IMPROVING THE GENERATION OF GROUND TRUTHS BASED ON PARTIALLY ORDERED LISTS

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

ABSTRACT

Ground truths based on partially ordered lists have been used for some years now to evaluate the effectiveness of Music Information Retrieval systems, especially in tasks related to symbolic melodic similarity. However, there has been practically no meta-evaluation to measure or improve the correctness of these evaluations. In this paper we revise the methodology used to generate these ground truths and disclose some issues that need to be addressed. In particular, we focus on the arrangement and aggregation of the relevant results, and show that it is not possible to ensure lists completely consistent. We develop a measure of consistency based on Average Dynamic Recall and propose several alternatives to arrange the lists, all of which prove to be more consistent than the original method. The results of the MIREX 2005 evaluation are revisited using these alternative ground truths.

1. INTRODUCTION

Information Retrieval (IR) is known for having evolved as a highly experimental discipline. New techniques appear every year, and it is necessary to perform an exhaustive and methodological evaluation to figure out which of these techniques really mean a step forward in the field. These evaluations have been carried out since the late 50's in what has come to be known as the Cranfield paradigm. Given a fixed document collection, IR systems provide their results for certain information needs. Then, they are evaluated against the so called ground truths, which contain information about the documents that should ideally be retrieved by a system. Usually, these ground truths take the form of a matrix, containing the relevance, assessed by humans, for each document to an information need (traditional values are "irrelevant", "relevant" and "highly relevant").

These evaluations have been carried out mostly in Text Information Retrieval, with the TREC conferences as its flagship [1]. Music Information Retrieval (MIR), on the other hand, is a relatively young discipline, and this kind of evaluations has been somewhat scarce until the arrival of MIREX in 2005 as a first attempt to perform TREC-like evaluations in the musical domain [2]. Music IR differs from Text IR in many aspects [3], making the construction and maintenance of such test collections very difficult. In particular, it is unclear what relevance level to assign to a document for a given information need.

In the case of melodic similarity, some studies indicate that relevance is continuous [4]. Single melodic changes such as moving a note up or down in pitch, or extending or shortening its duration, are not perceived to change the overall melody. Nonetheless, the relationship with the original melody is gradually weaker as more changes are applied to it. There does not seem to be common criteria to split the degree of relevance into different levels, so assessments with a fixed scale seem inappropriate.

Ground truths based on partially ordered lists attempted to handle this problem with relevance assessment by the beginning of 2005 [5]. Instead of having documents with a fixed relevance level, these ground truths are lists with ordered groups of documents. The earlier a group appears in the list, the more relevant its documents are, and documents within the same group are assumed to be equally relevant. That way, the ideal retrieval should return these documents in order of relevance, although permutations within the same group are allowed. Because traditional effectiveness measures such as precision or recall need relevance assessments with a fixed scale, a new measure, called Average Dynamic Recall (ADR) [6], was developed also in 2005 to evaluate retrieval systems with ground truths based on partially ordered lists.

The first edition of MIREX had a task for symbolic melodic similarity [7], where 11 ground truths based on partially ordered lists were used along with ADR to evaluate state-of-the-art retrieval systems. Similar methods were used in the 2006 and 2007 editions, as well as in private evaluations external to MIREX, such as [8] [9] [10] and [11]. However, we are not aware of any meta-evaluation work addressing the correctness or improvement of these ground truths. Indeed, a thorough examination shows that the lists have some inconsistencies as to the arrangement and aggregation of documents in groups.

The paper is organized as follows. In Section 2 we review the methodology followed to create these ground truths. Section 3 unveils some inconsistencies and shows that it is not possible to ensure fully-consistent lists. In Section 4 we propose several alternatives to set up the groups, and present a measure to quantify consistency. Section 5 shows the results of the alternatives proposed and revise the MIREX 2005 evaluation using them. The paper ends with conclusions and lines for future work.

2. CURRENT METHODOLOGY

The original method to create ground truths based on partially ordered lists, as described in [5], was used with the RISM A/II collection [12], which at the time contained about half a million musical incipits. The methodology followed may be divided in four steps: filtering, ranking, arranging and aggregating.

