In general, we can see that the statistical significance of the differences can be estimated fairly well despite the incompleteness of judgments and the uncertainty on the true $\Delta AG@k$ scores.

6. CONCLUSIONS

We have shown how to adapt the Minimal Test Collections (MTC) family of algorithms for the evaluation of the Audio Music Similarity and Retrieval task. Assuming a uniform distribution of similarity judgments, we showed that the distribution of $AG@k$ scores is normally distributed, which allows us to look at it as a random variable whose expectation may be estimated with a certain level of confidence. This confidence is proportional to the number of similarity judgments available, and MTC ensures that the set of judgments we make to reach some confidence level is minimal.

Using the data from the MIREX 2011 AMS evaluation, we simulated MTC and found that with just one third of the judgments the correct ranking of systems can be estimated with 95% confidence, and with no swap between significantly different systems. We also showed some simple rules to estimate the significance of the differences, which work reasonably well for 95% confidence but tend to overestimate significance when higher confidence is achieved in the ranking of systems.

Three clear lines for future work can be identified. First, this paper assumed that the distribution of similarity judgments was uniform, that is, that all assignments of similarity were equally likely. However, it is clear this assumption does not hold in reality: the distribution of similarity judgments is skewed towards the highly similar or the not similar side, depending on how well the system performs. Having better estimates of the true distribution would make the algorithm perform better in terms of effort and accuracy of the estimates. These estimations of the true distribution could be approximated from previous MIREX AMS data, or with a model fitted with the judgments we make for the very systems we are evaluating, learning the true distribution on a per system or per query basis. Second, our estimates of the significance of the differences were based on very simple rules that assumed the (very unrealistic) best cases. Developing a comprehensive mathematical framework for testing significance is another clear line for future work.

The most important direction for further research is the study of low-cost evaluation methodologies for other MIR tasks. In accordance with previous work [7], we have shown that the effort in evaluating a set of AMS systems can be greatly reduced, leaving open the possibility of building brand new test collections for other tasks for which creating annotations is very expensive. For instance, the group of volunteers requested by MIREX for the yearly evaluation of the AMS and SMS tasks could be better employed if some of them were instead dedicated to incrementally add new annotations for the other tasks.

Another clear setting for the application of low-cost methodologies is that of a researcher evaluating a set of systems with a private document collection, a scenario very common in MIR given the legal restrictions on sharing music corpora. Those researchers, and in most cases public forums too, do not have the possibility of requesting large pools of external volunteers for annotating their collections. Thus, being able to evaluate systems with the minimal effort is paramount. To this end, low-cost evaluation methodologies must be investigated for the wealth of MIR tasks.

In most of these tasks researchers rely on test collections annotated *a priori*, which can be very expensive and time consuming to build. However, we have seen that not all annotations are necessary to evaluate systems. For instance, if two Audio Melody Extraction algorithms predict the same F0 (fundamental frequency) in a given frame, whether that F0 prediction is correct or not is not useful to know which of the two systems is better. The adoption of *a posteriori* evaluation methodologies such as MTC can take advantage of this to greatly reduce the annotation cost or allow the use of larger collections. Getting to that point, though, requires a shift in the current evaluation practices. But given the benefits of doing so, both in terms of cost and reliability, we strongly encourage the MIR community to study these evaluation alternatives and progressively adopt them for a more rapid and stable development of the field.

ACKNOWLEDGEMENTS

This research is supported by the Spanish National Plan of Scientific Research, Development and Technological Innovation through grants TSI-020110-2009-439 and HAR2011-27540 and the Austrian Science Funds (FWF): P22856-N23.

REFERENCES

[1] B. Carterette. *Low-Cost and Robust Evaluation of Information Retrieval Systems*. Ph.D. dissertation, Department of Computer Science, University of Massachusetts Amherst, 2008.

[2] B. Carterette. Robust Test Collections for Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 55-62, 2007.

[3] B. Carterette, J. Allan, and R. Sitaraman. Minimal Test Collections for Retrieval Evaluation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268-275, 2006.

[4] J.S. Downie. The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future. *Computer Music Journal*. 28(2): 12-23, 2004.

[5] J.S. Downie, A.F. Ehmann, M. Bay, and M.C. Jones. . The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. In *Advances in Music Information Retrieval*, W.R. Zbigniew and A.A. Wieczorkowska, eds. Springer. 2010, 93-115.

[6] J. Urbano. Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain. In *International Society for Music Information Retrieval Conference*, pages 609-614, 2011.

[7] J. Urbano, D. Martín, M. Marrero, and J. Morato. Audio Music Similarity and Retrieval: Evaluation Power and Stability. In *International Society for Music Information Retrieval Conference*, pages 597-602, 2011.

[8] E.M. Voorhees. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management*. 36(5): 697-716, 2000.

[9] E.M. Voorhees and D.K. Harman. . *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, 2005.