A Ground Truth For Half A Million Musical Incipits

Rainer Typke, Marc den Hoed, Justin de Nooijer, Frans Wiering, Remco C. Veltkamp
Utrecht University, ICS
Padualaan 14 3584 CH Utrecht, The Netherlands
{rainer.typke|mhoed|jnooijer|frans.wiering|remco.veltkamp}@cs.uu.nl

ABSTRACT: *Musical incipits are short extracts of scores, taken from the beginning. The RISM A/II collection contains about half a million of them. This large collection size makes a ground truth very interesting for the development of music retrieval methods, but at the same time makes it very difficult to establish one. Human experts cannot be expected to sift through half a million melodies to find the best matches for a given query. For 11 queries, we filtered the collection so that about 50 candidates per query were left, which we then presented to 35 human experts for a final ranking. We present our filtering methods, the experiment design, and the resulting ground truth.To obtain ground truths, we ordered the incipits by the median ranks assigned to them by the human experts. For every incipit, we used the Wilcoxon rank sum test to compare the list of ranks assigned to it with the lists of ranks assigned to its predecessors. As a result, we know which rank differences are statistically significant, which gives us groups of incipits whose correct ranking we know. This ground truth can be used for evaluating music information retrieval systems. A good retrieval system should order the incipits in a way that the order of the groups we identified is not violated, and it should include all high-ranking melodies that we found. It might, however, find additional good matches since our filtering process is not guaranteed to be perfect.*

Categories and Subject Descriptors

H.3.1 **[Information Storage and Retrieval]:** Content Analysis and Indexing—*Symbolic music representation;* J.5 [Arts **and Humanities]:** Performing arts (music)

General Terms

Algorithms, Measurement, Performance

Keywords: Music information retrieval, RISM, ground truth, filtering

Reviewed and accepted 15 Dec. 2004

1. INTRODUCTION

For evaluating the performance of a music retrieval system, one needs a ground truth for its data collection and some given queries. The music retrieval systems we have in mind serve the information need for music that is melodically similar to a given query. Such search engines can be useful not only for retrieving sheet music or recordings in libraries or stores, but also for investigating how composers have influenced one another, or for raising or resolving copyright disputes.

The RISM A/II collection [8] contains 476,600 incipits, short excerpts of notated music from the beginnings of manuscripts in libraries, archives, cloisters, schools, and private collections worldwide. This collection is useful for content-based music retrieval because of its size and the fact that it contains real music written by human composers. A music retrieval system that works well with this collection should also perform well for real-world applications in general. Our ground truth can serve as a benchmark for deciding how well a music retrieval system works with the RISM A/II collection.

In TREC [12], relevance assessments are mostly binary ("relevant" or "not relevant"). Only in more recent TREC web tracks such as at TREC-9 [3], this was extended to ternary ("irrelevant"/"relevant"/ "highly relevant"). Studies such as *Selfridge-Field* [9] show that melodic similarity is continuous. Local melodic changes such as lengthening a note or moving it up or down a step are usually not perceived as changing the identity of a melody, and by applying more and more changes, the perceived relationship to the original becomes only gradually weaker. Also, melodies are generally quite resistant to the insertion of all sorts of ornamentation. Because of the continuity of melodic similarity, there are no sensible criteria for assigning one out of a few distinct degrees of relevance to a melody, so any relevance assessment with a given scale length seems inappropriate. Instead, we asked human experts to rank all incipits where they saw any similarity to the query. Our ground truth therefore does not consist of sets of highly relevant, relevant and irrelevant documents, but of ranking lists of documents.

A valid way of establishing such a ground truth would be to ask a number of human experts to look at all possible matches for a given query (carefully making sure that they stay concentrated long enough) and order them by similarity. Since we cannot expect our human experts to sift through half a million melodies, we needed to filter out incipits of which we can be reasonably sure that they do not resemble the query.

This paper describes how we filtered the collection for a list of 11 queries, how we employed 35 human experts for ranking the remaining incipits by their similarity to the queries, and how we established a ground truth as a result. For each of the 11 queries, we built a list of groups of incip-its that is ordered by similarity to the query. The *Wilcoxon* rank sum test [13] gives us a measure for the statistical significance of the differences between the groups of incipits.

Related Work. We are not aware of any previous efforts to establish a ground truth for the RISM A/II collection or a similarly large collection of musical incipits, themes, or scores. However, there is high interest in establishing a systematic, TREC-like paradigm for the music information retrieval research community [2], so that having a ground truth could be very helpful.

2. FILTERING MELODIES

To be able to exclude incipits that are very different from our selected queries, we calculated some features for every incipit in the database. Filtering could then easily be done by issuing SQL statements with selections based on those features.

•**Pitch range:** the interval between the highest and lowest note in the incipit.

•**Duration ratio:** the duration of the shortest note (not rest), divided by the duration of the longest note (not rest). The result is a number in the interval (0,1], where 1 means that all notes have the same duration, while a very small number means a very high contrast in durations.

