Generating Ground Truth for Music Mood Classification Using Mechanical Turk

Jin Ha Lee
University of Washington
Mary Gates Hall, Suite 370
Seattle, WA 98195
+1 206.685.0153
jinhalee@uw.edu

Xiao Hu
University of Denver
1999 E. Evans Ave. Room 249
Denver, CO 80208
+1 303.871.3352
xiao.hu@du.edu

ABSTRACT
Mood is an important access point in music digital libraries and online music repositories, but generating ground truth for evaluating various music mood classification algorithms is a challenging problem. This is because collecting enough human judgments is time-consuming and costly due to the subjectivity of music mood. In this study, we explore the viability of crowdsourcing music mood classification judgments using Amazon Mechanical Turk (MTurk). Specifically, we compare the mood classification judgments collected for the annual Music Information Retrieval Evaluation eXchange (MIREX) with judgments collected using MTurk. Our data show that the overall distribution of mood clusters and agreement rates from MIREX and MTurk were comparable. However, Turkers tended to agree less with the pre-labeled mood clusters than MIREX evaluators. The system evaluation results generated using both sets of data were mostly the same except for detecting one statistically significant pair using Friedman's test. We conclude that MTurk can potentially serve as a viable alternative for ground truth collection, with some reservation with regards to particular mood clusters.

Categories and Subject Descriptors
H.3.4. [**Information Systems**]: Information Storage and Retrieval – Systems and Software – *Performance evaluation.* J.5 [**Arts and Humanities**]: *Music*

General Terms
Measurement, Human Factors, Performance

Keywords
Music Information Retrieval, Evaluation, Ground Truth, Gold Standard, Mood, Mechanical Turk, Crowdsourcing

1. INTRODUCTION
Generating ground truth for evaluating various music information retrieval (MIR) systems is an essential process for the improvement of MIR and music digital libraries (MDL) research, yet it often tends to be a challenging problem. This is especially true for generating ground truth based on human input since collecting human judgments tends to be expensive and time consuming [1]. In the MIR/MDL community, there are several evaluation tasks that use human input as the basis for evaluating the performance of algorithms, namely audio music similarity, symbolic music similarity, and audio music mood classification tasks. Getting enough people to verify evaluation results for these tasks is a tedious and long process since it can require tens of thousands of responses for a modest collection of several hundred songs [14], [22]. In order to alleviate the difficulty of collecting human responses, two studies [14], [24] have explored the viability of using Amazon Mechanical Turk (MTurk) for collecting human judgments on music similarity evaluation tasks. Both studies have compared the human music similarity judgments obtained from music experts in Music Information Retrieval Evaluation eXchange (MIREX)[2] with the ones obtained from MTurk. The authors from both studies suggest MTurk as a viable alternative for collecting human judgments rather than relying on music experts from the MIR community.

The motivation for our study is to explore if MTurk works well for the evaluation task of music mood classification. Mood has become an important access point for MDL and online music repositories [25] (e.g., allmusic.com, stereomood.com). A number of algorithms have been proposed to classify music by its mood in the MIR/MDL domain (e.g., [2], [8]). Since 2007, the MIREX has been hosting an Audio Mood Classification (AMC) task to evaluate and compare mood classification algorithms [9] (also see Section 2.1). However, generating ground truth for evaluating music mood classification remains a difficult problem [11]. In terms of collecting human judgments, the task of music mood classification differs from music similarity tasks such as Audio

JCDL'12, June 10–14, 2012, Washington, DC, USA.

[1] The authors recognize that the term "gold standard" may be more suitable for describing the data set generated by human judgments since "ground truth" is usually assumed to be the objective measure of reality. However, in order to maintain consistency with the terms that have been used in the MIREX domain, we use the term "ground truth" in this work.

[2] Music Information Retrieval Evaluation eXhange (MIREX) is the annual evaluation campaign for various music information retrieval algorithms hosted by the International Music Information Retrieval Systems Evaluation Lab (IMIRSEL) at the University of Illinois at Urbana-Champaign.

Music Similarity (AMS) and Symbolic Music Similarity (SMS) in MIREX in the following aspects:

1) Compared to music similarity, music mood is even more subjective as there is very little objective reference, if any, on how music carrying a particular mood should sound;

2) Assessment of music mood requires a human evaluator to select one mood label from a set of options whereas for judging music similarity, a human judge only needs to answer whether two music pieces are similar or not similar [14], or select one out of two variations that sound more similar to the original piece [24]. Therefore, the cognitive load for evaluators may be heavier for the task of assigning mood labels than judging music similarity.

3) There does not exist an authoritative taxonomy of music mood, meaning there is no formal agreement on what kinds of music moods exist and how to define them. This further complicates music mood judgment.

Due to these unique difficulties of labeling music mood, the results of previous studies on applying MTurk to music similarity tasks may not be applicable to the task of mood classification. Therefore, it is necessary to specifically study the viability of crowdsourcing music mood classification judgments using MTurk. In order to do this, we test the effectiveness of MTurk for Audio Music Mood Classification (AMC) task in MIREX. The findings from this study will also help us further our general understanding on the viability of relying on crowdsourcing for generating ground truth for music related evaluation tasks.

2. BACKGROUND

2.1 MIREX Audio Mood Classification

The annual MIREX is hosted by the International Music Information Retrieval Systems Evaluation Lab (IMIRSEL) where a variety of evaluation tasks are carried out in order to test the performance of different MIR algorithms. Some of these evaluation tasks, such as rating music similarity or classifying music mood, attempt to model human perception or understanding of music by relying on human judgments as the basis of the evaluation.

Audio Music Mood Classification (AMC) task first started in 2007 and has been run every year since then. The objective of this task is to test whether automatic algorithms can correctly predict mood labels for music clips based on their audio characteristics. The ground truth data set for this task requires a set of songs with mood labels that are agreed upon by multiple human evaluators. In other words, if song A were to be used as part of the ground truth for mood cluster 1 (see Table 1), then we want to make sure that multiple people agree that song A in fact carries "passionate, rousing, confident, boisterous, and rowdy" moods. This means that multiple human evaluators are needed to judge the mood of the songs in the test collection. The task organizers ended up using songs that received the same mood classification judgment from two or more people for generating the ground truth (more discussion in Section 4).

IMIRSEL built a web-based survey system called E6K in order to collect responses from human evaluators for the tasks that needed human input. The evaluators are mostly volunteers from the MIR community who have some background in music and/or music related research, thus considered music experts. Every year when new ground truth data are needed, IMIRSEL sends out a series of emails seeking volunteers and it often takes weeks to collect all the responses necessary [14].

