DO ONLINE SOCIAL TAGS PREDICT PERCEIVED OR INDUCED EMOTIONAL RESPONSES TO MUSIC?

Yading Song, Simon Dixon, Marcus Pearce
Centre for Digital Music
Queen Mary University of London
firstname.lastname@eecs.qmul.ac.uk

Andrea Halpern
Department of Psychology
Bucknell University
ahalpern@bucknell.edu

ABSTRACT

Music provides a powerful means of communication and self-expression. A wealth of research has been performed on the study of music and emotion, including emotion modelling and emotion classification. The emergence of online social tags (OST) has provided highly relevant information for the study of mood, as well as an important impetus for using discrete emotion terms in the study of continuous models of affect. Yet, the extent to which human annotation reveals either perceived emotion or induced emotion remains unknown. 80 musical excerpts were randomly selected from a collection of 2904 songs labelled with the Last.fm tags "happy", "sad", "angry" and "relax". Forty-seven participants provided emotion ratings on the two continuous dimensions of valence and arousal for both perceived and induced emotion. Analysis of variance did not reveal significant differences in ratings between perceived emotion and induced emotion. Moreover, the results indicated that, regardless of the discrete type of emotion experienced, listeners' ratings of perceived and induced emotion were highly positively correlated. Finally, the emotion tags "happy", "sad" and "angry" but not "relax" predicted the corresponding experimentally provided emotion categories.

1. INTRODUCTION

Music provides a powerful means of conveying and evoking feeling, and has attracted increasingly significant research interest in the past decades [5]. People report that their primary motivation for listening to music lies in its emotional effects [10]. A study of recreational activities (watching television, listening to music, reading books, and watching movies) indicated that people listen to music more often than any of the other activities [18]. The ability of identifying emotional content is established at very early age; a 5-year-old child can discriminate happiness and sadness by tempo and mode [2]. Although a wealth of research has been performed on the study of music and emotion [5], many problems remain unsolved.

In the study of emotion and music listening, one fundamental distinction is between induced emotion (also known as felt emotion), which is the emotion experienced by the listener, and perceived emotion (also known as expressed emotion), which is the emotion recognised in the music [8]. However, separating induced emotion from perceived emotion is not always straightforward. Previous studies have suggested that music induces emotions similar to the emotional quality perceived in the music. Generally, induced emotion is more subjective and perceived emotion tends to be more objective [7, 11] and it was found that the agreement of listener ratings of joy and sadness was higher than those for anger and fear [22].

To describe musical emotions, two well-known and dominant models have arisen: the discrete model (also known as the categorical approach) and the dimensional model (also known as the valence-arousal model). The discrete, or categorical, model describes all emotions as being derived from a limited number of universal and innate basic emotions such as anger, happiness, sadness and fear [6,17]. In contrast, the dimensional model considers all affective terms as arising from independent neurophysiological systems: one related to valence (a pleasure-displeasure continuum) and the other to arousal (activation-deactivation) [19]. On one hand, the dimensional model has been criticised for its lack of differentiation when it comes to emotions that are close neighbours in the valence-activation space, and for the limitation that participants can express their responses in terms of only two dimensions. On the other hand, the discrete model is criticised as being inadequate to describe the richness of emotional effects. Both theoretical frameworks, the categorical model and dimensional model, have however received empirical support [12, 24], and a comparison of the two models was presented by Eerola [4]. In recent years, a novel music-specific model derived from the Geneva Emotion Music Scale (GEMS), has been developed for specifically music-induced emotions, which consists of 9 emotional scales - wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension and sadness [28].

Alongside the emergence of music discovery websites such as Last.FM [1] in the past decade, social tags have received increasing interest for the study of music and emotion [3, 15, 25]. Social tags are words or groups of words

[1] http://www.last.fm

supplied by a community of internet users. They are more and more commonly used to aid navigation through large media collections [27], allowing users to get a sense of what qualities characterise a song at a glance [9]. Compared with traditional human annotation by experts, semantic tags provide large-scale, cost-efficient, rich and easily accessible source of metadata [23]. In addition, the information they provide is highly relevant to music information retrieval, including genre, mood and instrument, which account for 70% of the tags [13].

