HUMAN SIMILARITY JUDGMENTS: IMPLICATIONS FOR THE DESIGN OF FORMAL EVALUATIONS

M. Cameron Jones **J. Stephen Downie** **Andreas F. Ehmann**

International Music Information Retrieval Systems Evaluation Laboratory
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

ABSTRACT

This paper presents findings of a series of analyses of human similarity judgments from the Symbolic Melodic Similarity, and Audio Music Similarity tasks from the Music Information Retrieval Evaluation Exchange (MIREX) 2006. The categorical judgment data generated by the evaluators is analyzed with regard to judgment stability, inter-grader reliability, and patterns of disagreement, both within and between the two tasks. An exploration of this space yields implications for the design of MIREX-like evaluations.

1. INTRODUCTION

The International Music Information Retrieval Systems Evaluation Laboratory (IMIRSEL) at the University of Illinois at Urbana-Champaign has been hosting and running the Annual Music Information Retrieval Evaluation eXchange (MIREX) since 2005. Inspired by TREC, the goal of MIREX is to formally evaluate state-of-the-art algorithms for Music Information Retrieval (MIR) systems [2].

MIREX 2006 comprised nine separate evaluation tasks which were defined by community input [5]. Two of these tasks, "Symbolic Melodic Similarity" (SMS) and "Audio Music Similarity and Retrieval" (AMS), called for human judgments of similarity in order to establish ground truth for the evaluation of the submitted algorithms. In order to capture these similarity judgments we created a new web-based tool called the "Evalutron 6000" (E6K).

In this paper, we present findings from our analysis of categorical human similarity judgment data collected using the E6K. We explore the consistency of the graders' scoring, measuring the amount of disagreement among graders. We discuss the implications of our findings for the design of future tasks which utilize human judgements of similarity in the MIR domain.

2. DATA CAPTURE: EVALUTRON 6000

The SMS and AMS tasks shared a common structure. Each task participant's algorithm was run against a collection of either symbolic or audio music files. For each query, each algorithm returned a list of the top-ranked "candidate" songs, the length *n* of the candidate lists was ten for SMS and five for AMS. All resulting candidates for each query were merged and then evaluated by graders using the E6K.

In the E6K, graders score the anonymized set of candidates for each query anonymously. Individual graders are tracked, but their scores are kept independent of their identities. This tracking allowed us to log each grader's interactions with the E6K. Events logged include: score inputs, score modifications, auditions, etc. Table 1 provides descriptive statistics for each the two evaluation tasks.

	SMS	AMS
No. of events logged	23,491	46,254
No. of submitted algorithms	8	6
Total no. of queries	17	60
Total no. of query-candidate pairs	905	1,629
No. of graders	21	24
No. of queries per grader	15	7-8
Avg. size of candidate lists	15	27
Avg. no. of evaluations per grader	225	205

Table 1. Summary of Evalutron 6000 statistics.

After listening to each query-candidate pair, graders were asked to rate the degree of similarity of the candidate to the query in two ways: 1) by selecting one of the three BROAD categories of similarity: Not Similar (NS), Somewhat Similar (SS), and Very Similar (VS); and, 2) by assigning a FINE score between 0.0 (Least similar) and 10.0 (Most similar). Each query-candidate pair was evaluated by three different graders. Data were collected between 5 Sept. and 20 Sept., 2006, from volunteer graders from the MIR/MDL research community, representing 11 different countries.

3. MEASURING DISAGREEMENT

Understanding the consistency of the graders' judgments is essential to interpreting human judgments of similarity in contexts such as MIREX. Previous studies [1,6] have analyzed the consistency of judgments between BROAD scores and FINE scores. Figure 1 shows the consistency of assignment of FINE scores within BROAD categories for both SMS and AMS tasks. The variation of FINE scores within the BROAD SS category for AMS is particularly interesting, indicating that graders were not very consistent in assigning FINE scores to items they had marked as Somewhat Similar (SS) for this task. The differences between the tasks and of the consistency of FINE scores are discussed in more detail in [1].