2.2 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) [3] is a crowdsourcing platform that provides 24/7, on-demand access to workers able to complete tasks requiring human-intervention. For task requesters, MTurk makes it possible to quickly collect human data in a cost-effective manner. The task requester can set up a "HIT" (Human Intelligence Task, the name for a task in MTurk) and set a fee for the completion of that HIT. The HITs are then worked on by human workers (called "Turkers" in MTurk) who are recruited by Amazon and may come from anywhere in the world. The task requester can also create a qualification test that needs to be passed before Turkers are able to work on particular tasks. When the HIT is completed and submitted, the requester can review the HITs and either approve or reject them.

3. RELATED WORK

3.1 MTurk used in Text IR Studies

Previous information retrieval (IR) literature proposed using MTurk for generating relevance judgments in TREC-like evaluations [1], [12], [23]. Alonso and Mizzaro [1] compared MTurk to TREC experts and found the results from MTurk to be comparable to TREC's ground truth generated by experts. Kitter et al. [12] applied MTurk to rating Wikipedia articles and found that the consistency of results provided by MTurk can be greatly improved by adding verifiable questions into the HITs and filtering out the bad results. An important lesson learned by Kitter et al. [12] is that some sort of verification questions are necessary to ensure the quality of answers gathered via Internet crowdsourcing platforms such as MTurk. In our study, we also employed multiple verification questions to filter out inconsistent answers [see Section 4.3].

With regards to affect, Snow et al. [23] used MTurk for generating ground truth data for Natural Language Processing tasks including determination of valence and emotion in text. They found that MTurk provided results on par with those obtained by domain experts. Determining the mood of music is potentially a more complicated task than determining the mood of text. As reported in Lee et al. [15], when people assess music mood, a range of factors come into consideration: lyrics, tempo, instrumentation, genre, delivery, and even cultural context. In addition, our study differs from Snow et al.'s [23] in that the mood space used in our study is a *categorical* model (five distinct mood clusters, Table 1) while Snow et al. [23] used a *dimensional* model where a continuous score was given to each of the seven dimensions of emotions (e.g., valence, joy, anger)[11].

3.2 MTurk used in MIR Studies

Previous MIR research has shown that MTurk can be used fairly successfully for rating music similarity [14], [24] and providing user tags [15], [17]. Lee [14] used MTurk for collecting human judgments for the past MIREX Audio Music Similarity task and found that the results from using MTurk were comparable to the ones from music experts. Urbano et al. [24] used MTurk to gather music preference judgments for the Symbolic Music Similarity

[3] http://www.mturk.com/mturk/welcome

task in MIREX. They reported that their results were very similar to the expert generated ones and recommended crowdsourcing as "a perfectly viable alternative to evaluate music systems without the need for experts" (p.9). Mandel et al. [17] used MTurk to collect user tags on different parts of the same songs and compared them to tags collected from a music tagging game. They found MTurk was a viable means to collect ground truth although the inter-rater agreement among the Turks was slightly lower than that of the gamers. It is noteworthy that the user tags collected in [17] were free terms that could be related to music genre, instrument, user's opinions or simply noisy text. This is different from our study in that we are asking users to select one of the five mood clusters that is most appropriate for the given music clip. More recently, Lee et al. [15] used MTurk to collect mood tags for various cover versions of the same song, with the purpose of investigating the factors affecting how end-users determine and describe music mood in their own terms. Although they also emphasized the mood aspect of music, their research focus was not on the viability of MTurk on generating ground truth for evaluation of mood classification algorithms, which is the main objective of our study.

4. RESEARCH QUESTIONS AND STUDY DESIGN

4.1 Research Questions

We explore the following two research questions in this paper:

I) How do music mood classification results obtained from Mechanical Turk compare to those collected from music experts in MIREX?

II) How different or similar are the evaluation outcomes for MIREX AMC task when based on ground truth collected from Mechanical Turk as compared to ground truth collected from E6K in MIREX?

In order to address these questions, we reproduced the human judgment collection measures like those in E6K for MIREX on MTurk. Specifically, we created an online survey, which asked Turkers to listen to the same music clips used in E6K and select appropriate mood clusters that reflect the mood of the songs. The results will inform us as to the viability of crowdsourcing mood classification judgments.

4.2 Test Collection

We used the same test collection that was used for collecting human judgments and generating ground truth in the MIREX 2007 AMC task[4] with the help of IMIRSEL. This test collection was created based on the APM (Associated Production Music)[5] collection [9], covering a variety of different music genres. This collection consists of 1250 tracks with 250 pieces from each of the five mood clusters. The mood clusters were derived from the 179 mood labels on allmusic.com, a major online music repository. These mood labels were generated by professional editors hired by allmusic.com. In order to identify a more general and clear way to describe the mood space, Hu and Downie [7] applied a hierarchical clustering algorithm to the most popular mood labels and their representative albums and songs on allmusic.com, which resulted in the five mood clusters as shown in Table 1. Further analysis revealed that the clusters had consistent relationships with more traditional music metadata types including genre and artist [7]. The five mood clusters have been used in the AMC task in MIREX as well as a number of other studies on music mood (e.g., [11], [13]).

Before presenting the music clips to human evaluators for mood judgment, each of the clips in the test collection was assigned an initial mood cluster based on the metadata and short descriptions of the songs provided by APM (part of the motivation for designing the AMC task in MIREX in this way was to have a balanced distribution of songs across mood clusters in the resultant ground truth data set). A 30-second clip was sampled from the middle of each of the tracks to be used in the classification task [9]. 30-second clips were used to reduce the burden on the human evaluators, as well as to minimize the potential for variation caused by changes in mood over the duration of a track (i.e., a song may start out with soft and sweet melody, but become loud and aggressive).

4.3 Task Design

In order to ensure that the task was similarly carried out as in MIREX, we included a qualification task that Turkers needed to pass in order to work on our HITs. The AMC task requires that the evaluators understand the meaning of the five mood clusters. In order to ensure that human evaluators can identify the kinds of songs that would fit into each cluster, in MIREX, they were asked to listen to three representative sample songs per cluster. In our qualification task on MTurk, we asked the Turkers to listen to the same sample songs for the clusters. After that, they had to provide answers to a short survey asking which mood clusters would be most appropriate for five given songs. We randomly selected one of the sample songs given for each cluster in order to set up this survey. If they provided the correct answers for all five questions, they were qualified to work on the HITs.

