# Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks

Julián Urbano, Jorge Morato, Mónica Marrero and Diego Martín

University Carlos III of Madrid
Department of Computer Science
Leganés, Madrid, Spain

{jurbano, jmorato, mmarrero, dmandres}@inf.uc3m.es

## ABSTRACT

Music similarity tasks, where musical pieces similar to a query should be retrieved, are quite troublesome to evaluate. Ground truths based on partially ordered lists were developed to cope with problems regarding relevance judgment, but they require such man-power to generate that the official MIREX evaluations had to turn over more affordable alternatives. However, in house evaluations keep using these partially ordered lists because they are still more suitable for similarity tasks. In this paper we propose a cheaper alternative to generate these lists by using crowdsourcing to gather music preference judgments. We show that our method produces lists very similar to the original ones, while dealing with some defects of the original methodology. With this study, we show that crowdsourcing is a perfectly viable alternative to evaluate music systems without the need for experts.

## Categories and Subject Descriptors

H.5.1 [Multimedia Information Systems]: Evaluation/ methodology; H.3.3 [Information Search and Retrieval]; H.3.4 [Systems and Software]: Performance evaluation (efficiency and effectiveness).

## Keywords

Crowdsourcing, relevance judgment, music information retrieval.

## 1. INTRODUCTION

Evaluation experiments are the corner stone of Information Retrieval (IR), as they are the main research tool for scientifically comparing retrieval techniques and figuring out which improve the state-of-the-art and which do not [1]. These evaluations have traditionally followed the so called Cranfield paradigm, where the set of relevance judgments are the most important and most expensive part of test collections. Usually, these ground truths take the form of a matrix containing information, assessed by humans, about the relevance of each document for each information need.

Music Information Retrieval (MIR) is a relatively young discipline, and this kind of evaluations has been somewhat scarce until the arrival of the Music Information Retrieval Evaluation eXchange (MIREX) in 2005, as a first attempt to perform TREC-like evaluations in the musical domain [2]. Evaluation in Music IR differs greatly from evaluation in Text IR, mainly with regard to the construction and maintenance of test collections [3]. On the one hand, MIR has been traditionally biased toward classical music because of many issues concerning copyright laws and royalties. On the other hand, many retrieval tasks defined for the music domain are inherently more complex to evaluate. This is the case of the Symbolic Melodic Similarity (SMS) and Audio Music Similarity (AMS) tasks, as defined in MIREX, in which systems are asked to retrieve a ranked list of musical pieces deemed similar to some piece of music acting as query. In particular, it is unclear how to assess the relevance of a document for a given query.


SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Ground truths are traditionally based on a fixed scale of relevance with levels such as "relevant" and "not relevant". However, several studies indicate that relevance is continuous for information needs involving music similarity [4][5][6]. 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. However, the relationship with the original melody is gradually weaker as more changes are applied to it. There are no common criteria to split the degree of relevance into different levels, so assessments based on a fixed scale do not seem suitable as it would be difficult to draw the line between levels.

Major advancements in this matter were achieved by Typke et al. by the beginning of 2005. They developed a methodology to create ground truths where the relevance of a document does not belong to any prefixed scale, but it is rather implied by its relative position in a partially ordered list [5]. These lists have ordered groups of candidates assumed to be equally relevant to the query, so that the earlier a group appears in the list, the more relevant its documents are (see Figure 2). That way, the ideal retrieval technique should return these documents in order of relevance, and permutations within the same group are not penalized. With this new form of ground truth, there does not need to be any prefixed scale of relevance, and human assessors only need to be sure that any pair of documents is well ordered according to their similarity to the query.

In the first edition of MIREX, a Symbolic Melodic Similarity task was run using ground truths based on partially ordered lists [7]. These lists have also been widely accepted by the research community as the most comprehensive means to evaluate new retrieval techniques, such as [8][9][10] and [11]. However, they have proven to be expensive to generate, which forced the MIREX evaluations to move to traditional level-based relevance judgments in the 2006, 2007 and 2010 editions.

In this paper we propose a modification of the original methodology followed to create these lists, by means of crowdsourced preference judgments that allow the candidate documents to arrange and aggregate themselves into relevance groups [12]. We implemented it with Amazon Mechanical Turk (AMT), as an attempt to explore its suitability for music tasks. Indeed, we show that our method generates lists very similar to the original ones with far less cost and no need for music experts.

The rest of the paper is organized as follows. In Section 2 we describe the issues that motivate our work, reviewing the current methodology followed to create these ground truths and some of its problems. Section 3 presents our alternative methodology, and Section 4 shows how we implemented it with Mechanical Turk. In Section 5 we summarize the results obtained, showing that our alternative leads to very good results in terms of cost, consistency and agreement between assessors. The paper finishes in Section 6 with conclusions and lines for future research.

# 2. MOTIVATION

Ground truths based on partially ordered lists have two main drawbacks: they are hard to replicate and expensive to generate in terms of man-power, and they have several inconsistency problems where equivalent music pieces are judged differently.