Though the use of social tags is a powerful tool which can assist searching and the exploration of music [14], several problems with tags have been identified, such as the "cold start" problem (new or unknown music has no tags), noise, malicious tagging, and bias towards popular artists or genres [13]. There are a number of incentives and motivations for tagging, such as to aid memory, provide context for task organisation, social signalling, social contribution, play and competition, and opinion expression [1]. However, we know very little about the criteria on which tagging is based.

To our knowledge, the two facets of emotion communication (perceived emotion and induced emotion) in music have rarely been studied in combination with semantic tags. The purpose of this paper is to explore the association between human-annotated tags and emotional judgements in perceived emotion and induced emotion based on the dimensional model. This study also helps the mapping and modelling of mood tags on the valence-arousal space. In this paper, the following research questions are examined: (1) How do induced emotion and perceived emotion differ from each other in the ratings of valence and arousal for a 2-dimensional model of emotion? (2) How well do semantic emotional tags reflect listeners' perceived emotion and induced emotion? (3) To what degree can the emotional tags be used to select stimuli for the study of music and emotion?

2. METHOD

2.1 Participants

Forty-seven English-speaking participants (male: 20; female: 27) took part in this study. They were recruited through various email lists (e.g. school lists, professional lists, and social media), and had ages ranging from 15 to 54 years (age <18: 1; age 18-24: 22; age 25-34: 21; age 35-44: 1; age: 45-54: 2) with various educational (e.g. undergraduate, postgraduate), cultural (e.g. British, American, French, Chinese, Canadian, Italian, Greek, Sri Lankan) and musical training backgrounds (musician and non-musician).

To assess the participants' musical expertise, the Goldsmiths Musical Sophistication Index questionnaire (GOLD-MSI) was given [16] (see section 2.3). Participants' *musical training* (life history of formal musical training) was calculated using a provided template [2] giving a scale from 9 to 63 (no formal training to formal training). A summary of the responses can be found in Table 1.

[2] http://www.gold.ac.uk/music-mind-brain/gold-msi/

Skill	Min	Max	Median	SD
Musical Training	17	42	29	6.2915

Table 1. Summary of musical training

2.2 Stimuli

The stimuli were selected from a collection of 2904 excerpts retrieved from Last.FM and 7Digital [3] which have been used previously in music and emotion studies [20,21].

Each excerpt had been tagged on Last.FM with one of the four words "*happy*", "*sad*", "*angry*" and "*relax*". We randomly chose a total of 80 excerpts from these four categories (n=20 from each category). The musical excerpts ranged from recent releases back to 1960s, and covered a range of Western popular music styles such as pop, rock, country, metal and instrumental. Each excerpt was either 30 seconds or 60 seconds long (as provided by 7Digital), and it was played from a standard mp3 format file (bitrate: 128 kbps or 64 kbps; sample rate: 22050 kHz or 44100 kHz). This 80-excerpt dataset will be made available [4] , to enable further studies with this data and comparisons with the current work.

In order to minimise the effect of song sequence and rating conditions (perceived and induced emotion), four different list conditions were constructed. The order of presentation of the two rating conditions and two song blocks (m=40, 10 for each emotion category) was counterbalanced across subjects. The songs in each block were randomly distributed across participants [26]. See Table 2 for the group allocation. To remind the subjects of two different rating conditions, the questions were highlighted in different colours (blue and red) and they were also counterbalanced across groups.

