Crowdsourcing Music Similarity Judgments using Mechanical Turk

Jin Ha Lee
University of Washington
jinhalee@uw.edu

ABSTRACT

Collecting human judgments for music similarity evaluation has always been a difficult and time consuming task. This paper explores the viability of Amazon Mechanical Turk (MTurk) for collecting human judgments for audio music similarity evaluation tasks. We compared the similarity judgments collected from Evalutron6000 (E6K) and MTurk using the Music Information Retrieval Evaluation eXchange 2009 Audio Music Similarity and Retrieval task dataset. Our data show that the results are highly comparable, and MTurk may be a useful method for collecting subjective ground truth data. Furthermore, there are several benefits to using MTurk over the traditional E6K infrastructure. We conclude that using MTurk is a practical alternative of music similarity when it is used with some precautions.

1. INTRODUCTION

A constant source of frustration for designers and developers of music information retrieval systems is finding users to generate ground truth for evaluation. This is particularly true in music similarity tasks where algorithms are attempting to model some aspect of human intuition or understanding and predict the similarity among a set of songs. Getting humans to verify the results of these algorithms is tedious as a modest collection of several hundred tracks can require tens of thousands of pair-wise comparisons which potentially need to be evaluated.

Our motivation for this study is to explore the usefulness of Amazon Mechanical Turk (MTurk) (http://mturk.com) for collecting the human judgments necessary for evaluating music similarity tasks like the Audio Music Similarity (AMS) and Symbolic Melodic Similarity (SMS) tasks in the Music Information Retrieval Evaluation eXchange (MIREX). In this paper, we compare the similarity judgments obtained from MTurk and Evalutron6000 (E6K) on the same data set used in the MIREX 2009 AMS task. We also compare how these judgments affect the ultimate evaluation outcomes as published by the International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) in the annual MIREX evaluation. Additionally, we were interested in exploring how MTurk could be used to supplement or replace E6K in future music similarity evaluations, opening the possibility for continuous evaluation without incurring the overhead of a full MIREX/E6K-based evaluation.

2. BACKGROUND

The AMS and SMS tasks were carried out as part of MIREX using the E6K infrastructure. Both tasks rely on human judgments of music similarity as ground truth for evaluation of algorithm performance. Every year the IMIRSEL group at the University of Illinois seeks volunteers from the ISMIR community to complete a set of similarity judgments. In addition to the MIREX AMS and SMS tasks, a number of studies have looked at the human judgments of music similarity; to name a few, Aucouturier & Pachet [2], Ellis et al. [5], Berenzweigh et al. [4], Timmers [13], Schubert & Stevens [11], and Novello & McKinney [10].

In these studies, the human judgments were collected by a web survey or by recruiting a number of subjects including musicians and non-experts. Two typical methods were used. In some studies, the users were presented a set of three song excerpts (triads) and were asked to choose the most similar and most dissimilar of the three possible pairs. In other studies, the users were presented with pairs of song excerpts and were asked to rate the similarity between the pairs. Regardless of which method was used, collecting human similarity judgments has always been a challenging, expensive, and time-consuming process. Searching for a better model for obtaining human similarity judgments is especially important considering the fact that the general trend in recent MIREX AMS submissions is to submit multiple variations of an algorithm; there were a total of 15 submissions from 9 participants in 2009 compared to 6 submissions from 5 participants in 2006. There is also a trend towards larger datasets, and evaluating more queries [9].

2.1 Evalutron6000 (E6K)

IMIRSEL collect similarity judgments from human graders using E6K which is in the form of a web-based survey. The graders are supposed to be music experts since they are volunteers from the ISMIR community who have backgrounds in music or music-related research. Collecting human judgments is a long and arduous process every year since the organizers must rely on volunteer labor. Every year it takes days to weeks to complete the evaluation. Table 1 shows the number of days it took to collect the human similarity judgments for the AMS and SMS tasks in past MIREX cycles.

	AMS	SMS
2006	15 days	18 days
2007	8 days	4 days
2009	14 days	n/a

Table 1. Number of days needed to collect human similarity judgments in past MIREX cycles.

2.2 Amazon Mechanical Turk

Amazon Mechanical Turk (MTurk) is a service which allows people to leverage human-computational power at scale to complete large numbers of tasks requiring human-intervention, cheaply and efficiently. Requesters upload tasks to the service, where they are matched with willing workers. Payment is mediated by Amazon, with a small per-task fee charged to the requester.

