear by one third of the participants), which might be the second cause of low agreement. On the other hand, we found that for the rest of the categories, such as *tenderness*, *joyful activation*, *power*, and *calmness*, the inter-rater agreement is high, and these categories also are comprehensible enough according to feedback questionnaire.

When conducting factor analysis on our data, we found three factors similar to Zentner et al. (2008), but their structure could not be replicated. We did not observe that *tension* and *sadness* jointly load on any of the factors, and *amazement* and *solemnity* were not loading on the same factor with other emotions that contributed to factor **Sublimity** in Zentner et al. (2008). The reason may be that our experiment was conducted in a different language, and with different music.

We did not find any significant differences in inter-rater agreement across genres, and therefore we conclude that GEMS is equally suitable for describing all the four studied genres.

In this paper we also studied factors external to music. We found that the most important factor that should be taken into account when predicting induced emotion is liking or disliking the music. However, for our particular selection of broad music genres, the self-reported genre preferences failed to predict liking of the music with any accuracy. In our study we

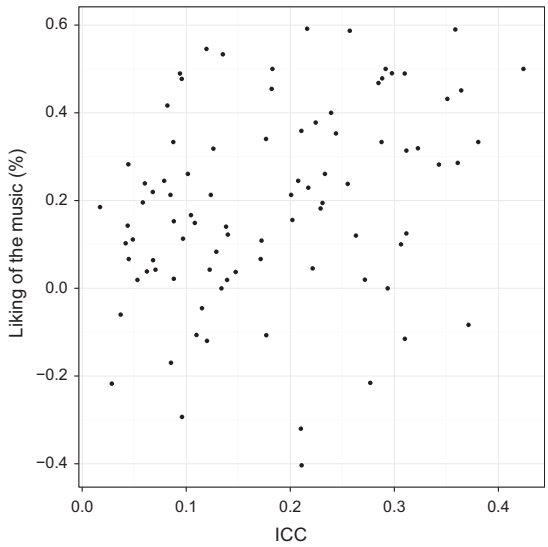

Fig. 3. Scatterplot of song ICC vs. its liking. Pearson's $r = 0.31$.

did not intend to control participants' liking of the music using their self-reported genre preferences, but this finding might be important for designing further experiments. We also found that participants' induced emotions were affected by their mood (to a considerable extent) and gender (to a lesser extent).

8. Conclusions and future work

One of the open research questions that we addressed with this study was whether music can express and induce a complex fine-grained range of emotions, or it is only possible to find crude counterparts of verbally expressible emotions in music. On basis of our study we conclude that there indeed is enough variety and expressive power in music to convey and induce such emotions as tenderness, nostalgia or peacefulness in such a way, that they can be distinguished by participants with sufficient inter-rater agreement. We also concluded that the GEMS model can be successfully used by participants from various linguistic backgrounds, though there obviously exists a lack of understanding concerning categories *wonder* and *transcendence*. It is a direction for future research to find how these categories could be modified.

Apart from this modification, it might also be worthwhile to study whether the GEMS could be augmented. Our study suggests that some of the nuances of emotional experience might be absent from GEMS model (8% of our participants were not able to use GEMS to describe their induced emotions). We agree with Coutinho and Scherer (2012), that feelings of boredom and interest must be added to the model, but also suggest that more semantic categories are lacking from it. Such semantic groups as impetus (call to action), humor and contentment were repeatedly named by the players of our game.

Another motivation for our study was collecting a dataset of music annotated with induced musical emotion which could be used as a ground-truth for MER research. The size of the dataset makes it possible to apply computational methods to explore the mechanisms underlying music emotional expressiveness, and to use these methods for automatic music classification and retrieval. A first study using our dataset has already been conducted (Aljanaki, Wiering, & Veltkamp, 2014b).

We hope that this work will contribute to solving the problem of finding the most appropriate model of musical emotion. Though this problem is important both for research on music psychology and music industry, currently it is far from being solved. We also hope that ground-truth data for such a rich emotional model like GEMS will be useful for MER research.

Acknowledgements

We would like to thank our colleagues Jan van Balen, Dimitrios Bountouridis, Bas de Haas, Marcelo Rodriguez Lopez and Anja Volk for the valuable discussion and advice on the design of the experiment. We are also very grateful to all the participants who devoted their time to playing the game, sent invitations to others, and otherwise helped with the data collection. We also thank the reviewers for their valuable comments that helped to improve the article.

This work was financially supported by the FES project COMMIT/.