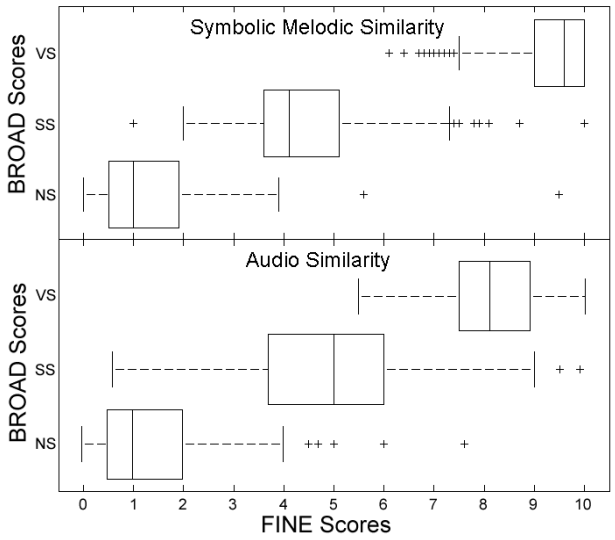

Figure 1. Distribution of FINE scores within BROAD score categories for SMS and AMS tasks.

The open, web-based nature of the E6K allowed graders to enter and leave the system at their convenience. This had the consequence of allowing graders to revisit and revise their similarity judgments. In assigning BROAD judgments to query-candidate pairs, graders did not tend to change their judgments (see Table 2). The majority of graders in both SMS and AMS made only a single BROAD category judgment, not changing it. Overall, only 5.86% of SMS graders and 9.04% of AMS graders changed their judgment at some point during the grading process.

	# Grading Opportunities	# Grading Events	Mean	Max	Mode
SMS	2715	2907	1.07	5	1 (94.14%)
AMS	4887	5450	1.12	14	1 (90.96%)

Table 2. Description of grader judgment changes.

3.1. Inter-grader Reliability

Inter-grader reliability measures the relative objectivity (or inter-subjectivity) of judgments; a necessary condition for measuring the validity of the E6K framework and, more generally, the design of future projects which will utilize similar mechanisms for evaluation. Several metrics have been developed to measure inter-grader reliability.

Fleiss's Kappa is a measure of inter-grader reliability for nominal data, and is based on Cohen's two-grader reliability Kappa, but measures reliability among an arbitrary number of graders [3]. The equation for Fleiss's Kappa is given in (3.1.1).

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \qquad (3.1.1)$$

where:

$$\bar{P} = \frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}\left(n_{ij} - 1\right) \qquad (3.1.2)$$

$$\bar{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 \qquad (3.1.3)$$

In the Fleiss Kappa equation, N is total number of query-candidate pairs to be graded; n is the number of judgments per query-candidate pair; k is the number of response categories; and n_{ij} is the number of graders who assigned the i-th query-candidate pair to the j-th category.

We measured the inter-grader reliability of the E6K judgments from the SMS and AMS tasks. In addition to computing the reliability score using all three BROAD judgment scores (VS, SS, NS), we also combined the (VS, SS) categories to create a general "Similar" (S) category and measured inter-grader reliability using the resulting 2-level judgment scores (S, NS). The resulting Kappa scores for the 3-level and 2-level judgments for both SMS and AMS are given in Table 3.

	3-level (VS, SS, NS)	2-level (S, NS)
SMS	0.3664	0.3201
AMS	0.2141	0.2989

Table 3. Fleiss's Kappa for AMS and SMS contests at 3-levels and 2-levels of judgment.

Fleiss's Kappa scores can range from 0.0 (no agreement) to 1.0 (perfect agreement). Landis and Koch [4] studied the consistency of physician diagnoses of patient illnesses and derived a scale for interpreting the strength of agreement indicated by Fleiss's Kappa score. According to this scale, the resulting Kappa scores reported in Table 3 indicate a "Fair" level of agreement (within the range of 0.21 – 0.40) for all tasks at both 3- and 2-levels of judgment. While greater levels of agreement are possible, and indeed desirable, the assessment of a "fair" level of agreement is encouraging given that the graders in the E6K are drawn from a heterogeneous pool of candidates with a wide variety of skills and backgrounds compared to the relatively highly trained physicians studied by Landis and Koch.

In order to better understand the nature of the disagreement indicated by the Kappa scores, we conducted further analysis of the data. To start, we looked at the distribution of scores given by each grader, i.e., how many VS, SS, and NS scores did each grader assign globally. A Chi-squared analysis (Table 4) shows that there is a significant difference between graders with respect to each grader's frequency distribution of VS, SS, and NS (or S, NS) scores. This might be a result of the fact that each grader evaluated a different subset of queries from the total set.

Looking at each query-candidate pair, the set of scores assigned by the three graders were counted (Tables 5 and 6). Within the set, triple assignment of the same judgment (same score given by each grader) indicates total agreement. Cases of partial agreement are

counted as "doubles" (i.e., where two graders agreed and a third dissented). No agreement is indicated by three different scores.