Tasks in MTurk are called HITs (Human Intelligence Tasks). Requesters define their HITs using an HTML-based template language and a data source which is used to populate the templates and generate the individual HITs. The requester also offers a payment amount and time limit for each HIT, and can limit who is able to complete the HITs, such as requiring workers to have a minimum percentage of previously accepted HITs.

Requesters can create a qualification test and require workers to have to pass it before being eligible to work on their HITs. Upon completion of a HIT, the requester can review the work and approve or reject payment on HITs individually. Furthermore, workers can be blocked by the requester which would reject all their un-approved HITs and prevent them from completing and submitting additional HITs for that same task.

Workers in MTurk call themselves Turkers. There are over 200,000 Turkers, from all parts of the globe. Ipeirotis [6] conducted a survey of 1,000 Turkers in February 2010. In total, Turkers represented 66 different countries, with 46.8% from the United States, and 34% from India. Among US workers, most (65.6%) are women; however, among Indian workers, most (70%) are men. 62.8% of respondents had a Bachelor's degree or higher.

MTurk has an API through which HITs can be created and results approved and downloaded, making it readily integratable into automated processes. For example, MTurk is successfully used to complete tasks such as filtering user-generated web content, searching through satellite photos for missing aircraft [3], and is gaining traction as a resource in research. Within the SIGIR community, MTurk has been proposed for use in generating relevance judgments in TREC-like evaluations [1]. Alonso and Mizzaro [1] compared MTurk to TREC experts and found the results from MTurk to be comparable to TREC's expert-generated ground truth data. They even claim Turkers found several errors in the TREC data.

Snow, et al. [12] have explored using MTurk for generating ground truth data for several kinds of Natural Language Processing tasks, including determining valence and affect in text, and assigning similarity scores. They found it possible to obtain results on par with those of domain experts.

Kittur et al [8] used MTurk to rate the quality of Wikipedia articles. Among their experiments, they found a naïve approach of simply asking Turkers to rate an article led to inconsistent responses which did not correlate strongly with ratings given by experts. However, when they redesigned the HITs to include several verifiable questions which could be used to filter out "bad" responses, the results improved significantly. They argue that the verification questions serve two purposes: first, they allow the requester to assess the quality of the response; and second, they signal to the Turkers that their responses are being scrutinized.

3. RESEARCH QUESTIONS & STUDY DESIGN

In this paper, we explore two main research questions:

I) How do music similarity judgments obtained from Mechanical Turk compare to those collected from music experts in the Evalutron6000?; and

II) How do evaluation outcomes for tasks like MIREX's Audio Music Similarity evaluation differ when based on similarity judgments collected from Mechanical Turk as compared to Evalutron6000?

To study these questions, we replicated the E6K similarity assessment and subsequent evaluation of the AMS task in the 2009 MIREX evaluation using MTurk. We obtained the query-candidate (QC) results lists from the IMIRSEL lab, consisting of 100 queries, and the top 5 candidates per query returned from the 15 participating algorithms in MIREX 2009. There were a total of 6,732 unique QC pairs which needed to be judged.

In order to keep the amount of work in each HIT reasonable, we limited the number of similarity judgments per HIT to 15 QC pairs, and all QC pairs in a HIT shared the same query. Among the 15 QC pairs in a HIT, two candidates were included for checking the quality of the ratings. One was an identity check and it asked the Turker to rate the similarity of the query to itself. The Turker should indicate that this candidate is "Very Similar (VS)" to the query as they are identical. This was also done in E6K in 2009; however, we were unable to locate those data published for comparison.

The other quality check was a consistency check; the same candidate was included twice in a single HIT, once towards the beginning, and again towards the end of the list of candidates. The expectation here was that the Turker should provide the same response for both instances since they are the same candidate. Excluding these two QC pairs for checking the quality of the results, there were 13 unique QC pairs in an individual HIT. The quality checks were mixed among the other candidates and were not specially demarcated in any way.

