Toward an Understanding of Similarity Judgments for Music Digital Library Evaluation

J. Stephen Downie, Jin Ha Lee, Anatoliy A. Gruzd, M. Cameron Jones
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
+1 (217) 333-3280

{jdownie, jinlee1, agruzd2, mjones2}@uiuc.edu

ABSTRACT
This paper presents an analysis of 7,602 similarity judgments collected for the Symbolic Melodic Similarity (SMS) and Audio Music Similarity and Retrieval (AMS) evaluation tasks in the 2006 Music Information Retrieval Evaluation eXchange (MIREX). We discuss the influence of task definitions, as well as evaluation metrics on user perceptions of music similarity, and provide recommendations for future Music Digital Library/Music Information Retrieval research pertaining to music similarity.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: User issues

General Terms
Measurement, Performance, Human Factors.

Keywords
MIREX, Music Digital Libraries, Music Information Retrieval, Music Similarity. Evaluation, Evalutron 6000

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 Digital Library (MDL) and Music Information Retrieval (MIR) systems [1].

MIREX 2006 comprised nine separate evaluation tasks which were defined by community input [3]. 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 these human similarity judgment data. We present a descriptive comparison of the observed behaviors of human evaluators with an emphasis on the similarities and differences between the AMS and SMS tasks. We conclude with considerations and recommendations for interpreting human similarity judgments.

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.

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 top-ranked "candidate" songs. All resulting query-candidate pairs were consolidated and then evaluated by "graders" using the E6K. In the E6K, graders score the anonymized query-candidate pairs 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.

Table 1. Evalutron 6000 Descriptive Statistics

	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

Graders were given the community-defined instructions found in Figure 1 prior to evaluating. 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.

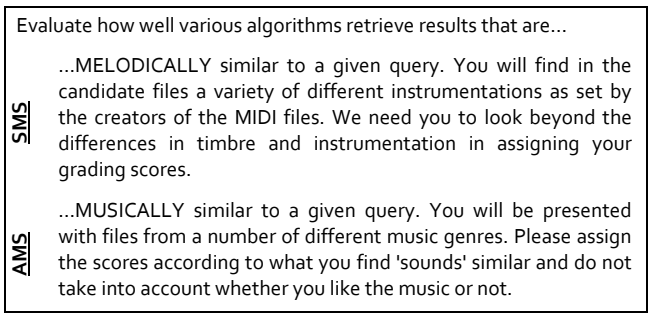

Figure 1. Excerpt of grader instructions for SMS and AMS.

3. DATA & DISCUSSION

To measure consistency between the BROAD category and FINE scores, we calculated the distribution of FINE scores within each BROAD category. Figure 2 shows box-and-whisker plots for both SMS and AMS tasks. The boxes have vertical lines at the 1st, 2nd, and 3rd quartiles. The whiskers bound the minimum and maximum values which fall within 1.5 times the inter-quartile range (IQR), outliers are denoted by + symbols.

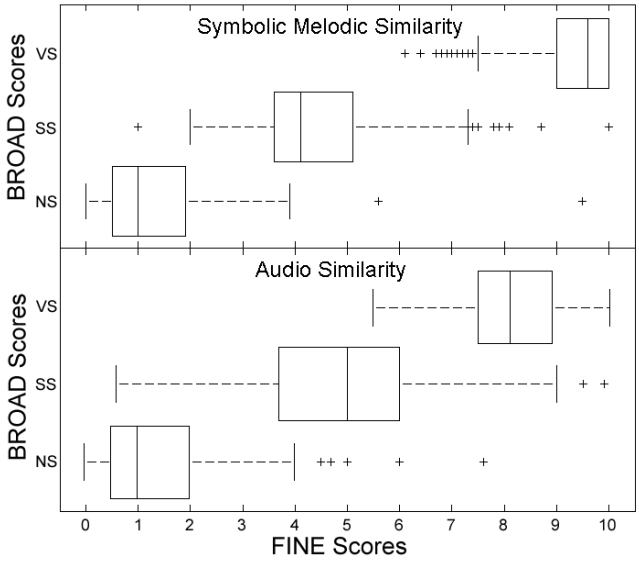

Figure 2. FINE Score Distributions for SMS and AMS.