## 2.1 Expensiveness

In the original lists created by Typke et al. [5], 35 music experts were needed for 2 hours to generate the ground truth for just 11 queries, and only 11 of them were able to work on all queries. This exceeds MIREX's human resources for a single task [2]. In part because of this restriction, the official MIREX evaluations were forced to move to traditional level-based relevance judgments from 2006 on. Two different scales were used: BROAD and FINE. The BROAD scale contained 3 levels: not similar (NS), somewhat similar (SS) and very similar (VS). The FINE scale was numerical, ranging from 0.0 to 10.0 with one decimal digit (note that this is not different than an ordinal scale with 101 levels). This choice of relevance scales presented several issues concerning assessor agreement, and the line between levels was again found to be very diffuse [6][2].

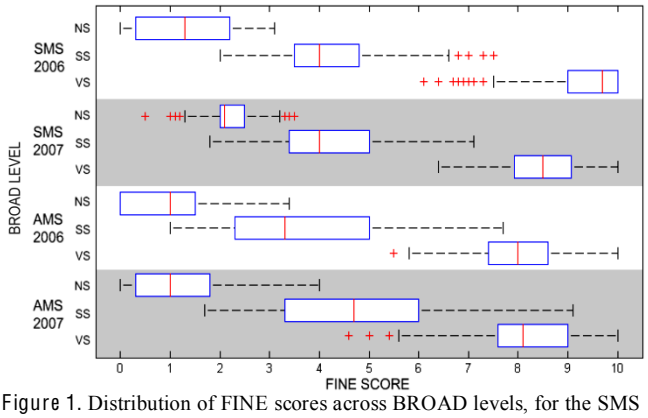


Figure 1. Distribution of FINE scores across BROAD levels, for the SMS and AMS tasks in 2006 and 2007. Taken from [2].

Figure 1 shows the distribution of FINE scores given across BROAD levels, both for the 2006 and 2007 editions of MIREX's Symbolic Melodic Similarity and Audio Music Similarity tasks. It can be seen that there was a great overlap between the FINE scores corresponding to the SS BROAD level and the NS and VS levels, as well as a large number of outliers, indicating that assessors were not very consistent when facing two different relevance scales. This is, again, evidence on the difficulty that relevance assessment posses for these tasks.

## 2.2 Inconsistencies Due to Ranking

The original method to generate ground truths based on partially ordered lists, as described in [5], was used with the RISM A/II collection [13], which at the time contained about half a million musical incipits (short excerpts from the beginning of musical pieces). The methodology followed may be divided in four steps: filtering, ranking, arranging and aggregating:

1. *Filtering*. Several musical features were calculated for each document (musical incipits in this case). Filtering by these features and using several melodic similarity algorithms, the initial collection was gradually narrowed down to about 50 candidate incipits per query.
2. *Ranking*. For each query, 35 experts ranked its candidates in terms of melodic similarity to the corresponding query. Incipits that seemed very different from the query could be left unranked. A limit of 2 hours per expert was imposed, so not every expert could work on every list.
3. *Arrangement*. Incipits were arranged according to the median of their rank sample, using the means to solve possible ties. Therefore, the incipits that on average were ranked higher by the experts appeared with higher ranks in the ordered list.
4. *Aggregation*. Incipits with similar rank samples 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 ranks swapped and still be considered correct. The Mann-Whitney U test (also known as Wilcoxon Rank-Sum test) [14] was used between the rank samples of two incipits to tell whether they were similar or not.

Several works have noted the presence of odd results in these lists [5][10][15]. The experts were instructed to disregard changes that do not alter the actual music perception, such as changes in clef or in key and time signatures. To compare, the textual counterpart of these changes would be something like changing the language of the text or replace some words with their synonyms, which do not change the actual contents but only its form [8]. Experts were also told to consider two incipits as equally relevant if one of them was part of the other.

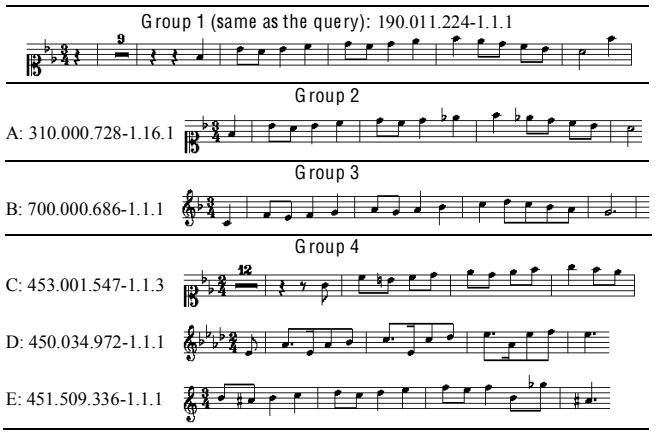


Figure 2. Excerpt of the ground truth for query 190.011.224-1.1.1.

However, incipits with such irrelevant differences ended up in different groups. For example, the second result (incipit *A*) expected for query 190.011.224-1.1.1 is like the query itself, but with the key signature changed (see Figure 2). Ignoring the leading silence, no listener would be able to tell the difference between this melody and the query, because they are the same note by note. Nonetheless, it was judged as less similar when compared to the query itself. The third result (incipit *B*) is like the second one, but with a change both in clef and key signature (see Figure 2). Again, these two melodies should be considered as equally similar to the query, but they ended up in different groups of relevance anyway.