First, several features were calculated for each document (musical incipits in this case), such as pitch range, interval histogram or motive repetitions. Filtering by these features, the initial collection was gradually narrowed down to under 300 incipits per query. Then, clearly irrelevant incipits were manually excluded, and several melodic similarity algorithms were used to add supposedly relevant incipits. Second, and once the lists had about 50 candidate incipits each, 35 experts ranked them in terms of melodic similarity to the corresponding query: the more similar a candidate was to the query, the higher it had to be ranked in the list. Incipits that seemed very different from the query could be left unranked. Third, incipits were arranged according to the median of their rank sample. If two incipits had the same median rank, the means were used to resolve the tie. Therefore, the incipits that on average were ranked higher by the experts appeared with higher ranks in the ordered list. Fourth, incipits whose rank samples were similar were aggregated within a group, so as to indicate that they were similarly relevant to the query. Thus, a retrieval system could return them with their rank swapped and still be considered correct. The Mann-Whitney U test (also known as Wilcoxon Rank Sum test, see Appendix) [13], was used to tell whether two incipits had similar ranks or not.

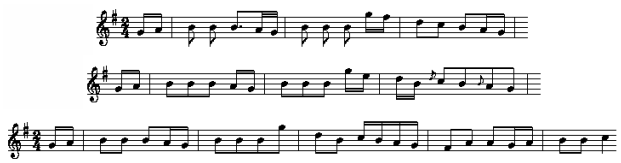

Figure 1. First three results for query 000.111.706-1.1.1. Top to bottom: same incipit as the query, incipit 000.113.506-1.1.1 and incipit 000.116.073-1.1.1.

The ground truths generated have some odd results, as already noted in [5] and [9]. For example, in the list for the query incipit 270.000.749-1.19.1, the first result is the same as the query; the second one (incipit 270.000.746-1.41.1) is written with a different clef, but otherwise identical to the query; and the third result (incipit 270.000.748-1.19.1) is the same as the first half of the query. Although the experts were told to disregard these kinds of changes in the melody, these three results ended up in different groups, indicating that their relevances to the query were significantly different. Also, incipits with virtually the same changes in the melody were sometimes placed in different groups, as it occurs with incipits 000.113.506-1.1.1 and 000.116.073-1.1.1 with respect to the query 000.111.706-1.1.1 (see Figure 1).

These rare results seem to be caused by the second step of the methodology, when experts ranked the results. Though important, we will not focus on them in this paper. There are other problems with this kind of ground truths that have not been addressed yet and lead to inconsistent result lists and incorrect evaluation. These inconsistencies arise at steps three and four, and they are the ones we address here.

3. INCONSISTENCIES DUE TO ARRANGEMENT AND AGGREGATION

We thoroughly examined the 11 ground truth lists used in the evaluation of the symbolic melodic similarity task in MIREX 2005 (*Eval05* for short), and found that there are pairs of incipits contained in the same group of relevance although there is a significant difference between the ranks the experts gave them (i.e. an intra-group inconsistency). For example, incipits 453.001.547-1.1.3 and 451.509.336-1.1.1, for query 190.011.224-1.1.1, are in the same group (see Figure 2), but their difference is significant. That means that if a retrieval system returned them in reverse order it would be considered correct, despite the experts clearly ranked them differently. On the other hand, incipits for which no significant difference could be found form part of different groups (i.e. an inter-group inconsistency). Incipits 700.000.686-1.1.1 and 450.034.972-1.1.1 for the previous query are an example (see Figure 2). Similarly, if a retrieval system returned them in reverse order, it would not be considered correct, despite no difference was found in the experts rankings.

These inconsistencies appear throughout the lists, and they are caused by the initial arrangement and the aggregation function used in the third and fourth steps.

3.1 Arrangement

In the third step of the methodology, incipits are arranged according to the median and mean ranks they were given by the experts. Because the Mann-Whitney U test is used later on to find statistically significant differences between the incipits' ranks, using central-tendency measures such as the median and the mean might not be appropriate to arrange the results, because they do not account for the dispersion in the samples.

Although rare, this phenomenon may happen: we examined the 11 ground truths of the *Eval05* collection and found it. For example, incipits 850.014.902-1.1.1 and 451.002.538-1.1.1 are ranked 20th and 22nd, respectively, for query 400.065.784-1.1.1. Their sample median ranks are 12 and 12.5, so the first one is ranked higher. However, a 1-tailed Mann-Whitney U test shows that it is highly probable for the true medians to be ordered the other way around, so the second incipit should be ranked higher than the first one.