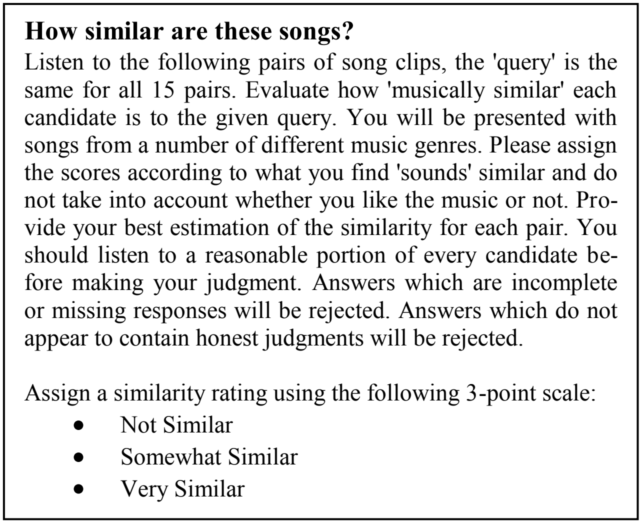
The list of all candidates for each query was broken down into multiple HITs containing 13 unique QC pairs. In the event that the last HIT contained less than 13 candidates, the list was padded to 13 with additional candidates selected from that query's other HITs. These padded judgments were not used in the evaluation. Each HIT was completed by single Turker, with the possibility that a single query could be evaluated by multiple Turkers. This was different from MIREX 2009 AMS where a single grader was responsible for judging all candidates for a single query, but is similar to MIREX 2006 AMS, where candidate lists were divided among multiple graders.

A total of 583 HITs were created, and we offered $0.20 per completed HIT. Instructions similar to what are given in E6K were given to Turkers as shown in Figure 1. Figure 2 shows a partial screenshot of MTurk's evaluation page. We tried to reproduce the E6K interface as much as possible. The Turkers were asked to rate the similarity on the E6K BROAD scale (Not Similar, Somewhat Similar, and Very Similar) and were not asked to provide a FINE score (0-10) in order to simplify the task. Additionally, we used the Yahoo! Media Player in the MTurk version of the interface, rather than the E6K player because it was much simpler to use and has much better cross-browser compatibility.

In addition to the HITs described above, we created 4 more HITs in order to see how much variability there was among the Turkers' responses. In these HITs, we took one candidate from each query, put 15 QC pairs into each HIT and had 3 different Turkers complete each HIT. This gave us multiple ratings for the same QC pairs, and allowed us to test the inter-rater agreement. We paid $0.20 for each of these 12 HITs. The total cost for all 595 HITs, including Amazon's administrative fees, was $130.90. Ultimately, we paid less than $0.02 per usable judgment, for a rate of approximately 53 judgments per US dollar.

4. DATA & OBSERVATIONS

In total we collected 15,705 similarity judgments from 1,047 submitted HITs, plus 180 additional judgments created to test agreement among the Turkers. Of the 1,047 HITs submitted, we approved 583 (55.7%), and rejected 464 (44.3%). We rejected HITs which were missing responses, those which were completed too quickly (less than 45 seconds), those in which Turkers failed to assign a Very Similar score to the identity case of a query compared to itself, and those in which Turkers assigned two different scores to the same candidate repeated in the list. Accounting for the rejected HITs, the integrity-check judgments, and for list padding, we ended up with 6,732 unique judgments. Even having to discard almost half the judgments, we were still able to obtain the needed results in less than 12 hours, an order of magnitude faster than the average E6K cycle (see Table 1).



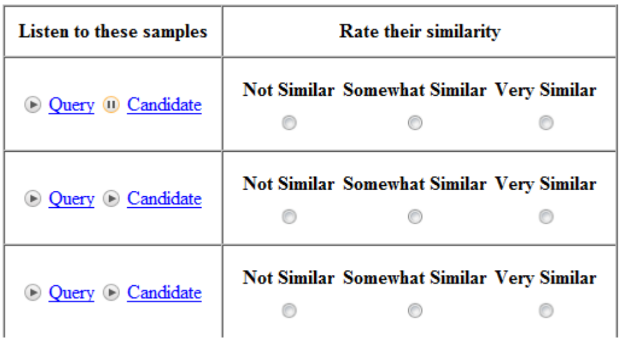
How similar are these songs?

