AUDIO MUSIC SIMILARITY AND RETRIEVAL: EVALUATION POWER AND STABILITY

Julián Urbano, Diego Martín, Mónica Marrero and Jorge Morato
University Carlos III of Madrid
Department of Computer Science
{jurbano, dmandres, mmarrero, jmorato}@inf.uc3m.es

ABSTRACT

In this paper we analyze the reliability of the results in the evaluation of Audio Music Similarity and Retrieval systems. We focus on the power and stability of the evaluation, that is, how often a significant difference is found between systems and how often these significant differences are incorrect. We study the effect of using different effectiveness measures with different sets of relevance judgments, for varying number of queries and alternative statistical procedures. Different measures are shown to behave similarly overall, though some are much more sensitive and stable than others. The use of different statistical procedures does improve the reliability of the results, and it allows using as little as half the number of queries currently used in MIREX evaluations while still offering very similar reliability levels. We also conclude that experimenters can be very confident that if a significant difference is found between two systems, the difference is indeed real.

1. INTRODUCTION

One of the most important tasks in Music Information Retrieval is Audio Music Similarity and Retrieval (AMS). Along with Symbolic Melodic Similarity (SMS), AMS is one of the traditional tasks evaluated in the annual Music Information Retrieval Evaluation eXchange (MIREX) [3], and one of the tasks that most closely resemble a real-world music retrieval scenario. A music similarity retrieval system returns a ranked list of music pieces deemed to be similar to a music piece given as a query. In the case of the MIREX evaluation of AMS, these music pieces are 30 second audio clips of music material.

As of the writing of this paper, a total of 41 AMS systems have been evaluated in 4 editions of MIREX from 2006 to 2010, and it is again planned for 2011. In these evaluations, a set of queries is randomly selected and provided to the participating systems, which then return the corresponding 5 most similar music pieces in a music collection. To evaluate the effectiveness of the systems two things are needed: relevance judgments and effectiveness measures. The relevance judgments are scores given to each query-candidate pair, representing their similarity. Two relevance scales are used in MIREX for both the AMS and SMS tasks. The Broad scale has three levels: not similar (NS = 0), somewhat similar (SS = 1) and very similar (VS = 2). The Fine scale uses real valued scores between 0.0 (not similar at all) and 10.0 (identical). As to the effectiveness measures, in AMS the so-called Sum measure is used, while more complex measures were developed for the SMS task [11].

The grand results of these evaluations are pairwise comparisons between the participating systems, indicating which is better and whether the difference is statistically significant or not. When drawing such conclusions, two characteristics of the evaluation must be kept in mind: power and stability. Power refers to how powerful the evaluation is to establish a significant difference between any two systems (i.e. it is concerned with Type II errors). If A is concluded to perform significantly better than B, the evaluation is considered powerful. If the difference were not statistically significant, no clear conclusion could be drawn from the experiment: A and B could actually perform identically (very unlikely), or the evaluation conditions might have not been sufficient to observe a difference large enough (most likely). Assuming two systems A and B are never exactly the same, an option to achieve significance is to increase the number of queries, though this has obvious limitations in terms of effort and cost [16][8]. The difference between practical and statistically significant differences must be considered if doing so.

Stability refers to how reliable a result is when claiming a statistically significant difference between two systems (i.e. it is concerned with Type I errors). If A and B were evaluated with a set of queries and the result were that A is significantly better than B, the expected result with a completely different (and independent) query set would therefore be that A is again significantly better than B. If it were not, it would be an indication that the evaluation is not stable when differentiating between systems. These conflicts do appear in IR evaluation experiments, and if the query set used is too small, the effectiveness measures not appropriate or the statistical procedures not suitable, they can be frequent [1]; even when statistical significance is involved [15].

In this paper we analyze the power and stability of the AMS evaluation methodology when concluding that a system A is significantly better than a system B. We analyze

the effect that different relevance judgment sets, effectiveness measures, query set sizes and statistical procedures have on the reliability of the AMS results. For this study we decided to use the MIREX 2009 Audio Music Similarity and Retrieval data, as it is the largest dataset available to date [4]. A total of 15 systems by 9 different research groups were evaluated with a total of 100 queries. The top 5 documents retrieved by each system were evaluated for each query using the Broad and Fine scales, and the Sum measure was used with these two sets of relevance judgments to assess the effectiveness of the systems. The Friedman test was ran with a Tukey's HSD post-hoc correction procedure to look for significant differences. The grand results of the evaluation are thus 105 pairwise comparisons between systems, some of which are statistically significant.