Group	Block 1	Block 2
Group 1	Induced emotion	Perceived emotion
Group 2	Perceived emotion	Induced emotion
	Block 2	**Block 1**
Group 3	Induced emotion	Perceived emotion
Group 4	Perceived emotion	Induced emotion

Table 2. Group allocation among participants

2.3 Procedure

The study was approved by the Research Ethics Committee (REF: QMREC1019). The listening test was conducted online [5] ; participants only required internet access and a speaker or headphones for this experiment. First, the participants were asked to read the instruction page:

1. Listen to the songs (they will last either 30 or 60 seconds)

[3] http://www.7digital.com/
[4] https://code.soundsoftware.ac.uk/projects/emotion-recognition
[5] http://isophonics.net/dimensional/test/

2. After listening, please rate each piece on two dimensions: Valence (happy-sad continuum) and Arousal (excited-relaxed continuum)

3. For each track, you may click the "stop" button of the audio player if required

4. Be careful, do not press too quickly, since you can only listen to each song once

5. Please answer all the questions; the test will take about 40 mins to complete

The participants filled in a demographic form including name, age, gender, "type of music they are most familiar with", nationality, and "music culture they grew up with" as well as a selected Goldsmiths Musical Sophistication Index (GOLD-MSI) questionnaire (9 questions) to measure participants' level of musical training. They responded to each excerpt (n=10 per page) and rated them on two dimensions, valence (sad to happy) and arousal (relaxed to excited). According to the session they chose, a pop-up window would appear reminding them, based on which condition they needed to answer, "*How would you describe the emotional content of the music itself? (expressed emotion)*" , and "*What emotion do you feel in response to the music? (felt emotion)*". The whole experiment lasted about 40 minutes without any breaks. However, the participants were able to stop whenever they wanted.

3. RESULTS

The data analysis was conducted using the Matlab 2012 Statistics Toolbox. The results were aggregated across people for song level analysis or aggregated across item for individual level analysis.

3.1 Song level analysis

3.1.1 Comparison of valence and arousal ratings for perceived and induced emotion

To understand the effects of rating conditions (perceived emotion and induced emotion) and emotions (happy, sad, relax[6] and angry) on the ratings of valence and arousal, a two-way analysis of variance (ANOVA) was conducted. No significant difference was found for the ratings for the two conditions and the interaction of emotion and condition. On the other hand, the emotion tag had a significant effect on the ratings for valence and arousal (see Table 3).

3.1.2 Correlation between the ratings for perceived emotion and induced emotion

Section 3.1.1 showed that there was no significant difference between ratings for perceived and induced emotion. Correlation analyses were performed to study the relationship between valence (respectively arousal) ratings for perceived and induced emotion. Regardless of the discrete emotion tag, the listeners' valence and arousal ratings were highly positively correlated between perceived and induced emotion cases (valence: $r = 0.9357$, $p < 0.0001$; arousal: $r = 0.9590$, $p < 0.0001$). The correlation of valence is shown in Figure 1.

[6] The term "relax" was used in the data collection, which represented the emotion "relaxed"

Source	SS	df	MS	F	Prob >F
Valence					
Condition	1.93	1	1.93	1.14	0.29
Emotion Tag	163.26	3	54.42	31.97	**0**
Interaction	0.27	3	0.09	0.05	0.98
Arousal					
Condition	0	1	0.0002	0	0.99
Emotion Tag	231.55	3	77.18	25.73	**0**
Interaction	0.99	3	0.33	0.11	0.95

Note: SS - the sums of squares, df - degrees-of-freedom MS - mean squares (SS/df), F - F statistics

Table 3. Two-way analysis of variance on the ratings of valence and arousal

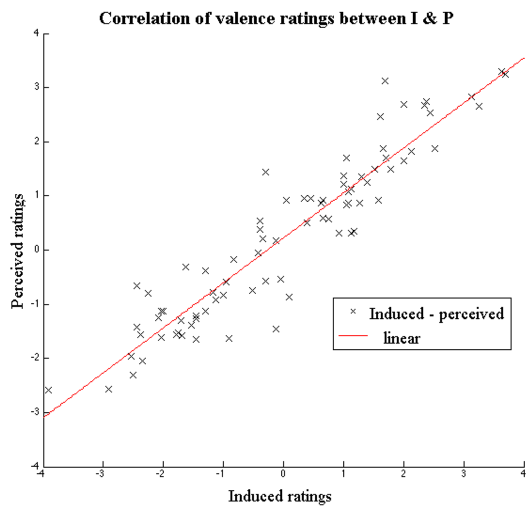

Figure 1. Correlation between the valence ratings for induced and perceived emotion