References

Aljanaki, A., Bountouridis, D., Burgoyne, J. A., van Balen, J., Wiering, F., Honing, H., et al. (2014). Designing games with a purpose for data collection in music research. Emotify and Hooked: Two case studies. In *Lecture notes in computer science*, pages 29–44, 2014.

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2014). Computational modeling of induced emotion using GEMS. In *Proceedings of the 15th international society for music information retrieval* (pp. 373–378).
Baltes, F. R., Avram, J., Miclea, M., & Miu, A. C. (2011). Emotions induced by operatic music: Psychophysiological effects of music, plot, and acting: A scientist's tribute to Maria Callas. *Brain and Cognition, 76*(1), 146–157.
Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion, 19*, 1113–11139.
Collier, G. L. (2007). Beyond valence and activity in the emotional connotations of music. *Psychology of Music, 35*(1), 110–131.
Coutinho, E., & Scherer, K. R. (2012). Towards a brief domain-specific self-report scale for the rapid assessment of musically induced emotions. In *12th International conference of music perception and cognition (ICMPC12)*.
Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.
Dibben, N. J. (2004). The role of peripheral feedback in emotional experience with music. *Music Perception, 22*(1), 79–115.
Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different? *Musicae scientiae, 5*(1), 123–147.
Gabrielsson, A. (2011). *Strong experiences with music: Music is much more than just music*. Oxford University Press.
Hevner, K. (1936). Experimental studies of the elements of expression in music. *American Journal of Psychology, 48*, 246–268.
Hu, X., Downie, S. J., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 MIREX audio mood classification task: Lessons learned. In *Proceedings of the 9th international conference on music information retrieval* (pp. 462–268).
Hunter, P. G., Schellenberg, E. G., & Schimmack, U. (2008). Mixed affective responses to music with conflicting cues. *Cognition & Emotion, 22*(2), 327–352.
Ilie, G., & Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception, 23*(4), 319–329.
Jaimovich, J. (2013). *Emotion recognition from physiological indicators for musical applications*. PhD thesis, Queen's University Belfast, Belfast, United Kingdom.
Jaimovich, J., Coghlan, N., & Knapp, R. B. (2012). Emotion in motion: A study of music and affective response. In *Proceedings of the 9th international symposium on computer music modeling and retrieval* (pp. 29–44).
Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research, 33*(3), 217–238.
Juslin, P. N., Liljestrom, S., Vastfjall, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: Listener, music, and situation. *Emotion, 8*(5), 668–683.
Juslin, P. N., & Sloboda, J. (2011). *Handbook of music and emotion: Theory, research, applications*. Oxford University Press.
Kim, Y., Schmidt, E., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection. In *Proceedings of the international conference on music information retrieval* (pp. 231–236).
Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology, 51*(4), 336–352.
Law, E. L. M., von Ahn, L., Dannenberg, R. B., & Crawford, M. (2007). TagATune: A game for music and sound annotation. In *Proceedings of the 8th international conference on music information retrieval* (pp. 361–364).
Mandel, M., & Ellis, D. (2008). A web-based game for collecting music metadata. *Journal of New Music Research, 37*(2), 151–165.
Rickard, N. S. (2004). Intense emotional responses to music: A test of the physiological arousal hypothesis. *Psychology of Music, 32*(4), 371–388.
Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178.
Scherer, K. R. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *The Journal of New Music Research, 33*(3), 239–251.
Schubert, E. (2007). The influence of emotion, locus of emotion and familiarity upon preference in music. *Psychology of Music, 35*(3), 499–515.
Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica, 156*, 177–201.
Torres-Eliard, K., Labbe, C., & Grandjean, D. (2011). Towards a dynamic approach to the study of emotions expressed by music. In *Proceedings 4th international ICST conference on intelligent technologies for interactive entertainment* (pp. 252–259).
Vuoskoski, J. K., & Eerola, T. (2010). Domain-specific or not? The applicability of different emotion models in the assessment of music-induced emotions. In *Proceedings of the 10th international conference on music perception and cognition* (pp. 196–199).
Vuoskoski, J. K., & Eerola, T. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception, 30*(3), 307–340.
Wedin, L. (1972). A multidimensional study of perceptual–emotional qualities in music. *Scandinavian Journal of Psychology, 13*(1), 241–257.
Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: a review. *ACM Transactions on Intelligent Systems and Technology, 3*(3), 1–30.
Zagrodski, M. (2013). Influence of musical context on the perception of emotional expression of music. In *Proceedings of the 3rd international conference on music & emotion*.
Zentner, M., & Eerola, T. (2011). *Handbook of music and emotion: Theory, research, applications, chapter self-report measures and models*. Oxford University Press.
Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion, 8*(4), 494–521.