The rest of the paper is organized as follows. Section 2 reviews previous work on the analysis of power and stability in TREC and related studies on the evaluation of music similarity tasks. Next we discuss the effectiveness measures considered, and Sections 4 and 5 present the results of the power and stability analysis. Section 6 argues and analyzes the use of different statistical procedures. Finally, Section 7 presents a discussion of the results and the paper then finishes with the conclusions and lines for further work.

2. RELATED WORK

The stability of effectiveness measures has been extensively studied in the context of the Text REtrieval Conference (TREC). Buckley and Voorhees first studied the stability of several measures, observing conflicts between 1% and 14% of the times, depending on the measure, when comparing any two systems [1]. They then studied the sensitivity of several measures as a function of the query set size, and they concluded that absolute differences larger than 0.05 (about 25% relative difference) are necessary for sets of 50 queries to assure a conflict ratio below 5% [16], confirming the reliability of TREC evaluations for using 50 queries as a minimum. However, none of these studies considered the effect of using statistical significance techniques when comparing two systems. Sanderson and Zobel somehow filled this gap by studying the effect of several statistical procedures on the sensitivity, and they concluded that virtually any relative difference of 10% or more, coupled with statistical significance, will not produce a conflict in other experiments [8]. Sakai reviewed most of this work with different data sets and with other, more recent measures [7]. With larger query sets, Voorhees found that even significant differences could still be conflictive [15]. However, the study of post-hoc statistical procedures was not part of any of these studies.

Meta-evaluation studies are very rare in Music IR [12], and to our knowledge the power and stability issues have not yet been studied for MIREX data. Nonetheless, some works have addressed similar problems with Music IR evaluation experiments concerning the similarity tasks. Typke et al. studied alternative forms of relevance judging for the SMS task [10], and they came up with a specific effectiveness measure to be used with them [11]. Urbano et al. then showed how to make the evaluation more reliable when using those relevance judgments [13]. Jones et al. studied the relevance judgments made for the SMS and AMS tasks, focusing on the effect of having different people do the judgments and with different scales. To reduce the cost of judging, the use of crowdsourcing platforms such as Amazon Mechanical Turk has been studied by Urbano et al. for the SMS task [14], and by Lee for the AMS task [6]. In this paper we focus on the power and stability of the MIREX AMS evaluations, employing techniques similar to Buckley's and Voorhees', but with some modifications specific to the AMS task and the post-hoc analysis used in MIREX.

3. EFFECTIVENESS MEASURES

The MIREX AMS evaluation campaigns use just one measure to assess the effectiveness of the participating systems. This is the so-called Sum measure, which is the average relevance of the retrieved results. When used with the Broad judgments, this measure is often called PSum; and when used with the Fine judgments, it is called FINE [3].

The Audio community has traditionally been reluctant to adopt more complex measures, even some specifically designed for this type of tasks [3]. In this paper we study the use of several of these measures in the Audio Music Similarity task, and their impact on the power and stability of the evaluation. First, we review the measures considered.

3.1 Average Gain

This measure is based on the concept of information gain provided by the retrieved documents. This information gain is usually represented by the relevance level assigned to the document, assuming that the larger the score, the more information is gained by the user.

G@k is the Gain of the *k*-th document retrieved, and $CG@k = \sum_{i=1}^{k} G@i$ is the Cumulated Gain of the first *k* documents retrieved [5]. Thus, the Average Gain of the top-*k* documents is calculated as the mean:

$$AG@k = \frac{CG@k}{k} = \frac{1}{k}\sum_{i=1}^{k} G@i \qquad (1)$$

This is the official measure used in the MIREX AMS task. We prefer to use this definition based on information gain for consistency with the other measures.

The problem of G, CG and AG is that they do not have a fixed upper bound, which causes some problems when averaging the results across queries. Consider a query `q1` for which there are 7 VS documents and another query `q2` with 2 VS and 5 SS documents. For `q1` a perfect system can achieve a total CG@5 score of 10, while for `q2` the maximum possible is 7. Apparently, the system performs better for `q1`, when in reality it returns ideal results for both queries. As with other simpler measures such as Precision, this lack of fixed upper bound makes them less stable [1][7].