In the HIT, Turkers were instructed to select the most appropriate mood cluster that reflects the mood of each song out of the five given clusters [Table 1]. They were given an option to select the "other" cluster if they think the song does not fit into any of the five given mood clusters. Each HIT consisted of 25 different clips. The songs were randomly assigned in order to minimize the ordering bias. Instruction for the task was given as shown in Figure 1. When the link of "Song X" is clicked, the mp3 file of the clip is played by Yahoo! webplayer[6] embedded in the survey page so that the Turkers do not need to open an external player or download the mp3 file. We wanted to collect two responses for each of the 1250 music tracks in order to check the agreement rate, thus needing 2500 responses in total. We paid $0.55 for completing each of these HITs. As each HIT asked for 27 judgments (25 unique songs + 2 repeated songs for verification questions), the payment for each judgment was $0.02. In [14], $0.20 was paid for a HIT consisting of 13 query-candidate pairs (i.e., 13 similarity judgments), resulting in $0.015 paid for each

[4] http://www.music-ir.org/mirex/wiki/2007:Audio_Music_Mood_Classification; the same judgments were used for the AMC task in the following years.

[5] http://www.apmmusic.com/pages/aboutapm.html

[6] http://webplayer.yahoo.com/

judgment. Also in [24], each HIT contained one judgment on melody similarity and paid for $0.02. Therefore, the amount of payment in our study is in line with previous studies. The total cost for administering 100 HITs, including Amazon's fee was $60.50.

Table 1. Five Mood Clusters

Cluster1	passionate, rousing, confident, boisterous, rowdy
Cluster2	rollicking, cheerful, fun, sweet, amiable/good natured
Cluster3	literate, poignant, wistful, bittersweet, autumnal, brooding
Cluster4	humorous, silly, campy, quirky, whimsical, witty, wry
Cluster5	aggressive, fiery, tense/anxious, intense, volatile, visceral

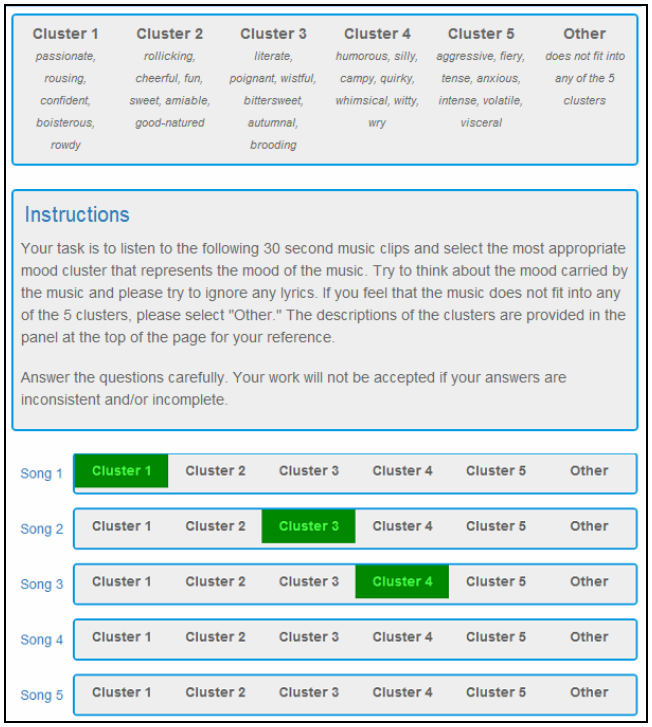

Figure 1. Screenshot of the MTurk HIT

As previously explained, on MTurk, task requesters are allowed to reject any responses that do not meet the requirements set by the requester. Previous studies have reported high proportions of bad responses in MTurk results [14], [15], [24]. Therefore having some kind of filtering mechanism is essential. Lee [14] used two different kinds of verification questions for the music similarity task: 1) inserting the same question twice to check the consistency of the answers, and 2) inserting a question in which they asked users to rate the similarity of the music clip compared to itself. In our study, we also employed a consistency check. We randomly selected two songs from the HIT and repeated them in the HIT. In other words, each Turker selected mood clusters for 27 songs, but there were only 25 unique songs in a single HIT. Our expectation was that the Turker should provide the same mood cluster to both instances of the same song. The submitted HITs which did not meet this consistency check were discarded. Among all the HITs submitted, 46.2% were rejected for inconsistent answers to the verification questions which is higher than [15] (26%), but comparable with [14] (44.3%).

5. DATA AND DISCUSSION

5.1 Comparison of E6K and MTurk Data

5.1.1 Number of Judgments and Distribution across Clusters

In E6K (for MIREX), each human evaluator was assigned 250 music clips. The initial plan was to have three different evaluators for each music clip, however, due to limited participation, only some clips received all three judgments, some received two or only one judgment. As a result, they were only able to collect 2468 judgments in total, with 1180 music clips that received at least one judgment.

On MTurk, we collected a total of 5022 mood classification judgments from 186 HITs that were submitted. Of 186 HITs that were submitted, we accepted 100 HITs (53.8%) and rejected 86 (46.2%) that did not pass the consistency check. After filtering out the rejected HITs and the consistency check judgments, we ended up with 2500 unique mood classification judgments. One benefit of using MTurk is that we do not have to rely on volunteers, which is an unknown variable. In E6K, there is no guarantee how many people will volunteer to provide classification judgments as there is no direct incentive for participation. Many volunteers started the task but never finished it, resulting in a large amount of incomplete data. This also resulted in complication for assignment of music clips; for example, one evaluator may take music clip 1 through 250 and only provide judgments on the first 100 clips which means that the evaluation task organizers have to ensure that clip 101 to 250 are reissued to another evaluator. On MTurk, this issue is non-existent due to the large number of available Turkers, who are compensated for completing tasks.

Table 2. Comparison of judgment distribution in E6K and MTurk

Cluster	E6K	MTurk	Diff. in % (E6K-MTurk)
Cluster1	405 (16.4%)	450 (18.0%)	-1.6%
Cluster2	472 (19.1%)	536 (21.4%)	-2.3%
Cluster3	542 (22.0%)	622 (24.9%)	-2.9%
Cluster4	412 (16.7%)	367 (14.7%)	2.0%
Cluster5	400 (16.2%)	403 (16.1%)	0.1%
Other	237 (9.6%)	122 (4.9%)	4.7%
Total	**2468**	**2500**	-

Table 2 shows the judgment distribution across the five mood clusters in E6K and MTurk. We can observe that Cluster 3 received most votes in both E6K and MTurk, followed by Cluster 2. E6K data contained a slightly larger proportion of votes for Cluster 4, and MTurk data had more votes in Cluster 1, 2, and 3. The proportions for Cluster 5 were almost the same. Since the total numbers of judgments are not equal, this is technically not an exact comparison. However, we can observe that overall proportions are fairly comparable when we take into account the difference in the total numbers (32, which counts for a difference of 1.3%).