3-level (VS, SS, NS)			
	χ^2	df	P
AMS	765.98	46	0*
SMS	414.01	38	0*
2-level (S, NS)			
	χ^2	df	P
AMS	413.22	23	0*
SMS	323.89	19	0*

* denotes significance

Table 4. Chi-Squared statistics for the variation of agreement across queries in AMS and SMS contests at 3-levels and 2-levels of judgment.

Comparing grader judgments at 3-levels of similarity (Table 5), only 2.2% of query-candidate pairs in SMS (7.1% in AMS) resulted in no agreement. There was partial disagreement in a majority of the cases (51.9% for SMS, 62.8% for AMS).

Judgments	3-level SMS		3-level AMS	
VS,VS,VS	114	12.6%	61	3.7%
SS,SS,SS	38	4.2%	137	8.4%
NS,NS,NS	263	29.1%	293	18.0%
Total triples	**415**	**45.9%**	**491**	**30.1%**
VS,VS	24	2.7%	150	9.2%
SS,SS	158	17.5%	469	28.8%
NS,NS	288	31.8%	404	24.8%
Total doubles	**470**	**51.9%**	**1023**	**62.8%**
VS,SS,NS	20	2.2%	115	7.1%
Total	**905**	**100.0%**	**1629**	**100.0%**

Table 5. Distribution of 3-level judgment triples (total agreement), doubles (partial agreement), and cases of no agreement.

When the SS and VS categories are combined into a broader "Similar" category (S) the amount of total agreement increased, as expected (Table 6).

Judgments	2-level SMS		2-level AMS	
S,S,S	188	20.8%	494	30.3%
NS,NS,NS	263	29.1%	293	18.0%
Total triples	**451**	**49.8%**	**787**	**48.3%**
S,S	166	18.3%	438	26.9%
NS,NS	288	31.8%	404	24.8%
Total doubles	**454**	**50.2%**	**842**	**51.7%**
Total	**905**	**100.0%**	**1629**	**100.0%**

Table 6. Distribution of 2-level judgment triples (total agreement) and doubles (partial agreement).

Within each query, the total number of cases of partial agreement and no agreement were counted and compared. A Chi-Squared test was used to measure the differences in the variance of this disagreement between queries. At 3-levels of judgment, there were no significant differences in the variance of disagreement across queries, meaning that graders tended to disagree fairly consistently across queries. When VS and SS were combined, the variance of disagreement across queries did reach significance for both AMS and SMS, indicating that there were some queries which had significantly more or less amounts of disagreement than others (Table 7).

3-level (VS, SS, NS)			
	χ^2	df	P
AMS	44.25	59	0.92
SMS-Mixed	10.92	5	0.053
SMS-RISM	7.94	5	0.159
SMS-Karaoke	1.03	4	0.905
2-level (S, NS)			
	χ^2	df	P
AMS	124.98	59	0*
SMS-Mixed	11.3	5	0.046*
SMS-RISM	10.09	5	0.073
SMS-Karaoke	1.17	4	0.882

* denotes significance

Table 7. Chi-Squared statistics for the variation of agreement across queries in AMS and SMS contests at 3-levels and 2-levels of judgment.

4. DISCUSSION

Graders' judgments tended to be fairly stable. The raw event log counts presented in Table 2 represent the total number of times BROAD score categories were changed. In SMS 39 (20.3%) of the 192 changes and in AMS 125 (22.2%) of the 563 changes to BROAD score resulted in the grader reverting to their original judgment, meaning that the final score recorded was the same as the initial score given. This may represent an artifact of the logging mechanism whereby the grader may have clicked on the interface, generating an event, which was not an intentional judgment action, or was essentially a duplicate event for the same judgment. However, even considering this possibility the majority of graders only made a single judgment for each query-candidate pair.

Additionally, the amount of disagreement observed among graders differed between the SMS and AMS tasks when considering 3-level judgments (Table 5). The raw proportions of cases resulting in no agreement were different, with AMS having slightly more disagreement among graders than SMS. This difference between the tasks may be attributable to the nature of the task and data being analyzed or it may be a consequence of the task definition. In SMS, graders were asked to evaluate the melodic similarity of two works and were explicitly instructed to look beyond differences in timbre and instrumentation [5]. In AMS, however, the task was less well-defined, asking graders to evaluate the musical similarity of two works, only specifying that the works should "sound" similar. [5]. This highly subjective definition of musical similarity may explain the observed differences between the tasks.

The likelihood of graders to assign a particular score did differ significantly across graders. This tells us that multiple graders are needed to temper the peculiarities of any individual. For example, one grader in the AMS evaluation did not assign a single query-candidate pair a Very Similar (VS) rating; another grader gave over 78% of judgments a Somewhat Similar (SS) rating. These very different judgment profiles imply that relying on a

single grader may skew the results significantly. However, the relatively small number of cases with no agreement may indicate that only two graders are needed per query-candidate pair, instead of three.