3.2 Normalized Discounted Cumulated Gain

AG does not consider the rank at which documents appear down the results list: a document at rank 3 provides as much gain as if it were at ranks 1 or 5. However, a highly relevant document is clearly more useful to the user if it appeared toward the top of the list. To model this usefulness, the gain scores are discounted as they appear later in the results list. A logarithm function with base *b* is used, and so the Discounted Cumulated Gain is defined recursively as:

$$DCG@k = \begin{cases} CG@k & k < b \\ DCG@(k-1) + \dfrac{G@k}{\log_b k} & k \geq b \end{cases} \quad (2)$$

Also, to avoid the lack of fixed upper bound problem, it is considered what the ideal ranking of documents would be: $IDCG@k = DCG@k \ \ s.t. \forall i < k : G@i \geq G@(i+1)$. Dividing the DCG@k score of the system by the ideal IDCG@k, the upper bound is always 1, meaning perfect retrieval:

$$NDCG@k = \frac{DCG@k}{IDCG@k} \quad (3)$$

This measure is called Normalized Discounted Cumulated Gain [5], which has been shown to be particularly stable and sensitive [7][17]. For this study we set the logarithm base to the standard *b=2*.

3.3 Average Normalized Discounted Cumulated Gain

The last measure of the information gain family we consider here is the Average Normalized Discounted Cumulated Gain, which is calculated as the average NDCG score throughout the retrieved list:

$$ANDCG@k = \frac{1}{k}\sum_{i=1}^{k} NDCG@i \quad (4)$$

ANDCG provides more information about the ranking of the retrieved documents, as still quite large NDCG scores could be achieved just by highly relevant documents towards the end of the list. Like NDCG, this measure has been shown to be particularly stable and sensitive [7].

3.4 Average Dynamic Recall

The last measure we consider originated in the context of the MIREX 2005 SMS task and the evaluation with relevance judgments in the form of partially ordered lists [10][13]: Average Dynamic Recall [11]. ADR was specifically designed for level-based relevance judgments without a scale fixed beforehand, and ever since it is one of the main measures used in MIREX SMS with the Broad judgments.

We also define ADR in terms of information gain. Let $I = \langle I_1, ..., I_n \rangle$ be the list of *n* judged documents ordered by descending relevance level (i.e. an ideal ordering), and let $R = \langle R_1, ..., R_k \rangle$ be the list of the top $k \leq n$ retrieved documents ordered by rank. The set A_i of allowed relevant documents at rank *i* is defined as:

$$A_i = \{I_1, ..., I_i\} \cup \{I_j : j > i \wedge G@j = G@i\} \quad (5)$$

that is, the union of all previous ideal documents and those with lower rank but equal information gain (i.e. same relevance level). The final score is then calculated as:

$$ADR@k = \frac{1}{k}\sum_{i=1}^{k} \frac{\left|A_i \cap \{R_1, ..., R_i\}\right|}{i} \quad (6)$$

which is the average across ranks of the ratio of documents retrieved that are actually in the ideal ranking. This measure is widely used by the SMS community, but it has never been used in the AMS task, nor has it been analyzed in terms of power or stability. In this paper we do so.

4. EVALUATION POWER

To assess the effect of different effectiveness measures and relevance scales on the power of the evaluation, we compute the number of pairwise system comparisons that result significant according to the Friedman-Tukey's HSD (FT) procedure used in MIREX. We evaluate the original measure, AG, as well as NDCG, ANDCG and ADR; both with the Broad and Fine set of relevance judgments, for a total of 8 distinct measures.

We study the trend for increasing query sets of sizes 5 to 100, with increments of 5 queries each. To diminish random effects when selecting a subset of queries for the 5 to 95 sizes, we choose 500 random samples in each case. Thus, there are 52,500 system pairwise comparisons for each measure and query subset size. Also, the queries in MIREX were balanced across music genres: the 100 original queries were selected from 10 different genres, with 10 queries per genre. We also reproduce this balance, using stratified sampling with equal priors when making query subsets. Therefore, our samples are also balanced across music genres, emulating as closely as possible a real MIREX evaluation.