5.1.2 Time for Collecting Judgments

On MTurk, the average time Turkers spent for completing a HIT (containing 27 songs) was 471.36 seconds, meaning they spent

about an average of 17.46 seconds for listening to each music clip and making judgment. By comparison, the expert evaluators in E6K spent 21.54 seconds on each music clip on average. The 4 more seconds spent on each music clip on average may indicate a more serious attitude of the expert evaluators towards the task, and/or simply more interest in listening to music. It is also possible that Turkers did not spend a long time listening to the two repeated music clips for the verification questions. We were not able to test the statistical significance of the time difference as MTurk does not provide time spent on each question in a HIT. However, it is clear that both groups of evaluators tended not to finish listening to the entire clips available (i.e., 30 seconds).

In order to collect all the judgments needed, it took approximately 19 days on MTurk. On the other hand, for MIREX, it took about 38 days in order to collect all the judgments needed. Hu et al. [9] explain that their data were collected from 18 volunteers starting 1 Aug. until 19 Aug. In order to build the ground truth, they needed music clips with at least two agreed judgments. There were not enough responses from volunteers, however, and the IMIRSEL lab had to do additional in-house assessments on clips that only received one judgment. The whole process ended up taking 38 days.

The 19 days taken on MTurk, when compared to studies that tested music similarity tests ([14]: 12 hours, and [24]: a day and a half), is significantly longer. We suspect that this is due to the qualification task that the Turkers were asked to complete in order to work on our HITs, which was not required in the other studies. Still, it took half the time taken by MIREX, which is a significant improvement. In addition, using MTurk, we were able to obtain the complete data set. In other words, there was no need for additional in-house assessment.

5.1.3 Inter-rater Reliability

In MTurk, every two Turkers were assigned the same set of clips, and thus Cohen's Kappa (κ) [4] was calculated for the 50 pairs of Turkers who made judgments on the same set of 25 clips. The E6K data is more complex in that each expert evaluator judged different number of clips. As Hu et al. [9] reported, only 8 evaluators finished all 250 clips assigned to each of them while the remaining 10 evaluators completed 6 to 140 clips. In order to calculate the inter-rater reliability on the E6K data, we paired up the 18 assessors and the number of common clips judged by each pair ranging from 0 to 60. For comparison to the MTurk data, we calculated κ values on the E6K pairs sharing 25 or more clips. The results are shown in Table 3.

Table 3. Inter-rater reliability in E6K and MTurk

	E6K	MTurk
Pairs of assessors	28	50
Common clips each pair	26-60	25
κ : Maximum	0.68	0.71
κ : Minimum	0.31	0.03
κ : Mean	0.50	0.40
κ : Standard deviation	0.09	0.17

The average κ values of 0.50 and 0.40 are comparable to those reported by Schuller et al. [21] (κ values were 0.40 and 0.44 for music mood judgments with regard to *valence* and *arousal* respectively). According to [21], such agreement level is "moderate to good" (p.17) for the task of music mood assessment, although it seems lower than what is expected in text analysis [3]. Table 3 also shows that the experts (E6K) seemed to reach a slightly higher inter-rater agreement than the Turkers, and the Turkers had a larger standard deviation on κ values. However, such comparison is not conclusive as there were more comparable pairs among the Turkers (50) than the E6K experts (28).

5.1.4 Agreement Rate

Table 4 shows how the distributions of clips with agreed judgments compare between E6K and MTurk data. The number in each cell indicates the number of clips that received the same mood classification judgment from at least two evaluators. The overall agreement rate on the music clips was 57.7% (681/1180) in E6K compared to 52.4% (655/1250) in MTurk. A chi-square test was conducted on the frequency counts in Table 4 and the result showed the agreement distributions across E6K and MTurk were not identical ($\chi^2 = 19.67$, df = 5, $p < 0.01$). In other words, there were disagreements among users in E6K and MTurk, as to how to classify the mood of these music clips. However, for both data sets, Cluster 3 had the highest number of agreed music clips whereas Cluster 1 and 4 were the least agreed clusters. Interestingly, the agreement among Turkers was much less than E6K evaluators for Cluster 1 and 4.

Table 4. Comparison of the agreed clips distribution in E6K and MTurk

Cluster	E6K	MTurk
Cluster1	121	89
Cluster2	130	131
Cluster3	163	216
Cluster4	121	85
Cluster5	126	121
Other	20	13
Total	**681**	**655**

Our data show that among the 2468 judgments in E6K, 1535 are the same in MTurk (1535/2468 = 62.20%). Among the 2500 judgments in MTurk, 1713 are the same in E6K (1713/2500 = 68.52%). The numbers are different because in E6K data, music clips can have a range of 1 to 3 judges while the in MTurk each clip has exactly 2 judges.

Table 5 shows the agreement between MTurk and E6K judgments across different clusters. The number in each cell indicates the number of judgments agreed by those in the other data set. For example, among the 1535 E6K judgments that are the same in MTurk, 196 have the value Cluster 1 and 457 have the value Cluster 3. Of the five clusters, Cluster 3 has the highest agreement and Cluster 1 has the lowest agreement. Overall, Cluster 3 seems to have the highest agreement: it had the highest agreement among the E6K evaluators, among the Turkers [Table 4], between the E6K evaluators and the Turkers [Table 5], between the E6K evaluators and the APM music experts [Table 8], and between the Turkers and APM music experts [Table 9].

Table 5. Agreement between E6K and MTurk judgments across clusters (percentages in parentheses)

E6K						
C1	**C2**	**C3**	**C4**	**C5**	**Other**	**Total**
196 (12.8)	304 (19.8)	457 (29.8)	230 (15.0)	302 (19.7)	46 (3.0)	1535 (100)
MTurk						
C1	**C2**	**C3**	**C4**	**C5**	**Other**	**Total**
242 (14.1)	321 (18.7)	485 (28.3)	280 (16.3)	338 (19.7)	47 (2.7)	1713 (100)

5.1.5 *Confusions between the Clusters*

We were also interested in knowing which cluster pairs are most confused by E6K evaluators and Turkers. Table 6 shows the distribution of disagreed judgments across different combinations of clusters sorted by the highest to the lowest number of clips receiving disagreements between the given clusters in the E6K data. In this data set, there were different number of judgments assigned for each clip (ranging from 1 to 3), thus there are more combinations than in the MTurk data [Table 7]. The sorted list in Table 6 indicates that Cluster 3 and Other were the mostly confused clusters followed by Cluster 2 and 4, and Cluster 4 and Other in E6K. Table 7 shows the distribution of disagreed judgments across different combinations of clusters in the MTurk data. Here we observe that Cluster 1 and 2 were the most confused clusters by the Turkers, followed by Cluster 2 and 4, and Cluster 1 and 5.