3.2 Aggregation

In the fourth step of the methodology, incipits are aggregated in groups according to their relevance to the query. The rationale originally used by the aggregation function is as follows: traverse from top to bottom the list of incipits already arranged by median and mean, and begin a

new relevance group if the pivot incipit is significantly different from all incipits in the current group [5]. Therefore, it will probably allow significantly different incipits in the beginning and the end of the same group, just because they are not different from a third one. A new group will begin only when an incipit is very significantly different from all the previous ones, so the group is likely to grow a lot. We looked for this kind of inconsistency in the 11 lists of the *Eval05* collection, and found that out of the total 509 ordered pairs of incipits in the same relevance group, 178 (35%) are significantly different. All these intra-group inconsistent pairs translate into incorrect evaluation when allowing an incipit to appear earlier in the results list just for being misplaced in the same group as another one ranked a little higher.

There can also be cases where the aggregation function places an incipit in a new group, but the next one is not significantly different from some others in the group just finished. This next incipit should be in the previous group, but that is not possible since it has been already closed because of the previous one. For example, incipit 453.001.547-1.1.3 started group 4 for the query 190.011.224-1.1.1, because it was different from all incipits in group 3. However, the incipit 700.000.686-1.1.1, in group 3, is not significantly different from incipit 450.034.972-1.1.1, which is in group 4 (see Figure 2). All these inter-group inconsistent pairs also translate to incorrect evaluation when not permitting an incipit to appear earlier in the results for being misplaced in a later group started by another incipit that was sufficiently different.

3.3 Fully Consistent Lists

Inconsistencies come from two different sources: the initial arrangement by median and mean may arrange pairs of documents in the wrong order, and the aggregation function may combine significantly different incipits or set apart similar ones. The aggregation function can mitigate these problems, but there is a more profound problem: hypothesis testing is not transitive.

Let *X*, *Y* and *Z* be the rank samples given to three different incipits. The Mann-Whitney U test may suggest that the median of *Y* is less than the median of *Z*, and that the median of *X* is less than the median of *Y*. Still, it may suggest that the medians of *X* and *Z* are not different (i.e. $X < Y$, $Y < Z$ but $X = Z$). Although this might seem completely paradoxical, it actually happens, for example in the ground truth list for query 400.065.784-1.1.1. Let *X*, *Y* and *Z* be the rank samples of incipits 702.002.512-1.1.1, 804.002.648-1.1.2 and 450.021.643-1.1.1, ranked 6th, 8th and 16th respectively. A 1-tailed Mann-Whitney U test shows that the median of *X* is significantly smaller than the one of *Y* (p-value = 0.238) and that the median rank of *Y* is significantly smaller than the one of *Z* (p-value = 0.239), but the median rank of *X* does not seem significantly smaller than the one of *Z* (p-value = 0.272). Although p-values this large would not usually be accepted to reject a null hypothesis, they are valid in our case, since the significance level originally used was 0.25 [5].

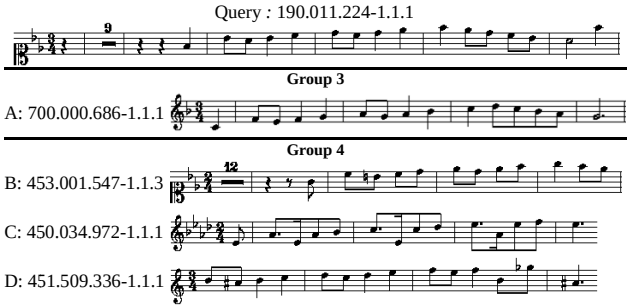

Figure 2. Excerpt of the ground truth for query 190.011.224-1.1.1. According to the experts: $B \neq D$ (intra-group inconsistency), $A = C$ (inter-group inconsistency), $A \neq B$ and $B = C$ (2-tailed non-transitivity as $A = C$).

Therefore, it is not possible to ensure fully consistent lists with this method. In the example above, incipit *X* should be in a group ranked higher than *Y*, which should be in a group ranked higher than *Z*. However, *X* and *Z* should be in the same group, which is clearly impossible. Similar cases can be found with 2-tailed tests, such as the example in Figure 2 ($X \neq Y$, $X = Z$ but $Y = Z$).

4. ALTERNATIVE AGGREGATION FUNCTIONS

The number of intra- and inter-group inconsistencies depend on the aggregation function used. A function too permissive, like the original one, leads to larger groups with more likelihood of intra-group inconsistencies, but a function too restrictive leads to smaller groups with more likelihood of inter-group inconsistencies. The aggregation function should minimize these problems and generate lists as consistent as possible.