Listen to the following pairs of song clips, the 'query' is the same for all 15 pairs. Evaluate how 'musically similar' each candidate is to the given query. You will be presented with songs from a number of different music genres. Please assign the scores according to what you find 'sounds' similar and do not take into account whether you like the music or not. Provide your best estimation of the similarity for each pair. You should listen to a reasonable portion of every candidate before making your judgment. Answers which are incomplete or missing responses will be rejected. Answers which do not appear to contain honest judgments will be rejected.

Assign a similarity rating using the following 3-point scale:

- Not Similar
- Somewhat Similar
- Very Similar

Figure 1. Instructions given to Turkers are based on the instructions given to E6K graders.



Listen to these samples	Rate their similarity
Query Candidate	Not Similar ○ Somewhat Similar ○ Very Similar ○
Query Candidate	Not Similar ○ Somewhat Similar ○ Very Similar ○
Query Candidate	Not Similar ○ Somewhat Similar ○ Very Similar ○

Figure 2. Partial screenshot of an MTurk HIT.

Research Question I: *How do music similarity judgments obtained from Mechanical Turk compare to those collected from music experts in the Evalutron6000?*

The similarity scores derived from MTurk are different from those obtained from E6K, but they are not entirely incomparable. We compared the 6,732 similarity judgments obtained from MTurk to the 6,732 judgments obtained via E6K in AMS 2009 for the same set of QC pairs. We measured the percent-agreement between the MTurk results and the E6K results, and found that 54.6% of the pair-wise ratings were the same. Agreement increases to 72.4% when we consider similarity as a binary decision (Very Similar & Somewhat Similar vs. Not Similar). To our knowledge, E6K has not been used to do multiple evaluations of the same data set in this way, so we do not have a basis for comparison. However, the relatively low agreement between MTurk and E6K does underscore the subjective nature of similarity ratings and similarity-based tasks in general.

The two sets of similarity judgments (MTurk & E6K) have a Pearson's correlation of r=0.495, which while not particularly strong, is comparable to the correlation (r=0.433) found by Snow [12] between NLP experts and Turkers in an affect annotation study.

Looking at the data in aggregate, Figure 3 shows the similarity judgments derived from MTurk tended to skew towards Not Similar (NS=3,605), where as E6K graders tended to assign similarity scores more uniformly across the categories. In AMS 2009, as a requirement of participation each team had to provide an E6K volunteer to help with the judging for each algorithm they submitted. These volunteers might have had some stake in the outcome of the evaluation which might explain the greater proportion of VS scores in 2009 compared to 2006 and MTurk where the ratings were generated by independent volunteers. However, Figure 3 also shows the distribution of scores from previous MIREX cycles, and while the underlying queries and candidates differ across the years, the distribution of scores from MTurk is not dissimilar to other distributions from previous E6K results.

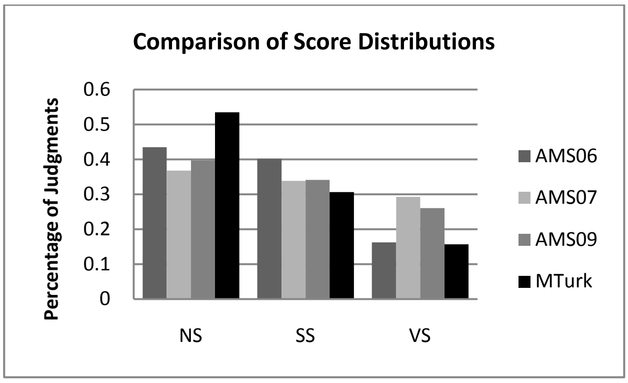


Figure 3. Distributions of scores from the 3 years of MIREX AMS evaluation using E6K compared to the distribution of scores derived from MTurk.

In order to further investigate the similarity of MTurk judgments to E6K judgments, we collected multiple similarity judgments over the same set of QC pairs from several different Turkers. Specifically, we randomly selected one candidate from each candidate list for each of the 60 queries used in MIREX 2006 AMS, and created 4 lists of 15 QC pairs to be evaluated by three different Turkers. This setup is very similar to how E6K was configured in 2006 (multiple graders rating portions of candidate lists), and given we were working with a subset of the 2006 data, we are able to compare the inter-Turker agreement to that found by Jones, et al. [7]. While we did not test SMS in MTurk, we have provided the data for comparison.