In both AMS and SMS, the NS and VS categories have the most compact distributions of FINE scores. Similarly, the SS category has the largest IQR in both SMS and AMS. SS scores overlap with the NS values from both tasks. In AMS, however, note how SS greatly overlaps ***both*** the NS and VS values. In Figure 3 we see a related pattern: the SS category has the greatest average noteworthy is the consistently lower number of auditions for SMS across all BROAD categories.

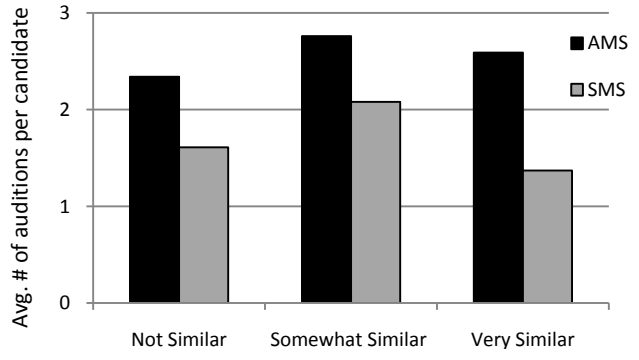

Figure 3. Auditioning of candidates by BROAD category.

The data presented in Figures 2 and 3 lead us to two important observations. First, there is a fundamental difference in the interpretations of "similarity" between the SMS and AMS tasks. Second, the SS category, regardless of task, appears problematic.

3.1 Ambiguous Notions of Music Similarity

Recall from Figure 1, AMS graders were instructed to evaluate "musically similar" items, while SMS graders were instructed to evaluate "melodically similar" items. These differing definitions of similarity are reflected in the data. In AMS, only 9 out of 24 graders used the maximum FINE score (10.0), whereas in SMS, 16 out of 21 used it, suggesting that the graders have a more concrete understanding of "melodic" similarity than "musical" similarity. This observation is reinforced by the fact that graders tended to audition SMS candidates *less* often than AMS candidates. They also modified their BROAD category assessment of similarity in SMS (5.9%) *less* often than AMS (9.0%).

3.2 Ordinal Relevance/Similarity Judgments

Previous research on relevance indicates that binary relevance measures are not adequate for representing relevance judgments [2]. The addition of the SS category, at the request of the community, appears to have caused confusion among the graders as to what "somewhat similar" really means. That FINE scores assigned in the SS category spanned both NS and VS scores for both tasks, and that graders listened to candidates assigned SS scores more often, imply that assigning scores to "somewhat similar" items requires greater cognitive overhead. The vague and problematic SS category may be unreliable as a measure of system performance. The clear division between the NS and VS categories, however, suggests that a binary measure may be preferred for obtaining more consistent and reliable data for MDL evaluation purposes.

4. CONCLUSION & FUTURE WORK

The MDL/MIR community should establish more focused definitions of similarity which are not only useful to, but also readily understandable by users (and their surrogates, the graders). Expecting graders to consistently comprehend what is meant by a query-candidate pair being "musically" similar opens the door to misinterpretation which, in turn, generates inconclusive evaluation data. Furthermore, the evaluation of MDL/MIR systems, regardless of the definition of similarity being used, requires clearer metrics for measurement: We would now argue for a simple, binary, Similar/Not Similar system. We plan on working with the MDL/MIR community to address these important issues in preparation for MIREX 2007. The E6K system will be modified to reflect the outcome of the community decisions on these topics. We will continue to analyze the MIREX 2006 data as well as collect more data from future MIREX runs.

5. ACKNOWLEDGMENTS

Special thanks to: The Andrew W. Mellon Foundation, the National Science Foundation (Grant No. NSF IIS-0327371), Dr. Ellen Voorhees, and all the MIREX 2006 graders.

6. REFERENCES

[1] 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.

[2] Mea, V. D., and Mizzaro, S. Measuring retrieval effectiveness: A new proposal and a first experiment validation. *Journal of the American Society for Information Science and Technology, 55*, 6 (April 2004), 530-543.

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