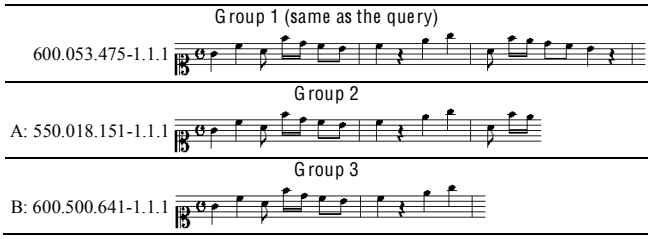


Figure 3. Excerpt of the ground truth for query 600.053.475-1.1.1

The top three results for query 600.053.475-1.1.1 show similar problems (see Figure 3). The third one (incipit *B*) is just like the second one (incipit *A*) but with 3 notes missing at the end, and this one is just like the first one (the query itself) with 3 notes missing at the end too. These three results ended up in different groups of relevance according to the rankings they were given by the experts, when they were instructed to judge them as equally similar. There are also cases were incipits with virtually the same changes in the melody were placed in different groups, as with the second and third results for query 000.111.706-1.1.1.

Despite they are no longer used in MIREX, ground truths based on partially ordered lists are still used to date for the evaluation of new retrieval techniques, as they are clearly more suitable for similarity tasks than traditional assessments. However, as no new lists have been generated since 2005, in house evaluations may be overfitting to this single collection. Therefore, we strongly believe that partially ordered lists should be brought back to the official MIREX evaluations so that new test collections are adopted. For that, further research should focus on new and more affordable ways to generate them. In a previous work we analyzed and dealt with inconsistencies originated in the latter steps of the methodology to generate the lists [15], and in this paper we deal with inconsistencies originated by the experts from the very beginning, while cheapening the whole process.

It has been hypothesized that with sufficient description of the information need sought by these tasks, any reasonable person should concur as to whether a given returned item satisfies the intention of the query (in our case, whether a returned piece is similar to another one). This is called the "reasonable person assumption" [3]. We decided to use Amazon Mechanical Turk to examine whether crowdsourcing alternatives can be used to gather accurate relevance judgments without the need for experts [16][17]. Doing so, we review the reasonable person assumption, evaluate crowdsourcing for a task very different from the usual ones focused on text, and study whether this alternative is doable and produces reliable results to evaluate music similarity tasks with partially ordered lists.

# 3. ALTERNATIVE METHODOLOGY

In a first attempt to bring partially ordered lists back to the evaluation of music similarity tasks, we explored alternatives in the current methodology to make the process more affordable and work toward large-scale evaluations, while trying to minimize inconsistencies. We opted for two changes: allow assessors to indicate that certain incipits are equally relevant, and have them perform simple preference judgments [12].

## 3.1 Equally Relevant Incipits

Reviewing the inconsistencies due to ranking (see Section 2.2), the reason seems to be clear: experts were not allowed to judge two incipits as equally relevant in the first place, they were only able to rank one above or below the other. Under this condition, for the example list in Figure 2 they will rank first the same incipit as the query, as it is identical. Even though incipit *A* is perceived as the same melody, they will surely rank it below and not above, as it has a change in key signature, even if they are told to ignore it. Same thing happens with incipit *B* (a change in clef). One would expect the experts to randomly assign opposite orders to such pairs of incipits for their medians to average out, but that is hardly the case. For instance, half the experts might rank incipit *A* as the second most similar, and incipit *B* right after, while the other half might rank them the other way around. However, any person looking at the staves would rank *A* before *B* because its image is more similar to the query's. In the example of Figure 3, the three incipits should be equally ranked, but the experts ranked them according to the number of notes missed. In no case should we expect such incipits to have similar ranks if we do not allow the experts to give them similar ranks in the first place.

The immediate solution to this problem would be to allow experts to specify groups of relevance from the very beginning. Also, the query-candidate pairs could be given as audio files to listen instead of as images of the corresponding staves. That way, the irrelevant changes indicated in Section 2.2 would be undistinguishable to the assessors, besides other misleading changes such as different arrangements of the stems of a group of eighth notes (quavers).

## 3.2 Preference Judgments

It is also important to note that the experts had to judge all candidates at once for each query. That is, they had to return a list of relevant candidates ranked by similarity. It is normal to guess that they would have more problems to set up a new incipit as the list grows: the first two candidates can be easily ordered, but once the list has, say, 15 incipits, it is clearly more difficult to decide where between those 15 should the next one be placed. This phenomenon could clearly accelerate assessor fatigue, and it was already observed for the level-based relevance judgments gathered in the 2006 and 2007 editions of MIREX [2][6]. Some experts had to go back and re-judge some documents, surely after they were presented a candidate which made them realize that a previous judgment was not very congruent. This agrees with the overlapping of FINE scores across BROAD levels shown in Figure 1, and indicates, again, that the relevance for music similarity is rather continuous and the differences between levels is certainly not clear.

To alleviate this problem we propose to ask for preference judgments of the form "incipit A is more similar to the query than incipit B" ($A < B$ for short). Carterette et al. studied the use of preference judgments for text IR and showed that they are better than traditional level-based judgments, both in terms of agreement and time to answer [12]. However, in their study they decided not to allow an option like "A and B are equally relevant" ($A = B$ for short), which we must permit in our case to form groups (see Section 4.1). Using preference judgments, we could implement a modified QuickSort algorithm to make the incipits auto-organize themselves following the preferences of the assessors. Such an algorithm has been shown to reduce dramatically the number of judgments needed to fully order a list, as the rate of growth in the number of comparisons is $O(n \cdot lg\ n)$, much slower than the $O(n^2)$ growth rate of all comparisons [12]. Table 1 shows an example.