As Figure 1 shows, 57% of the results were significant using AG_{Broad} and 54% using AG_{Fine} (horizontal dotted lines). We omitted query subset sizes below 40 for clarity: the curves follow a somewhat logarithmic trend (see the thumbnails for the whole plot). Indeed, it can be seen that the increment in significant pairwise comparisons is very soft and quite similar for all measures but ADR_{Fine}.

The right figure also shows that for larger query sets $(A)NDCG_{Fine}$ clearly outperform AG_{Fine}, which seems to converge. ADR_{Fine} performs quite poorly, following a somewhat linear trend. This is expected though, as the contribution of each document retrieved is here binary: if a document is allowed at rank *i* it contributes $+1/i \cdot k$ to the score, 0 otherwise. In the $(A)NDCG_{Fine}$ measures the contribution is discounted, but it is never binary. This makes ADR_{Fine} perform significantly worse. Nonetheless, it is important to note that ADR was not intended for real valued relevance judgments, which make it very difficult for two documents to have the same relevance score (right term in Equation 5) and thus it requires systems to obtain a nearly ideal ranking.

Most importantly, it can be seen that the query set size could be significantly reduced to lower the cost of the eval-

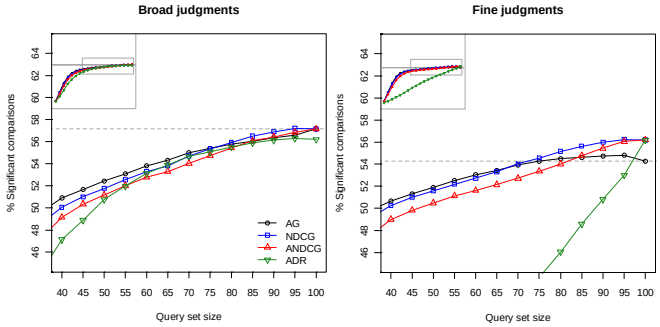

Figure 1. Evaluation power (larger is better) with FT, for all measures with the Broad (left) and Fine judgments (right).

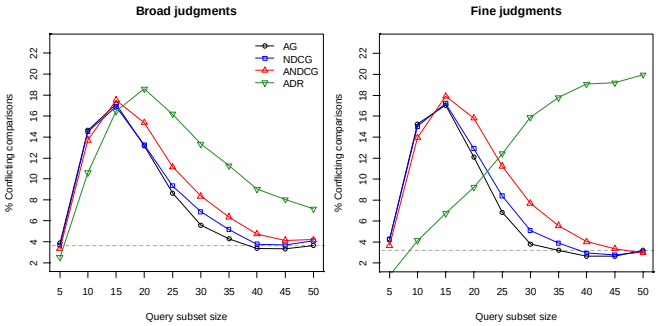

Figure 2. Evaluation stability (lower is better) with FT, for all measures with the Broad (left) and Fine judgments (right).

uation in terms of relevance judging effort. For example, having reduced the query set to 70 queries (70%) only 2 significant differences would have been missed if using AG_{Broad}, none if using AG_{Fine}.

5. EVALUATION STABILITY

To assess the effect that different effectiveness measures and relevance scales have on the stability of the evaluation, we need two different query sets, as if we were evaluating the systems with two completely different collections. Unfortunately, having the same 15 systems with another 100 completely different queries is not yet feasible. Nonetheless, we can use smaller query sets and then observe the trends to extrapolate to larger sets. We start with the 5,000 random query subsets of sizes 5 to 50 used before. Then, for each of these we sample another query subset of the same size, again stratified, but also without replacement. That is, for each of the 500 trials of each of the 10 query subsets, there are two query samples with no common query. Because they are disjoint, we can treat them as if coming from two different evaluation experiments. Note also that having a total of 100 queries limits the query subsets to 50 queries at most, as the paired subset samples would contain the remaining 50 queries in each case.

We re-evaluate the 15 systems for with each pair of query samples, and then compare the 105 system pairwise results from both samples. We count the number of times there is a significant difference with one sample but not with the other one, again according to the original Friedman-Tukey's HSD procedure. These would represent stability conflicts across two real evaluations.

As Figure 2 shows, about 4% of the system pairwise comparisons are conflicting with the Broad judgments using 40 queries or more, and as few as 3% with the Fine judgments (dotted horizontal lines). This is consistent with the 5% significance level set for the statistical procedure (see Section 6). Indeed, the curves tend to converge toward the end. It is noticeable again that ADR performs significantly worse, especially for the Fine judgments, where the increasing conflict rates can be explained by the very low sensitivity of the measure, as explained before. The other measures behave remarkably similarly.