3.1.3 Correspondence between ratings and tags

The four emotion tags were chosen such that each occupies a unique quadrant of the valence-arousal plane, as shown in Figure 2. Considering that these four basic emotions are widely accepted across different cultures, we are able to assess the agreement between tags and participant ratings according to the extent that participants' ratings correspond with the quadrant belonging to the song's tag.

For each song, the average of participants' valence and arousal ratings were calculated for both perceived and induced emotion, to give a centroid for each song. The quadrant of this song centroid was then compared with the expected quadrant based on the emotion tag associated with the song. The fraction of songs for which the centroid quadrant corresponded with that of the tag is shown in Table 4. In addition, the standard deviations (SD) of the valence and arousal ratings for songs in each emotion category were calculated.

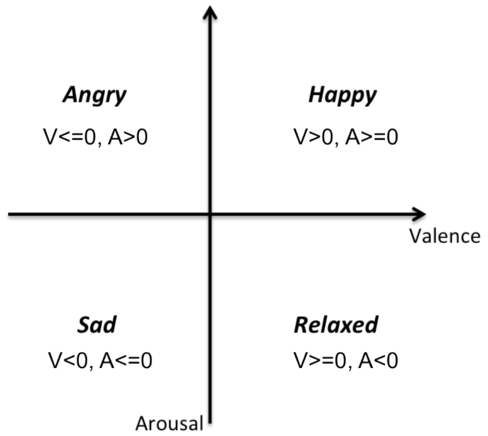

Figure 2. Valence-Arousal model showing the quadrants of the four emotion tags used in this experiment.

The results are shown with the highest values shown in bold in Table 4. Apart from the excerpts tagged with "relax", more than 55% of the average valence and arousal ratings lie in the song's corresponding tag quadrant. Fewer than 20% of mean ratings for songs tagged "relax" were located in the correct quadrant. Moreover, the high standard deviation of valence-arousal ratings for both perceived and induced emotion were found, indicating that the ratings of "relax" excerpts were not consistent across songs. For all emotion tag categories, and for both perceived and induced emotion, it was found that valence ratings were more consistent than arousal ratings.

	Happy	Sad	Relax	Angry
Induced emotion				
Rating=Tag	0.70	0.70	0.20	0.60
Valence SD	1.05	1.41	**1.81**	1.20
Arousal SD	1.56	1.45	**2.46**	1.82
Perceived emotion				
Rating=Tag	0.80	0.65	0.15	0.55
Valence SD	1.10	1.25	**1.47**	0.93
Arousal SD	1.34	1.32	**2.04**	1.56

Table 4. Agreement of valence-arousal ratings with tag quadrants, and spread of per-song ratings.

Figure 3 presents the average participant ratings of each song using the 2-dimensional model of emotion. The figure shows a high level of agreement between ratings for perceived and induced emotion (note that each participant rated each song for only one of perceived and induced emotion), compared to the spread of songs corresponding to an emotion tag.

3.1.4 Spread of participants' ratings

For each song, the spread of participant ratings was computed as the mean distance between points (centroid of the averaged ratings across participants and the ratings of each participant) in the valence-arousal plane. These values were then averaged across songs labelled with each emotion tag, as shown in Table 5. The spread of ratings for perceived and induced emotion were similar; likewise the spreads for each emotion category were similar, with a slightly higher spread found for "sad" songs.

	Happy	Sad	Relax	Angry
Induced emotion	2.59	**2.75**	2.45	2.65
Perceived Emotion	2.56	**2.76**	2.54	2.52

Table 5. The spread of the participant ratings across songs in each emotion category.