In the first iteration of the algorithm, we choose the last document as the pivot, which is *F* in this case. The assessors would have to

answer preference judgments between *F* and each of the other documents. In this case, every document was judged as more similar, except for *G*, which was judged equally similar (or dissimilar). Therefore, a new segment appears to the left of *F* with all the candidates judged more relevant, and *G* is set up in the same group as *F*. For the second iteration, in the rightmost segment no judgment is needed because *F* and *G* were already compared, and *B* would be the pivot for the leftmost segment. Incipits *A* and *C* are judged similar to *B*, but *D* and *E* are judged as less similar, so they are set up in a segment to the right of *B*. At the end, there are 3 ordered groups of relevance formed with preference judgments. Note that not all the 21 judgments were needed to arrange and aggregate every incipit (e.g. *G* is only compared with *F*).

Table 1. Example of self-organized partially ordered list. Pivots for each segment appear in bold face. Documents that have been pivots already appear underlined.

| Iteration | Segments | Preference Judgments |
|---|---|---|
| 1 | ⟨C, D, E, A, G, B, **F**⟩ | C<F, D<F, E<F, A<F, G=F, B<F |
| 2 | ⟨⟨C, D, E, A, **B**⟩, ⟨F, G⟩⟩ | C=B, D>B, E>B, A=B |
| 3 | ⟨⟨B, C, **A**⟩, ⟨D, **E**⟩, ⟨F, G⟩⟩ | C=A, D=E |
| 4 | ⟨(A, B, C), (E, D), (F, G)⟩ | - |

With preference judgments, the sample of rankings given to each candidate is less variable than with the original method. Whenever a candidate is preferred over another one, it would be given a rank of 1 and -1 otherwise. In case it was judged equally similar, a rank of 0 would be added to its sample. With the original methodology, on the other hand, the ranks given to an incipit could range from 1 to well beyond 20, which increases the variance of the samples. Note that, with our scheme, the two samples of rankings given to each pair of documents are the opposite and therefore have the same variance. Signed Mann-Whitney U tests can be used again to decide whether two rank samples are different or not. Because the samples are less variable, the effect size is larger, which increases the statistical power of the test and makes it more likely for it to find a true difference where there is one. As a consequence, fewer assessors are needed overall.

# 4. CROWDSOURCING PREFERENCES

The use of a crowdsourcing platform seems very appropriate for our purposes. If the reasonable person assumption holds, we could use non experts to generate a ground truth like these. Because we no longer show the image of the staves, but offer an audio file instead, no music expertise is needed. We have also seen how to use preference judgments to generate partially ordered lists instead of having assessors rank all candidates at once. Therefore, the whole process can be divided into very small and simple tasks where one incipit has to be preferred over the other, which seems perfectly doable for any non expert. Also, the number of judgments between pairs of documents can be smaller, and given that we use non experts, the overall cost should be much less.

We are not aware of any work examining the feasibility of music related tasks with crowdsourcing platforms like Amazon Mechanical Turk (AMT), so we decided to use it for our experiments. AMT has been widely used before for tasks related to Text IR evaluation. HITs (each of the single tasks assigned to a worker) have traditionally used the English language, but it has been shown recently that workers can also work in other languages such as Spanish [18]. Other multimedia tasks, such as image tagging, have also been proved to be feasible with crowdsourcing [19].

## 4.1 HIT Design

The use of preference judgments is prone to have a very simple HIT design (see Figure 4). We asked workers to listen to the query or "original melody", as we called it. Then, they had to listen to what we called "variations", that is, the two incipits to compare. Next, they were asked what variation was more similar to the original melody, allowing 3 options: *A* is more similar, *B* is more similar, and they are either equally similar or dissimilar. We indicated them that if one melody was part of another one, they had to be considered equally similar, so as to comply with the original guidelines. As optional questions, they were asked for their musical background, if any, and for comments or suggestions to give us some feedback.

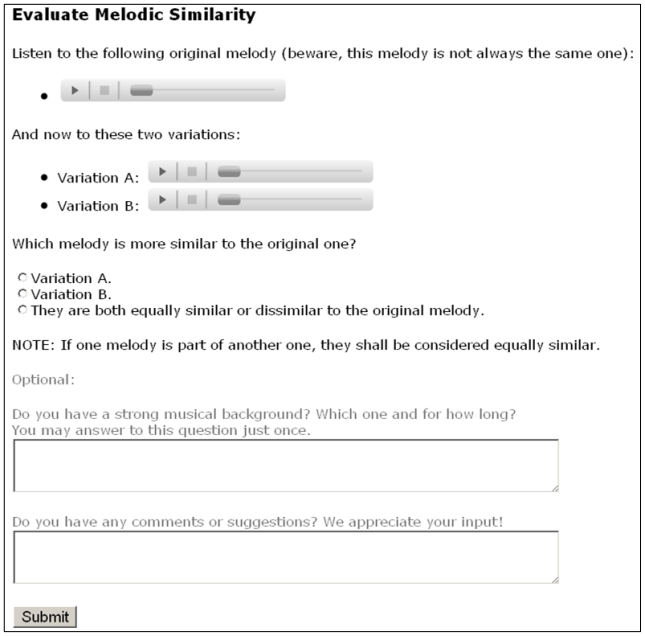


Figure 4. Example of HIT for music preference judgment.