We consider three different rationales to be followed:

- *All*: a new group is started if the pivot incipit is significantly different from every incipit in the current group. This should lead to larger groups.
- *Any*: a new group is started if the pivot incipit is significantly different from any incipit in the current group. This should lead to smaller groups.
- *Prev*: a new group is started if the pivot incipit is significantly different from the previous one.

At this point, we have to consider whether 2-tailed or 1-tailed tests should be used. Originally, 2-tailed tests were used, looking for 2-way differences in median ranks. But, because the incipits are already sorted by median and mean, we believe the tests should be 1-tailed, looking for 1-way differences in ranks. After arranging incipits in step three, we may assume that an incipit appearing after another one has a rank either lower or equal, but not higher. In these situations, 1-tailed tests are more powerful than their 2-tailed counterparts, so it is more probable for them to find a difference between two samples if there really is one (see Appendix).

Therefore, we obtain six different functions, combining each of the three rationales with each of the two statistical tests. We call these functions *All-2*, *All-1*, *Any-2*, *Any-1*, *Prev-2* and *Prev-1*. Note that *All-2* is the function originally used by Typke et.al. [5], while the other five are proposed in this paper.

4.1 Measure of list consistency

To evaluate the 5 alternative functions presented above, and compare them with the original one, we developed a measure of consistency based on Average Dynamic Recall [6]. ADR is the main effectiveness measure used to evaluate retrieval systems against ground truths based on partially ordered lists, so we followed its same idea to measure their consistency, and hence the correctness of the evaluation itself. ADR measures the average recall over the first *n* documents, where *n* is the number of documents in the ground truth. At each point, the set of relevant documents allowed comprises all previous documents in the list plus all those in the same group as the pivot, because they are supposed to be equally relevant.

With a ground truth list like *⟨(A, B), (C), (D, E, F)⟩*, and a retrieval list such as *⟨B, C, A, G, H, D⟩*. ADR would be calculated as in Table 1. In the first two positions, either document *A* or *B* is considered correct because they are in the same relevance group, so both of them can be expected. At position 3, both *A* and *B* are expected because they appear before in the list, and only *C* is added when expanding the second group. However, when position 4 is reached, every document in the third group may be expected. Recall is calculated at each position, and the overall ADR is the mean average of these recalls, 0.753 in this case.

Position	Retrieved	Expected	Correct	Recall
1	**B**	A,**B**	1	1
2	**B**,C	A,**B**	1	0.5
3	**B**,**C**,**A**	**A**,**B**,**C**	3	1
4	**B**,**C**,**A**,G	**A**,**B**,**C**,D,E,F	3	0.75
5	**B**,**C**,**A**,G,H	**A**,**B**,**C**,D,E,F	3	0.6
6	**B**,**C**,**A**,G,H,**D**	**A**,**B**,**C**,**D**,E,F	4	0.667

Table 1. Example of ADR calculation.

To measure the consistency, the list is traversed from top to bottom, expanding the group corresponding to the pivot incipit. At each position, it is calculated the percentage of incipits expanded that are actually correct according to the experts rankings. At the end, the mean of those percentages is calculated. Therefore, a final value of 1 means that every expansion is correct and hence the list is fully-consistent. A value of 0 means that every expansion is incorrect. The pivot incipit is never considered for the calculation, because it will always be correctly expanded.

There are two types of incorrect expansion: false positives (i.e. an incipit is included in the set of expected, but it is significantly different from the pivot) and false negatives (i.e. an incipit is not included in the set of expected, but it is not significantly different from the pivot). Note that false positives correspond to intra-group inconsistencies, and false negatives correspond to inter-group inconsistencies. In the example above, imagine *A* and *C* are not significantly different, but *D* and *F* are. In that case, the expansion at position 1 is missing incipit *C* (a false negative due to an inter-group inconsistency between groups 1 and 2). Also, the expansion at position 4 would incorrectly include incipit *F* (a false positive due to an intra-group inconsistency in group 3). Note that at position 2, *C* would not be correctly expanded, as it is still significantly different from *B*. Note also that position 6 is not considered, as there is actually no expansion at the end of the list. In this case, the overall list consistency would be 0.86 (see Table 2).

Position	Correct expansion	Actual expansion	% of correct expansions
1	B,**C**	B	0.5
2	A	A	1
3	A,B	A,B	1
4	A,B,C,E	A,B,C,E,**F**	0.8
5	A,B,C,D,F	A,B,C,D,F	1

Table 2. Example of list consistency calculation.