Table 2 shows the agreement using the 3-level and 2-level analysis used in [7]. The overall distribution of scores is fairly similar to the AMS 2006 results; the proportions of QC pairs across the various levels of agreement are comparable between the two data sources. The percentage of cases of total agreement are the same, and there is some shifting of cases between partial and total disagreement with slightly more cases of total disagreement (VS,SS,NS) among Turkers. This may be due to the nature of the QC pairs sampled for this evaluation which is plausible given the small sample size, or it may be inherent given what [7] describes the vague definition of "music similarity".

3-level	SMS 2006		AMS 2006		MTURK	
VS,VS,VS	114	12.6%	61	3.7%	4	6.7%
SS,SS,SS	38	4.3%	137	8.4%	1	1.7%
NS,NS,NS	263	29.1%	293	18.0%	13	21.7%
Triples	**415**	**45.9%**	**491**	**30.1%**	**18**	**30.0%**
VS,VS,*	24	2.7%	150	9.2%	3	5.0%
SS,SS,*	158	17.5%	469	28.8%	18	30.0%
NS,NS,*	288	31.8%	404	24.8%	11	18.3%
Doubles	**470**	**51.9%**	**1023**	**62.8%**	**32**	**53.3%**
VS,SS,NS	20	2.2%	115	7.1%	10	16.7%
2-level	**SMS 2006**		**AMS 2006**		**MTURK**	
S,S,S	188	20.8%	494	30.3%	19	31.7%
NS,NS,NS	263	29.1%	293	18.0%	13	21.7%
Triples	**451**	**49.8%**	**787**	**48.3%**	**32**	**53.3%**
S,S,N	166	18.3%	438	26.9%	17	28.3%
N,N,S	288	31.8%	404	24.8%	11	18.3%
Doubles	**454**	**50.2%**	**842**	**51.7%**	**28**	**46.7%**

Table 2. Comparison of disagreement among Turkers and E6K graders from MIREX 2006 AMS evaluation.

When we examine the results using a binary similarity measure (SS+VS against NS), we see greater similarity between the E6K graders and the Turkers. The distributions across the levels of agreement are nearly identical between the two sets. Jones [7] also found greater consensus when considering similarity on a binary scale, and suggest that the binary metric might be sufficient for the task of evaluation.

Research Question II: *How do evaluation outcomes for tasks like MIREX's Audio Music Similarity evaluation differ when based on similarity judgments collected from Mechanical Turk as compared to Evalutron6000?*

Given the similarity judgments derived from MTurk appear to be different from those generated via E6K, we wished to see if those differences have any substantial bearing on MIREX evaluations. It is possible that the individual ratings differ, but still produce similar outcomes in comparing the performance of individual algorithms in the audio music similarity task. Conversely, the differences may in fact be substantive and produce significantly different end results.

MIREX evaluates the performance of similarity algorithms using Friedman test with repeated-measures. Figure 4 shows a graphical depiction of the results of the MIREX 2009 AMS Friedman evaluation, comparing the average rankings among the different algorithms. There are clearly two distinct groupings to the data: ANO, BSWH1, BSWH2, CL2, GT, LR, PS1, PS2, SH1, SH2; and BF1, BF2, CL1, ME1, ME2. One way to interpret the figure is that all algorithms in one group are significantly different from all algorithms in the other group, but within the groups the algorithms are not all significantly different from each other. So, while PS2 does appear to lie slightly outside the rest of the larger group, it is not significantly different from all members of that group (it overlaps partially with PS1).

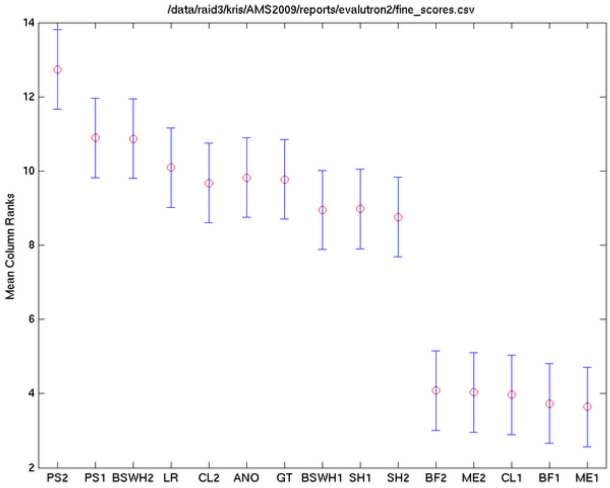


Figure 4. Friedman rank comparison for MIREX 2009 AMS based on judgments from E6K (from [9]).