The evaluation collection used in MIREX 2005 (*Eval05* for short) had about 550 short incipits in MIDI format, which we transformed to MP3 files as they are easier to play in a standard web browser. The average duration was 6 seconds, ranging from 1 to 57 seconds. However, many incipits start with rests (see query and incipit *C* in Figure 2), which would make workers lose a lot of time. Therefore, we trimmed the leading and tailing silence, which resulted in durations from 1 to 26 seconds, with an average of 4 seconds. With this cuts, the average time needed to listen to the 3 files in a HIT at least once was 13 seconds, ranging from 4 to 24 seconds. This decision agrees with the initial guidelines that were given to the experts, as two incipits should be considered equally relevant despite one of them having leading or tailing rests (i.e. one would be just part of the other). We uploaded all these trimmed MP3 files to a private web server, as well as the source of a very simple Flash player to play the queries and candidate incipits. Therefore, our HIT template was designed to display the MP3 players and stream the audio files from our server.

We created a batch of HITs for each of the iterations calculated with our methodology, and paid every answer with 2 cents of dollar (plus half a cent for Amazon's fee). After downloading the results and analyzing them, we calculated the next preference judgments to perform and uploaded a new batch to AMT,

Table 2. Summary of batches submitted to Mechanical Turk.

| Iteration | Pairs judged | Unique workers | Previous workers (%) | Median time per judgment (seconds) | Time to completion | Inter-agreement per pair | Cost (US $) |
|---|---|---|---|---|---|---|---|
| 1 | 107 | 32 | - | 26 | 13h 29m | 0.656 | 26.75 |
| 2 | 83 | 20 | 4 (20%) | 14 | 3h 2m | 0.822 | 20.75 |
| 3 | 51 | 15 | 11 (73%) | 19 | 3h | 0.72 | 12.75 |
| 4 | 19 | 17 | 10 (59%) | 30 | 10h 3m | 0.644 | 4.75 |
| 5 | 10 | 16 | 11 (69%) | 21 | 3h 29m | 0.663 | 2.5 |
| 6 | 5 | 12 | 8 (67%) | 24 | 2h 48m | 0.732 | 1.25 |
| 7 | 4 | 11 | 7 (64%) | 15.5 | 1h 21m | 0.569 | 1 |
| 8 | 2 | 11 | 4 (36%) | 24.5 | 28m | 0.506 | 0.5 |
| Total/Avg. | 281 | 79 | 55% | 21.75 | 37h 40m | 0.664 | 70.25 |

corresponding to the next iteration. Whenever all pairs of incipits within the same segment had been judged, we considered that segment closed, and whenever all segments were closed, the list was completed.

## 4.2 Threats to Validity

The initial order of candidates in the first iteration and the choice of the pivot element could clearly affect the results. If the pivot chosen were the query itself, most of the incipits would be judged less similar and go to the right segment, which would not provide much information. Therefore, we randomized the initial order of incipits in the first iteration. Moreover, we always chose the last incipit of a segment as the current pivot, and for the next iteration this element would be the first one of the equally-similar segment. See for example incipit *A* from iterations 3 to 4, in Table 1.

Workers could be tempted to stop listening to the original melody (i.e. the query) after a few HITs have been answered. Then, whenever the query changes as they start judging for another list, all answers given from that point on would be plainly useless. Even within the list of a single query, there will usually be several pivots, each of which will be compared with different incipits. Likewise, if the pivot is always kept as the first or second variation, workers could stop listening to them and just listen to the other variation, which would again make every answer useless after the pivot is changed when a new segment begins to be evaluated. See for example the 3rd iteration in Table 1, where both *A* and *E* are pivots. Again, we addressed this problem by randomizing the HITs: not all HITs from the same queries were presented together, and pivots were sometimes the variation *A* and some others the variation *B*. The HIT design explicitly warned the workers about this randomization anyway.

We also have to deal with carelessness of the workers. In first trials of our experiment we found that sometimes they judged some incipits as more similar to the query than the query itself, in cases where it was clearly different. We tried to alleviate this problem by accepting workers only with a 95% or higher rate of acceptance, and by using a sufficiently large number of answers per HIT. We chose to ask for 10 different workers per HIT, which we considered enough given that fewer answers are needed to begin with (see Section 3.2). This decision was also successfully taken by Alonso et al. when crowdsourcing relevance judgments for TREC documents [17]. We also found 2 workers that always gave the "Equal" answer in exactly 8 seconds. It seemed clear to us that we were dealing with some kind of robot, so we directly blocked them from our experiments and re-assigned their HITs.

# 5. RESULTS

The 11 lists in the *Eval05* collection account for a total of 119 candidate documents to judge for relevance, ranging from 4 to 23 documents per query. In order to complete the judgments, we had to submit 8 batches to Mechanical Turk, each corresponding to an iteration of the self-organizing algorithm. These batches were submitted from April 14th to April 17th, with some time taken between iterations to semi-automatically calculate what documents to compare for the next batch.