We can observe some similarities between the two lists: for example, Cluster 2 and 4, and Cluster 1 and 2 appear at the top of both lists, while Cluster 3 and 5, and Cluster 2 and 5 appear closer to the bottom of the lists (besides the confusion with the Other cluster). The results seem to suggest that the mood clusters may not be mutually exclusive, at least when they are perceived by real users. For instance, a song may perceived to be "cheerful" and "humorous" (Cluster 2 and 4) or "passionate" and "sweet" (Cluster 1 and 2) at the same time.

Table 6. Distribution of disagreed judgments across clusters (E6K)

Clusters	Disagreed judgments
Cluster 3 & Other	37
Cluster 2 & Cluster 4	31
Cluster 4 & Other	23
Cluster 1 & Cluster 2	20
Cluster 5 & Other	18
Cluster 2 & Other	17
Cluster 1 & Cluster 3	13
Cluster 1 & Cluster 5	13
Cluster 1 & Other	10
Cluster 2 & Cluster 3	9
Cluster 4 & Cluster 5	9
Cluster 1 & Cluster 2 & Cluster 4	7
Cluster 1 & Cluster 3 & Other	6
Cluster 1 & Cluster 4	6
Cluster 3 & Cluster 4	6
Cluster 1 & Cluster 2 & Cluster 3	5
Cluster 4 & Cluster 5 & Other	5
Cluster 1 & Cluster 2 & Other	4
Cluster 1 & Cluster 5 & Other	3
Cluster 2 & Cluster 4 & Other	3
Cluster 1 & Cluster 4 & Other	2
Cluster 1 & Cluster 2 & Cluster 5	1
Cluster 2 & Cluster 3 & Other	1
Cluster 2 & Cluster 4 & Cluster 5	1
Cluster 2 & Cluster 5	1
Cluster 3 & Cluster 5	1
Cluster 3 & Cluster 5 & Other	1
Total	**253**

Table 7. Distribution of disagreed judgments across clusters (MTurk)

Clusters	Disagreed judgments
Cluster 1 & Cluster 2	95
Cluster 2 & Cluster 4	86
Cluster 1 & Cluster 5	74
Cluster 2 & Cluster 3	61
Cluster 1 & Cluster 3	45
Cluster 1 & Cluster 4	41
Cluster 3 & Other	37
Cluster 4 & Cluster 5	28
Cluster 3 & Cluster 4	27
Cluster 2 & Cluster 5	22
Cluster 3 & Cluster 5	20
Cluster 1 & Other	17
Cluster 5 & Other	17
Cluster 4 & Other	15
Cluster 2 & Other	10
Total	**595**

Much of the previous research on music mood [5], [10], [16], [18], [19] confirmed the two dimensional model of valence and arousal for representing the mood space proposed by Russell [20] [Figure 2]. When we attempt to position the five mood clusters according to this two-dimensional model, both Clusters 2 (rollicking, cheerful, fun, sweet, amiable/good natured) and 4 (humorous, silly, campy, quirky, whimsical, witty, wry) seem to share positive valence. Additionally, Cluster 1 (passionate, rousing, confident, boisterous, rowdy) seems to consist of moods with positive valence as well. This may explain the reason for high confusion among these clusters by human evaluators. Then, it is interesting that Clusters 3 (literate, poignant, wistful, bittersweet, autumnal, brooding) and 5 (aggressive, fiery, tense/anxious, intense, volatile, visceral) are the least confused clusters, as they both consist of negative moods. We speculate that this may be because the difference between Clusters 3 and 5 in the arousal dimension is for some reason clearer, compared to Clusters 1, 2 and 4.

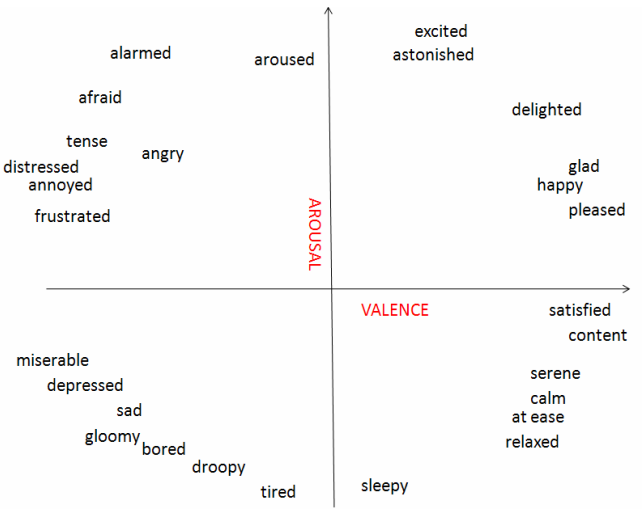

Figure 2. Russell's model of valence and arousal

The facts that these clusters in reality are not mutually exclusive and a song can carry multiple moods are good indications of why it is so challenging to evaluate music mood classification algorithms. Suppose there is a song that has a passionate mood as a dominant mood of the song, but it is delivered in a cheerful tone, can we really say it has to be classified in Cluster 1 and not 2? For the purpose of evaluation, there seems to be an underlying assumption that it is possible to classify songs into one type of mood cluster in MIREX. Based on this assumption, the songs that have agreed judgments are used for building the ground truth. Our data, however, suggest that maybe mood classification systems should be permitted to indicate multiple moods instead of classifying songs in to one particular "correct" mood cluster.

5.1.6 Reclassification

As previously explained, the AMC test collection had mood labels pre-assigned according to metadata provided by APM. We wanted to see to what extent human evaluators would agree with these pre-assigned labels in E6K and MTurk data sets. Table 8 and Table 9 show the numbers of clips that were reclassified according to the human judgments in EK6 and MTurk data set, respectively.