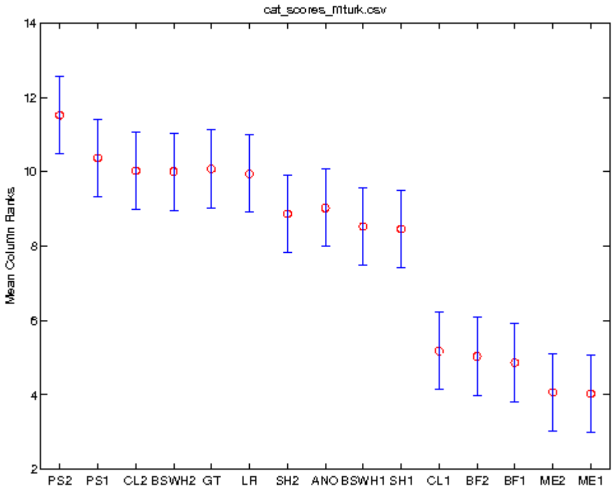


Figure 5. Friedman rank comparison for MIREX 2009 AMS based on judgments from MTurk.

Figure 5 shows the Friedman evaluation results performed using the similarity judgments derived from MTurk. As we can see, the two main groupings are still evident, with clearly significant differences in the performance of the algorithms between the groups. However, the gap between the groups is narrower, and the data are more compact. Furthermore, the global ordering of the algorithms has changed. Notwithstanding these differences, the overall results are remarkably similar.

The main significant differences between the groupings are still evident and significant differences are still preserved among most of the algorithm pairs. Table 3 summarizes the differences between the E6K and MTurk results. Out of the 105 possible pair-wise comparisons among the 15 algorithms submitted to MIREX 2009 AMS, only six (5.7%) algorithm pairings were determined to be significantly different based on E6K judgments and not found to be significantly different based on MTurk judgments. No algorithms which were not significantly different under E6K were found to be significantly different under MTurk. This discrepancy is not substantially different compared to the Friedman results computed using the E6K BROAD scores and FINE scores in AMS 2009. The Friedman test based on the FINE scores rates 3 (2.9%) algorithm-pairs differently than the BROAD score results [9].

Algorithm 1	Algorithm 2	Significant in E6K?	Significant in MTurk?
BSWH2	SH1	TRUE	FALSE
PS1	SH1	TRUE	FALSE
PS1	SH2	TRUE	FALSE
PS2	CL2	TRUE	FALSE
PS2	GT	TRUE	FALSE
PS2	LR	TRUE	FALSE

Table 3. Excerpt from Friedman table of differences in significance between E6K and MTurk.

5. IMPLICATIONS FOR MIR EVALUATION

Overall, we were quite impressed with the quality of the data we were able to obtain from MTurk. We can identify many benefits to using MTurk in MIR evaluation, some of which include:

1. The tasks were inexpensive to submit, costing approximately USD$0.02 per judgment, although other researchers have obtained quality results for far less payment (c.f., [1],[12]).
2. Evaluation was fast, taking less than 12 hours to complete. It took over 2 weeks to obtain the same number of judgments using E6K.
3. MTurk has a scriptable API, meaning it can be used for “on-demand” evaluation. This is especially attractive given the developments of the MIREX Do-It-Yourself infrastructure.
4. No judging fatigue. MTurk provides a nearly endless supply of willing labor, and it is not feasible to expect the ISMIR community to continuously provide input for use in MIREX DIY evaluations. Several Turkers sent us messages saying they found the HITs more fun than most other HITs on MTurk, one even expressed a willingness to work for free.
5. Using MTurk for similarity judgments avoids any conflict of interest inherent in asking participants or their labmates to evaluate the results.
6. MTurk provides a mechanism for compensating volunteers for their time and effort.
7. MTurk is considered “exempt” by human subjects review as it falls under the description of a web survey, or as paid workers.
8. Using MTurk does not preclude restricting participation in the evaluation to only ISMIR members. It is possible to require Turkers obtain a qualification prior to working on any HITs. Qualification credentials could include membership in the ISMIR Society.