## 5.1 Summary of Submissions

The 119 candidate documents in the 11 lists sum up 740 pairs of candidates (i.e. the $O(n^2)$ case), We only needed to judge a total of 281 (38%) pairs of documents to completely organize the 11 lists, which account for a total of 2810 preference judgments by the workers (see Table 2). A total of 79 unique workers performed those judgments, with an average of 55% of the workers in an iteration having worked in previous ones. It took for them almost 22 seconds in median to submit the judgments, although this time reflects only how long they took to complete the assignment since they accepted it, rather than since it was displayed to them. Summing up the time to complete all iterations, the 2810 judgments took about a day and a half.

For all the 2810 judgments the total cost of generating the ground truths was about 70 dollars. The original lists needed 35 music experts for 2 hours, and during this time only 11 of them were able to work on the 11 queries. This accounts for roughly 70 hours of the time of one single expert, which is about twice as much as we needed using non-expert workers from Mechanical Turk.

## 5.2 Worker Feedback and Music Background

Out of the 79 unique workers, 23 gave us feedback. Four of them reported very positive comments about the HITs, one asked for more money and two reported problems loading one of the MP3 files for two HITs (the other workers did not report to have such problems for the same HITs).

Five workers explicitly indicated not to have any musical background, but fourteen did. Six of them had formal musical education, mainly in college and high school, while nine reported to have been practitioners for several years. Nine played an instrument, mainly piano, and six others performed in a choir.

## 5.3 Agreement between Workers and Experts

For each of the 281 HITs (i.e. pairs of candidates) we have 10 judgments made by workers. We calculated their inter-agreement score for each HIT as follows. Consider the 45 pairs of answers given for a single HIT, adding 2 points to the score if the two workers agreed (complete agreement); adding 1 point if one judged "Equal" and the other judged either document (partial agreement), and adding nothing if they judged both documents (no agreement). The perfect agreement would sum up 90 points, so we divided the score obtained by 90 to normalize from 0 (no agreement at all) to 1 (perfect agreement). Table 2 shows the mean agreement for every HIT judged in each iteration. We can

see that the agreement among workers is very high, ranging from 0.506 to 0.822, averaging to 0.664.

It is also interesting to measure the agreement between the workers of AMT and the music experts that ranked the original lists. We compared each of the resulting 281 preference judgments (aggregating the 10 corresponding answers of the workers, see Section 3.2) with the rankings given by the experts, inferring their preference judgments as well with signed Mann-Whitney U tests over the rankings they gave to each document. Table 3 shows the results.

Table 3. Agreement between workers (columns) and experts (rows) for aggregated judgments. Percentages are calculated over the row total.

| | | Workers | | |
|---|---|---|---|---|
| | | Less (56) | Equal (110) | Greater (115) |
| Experts | Less (91) | 38 (42%) | 37 (41%) | 16 (18%) |
| | Equal (55) | 11 (20%) | 31 (56%) | 13 (24%) |
| | Greater (135) | 7 (5%) | 42 (31%) | 86 (64%) |

Not surprisingly, the agreements are fairly high. There were 155 (55%) cases of complete agreement, 102 (36%) cases of partial agreement and only 23 (8%) cases of no agreement at all. Computing a global score as before, rewarding complete agreements with 2 points and partial agreements with 1 point, the agreement between workers and experts results in 0.735. These figures serve as empirical verification of the reasonable person assumption, indicating that the notion of musical similarity, though not formally formulated, appears to be common between experts and non experts.

Table 4. Agreement among single workers with no music background and experts. Percentages are calculated over the row total.

| | | Workers with no music background | | |
|---|---|---|---|---|
| | | Less (81) | Equal (97) | Greater (193) |
| Experts | Less (100) | 55 (55%) | 27 (27%) | 18 (18%) |
| | Equal (92) | 16 (17%) | 35 (38%) | 41 (45%) |
| | Greater (179) | 10 (6%) | 35 (20%) | 134 (75%) |

We also calculated the agreement between the original experts and the 5 workers that explicitly reported no music background, the 14 that reported to have some background, and the other 60 that did not answer. The workers that reported no background fully agreed with the experts 60% of the times, partially agreed 32% and did not agree in 8% of the judgments, which accounts for a total agreement of 0.764 (see Table 4).

Table 5. Agreement among single workers with music background and experts. Percentages are calculated over the row total.

| | | Workers with music background | | |
|---|---|---|---|---|
| | | Less (70) | Equal (80) | Greater (116) |
| Experts | Less (70) | 45 (64%) | 18 (26%) | 7 (10%) |
| | Equal (67) | 15 (22%) | 32 (48%) | 20 (30%) |
| | Greater (129) | 10 (8%) | 30 (23%) | 89 (69%) |

When considering the workers that reported some background, the agreement rises to 0.78, having 62% cases of total agreement with the experts, 31% of partial agreement and 6% of no agreement at all (see Table 5).

Table 6. Agreement among single workers with unknown music background and experts. Percentages are calculated over the row total.

| | | Workers with unknown background | | |
|---|---|---|---|---|
| | | Less (426) | Equal (1230) | Greater (517) |
| Experts | Less (390) | 218 (56%) | 152 (39%) | 20 (5%) |
| | Equal (941) | 127 (13%) | 707 (75%) | 107 (11%) |
| | Greater (842) | 81 (10%) | 371 (44%) | 390 (46%) |

The 60 workers that did not report anything about musical background had an agreement score with the experts of 0.777, with 60% of total agreement, 34% of answers with partial agreement and 5% of no agreement (see Table 6). All these results support again the reasonable person assumption, as very similar agreement scores can be found not only between groups of workers, but also between single workers with and without music background. As a consequence, they also support the use of crowdsourcing platforms to gather music relevance judgments.