The peaks for small query subsets are explained by the power of the statistical procedures used: with that few queries the tests are not powerful enough to result significant, and when they happen to do for one query sample they still do not for the other one. This increment in conflicts starts decreasing and converges because the tests get more powerful with larger samples, so they are able to give significance with both query subsets. In fact, for AG with 50 queries all conflicts are caused by this lack of significance in one of the samples, 99.9% for NDCG and 99.7% for ANDCG; even 99.7% for ADR_{Fine}. Most importantly, there was no case whatsoever where the two system pairwise comparisons were significant but with opposite sign. As such, one can be quite confident about the difference between two systems when it comes up significant.

6. STATISTICAL ANALYSIS

The usual method to check whether two systems are significantly different or not is to run a statistical test such as the Wilcoxon test or the t-test. Each of these has an associated significance level, which is the maximum allowed probability of committing a Type I error. In our case, these errors occur when the test says there is a significant difference but there actually is none. This significance level uses to be set to α=0.05 or α=0.01. That is, a probability of 5% or 1% of incorrectly getting significant differences between systems.

In the case of MIREX 2009 AMS, 105 of these pairwise tests would need to be run. Unfortunately, if setting α=0.05 the probability of committing a Type I error in any of these would be $1-(1-\alpha)^{105}$=0.995. This is the experiment-wide significance level. Thus, almost certainly we would at least once be saying that two systems are significantly different when they actually are not. In MIREX, the Friedman test is run instead, with the Tukey's HSD post-hoc procedure for significance correction [3]. This compares all system pairs at once, with the difference that the experiment-wide significance level remains close to α=0.05. The test is thus much less likely to fail in one comparison, at the cost of being much more conservative and give fewer significant results in the first place [9]. Finally, we also note that while the Friedman test is used because it does not assume normality of the score distributions, Tukey's HSD does assume it.

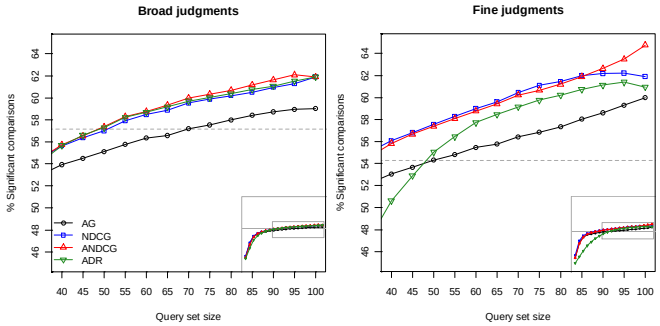

Figure 3. Evaluation power (larger is better) with W1, for all measures with the Broad (left) and Fine judgments (right).

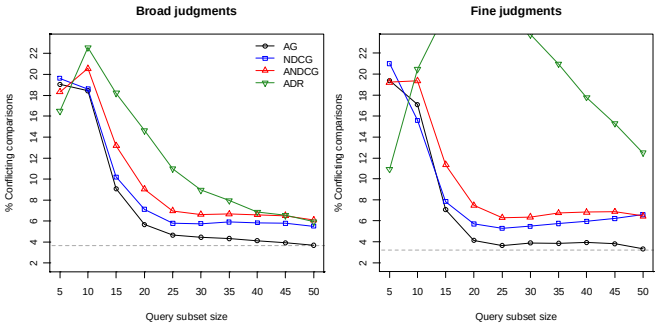

Figure 4. Evaluation stability (lower is better) with W1, for all measures with the Broad (left) and Fine judgments (right).

Tukey's HSD thus commits fewer Type I errors, but in the downside it is less powerful. We should at this point consider whether this is what we want. From the point of view of the participants, what they are interested in is the *subexperiment* of comparing their system with the other 14, and the remaining 91 pairwise system comparisons are rather uninteresting for them. Therefore, why not perform these simple 14 pairwise comparisons? The subexperiment-wise significance level would be $1-(1-\alpha)^{14}=0.512$. Most importantly, note that the number of pairwise system comparisons grows quadratically in the whole experiment but linearly in the subexperiments. As such, for evaluations with many more systems the power would decrease drastically if using Friedman-Tukey's HSD.