As before, we can measure the inconsistencies using both 2-tailed and 1-tailed tests. In the former, two incipits are expected to be in the same expanded set if a 2-tailed Mann-Whitney U test is not rejected. In the latter, they are expected to be in the same expanded set if none of the two 1-tailed tests is rejected (i.e. the true median rank of one incipit seems to be neither less nor greater than the other's). Note that both the 2-tailed and the 1-tailed measures account for inconsistencies originated by the aggregation function but only the 1-tailed version accounts for inconsistencies due to the simple arrangement by median and mean. We call these two measures ADR-2 and ADR-1 consistency.

Because of the non-transitivity problem, lists are not expected to have an overall consistency of 1. However, it could be maximized by changing the aggregation function, thus improving the correctness of the evaluation.

5. RESULTS

The five alternative aggregation functions proposed back in Section 4 were used to re-generate the 11 lists in the *Eval05* collection and compare them with the original function *All-2*. We used the ADR-1 consistency measure to calculate the overall consistency of each list. The results are in Figure 3 and in Table 3.

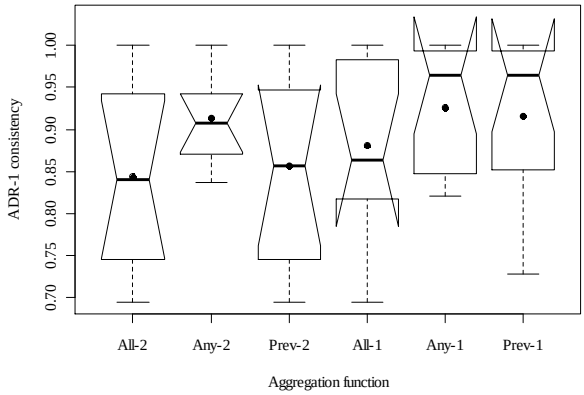

Figure 3. ADR-1 consistency for the six aggregation functions. Solid circles indicate the mean value. Notches mark the 95% confidence interval around the median.

As can be seen, the original function, *All-2*, is outperformed by all of the five alternatives proposed. *All-2* leads to an average consistency of 0.844, which is the

smallest of the six. However, *Prev-2* and *All-1* are not significantly better according to a 1-tailed t-test at the 0.10 significance level. Moreover, the *All* functions lead to results with more variability, while the *Any* functions are more stable in terms of consistency. These results indicate that if the lists were generated with the *Any-2*, *Any-1* or *Prev-1* aggregation functions, they would be more consistent, and so would be the evaluation with them.

Interestingly, the relative order for each of the three 2-tailed and 1-tailed functions is maintained. That is, the *All* functions perform the worst, followed by *Prev* and *Any*, which perform the best both in terms of average consistency and variability.

Our guess back in Section 3.2 was that the larger the sizes of the relevance groups are, the more inconsistent the lists are too. To examine this, we calculated the mean number of incipits per group for each of the 11 resultant lists. Figure 4 and Table 3 show the results.

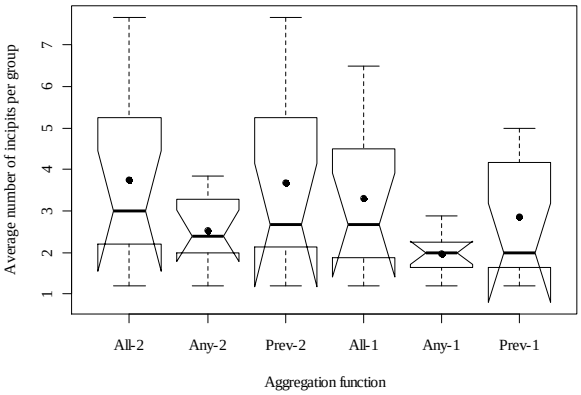

Figure 4. Mean number of incipits per group for each aggregation function. Solid circles indicate the mean value. Notches mark the 95% confidence interval around the median.

As expected, the *All* functions lead to larger groups, because an incipit goes to a new group only if it is different from all the previous ones. On the other hand, the *Any* functions generate smaller groups, because only one difference needs to be found to place the incipit in a new group. Similarly, the *Any-2* function leads to significantly smaller groups than *All-2* at the 0.10 significance level, and *Any-1* is significantly smaller at the 0.05 level.