## 5.4 Comparison with the Original Lists

Given the high agreement scores obtained by the workers of Mechanical Turk, one would expect to obtain lists very similar to the original ones generated with experts. To measure the similarity, we considered the original lists as ground truths and the crowdsourced lists as if they were the results of a system, evaluating the ADR score that would be obtained in a real evaluation [20]. Moreover, we considered the original lists as aggregated with the *Any-1* function we proposed in [15], as the resulting lists proved to be the most consistent. Finally, and to compare lists in both directions, we considered the crowdsourced lists as ground truths and the original ones as results.

There is one important detail to note, though: both the ground truth list and the results list have groups of relevance, but the latter will be considered as a fully ranked list (i.e. a sequence without groups) when computing the ADR score. For example, consider the list $L1=\langle(A, B, C), (D, E)\rangle$ is taken as ground truth and the list $L2=\langle(A, B), (D, E, C)\rangle$ as results. When evaluating $L2$, it would be considered as $\langle A, B, D, E, C\rangle$, which results in an ADR score of 0.933 because at position 3 the document retrieved is $D$, when $C$ was expected. However, $C$ and $D$ were judged as equally relevant. These cases depend directly on the order the documents were randomly arranged at the beginning. If the results list were $L3=\langle(A, B), (C, D, E)\rangle$, which is equivalent to $L2$, the ADR score would be 1. To account for the random effect of the initial arrangement, we generated 1000 random versions of the lists obtained with Mechanical Turk, by randomly permuting the order of documents within the same group. The results of the comparisons appear in Table 7, with the minimum, mean and maximum ADR scores obtained for the 1000 random sets of equivalent lists.

Table 7. Comparison between the original lists and the lists crowdsourced, in terms of average ADR score. Columns represent lists acting as ground truth, rows for lists acting as results. The numbers between square brackets indicate the minimum and maximum scores.

| | | Ground truth | | |
|---|---|---|---|---|
| | | All-2 | Any-1 | MTurk |
| Results | All-2 | 1 | 0.872 [0.830-0.927] | 0.824 [0.785-0.872] |
| | Any-1 | 1 | 1 | 0.850 [0.828-0.873] |
| | MTurk | 0.943 [0.915-0.977] | 0.840 [0.812-0.881] | 1 |

When compared to the original lists generated by Typke et al. (i.e. *All-2*), the crowdsourced lists performed exceptionally well, with very high ADR scores across the 11 queries, between 0.915 and 0.977. As expected, the *Any-1* lists reduce the scores because they are more restrictive than the *All-2* alternatives, although the averages are still high over 0.812. When using the crowdsourced lists as ground truth, the average across the 11 queries is still high. The *Any-1* lists would obtain a higher score than the *All-2*, showing that the crowdsourced lists are also more restrictive than

the original ones. These results confirm that the lists generated with Mechanical Turk workers are, in fact, very similar to the ones generated by experts, as already anticipated by the high agreement scores.

## 5.5 Judgments Consistency

We examined the crowdsourced lists to check whether inconsistent results like the ones described in Section 2.2 did still appear or not, and in several cases they did not. For example, the first two incipits in Figure 3 ended up in the same group of relevance, at the top of the list, as did the first three incipits in Figure 2. Other lists, like the one for query 600.054.278-1.1.1, also showed such correct variations.

## 5.6 MIREX 2005 Results Revisited

The question is whether those small variations in the lists would affect the evaluation of real systems or not [1]. We re-evaluated the 7 systems that participated in the MIREX 2005 Symbolic Melodic Similarity task with the crowdsourced ground truth lists. In addition, we also re-evaluated and compared the *Splines* method we proposed in [8] (see Table 8). Again, we compare also with the Any-1 version of the original lists.

Table 8. ADR results of the systems that participated in MIREX 2005 with the original and crowdsourced lists. 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. * for significant difference at the 0.05 level and ** at the 0.01 level.

| | Splines | GAM | O | US | TWV | L(P3) | L(DP) | FM |
|---|---|---|---|---|---|---|---|---|
| All-2 | 0.71 | 0.66 | 0.65 | 0.642 | 0.571 | 0.558 | 0.543 | 0.518 |
| Any-1 | **0.646*** | 0.583 | 0.593* | 0.594* | 0.556 | 0.515 | 0.494* | 0.483* |
| MTurk | **0.6**** | 0.574* | 0.572* | 0.546** | 0.517* | 0.51* | 0.467* | 0.462* |

As with the *Any-1* version, the crowdsourced lists seem to be more restrictive than the original *All-2*. All systems get reductions in average ADR score between 9% and 15%, and all these differences were statistically significant according to 1-tailed paired Mann-Whitney U tests. The important result is, however, that the ranking of systems is exactly the same as with the original lists. That is, the crowdsourced lists ranked the 7 systems in terms of average ADR score as the original lists did. This, again, supports the use of our methodology for evaluation of music similarity tasks.