6.1 Evaluation Power

Here we perform the same experiment as in Section 4 and with the same query subsets, but instead of using Friedman-Tukey's HSD we perform the 105 pairwise system comparisons using 1-tailed Wilcoxon tests at the $\alpha=0.01$ significance level (W1). Therefore, the probability of committing a Type I error for the complete experiment (in any of the 105 system comparisons) is $1-(1-\alpha)^{105}=0.652$, but for the subexperiments (14 system comparisons in each one) it is dramatically reduced to $1-(1-\alpha)^{14}=0.131$.

Figure 3 shows as expected that many more significant differences are found between systems: as much as 20% more (the horizontal dotted lines mark the power achieved by the original evaluation). Interestingly, the difference between AG and (A)NDCG is here more acute, and it gets larger as more queries are used. The plots also suggest that W1 with about half the queries can achieve the same or better power levels as the original evaluation with FT.

6.2 Evaluation Stability

As expected, with simple Wilcoxon tests there are many more significant differences, but how many of them are actually caused by mere Type I errors? We have shown that the probability of having at least one incorrect result is very high, so next we look into stability.

As Figure 4 shows, the stability levels are very similar. AG again converges at about 3.5% of stability conflicts, and it does so much earlier than in the original evaluation. Most notably, (A)NDCG show here more stability conflicts, converging to about 6%. Note that the peaks observed for small subsets are here narrower because the statistical tests are more powerful in the first place. Again, ADR performs worse, especially for the Fine judgments.

7. DISCUSSION

Taking a close look at the power and stability results, one may wonder whether it is necessary to use as many as 100 queries. From a pragmatic point of view, we have argued that simple 1-tailed Wilcoxon tests are more useful to the MIREX participants than Friedman-Tukey's HSD. Next, we show analytically that they are even more reliable and cheaper (see Table 1).

	50 queries			100 queries	
	Power	Conflicts	Stable	Power	Stable
AG_{Broad}(FT)	52.4%	3.6%	48.8%	57.1%	53.5%
AG_{Fine}(FT)	51.9%	3.2%	48.7%	54.3%	51.1%
AG_{Broad}(W1)	55.1%	3.7%	51.4%	59.0%	55.4%
AG_{Fine}(W1)	54.3%	3.3%	51.0%	60.0%	56.7%

Table 1. Power and stability for 50 and 100 query sets when using Friedman-Tukey's HSD (FT) or 1-tailed Wilcoxon tests (W1).

For instance, with AG_{Fine} and 50 queries 51.9% of the 105 pairwise comparisons are significant according to FT, but 3.2% have a stability conflict. Thus, 48.7% of the comparisons are both significant and stable. Assuming the apparent convergence of conflicting results, for 100 queries there would be 51.1% significant and stable results. But also with 100 queries, W1 is even more stable, and with as little as 50 queries it is as reliable as FT with the full query set, having 51.0% of significant and stable results. (A)NDCG show very similar results, with differences of about 2%.

We note again that very few of these conflicts are caused by a change in the sign of the difference between systems, and never is it found significant for both query samples. Indeed, 97.3% of the conflicts with AG were caused by mere lack of statistical power in one of the paired query samples, 96.7% with NDCG and 95.9% with ANDCG. Again, this indicates that if significance is found, it most probably is correct.

8. CONCLUSIONS AND FUTURE WORK

We have analyzed the MIREX Audio Music Similarity and Retrieval task in terms of power and stability of the evaluations, studying four effectiveness measures (AG, NDCG, ANDCG and ADR) with the two traditional sets of relevance judgments employed in MIREX (Broad and Fine). About 55% of the pairwise system comparisons come up statistically significant with current practices, with all measures but ADR behaving very similarly. The increase in power follows a logarithmic trend with the number of queries used, so merely using more queries to achieve significance does not pay off at some point. As to stability, we observed that about 4% of the pairwise system comparisons are unstable: with one test collection the difference would be significant, but with a different collection it would not. However, less than 0.14% of these conflicts had a swap in the sign of the difference, and in no case was a sign swap coupled with significance in both query samples: at worst, they were too small to observe significance in both evaluations. This indicates that if a significant difference is found between two systems, experimenters can be very confident that the result is indeed correct and general.