Aggregation function	ADR-1 consistency	Incipits per group	Pearson's r
All-2	0.844	3.752	-0.892***
Any-2	0.913**	2.539*	-0.862***
Prev-2	0.857	3.683	-0.937***
All-1	0.881	3.297	-0.954***
Any-1	0.926**	1.981**	-0.749***
Prev-1	0.916*	2.858	-0.939***

Table 3. Summary of results. * for significant difference at the 0.10 level, ** at the 0.05 level and *** at the 0.01 level.

Following these results, there seems to be a direct relationship between the size of the groups and the overall consistency of the lists. We checked this by calculating the Pearson's r correlation coefficient between the two variables and, as expected, there is a strong negative correlation, indicating that the size of the groups affects the consistency of the lists (see Table 3). This is why the *All* functions perform worse and the *Any* functions perform better: the *All* versions generate larger groups. Doing so, they allow for many incorrect expansions in the form of false positives due to intra-group inconsistencies.

5.1 MIREX 2005 Results Revisited

In the 2005 edition of the MIREX evaluations, there was a task for symbolic melodic similarity that used 11 ground truths based on partially ordered lists (what so far we have called the *Eval05* collection). In particular, 7 different systems were evaluated.

We calculated the ADR score of each system with the lists generated by the five alternative aggregation functions (see Table 4). Every alternative evaluation produces worse results than the original, except for *Prev-2*, which leads to the same scores. Indeed, every system performed worse for every alternative set of ground truths, with reductions in ADR score of up to 12%.

System	All-2	Any-2	Prev-2	All-1	Any-1	Prev-1
GAM	**0.66**	0.59	**0.66**	0.624	0.583	0.605
O	0.65	**0.607**	0.65	**0.643**	0.593	**0.639**
US	0.642	0.604	0.642	0.639	**0.594**	0.628
TWV	0.571	0.558	0.571	0.566	0.556	0.564
L(P3)	0.558	0.52	0.558	0.54	0.515	0.534
L(DP)	0.543	0.503	0.543	0.511	0.494	0.506
FM	0.518	0.498	0.518	0.507	0.483	0.507
τ	-	0.81	1	0.81	0.714	0.714

Table 4. ADR results of the systems that participated in MIREX 2005 with the lists resulting from the alternative aggregation functions. GAM = Grachten, Arcos and Mántaras; O = Orio; US = Uitdenbogerd and Suyoto; TWV = Typke, Wiering and Veltkamp; L(P3) = Lemström (P3), L(DP) = Lemström (DP); FM = Frieler and Müllensiefen. Best scores appear in bold face.

More importantly, the relative order of the systems, in terms of their mean ADR score, is also modified. For example, with the original lists GAM was the best system, followed by O and US. With the *Any-2* lists, O is ranked first, before US and GAM. However, with the *Any-1* lists the order is reversed: US, O and GAM. We calculated Kendall's τ correlation coefficient to measure the differences in the ranking of systems (see Table 4). A value of 1 means that two rankings are equal, and a value of -1 means that they are reversed. Except for *Prev-2*, which produces the same results as *All-2*, the correlation coefficients tell us that the resulting rankings are different.

6. CONCLUSIONS AND FUTURE WORK

With their appearance in 2005, ground truths based on partially ordered lists represented a big leap towards the scientific evaluation of Music Information Retrieval systems, particularly for melodic similarity tasks. They have been widely accepted and used by the community, both in MIREX and other private evaluations.

We have revised the methodology used to generate these lists, unveiling some unaddressed problems. We have shown that the lists generated have inconsistencies, and propose several alternatives to minimize them. Using ADR-1 consistency, we have shown that our alternatives

lead to better results. We have also seen how would have changed the evaluation of the symbolic melodic similarity task in MIREX 2005, showing that the absolute effectiveness figures would have changed notably, and the ranking of systems would have been different too.

More meta-evaluation work in this line has to be carried out to improve the evaluation in MIR. In this paper we have focused on the last two steps of the methodology, analyzing the evaluation collection used in MIREX 2005. Other test collections should be analyzed, and the first two steps of the methodology should be studied as well because they are known to produce odd results too. One of the reasons may be the subjectivity on the judgments that the loose definition of the task can lead to, as already noted in [2] and [3]. More precise definitions of the information need sought by these tasks would surely lead to more coherent judgments by the experts.