Interpreting the amount of consistency between graders is subjective. When considering three levels of responses, there is greater consistency between queries than when considering only two. Greater consistency is desired as it reflects a more objective measure of similarity. However, as these judgments are ultimately being used to evaluate the performance of algorithms and improve their design, the greater variation in agreement afforded by the two-level judgments may provide greater discrimination between queries. Identifying particular queries which caused unusual amounts of agreement or disagreement may help background the results, allowing MIREX organizers to weight those judgments accordingly, and help developers identify limitations of their algorithms.

5. CONCLUSIONS AND RECOMMENDATIONS

In this paper we have tried to unpack the issues of human judgment in evaluating music similarity systems. These analyses can be used to improve the design of future MIREX tasks, and inform the design of other projects which seek to use human judgments for evaluation. Towards this goal, we present several recommendations based on our understanding of the data.

The differences between the AMS and SMS tasks did result in differences in the judgments of graders. Building on the discussion started in [1], we advance that many of the differences observed in the reliability and consistency of the graders can be attributed to problems in the definition of the evaluation tasks. The lesser-specified instructions given to graders in the AMS task are likely contributing to the greater observer variation in the scores. To the extent that Inter-grader reliability tests measure the consistency of grader judgments, they also reflect the degree graders understood the instructions given to them. Tasks like AMS may need to be more narrowly defined to include more objective features, such as melody, instrumentation, style, genre, etc.

The number of response categories presented to graders does have an effect on their ability to reach consensus; fewer categories yields greater consensus. This might lead to the conclusion that only two categories should be used for evaluation (Similar or Not Similar). Indeed, this is a model which was discussed in the design of the AMS task, and reflects the user behavior of managing a playlist (i.e., songs are either included in the list or not). However, the additional costs of collecting the slightly more nuanced judgments is negligible, and based on anecdotal feedback, graders appear to prefer having three options.

As mentioned above, three graders may not be needed to achieve consensus. The amount of partial and total agreement observed indicates that two graders may suffice where three were previously used. Furthermore, the amount of disagreement is consistent across queries. Thus, a more advanced E6K system may adaptively allocate graders to queries, only assigning a third grader to queries which deviate from the expected level of agreement, or to query-candidate pairs which have not achieved consensus. This would allow for a greater number of query-candidate pairs to be evaluated by the same number of graders.

The major limiting factor in the design of evaluation schemes, like that employed by AMS and SMS, is the availability of graders. Ultimately, evaluating more queries and more candidates per query would more greatly benefit algorithm developers. We argue that the number of queries be increased in future iterations of MIREX.

The analyses presented in this paper have only scratched the surface of understanding the judgment data collected with the E6K. In the future we intend to do similar analyses of the FINE scores collected in the E6K, as well as explore the temporal dimension of the grader judgments in order to understand how presumed independent judgments may be interacting or interfering. Furthermore, future iterations of MIREX and use of the E6K will generate more data to be analyzed and compared towards the production of a model of human evaluation and interaction in the E6K.

6. ACKNOWLEDGEMENTS

Special thanks to: The Andrew W. Mellon Foundation, the National Science Foundation (#IIS-0327371), Dr. Ellen Voorhees, and the MIREX 2006 graders.

7. REFERENCES

[1] Downie, J. S., Lee, J. H., Gruzd, A. A., Jones, M. C. (2007). Toward an Understanding of Similarity Judgments for Music Digital Library Evaluation. In the Proceedings of the ACM/IEEE Joint Conference on Digital Libraries.

[2] Downie, J. S., West, K., Ehmann, A., and Vincent, E. The 2005 Music Information Retrieval Evaluation eXchange (MIREX 2005): Preliminary overview. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005),* Queen Mary, UK, 2005, 320-323.

[3] Fleiss, Joseph L. Measuring nominal scale agreement among many raters. Psychological Bulletin. 76(5):378-382, November 1971.

[4] Landis, J. R. and Koch, G. G. (1977) "The measurement of observer agreement for categorical data" in Biometrics. Vol. 33, pp. 159-174.

[5] MIREX Wiki. Available at: http://music-ir.org/mirexwiki/.

[6] Pampalk, E. (2006). Audio-Based Music Similarity and Retrieval: Combining a Spectral Similarity Model with Information Extracted from Fluctuation Patterns. Technical Report. http://staff.aist.go.jp/elias.pampalk/papers/pam_mirex06.pdf