Table 8. Reclassification among all the judgments in E6K data (percentages in parentheses)

	Judgments						
Pre-label	C1	C2	C3	C4	C5	Other	Total
Cluster1	292 (54.0)	68 (12.6)	74 (13.7)	34 (6.3)	16 (3.0)	57 (10.5)	541 (100)
Cluster2	35 (7.6)	327 (70.8)	40 (8.7)	29 (6.3)	1 (0.2)	30 (6.5)	462 (100)
Cluster3	23 (4.6)	16 (3.2)	408 (81.6)	5 (1.0)	5 (1.0)	43 (8.6)	500 (100)
Cluster4	15 (10.2)	56 (11.6)	13 (2.7)	321 (66.3)	21 (4.3)	58 (12.0)	484 (100)
Cluster5	105 (21.0)	5 (1.0)	7 (1.5)	23 (4.8)	357 (74.2)	49 (10.2)	481 (100)
Total	450	472	542	412	400	237	2468

Table 9. Reclassification among all the judgments in MTurk data (percentages in parentheses)

	Judgments						
Pre-label	C1	C2	C3	C4	C5	Other	Total
Cluster1	160 (32.0)	130 (26.0)	119 (23.8)	43 (8.6)	27 (5.4)	21 (4.2)	500 (100)
Cluster2	87 (17.4)	235 (47.0)	68 (13.6)	84 (16.8)	4 (0.8)	22 (4.4)	500 (100)
Cluster3	47 (9.4)	21 (4.2)	368 (73.6)	10 (2.0)	23 (4.6)	31 (6.2)	500 (100)
Cluster4	51 (10.2)	125 (25.0)	48 (9.6)	208 (41.6)	40 (8.0)	28 (5.6)	500 (100)
Cluster5	105 (21.0)	25 (5.0)	19 (3.8)	22 (4.4)	309 (61.8)	20 (4.0)	500 (100)
Total	450	536	622	367	403	122	2500

By comparing this pair of tables, we can observe that overall the Turkers agreed less with pre-assigned labels than the E6K evaluators. Both data sets show that the pre-labeled mood cluster with the greatest agreement is Cluster 3, in fact, much more so than the other clusters. The cluster with least agreement is Cluster 4, meaning that many clips pre-labeled as Cluster 4 got controversial judgments in both data sets. Additionally, the Turkers seem to agree much less with the pre-assigned mood cluster 1 and 2 than E6K evaluators.

We also wanted to see the effect of reclassification among the music clips that had 2 or more agreed judgments. Table 10 and Table 11 show these results for the E6K and MTurk data in that order. In these two tables, we see lower percentages on reclassifications than those in Table 8 and 9 (i.e., higher percentages on the diagonal cells in Table 10 and 11). This observation indicates that for both E6K and MTurk data, agreed judgments were more in accordance with the pre-assigned labels than non-agreed judgments. However, while the reclassification rates were low for the E6K data [Table 10], for the MTurk data a larger number of music clips still needed to be reclassified with the exception of Cluster 3 [Table 11]. The pre-assigned labels were based on metadata and textual descriptions in the APM data set, which were written by music experts employed by APM. E6K judges are also expected to be music experts in the sense that they either have some background in music or music related research. Thus, we suspect this to be part of the reason for seeing this discrepancy. Perhaps the meanings of the mood clusters (except Cluster 3) are not as clear to lay people as they are to music experts. However, it is difficult to reach a solid conclusion based solely on these data, since we do not know how much music expertise the Turkers have. Future studies in a more controlled environment with two groups of evaluators, one consisting of music experts and the other, non-experts, will help us understand if there is significant difference between the two groups with regards to mood judgments.

Table 10. Reclassification among the clips with 2 or 3 agreed judgments in E6K data (percentages in parentheses)

	Reclassified Clips						
Pre-label	C1	C2	C3	C4	C5	Other	Total
Cluster1	111 (73.0)	12 (7.9)	18 (11.8)	7 (4.6)	0 (0.0)	4 (2.6)	152 (100)
Cluster2	1 (0.8)	109 (90.8)	6 (5.0)	1 (0.8)	0 (0.0)	3 (2.5)	120 (100)
Cluster3	3 (2.0)	2 (1.4)	136 (92.5)	1 (0.7)	1 (0.7)	4 (2.7)	147 (100)
Cluster4	0 (0.0)	7 (5.6)	1 (0.8)	110 (88.0)	2 (1.6)	5 (4.0)	125 (100)
Cluster5	6 (4.4)	0 (0.0)	2 (1.5)	2 (1.5)	123 (89.8)	4 (2.9)	137 (100)
Total	121	130	163	121	126	20	681

Table 11. Reclassification among the agreed clips in MTurk data (percentages in parentheses)

	Reclassified Clips						
Pre-label	C1	C2	C3	C4	C5	Other	Total
Cluster1	41 (34.7)	26 (22.0)	38 (32.2)	9 (7.6)	3 (2.5)	1 (0.8)	118 (100)
Cluster2	18 (13.8)	72 (55.4)	15 (11.5)	20 (15.4)	1 (0.8)	4 (3.1)	130 (100)
Cluster3	5 (3.1)	2 (1.2)	149 (92.5)	0 (0.0)	2 (1.2)	3 (1.9)	161 (100)
Cluster4	7 (6.3)	29 (26.1)	8 (7.2)	56 (50.5)	8 (7.2)	3 (2.7)	111 (100)
Cluster5	18 (13.3)	2 (1.5)	6 (4.4)	0 (0.0)	107 (79.3)	2 (1.5)	135 (100)
Total	89	131	216	85	121	13	655

5.2 Comparison of System Performances

5.2.1 Generating Ground Truth

In E6K, the ground truth data set was selected from the 681 music clips that received the same classification judgments from two or more human evaluators. 20 of the 681 music clips were classified in the "other" mood cluster, thus were not included in this process. Each cluster contained 120 music clips, thus resulting in 600 clips in total [Table 12].

Table 12. Composition of the ground truth data set from E6K

Cluster	Doubles w/2 judges	Doubles w/3 judges	Triples	Total
Cluster1	58	41	21	120
Cluster2	61	35	24	120
Cluster3	46	18	56	120
Cluster4	73	26	21	120
Cluster5	75	14	31	120
Total	**313**	**134**	**153**	**600**

In Table 12, each column indicates the number of music clips used for generating the ground truth based on the degree of agreement. The first column shows the music clips that only had 2 evaluators (judges) whom both agreed upon a mood cluster. The second column indicates the number of music clips that had 3 evaluators and 2 of them agreed upon a mood cluster. The third column shows the music clips that had 3 evaluators whom all agreed upon a mood cluster. When selecting music clips for the ground truth data set, triples (i.e., clips with 3 agreed judgments) were all included, since they had a higher level of agreement than other clips. Then the doubles (i.e., clips with 2 agreed judgments) were randomly selected to make the balanced data set of 120 clips in each cluster. The number of 120 clips in each cluster was decided by the MIR community via a poll on the AMC task audio collection in MIREX[7].

We also wanted to find out how much overlap exists between the ground truth data set generated from E6K and MTurk. We compared the ground truth data set from E6K with the MTurk data in order to find a common set of music clips that had agreed judgments. We were able to find a new ground truth set of 343 music clips as an intersection of MTurk agreements and E6K ground truth set [Table 13]. The system performance comparison in the next section is based on this new ground truth set. We were not able to compare system performance on all 655 clips with MTurk agreement because the systems were run against the E6K ground truth data set and thus IMIRSEL only has system classification results on the 600 clips of E6K ground truth set.