From the pragmatic point of view of a MIREX participant, we argue that the Friedman-Tukey's HSD procedure used to measure significance is not appropriate. In fact, comparing all system pairs with simple 1-tailed Wilcoxon tests at the α=0.01 significance level we can obtain even more reliability. Most importantly, we have shown that with this procedure the query set can be cut in half, and yet the reliability of the results would be as good as if using all 100 queries and Friedman-Tukey's HSD. This effectively reduces to 50% the effort needed for relevance judging, which is especially appealing both for in-house evaluations with little resources and for the continuity of MIREX, given its recent funding issues [2]. Some of the spare effort could even be dedicated to the evaluation of more queries in the SMS task.

Future work will examine other test collections, used both in audio and symbolic similarity retrieval. We believe that the similar behavior observed for AG, NDCG and ANDCG is due to the small evaluation depth: only the top 5 results per system are judged for relevance. Using (A)NDCG with the standard logarithm base 2, as we did, takes advantage of the ranking only beyond the second document retrieved. Just the top 5 documents might be too few to note the difference, so we also plan to study the effect of evaluation depth in power and stability. The effect of the number of systems is also subject for further research, as it affects not only the statistical procedure but also the evaluation of other systems through the discovery of more relevant material. Indeed, we expect to find different patterns when evaluating systems by the same research group as opposed to systems by different groups. The ultimate goal of looking into these factors with more data is to come up with a model that allows us to draw some rules of thumb to guide experimenters in the tradeoff between reliability and cost.

REFERENCES

[1] C. Buckley and E.M. Voorhees, "Evaluating Evaluation Measure Stability," *ACM SIGIR*, pp. 33-34, 2000.

[2] J.S. Downie, "MIREX Next Generation," *music-ir email list*, 2011. Available at: http://listes.ircam.fr/wws/info/music-ir.

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

[4] IMIRSEL, "MIREX 2009 Audio Music Similarity and Retrieval Results," http://music-ir.org/mirex/wiki/2009:Audio_Music_Similarity_and_Retrieval_Results.

[5] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM Transactions on Information Systems*, 20:4, pp. 422-446, 2002.

[6] J.H. Lee, "Crowdsourcing Music Similarity Judgments using Mechanical Turk," *ISMIR*, pp. 183-188, 2010.

[7] T. Sakai, "On the Reliability of Information Retrieval Metrics Based on Graded Relevance," *Information Processing and Management*, 43:2, pp. 531-548, 2007.

[8] M. Sanderson and J. Zobel, "Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability," *ACM SIGIR*, pp. 162-169, 2005.

[9] M.A. Seaman, J.R. Levin, and R.C. Serlin, "New Developments in Pairwise Multiple Comparisons: Some Powerful and Practicable Procedures," *Psychological Bulletin*, 110:3, pp. 577-586, 1991.

[10] R. Typke, M. den Hoed, J. de Nooijer, F. Wiering, and R.C. Veltkamp, "A Ground Truth for Half a Million Musical Incipits," *Journal of Digital Information Management*, vol. 3, no. 1, pp. 34-39, 2005.

[11] R. Typke, R.C. Veltkamp, and F. Wiering, "A Measure for Evaluating Retrieval Techniques based on Partially Ordered Ground Truth Lists," *IEEE International Conference on Multimedia and Expo*, pp. 1793-1796, 2006.

[12] J. Urbano, "Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain," *ISMIR*, 2011.

[13] J. Urbano, M. Marrero, D. Martín, and J. Lloréns, "Improving the Generation of Ground Truths based on Partially Ordered Lists," *ISMIR*, pp. 285-290, 2010.

[14] J. Urbano, J. Morato, M. Marrero, and D. Martín, "Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks," *ACM SIGIR Workshop on Crowdsourcing for Search Evaluation*, pp. 9-16, 2010.

[15] E.M. Voorhees, "Topic Set Size Redux," *ACM SIGIR*, pp. 806-807, 2009.

[16] E.M. Voorhees and C. Buckley, "The Effect of Topic Set Size on Retrieval Experiment Error," *ACM SIGIR*, pp. 316-323, 2002.

[17] W. Webber, A. Moffat, J. Zobel, and T. Sakai, "Precision-At-Ten Considered Redundant," *ACM SIGIR*, pp. 695-696, 2008.