One point that has not been discussed in the literature either is the significance level used by the aggregation function, which was 0.25 for the original lists. Our measure of consistency also works with a significance level to decide whether incipits are correctly arranged or not, and though they should probably be the same, we should study what value is more appropriate in both cases.

Finally, the lists generated with the alternative aggregation functions show diverse characteristics, mainly in terms of group sizes and differences among incipits in the same group. Other effectiveness measures, besides ADR, could be proposed to exploit these characteristics, while accounting for the unavoidable inconsistencies.

ACKNOWLEDGEMENTS

We thank Carlos Gómez, Rainier Typke and the IMIRSEL group, especially Mert Bay and Stephen Downie, for providing us with the MIREX evaluation data. We also thank William Frakes and Gabriella Belli for their insight regarding statistics and hypothesis testing.

APPENDIX. THE MANN-WHITNEY U TEST

The Mann-Whitney U test [13], or Wilcoxon Rank-Sum test, is a non-parametric statistical test to assess whether the true medians of two independent samples, say *X* and *Y*, are significantly different or not. Consider *X* as the sample of ranks of 600.258.342-1.1.2 for query 600.053.481-1.1.1, and *Y* the ranks given to incipit 850.020.721-1.1.1. The test statistic *U* is calculated as:

$$U=|X|\cdot|Y|+\frac{|Y|(|Y|+1)}{2}-\sum_{i=1}^{|Y|}\text{rank}(y_i)$$

where *rank(y_i)* is the rank that the *i-th* number of *Y* would have in the set $X \cup Y$. In our example, $U = 131$. The critical value is calculated depending on the alternative hypothesis H_1. For a 2-tailed test, H_1 would be that the true medians are different, but if a 1-tailed is chosen instead H_1 would be that the true median of *X* is less than the true median of *Y* (or the other way around). In the 2-tailed case, the rejection region is spread around both sides of the critical value, while in the 1-tailed case it is only in one side. Therefore, the 2-tailed case accounts for 2-way differences ($X > Y$ or $X < Y$), while the 1-tailed case looks only for 1-way differences ($X < Y$ in our case).

With a significance level of 0.25, the critical value for the 2-tailed test is $U_2 = 121$, while for the 1-tailed test it is $U_1 = 136$. Thus, the 1-tailed null hypothesis would be rejected because $U < U_1$, but the 2-tailed would not because $U > U_2$. In this case, the 2-tailed test fails to detect that the medians are, in fact, different. Because the 1-tailed test looks for a signed difference, it is more powerful and rejects the null hypothesis ($H_0 = X \leq Y$ in our example).

REFERENCES

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

[2] 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, Springer, 2010, pp. 93-115.

[3] J.S. Downie, "The Scientific Evaluation of Music Information Retrieval Systems: Foundations and Future," *Computer Music Journal*, vol. 28, no. 2, 2004, pp. 12-23.

[4] E. Selfridge-Field, "Conceptual and Representational Issues in Melodic Comparison," *Computing in Musicology*, vol. 11, 1998, pp. 3-64.

[5] 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, 2005, pp. 34-39.

[6] 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*, 2006, pp. 1793-1796.

[7] J.S. Downie, K. West, A.F. Ehmann, and E. Vincent, "The 2005 Music Information Retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview," *International Conference on Music Information Retrieval*, 2005, pp. 320-323.

[8] M. Grachten, J. Arcos, and R. López, "A Case Based Approach to Expressivity-Aware Tempo Transformation," *Machine Learning*, vol. 65, no. 2, 2006, pp. 411-437.

[9] P. Hanna, P. Ferraro, and M. Robine, "On Optimizing the Editing Algorithms for Evaluating Similarity Between Monophonic Musical Sequences," *Journal of New Music Research*, vol. 36, no. 4, 2007, pp. 267-279.

[10] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado, "Using the Shape of Music to Compute the Similarity between Symbolic Musical Pieces," *International Conference on Computer Music Modeling and Retrieval*, 2010.

[11] A. Pinto and P. Tagliolato, "A Generalized Graph-Spectral Approach to Melodic Modeling and Retrieval," *International ACM Conference on Multimedia Information Retrieval*, 2008, pp. 89-96.

[12] K. Saur Verlag, "Répertoire International des Sources Musicales (RISM). Serie A/II, Manuscrits Musicaux après 1600," 2002.

[13] H.B. Mann and D.R. Whitney, "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other," *Annals of Mathematical Statistics*, vol. 18, no. 1, 1947, pp. 50-60.