The fact that the intersection of MTurk agreements and the E6K ground truth only led to a set of 343 music clips (57.2% of the original E6K ground truth data set) further suggests that music mood classification is a challenging problems not only for machines, but for humans as well. By increasing the number of evaluators from 2/3 to 4/5 (i.e., E6K evaluators + Turkers), the number of music clips with agreement dropped to 270 (sum of numbers in the diagonal cells in Table 13), which is a significant loss of 55%. In future work, it would be interesting to see how the distribution of agreement changes with more evaluators.

Table 13. Intersection of E6K ground truth set and MTurk agreements

MTurk agreements	E6K ground truth set					
	C1	C2	C3	C4	C5	Total
Cluster1	29	4	0	2	7	42
Cluster2	17	44	0	14	0	75
Cluster3	5	5	91	1	1	103
Cluster4	0	13	0	42	0	55
Cluster5	4	0	0	0	64	68
Total	55	66	91	59	72	**343**

5.2.2 Cross-validation for Accuracies

Just like in the MIREX AMC task, we used 3-fold cross validation to test the accuracies of the algorithms that participated in the MIREX 2007 AMC task. Table 14 compares the order of those algorithms ranked by their accuracies based on E6K data and MTurk data.

Table 14. AMC algorithms ranked by accuracy (3-fold cross validation)[8]

E6K	Average accuracy	MTurk	Average accuracy
CL	0.65	GT	0.66
GT	0.64	CL	0.63
TL	0.64	TL	0.63
ME1	0.61	ME1	0.57
ME2	0.61	ME2	0.57
IM2	0.57	IM2	0.57
KL1	0.56	KL1	0.55
IM1	0.53	IM1	0.54
KL2	0.29	KL2	0.29

In comparing system (algorithm) performances, Friedman's ANOVA is applied to determine whether there are significant differences between the systems. Friedman's ANOVA is a non-parametric test which does not require the data to be normally distributed, and accuracy data are rarely distributed normally [9]. If Friedman's ANOVA indicates a significant difference exists among the systems (at $p < 0.05$), a follow up Tukey-Kramer Honestly Significantly Different (TK-HSD) analyses is then employed to determine which pairs of the systems are significantly different. These are the same tests done in the MIREX AMC task.

[7] http://www.music-ir.org/mirex/wiki/2007:Audio_Music_Mood_Classification#Audio_collection_poll

[8] In MIREX, algorithms are identified by the names of their developers. If multiple algorithms are submitted by the same developers, a number is attached to differentiate them: ME = Michael I. Mandel, Daniel P. W. Ellis; TL = Thomas Lidy, Andreas Rauber, Antonio Pertusa, José Manuel Iñesta; GT = George Tzanetakis; CL = Cyril Laurier, Perfecto Herrera; IM = IMIRSEL M2K; KL = Kyogu Lee

The order of the algorithms is almost the same for the two data sets except the first two, which are reversed (i.e., CL & GT). The results of the TK-HSD analysis indicate that there is no statistically significant difference between these two algorithms, however. The order of CL and GT algorithms was controversial within the E6K data set itself. Hu et al. [9] reported that "CL has the best ranks across all mood clusters despite its average accuracy being the second highest" (p. 465).

Figure 3 and Figure 4 show the TK-HSD evaluation results performed using the E6K and MTurk data, respectively. The X axes in both figures represent the mean column rank. As shown in Figure 3 [E6K data], the only pair that had statistically significant difference was CL and KL2. In Figure 4 [MTurk data], however, the only pair with statistically significant difference was GT and KL2. Therefore the E6K and MTurk data did produce different results with regards to that aspect. As previously discussed, however, CL and GT algorithms were indeed very close in their performance that one can hardly tell the difference even using the original E6K data set. No other pairs of algorithms were significantly different in either E6K or MTurk data set.

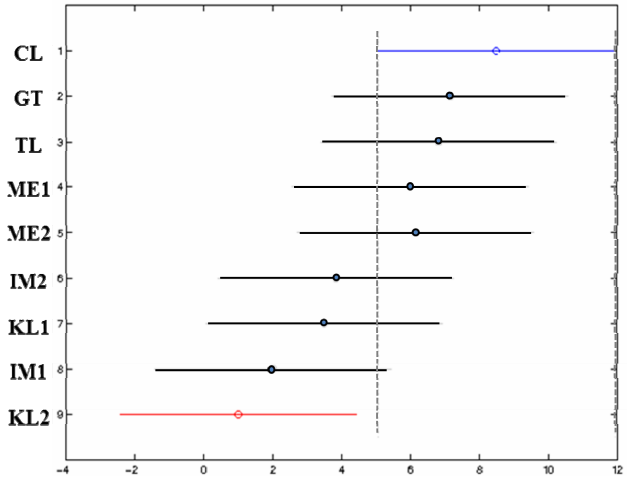

Figure 3. TK-HSD rank comparison for MIREX 2007 AMC based on judgments from E6K

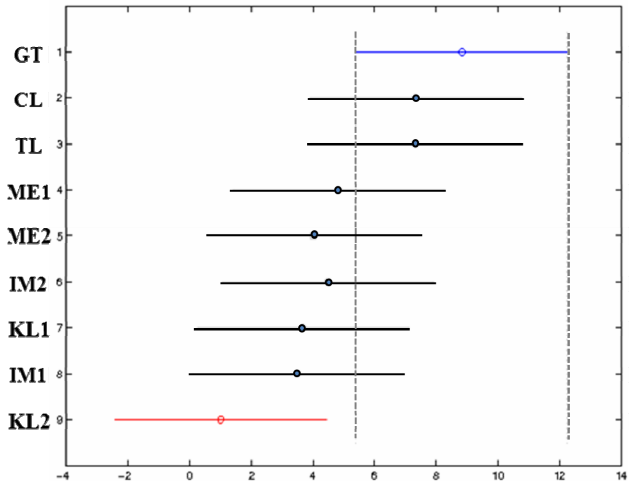

Figure 4. TK-HSD rank comparison for MIREX 2007 AMC based on judgments from MTurk

6. CONCLUSIONS AND FUTURE WORK

Overall the human judgments collected from E6K and MTurk showed similar distribution across the five mood clusters, and also comparable agreement rate overall. In general, Cluster 3 was the most agreed upon mood cluster among all user groups. We did observe some differences with the agreement between the E6K evaluators and APM music experts vs. Turkers and APM music experts for clusters, except for cluster 3. The comparison of confused mood clusters among E6K evaluators and among Turkers revealed a similar pattern. From the perspective of evaluating different algorithms, the results did slightly change the order of the ranked algorithms (i.e., the first and the second) although there was no statistical difference between these two algorithms. The rest of the order was preserved intact. This did change the statistically significant pair found in each data set, however.