3.2 Individual level analysis

3.2.1 Individual agreement of participant ratings with the emotion tags

The analysis and results in section 3.1.3 were based on ratings for each excerpt that were averaged across participants. To analyse the relationship between individual ratings and emotion tags, we compute the fraction of ratings that are in the same quadrant as the emotion tag for the song, and compare this with the baseline of 25% for random choice of quadrants. The results are shown in Table 6. Chi-square tests were used to test whether the agreement with the emotion tag was significantly above chance level.

	Happy	Sad	Relax	Angry
Induced emotion	*****0.56**	*0.47	0.25	**0.49
Perceived emotion	****0.57**	0.42	0.25	**0.48

Table 6. Agreement of participant ratings with the quadrant of the emotion tag for each category. Values above chance level according to χ^2 tests are shown for the following significance levels: *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

It was found that the songs labelled with "happy" had the highest agreement 55% (p <0.001). Significant results were also found for tags "sad" (p <0.05, induced emotion) and "angry" (p <0.01). However, the agreement of participant ratings and the expected quadrant for songs labelled with "relax" was at the level of chance.

3.2.2 Agreement among participants

Since the tags associated with a song are not an absolute ground truth, but are also generated by users, under unknown conditions, we also look at the agreement of rating quadrants among the participants. The level of participant agreement is defined as the fraction of participants whose ratings are in the quadrant with the highest number of participant ratings. This value has as a lower bound the agreement with the tag quadrant, but can be higher if a greater number of participants agree on a quadrant other than that of the tag. The agreement of the individual dimensions of valence and arousal was also computed. The Wilcoxon

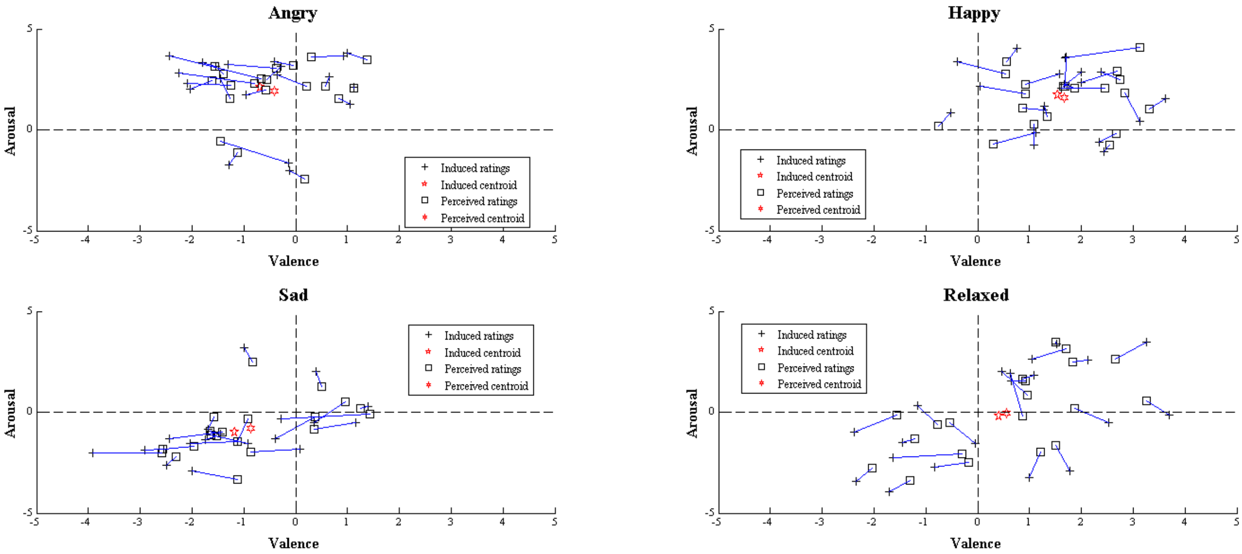

Figure 3. Map showing the average valence-arousal ratings for each excerpt. Each pane shows the ratings and centroid for one emotion tag. For each song, the points representing average ratings for perceived and induced emotion are connected by a line segment.