9. MTurk is very stable. It is built on Amazon's cloud infrastructure and is very robust.

However, there are several limitations which need to be kept in mind when using MTurk, including:

1. HITs need to contain validation questions which allow you to check responses for quality and consistency. Early in our exploration of MTurk we created HITs without validation questions and the data we collected was highly inconsistent.
2. The instructions we provided also needed clarification over several tests. We found it helpful to spell out the precise conditions why a HIT would be rejected in the instructions. Likewise, you need to provide clear explanations why HITs were rejected.
3. It does cost money to use MTurk. While each HIT is cheap, in aggregate the price adds up quickly. Amazon's overhead is either \$0.005 or 10% per HIT, whichever is greater. However, the total costs are small compared to the value of E6K volunteers' time.
4. Some university human subjects review boards might not be very familiar with MTurk and might be unsure how to handle it. This could slow applications.

6. CONCLUSIONS

Our data show that while the specific judgments may differ, using MTurk produces comparable results to using E6K for collecting human similarity judgments. The differences are not dissimilar to the findings of other studies of MTurk and do not significantly alter the MIREX evaluation outcomes, indicating that MTurk may be a used as a reliable source of similarity judgments for audio-based music similarity comparisons. Overall, the differences between MTurk and E6K judgments resulted in a 5.7% difference in the ultimate outcome of the Friedman test comparing the 15 algorithms submitted to AMS 2009.

We are excited by the possibilities MTurk offers to the development of future evaluation infrastructure, like the MIREX DIY, but are most excited by what MTurk can do for the community as a whole. There are many types of data which could be collected from Turkers, and it remains to be seen how well MTurk is suited for collecting those data. In future work we would like to explore other data types; for example, music mood labels, music tagging, onset annotation work, key-identification, humming or singing, tapping, transcribing, etc.

7. ACKNOWLEDGEMENTS

We would like to thank IMIRSEL for providing us with access to the MIREX data and helping us to reproduce the MIREX analyses.

8. REFERENCES

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

[2] J-J. Aucouturier and F. Pachet: "Music Similarity Measures: What's the Use?" *Proceedings of the 3rd International Society of Music Information Retrieval (ISMIR) Conference,* pp. 157-163, 2002.

[3] D. Axe; "Geeks Spot Fossett?" Wired, Retrieved from http://www.wired.com/dangerroom/2007/09/geeks-spot-foss/comment-page-3/, 2007.

[4] A. Berenzweig, B. Logan, D. P. W. Ellis, and B. Whitman: "A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures," *Proceedings of the 4th ISMIR,* pp. 99–105, 2003.

[5] D. P. W. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence: "The Quest For Ground Truth in Musical Artist Similarity," *Proceedings of the 3rd ISMIR Conference,* pp. 170-177, 2002.

[6] P. G. Ipeirotis: "Demographics of Mechanical Turk," CeDER Working Papers, Retrieved from http://hdl.handle.net/2451/29585, 2010.

[7] M. C. Jones, J. S. Downie, and A. F. Ehmann: "Human Similarity Judgments: Implications for the Design of Formal Evaluations," *Proceedings of the 8th ISMIR,* pp. 539-542, 2007.

[8] A. Kittur, E. H. Chi, and B. Suh: "Crowdsourcing User Studies With Mechanical Turk," *CHI 2008 Proceedings,* pp. 453-456, 2008.

[9] MIREX 2009 Wiki, Retrieved from http://music-ir.org/mirex/2009/, 2009.

[10] A. Novello and M. F. McKinney: "Assessment of Perceptual Music Similarity," *Proceedings of the 8th ISMIR,* pp. 111-112, 2007.

[11] E. Schubert and C. Stevens: "The Effect of Implied Harmony, Contour and Musical Expertise on Judgments of Similarity of Familiar Melodies," *Journal of New Music Research,* Vol. 35, No. 2, pp. 161-174, 2006.

[12] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng: "Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks," *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing,* pp. 254-263, 2008.

[13] R. Timmers: "Predicting the Similarity Between Expressive Performances of Music From Measurements of Tempo and Dynamics," *Journal of the Acoustical Society of America,* Vol. 117, No. 1, pp. 391-399, 2005.