Looking at the results, we would suggest using MTurk for Audio Music Mood Classification task with some reservation as there were data suggesting the differences between E6K evaluators and Turkers with regards to how they perceive certain mood clusters such as Cluster 1, in particular. Note that despite these differences, the evaluation outcome did not significantly change as the ranking of algorithms were mostly preserved. Additionally, MTurk does provide certain benefits over E6K, such as being able to collect the data much faster and in a more reliable way than having to continuously asking for help from volunteers from the MIR/MDL community. As a result, the MTurk data set was a lot easier to work with than the E6K data set, which contained music clips with different number of judgments.

More importantly, the confusion we observed among different mood clusters and the combined ground truth of 343 music clips which is only about 60% in size of the original E6K ground truth set all show how music mood is a challenging metadata to deal with. This is most likely due to the nature of music mood itself. Music mood is much more vague that other metadata such as artists, album/song titles [11], and even genre labels. As previously discussed, this is because music can carry multiple moods from mood clusters that are not mutually exclusive. Moreover, the mood can also change during the song [6]. We propose that better evaluation measures will have to take this into account permitting a song to be categorized in multiple mood clusters that perhaps can be weighted in some way.

One limitation of this study is that the E6K data were not as neat as the MTurk data due to the missing number of judgments, which inevitably made the comparison a bit challenging. We did, however, made our best effort to provide fair comparison of the results from E6K and MTurk. In future studies, we hope to conduct a more controlled study with different user groups (e.g., experts with music background, people who have no educational background in music) accompanied by in-depth interviews in order to find out more about how they perceive music mood and which factors play important role when they classify music clips in different mood clusters.

7. ACKOWLEDGEMENTS

We would like to thank the International Music Information Retrieval Systems Evaluation Lab (IMIRSEL) for providing us with access to the test collection for AMC task as well as past MIREX data. We also thank the anonymous reviewers for their helpful comments.

8. REFERENCES

[1] Alonso, O., and Mizzaro, S. 2009. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, 15-16.

[2] Bischoff, K., Firan, C. S., Nejdl, W., and Paiu, R. 2009. How do you feel about "Dancing Queen"? Deriving mood and theme annotations from user tags. In *Proceedings of THE 9th Joint Conference on Digital Libraries* (JCDL'09). ACM, New York, NY, USA, 285-294. DOI= http://doi.acm.org/10.1145/1555400.1555448

[3] Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22, 2, 249–254.

[4] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1, 37-46.

[5] Eerola, T, Lartillot, O., and Toivianen, P. 2009. Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proceedings of the 10th ISMIR (International Society for Music Information Retrieval) Conference*, 621-626.

[6] Hu, X. 2010. Music and mood: Where theory and reality meet. In *Proceedings of 2010 iConference.*

[7] Hu, X. and Downie, S. J. 2007. Exploring mood metadata: Relationships with genre, artist and usage metadata. In *Proceedings of the 8th ISMIR (International Society for Music Information Retrieval) Conference*, 67-72.

[8] Hu, X. and Downie, S. J. 2010. Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the 10th Joint Conference on Digital Libraries* (JCDL'10), ACM, New York, NY, USA, 159-168. DOI= http://doi.acm.org/10.1145/1816123. 1816146.

[9] Hu, X., Downie, S. J., Laurier, C., Bay, M., and Ehmann, A. F. 2008. The 2007 MIREX Audio Mood Classification task: Lessons learned. In *Proceedings of the 9th ISMIR (International Society for Music Information Retrieval) Conference*, 462-467.

[10] Kim, Y., Schmidt, E., and Emelle, L. 2008. Moodswings: A collaborative game for music mood label collection. In *Proceedings of the 9th ISMIR (International Society for Music Information Retrieval) Conference*, 231-236.

[11] Kim, Y., Schmidt, E., Migneco, R., Morton, B. G., Richardson, P., Scott, J., Speck, J. A. and Turnbull, D. 2010. Music emotion recognition: A state of the art review. In *Proceedings of the 11th ISMIR (International Society for Music Information Retrieval) Conference*, 255-266.

[12] Kitter, A., Chi, E. H., and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), ACM, New York, NY, USA, 453-456. DOI= http://doi.acm.org/10.1145/1357054.1357127

[13] Laurier, C., Sordo M., Serrà J., and Herrera P. 2009. Music mood representations from social tags. In *Proceedings of the 11th ISMIR (International Society for Music Information Retrieval) Conference*, 381-386.

[14] Lee, J. H. 2010. Crowdsourcing music similarity judgments using Mechanical Turk. In *Proceedings of the 11th ISMIR (International Society for Music Information Retrieval) Conference*, 183-188.

[15] Lee, J. H. , Hill, T., and Work, L. 2012. What does music mood mean for real users? In *Proceedings of the 2012 iConference*, ACM, New York, NY, USA, 112-119. DOI= http://doi.acm.org/10.1145/ 2132176.2132191

[16] Lu, L., Liu, D., and Zhang, H. 2006. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14, 1, 5-18. DOI=10.1109/TSA.2005.860344

[17] Mandel, M., Eck, D, and Bengio, Y. 2010. Learning tags that vary within a song. In *Proceedings of the 11th ISMIR (International Society for Music Information Retrieval) Conference*, 399-404.

[18] Meharabian, A. and Russell, J. A. 1974. *An Approach to Environmental Psychology*. MIT Press.

[19] Mion, L., and Poli, G. D. 2008. Score-independent audio features for description of music expression. *IEEE Transactions on Audio, Speech and Language Processing*, 16, 2, 458-466. DOI=10.1109/TASL.2007.913743

[20] Russell, J. 1980. A Circumplex model of affect. *Journal of Personality and Social Psychology*, 39, 1161-1178.

[21] Schuller, B., Hage, C., Schuller, D., and Rigoll, G. 2010. Mister D.J., Cheer me up!: Musical and textual features for automatic mood classification. *Journal of New Music Research*, 39, 1, 13-34.

[22] Skowronek, J., McKinney, M.F., & van de Par, S. 2006. Ground truth for automatic music mood classification. In *Proceedings of the 7th ISMIR (International Society for Music Information Retrieval) Conference*, 395-396.

[23] Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. 2008. Cheap and fast – But is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 254-263.

[24] Urbano, J. Morato, J., Marrero, M., and Martin, D. 2010. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, 9-16.

[25] Vignoli, F. 2004. Digital Music Interaction concepts: a user study. In *Proceedings of the 5th ISMIR (International Society for Music Information Retrieval) Conference,* 415-421.