Signed Rank test was used to compare the difference of agreement for both perceived and induced emotion. Results are shown in Table 7. In comparison with Table 6, the levels of participant agreement are higher, suggesting that at least some of the tags do not correspond either with participants' perceived or induced emotion.

	Mean	SD	P-value
Induced emotion	0.60	0.14	0.0032
Perceived emotion	0.57	0.15	
Induced Valence	0.68	0.14	0.0211
Perceived Valence	0.66	0.15	
Induced Arousal	0.77	0.14	0.0001
Perceived Arousal	0.73	0.16	

Table 7. Participant agreement among themselves

3.2.3 The spread of song ratings for each emotion category

Similar to section 3.1.4, the spread of the song ratings for each person was computed for each emotion tag. Once again, the highest values were found for the tag "relax", because the responses for songs in this category vary more than for songs in other categories. This also helps to explain the low agreement with tags in section 6.

	Happy	Sad	Relax	Angry
Induced emotion	2.67	2.82	**3.49**	2.82
Perceived emotion	2.67	2.61	**3.15**	2.55

Table 8. The spread of the song ratings for each emotion category.

4. DISCUSSION

The main purpose of the paper was to investigate the associations between categorical emotion tags (happy, sad, relax and angry) and musical judgements of perceived emotion and induced emotion using a 2-dimensional (valence-arousal) model of emotion. First, two-way analysis of variance was used to test the effects of rating conditions and emotion categories on the ratings of valence and arousal. No significant difference was found between rating conditions, perceived and induced emotion, nor for the interaction of emotion categories and conditions. A correlation analysis on the ratings for induced and perceived emotion showed strong positive correlations between perceived and induced emotion, for ratings of both valence ($p<0.0001$ $r=0.94$) and arousal ($p<0.0001$ $r=0.96$). This suggests that listeners will typically feel the emotions expressed by the song, so if a song expresses happiness, it is likely that the listener will feel happy as well [21].

Second, we studied the reliability of the mean value among participants' ratings for predicting the emotion tag (happy, sad, relax and angry). The average valence and arousal ratings for both induced and perceived emotion were consistent with more than 55% of the happy, sad and angry tagged songs. However, the ratings of the songs labelled with "relax" tended to be less consistent, and did not agree with the quadrant of the tag. This can be seen in the mapping of songs to the valence-arousal plane (Figure 3), which shows the average ratings for songs labelled "relax" spread across happy, sad and relax quadrants. Interestingly, although 60% of the songs labelled with "sad" lie in correct quadrant, the spread of participants' ratings was higher than for the other emotion categories.

Third, we analysed the relationship between emotion tags and individual participant ratings. The results were very similar to the average ratings in section 3.1.3; the songs labelled with “happy”, “sad” and “angry” had ratings in the corresponding quadrants of the valence-arousal plane at a level that was significantly above chance. For songs tagged “relax”, however, the agreement of ratings with the positive-valence, negative-arousal quadrant was at the level of chance for both perceived and induced emotion. Further, the spread of the ratings was also higher than for the other categories. Comparing these four tags, regardless of song or person, the excerpts tagged with “happy” are most likely to produce responses in the corresponding quadrant of the valence-arousal plane.

Finally, the agreement among participants were computed. Gabrielsson noted a higher agreement among listeners for perceived emotion, due to its objectivity, over induced emotion [8], and this was confirmed in a study using the categorical model and the same musical excerpts as the present study [21]. In this paper, we found that agreement among participants was significantly higher for induced emotion than for perceived emotion ($p<0.01$). However, we concede that the overall levels of agreement (60% for induced emotion and 57% for perceived emotion) are not high. As a partial explanation of discrepancies between emotion tags and user ratings, we observed lower levels of agreement among listeners for songs tagged “relax” than for the classes which had higher agreement between tags and participant ratings.