# 6. CONCLUSIONS AND FUTURE WORK

Ground truths based on partially ordered lists represented a big leap towards the scientific evaluation of music similarity tasks. They have been widely accepted by the community, but their use in the MIREX evaluations was interrupted mainly because of their expensiveness in terms of man-power and need for music experts.

In this paper we have proposed a modification of the methodology followed to generate these lists, and we have implemented it with Amazon Mechanical Turk to gather music relevance judgments, showing that crowdsourcing platforms are viable alternatives for the evaluation of music retrieval systems. This allowed us to review the reasonable person assumption, which may lead to more affordable and large-scale evaluations without the need for music experts. We provided empirical evidence supporting it, showing high agreement scores between workers and experts.

Our methodology has several advantages. Fewer assessors are needed to judge, so more queries can be evaluated with the same man-power. Preference judgments are easier to perform, and the number of actual judgments made by the assessors is far less, because they do not need to assess where between several candidates should a new incipit be placed. Allowing judgments of the form "A and B are equally similar", we avoid inconsistency problems where incipits equal for all purposes were judged differently. Offering the incipits as audio files instead of images, also helps in this matter, and it seems to avoid the necessity of having experts.

Further research should focus on the sorting algorithm used to organize incipits. The choice of good pivots is essential, and more empirical research should focus on the nature of music similarity to assess whether it is transitive or even symmetrical. That is, if $A$ is preferred over $B$ and $B$ is preferred over $C$, will $A$ be preferred over $C$? And if $A$ is preferred over $B$ for query $C$, will $C$ be preferred over $B$ for query $A$? So far it has been assumed that these properties hold, but such assumption should be subject of further experimental studies. In case it were valid, more work in the line of Carterette et al. should be carried out to minimize the number of judgments needed to sort all candidates and find true differences in the performance of retrieval systems [21].

# 7. ACKNOWLEDGMENTS

We would like to thank Omar Alonso for his thoughtful comments on Mechanical Turk and the paper itself. We also thank Carlos Gómez, Rainer Typke, and the IMIRSEL group, especially Stephen Downie and Mert Bay, for providing us with the MIREX 2005 evaluation data.

# 8. REFERENCES

[1] Voorhees, E.M. The Philosophy of Information Retrieval Evaluation. *Workshop of the Cross-Language Evaluation Forum* (2002), 355-370.

[2] Downie, J.S., Ehmann, A.F., et al. The Music Information Retrieval Evaluation eXchange: Some Observations and Insights. *Advances in Music Information Retrieval*. W.R. Zbigniew and A.A. Wieczorkowska. Springer. 2010. 93-115.

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

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

[5] Typke, R., den Hoed, M., et al. A Ground Truth for Half a Million Musical Incipits. *Journal of Digital Information Management*. 3, 1 (2005), 34-39.

[6] Jones, M.C., Downie, J.S., et al. Human Similarity Judgments: Implications for the Design of Formal Evaluations. *International Conference on Music Information Retrieval* (2007), 539-542.

[7] Downie, J.S., West, K., et al. The 2005 Music Information Retrieval Evaluation Exchange (MIREX 2005): Preliminary Overview. *International Conference on Music Information Retrieval* (2005), 320-323.

[8] Urbano, J., Lloréns, J., et al. Using the Shape of Music to Compute the Similarity between Symbolic Musical Pieces. *International Symposium on Computer Music Modeling and Retrieval* (2010), 385-396.

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

[10] Hanna, P., Ferraro, P., et al. On Optimizing the Editing Algorithms for Evaluating Similarity Between Monophonic Musical Sequences. *Journal of New Music Research*. 36, 4 (2007), 267-279.

[11] Grachten, M., Arcos, J., et al. A Case Based Approach to Expressivity-Aware Tempo Transformation. *Machine Learning*. 65, 2 (2006), 411-437.

[12] Carterette, B., Bennett, P.N., et al. Here or There: Preference Judgments for Relevance. *European Conference on Information Retrieval* (2008), 16-27.

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

[14] Mann, H.B. and Whitney, D.R. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*. 18, 1 (1947), 50-60.

[15] Urbano, J., Marrero, M., et al. Improving the Generation of Ground Truths based on Partially Ordered Lists. *International Society for Music Information Retrieval Conference* (2010).

[16] Alonso, O., Rose, D.E., et al. Crowdsourcing for Relevance Evaluation. *ACM SIGIR Forum*.

[17] Alonso, O. and Mizzaro, S. Can We Get Rid of TREC assessors? Using Mechanical Turk for Relevance Assessment. *SIGIR Workshop on the Future of IR Evaluation* (2009), 15-16.

[18] Alonso, O. and Baeza-Yates, R. An Analysis of Crowdsourcing Relevance Assessments in Spanish. *Spanish Conference on Information Retrieval* (2010).

[19] Nowak, S. and Rüger, S. How Reliable are Annotations via Crowdsourcing? A Study about Inter-annotator Agreement for Multi-label Image Annotation. *International ACM Conference on Multimedia Information Retrieval* (2010), 557-566.

[20] Typke, R., Veltkamp, R.C., et al. A Measure for Evaluating Retrieval Techniques based on Partially Ordered Ground Truth Lists. *IEEE International Conference on Multimedia and Expo* (2006), 1793-1796.

[21] Carterette, B. and Allan, J. Incremental Test Collections. *ACM International Conference on Information and Knowledge Management* (2005), 680-687.