In future studies we plan to investigate the different responses in perceived and induced emotion, and compare the use of the categorical and dimensional models of emotion.

5. ACKNOWLEDGEMENTS

We acknowledge the support of the China Scholarship Council. We would like to thank the listeners and reviewers for their participation and comments.

6. REFERENCES

[1] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *SIGCHI conference on Human Factors in Computing Systems*, 2007.

[2] S. Dalla Bella, I. Peretz, L. Rousseau, N. Gosselin. A developmental study of the affective value of tempo and mode in music. *Cognition*, 80(3):B1–10, 2001.

[3] D. Eck, P. Lamere, T. Bertin-Mahieux, S. Green. Automatic generation of social tags for music recommendation. *Advances in Neural Information Processing Systems*, 2007.

[4] T. Eerola and J.K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2010.

[5] T. Eerola and J.K. Vuoskoski. A review of music and emotion studies: approaches, emotion models, and stimuli. *Music Perception*, 30(3):307–340, 2013.

[6] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[7] P. Evans and E. Schubert. Relationships between expressed and felt emotions in music. *Musicae Scientiae*, 12(1):75–99, 2008.

[8] A. Gabrielsson. Emotion perceived and emotion felt: same or different? *Musicae Scientiae*, 123–147, 2002.

[9] M. Hoffman, D. Blei, et al. Easy as CBA: A simple probabilistic model for tagging music. In *10th International Society for Music Information Retrieval*, 2009.

[10] P.N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions. *Journal of New Music Research*, 33(3):217–238, 2004.

[11] K. Kallinen and N. Ravaja. Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 191–213, 2006.

[12] G. Kreutz, U. Ott, D. Teichmann, P. Osawa, D. Vaitl. Using music to induce emotions: Influences of musical preference and absorption. *Psychology of Music*, 36(1):101–126, 2007.

[13] P. Lamere. Social tagging and music information retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

[14] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *8th International Society for Music Information Retrieval*, 2007.

[15] M. Levy and M. Sandler. Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia*, 11(3):383–395, 2009.

[16] D. Müllensiefen, B. Gingras, L. Stewart. Goldsmiths Musical Sophistication Index (Gold-MSI) *Technical report*, 2012.

[17] J. Panksepp. Affective neuroscience: The foundation of human and animal emotions. *Oxford University Press*, 1998.

[18] P.J. Rentfrow and S.D. Gosling. The Do Re Mi’s of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology*, 84(6):1236–1256, 2003.

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

[20] Y. Song, S. Dixon, M. Pearce. Evaluation of musical features for emotion classification. In *13th International Society for Music Information Retrieval*, 2012.

[21] Y. Song, S. Dixon, M. Pearce, G. Fazekas. Using tags to select stimuli in the study of music and emotion. In *3rd International Conference on Music & Emotion*, 2013.

[22] M. Terwogt and F. Van Grinsven. Musical expression of moodstates. *Psychology of Music*, 90–109, 1991.

[23] D. Turnbull, L. Barrington, G. Lanckriet. Five approaches to collecting tags for music. In *9th International Society for Music Information Retrieval*, 2008.

[24] S. Vieillard, I. Peretz, et al. Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, 22(4):720–752, 2008.

[25] D. Wang, T. Li, M. Ogihara. Tags better than audio features? the effect of joint use of tags and audio content features for artistic style clustering. In *11th International Society for Music Information Retrieval*, 2010.

[26] N. Welch and J.H. Krantz. The World-Wide Web as a medium for psychoacoustical demonstrations and experiments: Experience and results. *Behavior Research Methods, Instruments,& Computers*, 28(2):192–196, 1996.

[27] X. Wu, L. Zhang, Y. Yu. Exploring social annotations for the semantic web. In *15th International Conference on World Wide Web*, 2006.

[28] M. Zentner, D. Grandjean, et al. Emotions evoked by the sound of music: characterization, classification, and measurement. *Emotion (Washington, D.C.)*, 8(4):494–521, 2008.