•**Maximum interval:** the largest interval between subsequent notes. Rests are ignored.

•**Editing distance between gross contours:** the editing distance between two character strings is the sum of the costs of the cheapest possible combination of character insertion, deletion, and replacement operations that transform one string into the other. We determined the gross contour as a string of characters from the alphabet U ("up"), D ("down"), and R.("repeat") and calculated the distance to every query for each incipit in the database, using the editing distance described by *Prechelt* and *Typke* in [7]. They had optimized the costs for the insertion, deletion, and replacement operations for gross contour strings such that the resulting similarity measure corresponds well with human perception.

- **Editing distance between rhythm strings:** we also represented the incipits as rhythm strings with one haracter from a three-character alphabet for each pair of subsequent notes: longer, shorter, and same duration.
- **Interval histogram:** the number of occurrences for each interval between subsequent notes, normalized with the total number of intervals. With this feature, we can base selections on things like "incipits with many thirds".

- **Interval strings:** one string of diatonic intervals and one string of chromatic intervals for every incipit. This makes it possible to select incipits that contain a certain sequence of intervals.
- **Motive repetitions:** in order to be able to select things like "all incipits with at least three repeated notes in two different places", we collected sequences of intervals that were repeated at least once, along with their number of occurrences, for every incipit.

We used different filtering steps and features for every query since every query has its own characteristic features. Every filtering step had the aim of reducing the number of candidates for matches for a given query by excluding incipits with features that make them very different from the query. As long as this holds for every filtering step, different people should arrive at similar candidate lists even if they apply different filtering steps. However, they need to have similar notions of melodic dissimilarity.

For example, we used the following filtering steps for the "White Cockade" incipit whose ground truth is shown in Table 3:

- Exclude incipits whose pitch range is less than an octave or greater than a minor tenth. This excluded 78 % of the incipits in the database.
- Exclude incipits whose maximum interval between subsequent notes is less than a minor sixth or greater than a diminished seventh. This excluded 79 % of the remaining incipits.
- Exclude incipits with a duration ratio greater than 0.51, i. e. incipits where all notes have quite similar durations. This excluded a further 4 % of incipits.
- Exclude incipits that do not contain at least one of the two interval sequences "fifth up, third down, unison, sixth up" or "third up, unison, unison, sixth up". This left us with 88 incipits.

Because of the dangers of filtering too strictly and thereby accidentally excluding incipits that are similar to the query, we stopped the filtering process once the number of remaining incipits had fallen below 300. To arrive at the desired number of about 50 candidates, we manually excluded remaining incipits that were very different from the query.

As an additional measure to limit the error introduced by accidentally filtering out similar incipits, we used our prototype of a search engine based on transportation distances (described in [10] and [11]) as well as two algorithms from [5] for finding incipits that are similar to the query. The latter two algorithms, called P2 and P3 by their authors, find incipits containing transpositions of the query where many onset time/pitch combinations match, and incipits containing transpositions of the query with maximum common duration with matching pitch. From these search results, we included candidates that we considered similar although they had been filtered out. Also, we used the metadata in the RISM A/II collection. For example, for "Roslin Castle" (see Table 1), we made sure that every incipit whose title contains the word "Roslin" was included. With these methods, we found between 0 and about 8 additional candidates for each query, with an average of about 4.

Once we had filtered out all incipits that are not similar to the query, we also removed incipits that were either identical to other incipits or parts of other incipits. Including identical incipits multiple times in the candidate list would have amounted to asking our experts the same question multiple times, and we wanted to put their time to a more productive use. As a result, only 6 versions of "Roslin Castle" occur in our ground truth in Table 1 although we list 16 known occurrences of this melody in our paper about using transportation distances for measuring melodic similarity [10], for which we used the same 2002 version of the RISM database.

3. EXPERIMENT DESIGN

3.1 Notated music, MIDI files

Our goal was to establish a ground truth for the incipits that are contained in the RISM A/II collection. These incipits can be exported from the database in the "Plaine & Easie" format [4] and then rendered in common music notation. In order to prevent differences in the rendition of the notated music from having an impact on the ground truth, we used the software that is included with the RISM A/II database [8] for rendering the music notation bitmaps and took screen shots of the results. Only in cases where the RISM software fails to show the whole incipit because it is too long for fitting on the screen, we rendered the notated music ourselves by converting the Plaine & Easie data into the Lilypond format. In addition to the notated music, we also provided MIDI files generated from the Plaine & Easie data as an illustration of the incipits. However, we told the experiment subjects that the definitive source for similarity judgements is the notated music, and that the MIDI files only serve as an illustration. The metadata from the RISM A/II collection (composer, work title, title of the movement, instrumentation etc.) was not shown to the human experts. They only saw the notated music of the incipits and could listen to a MIDI rendition, as can be seen in Figure 1.

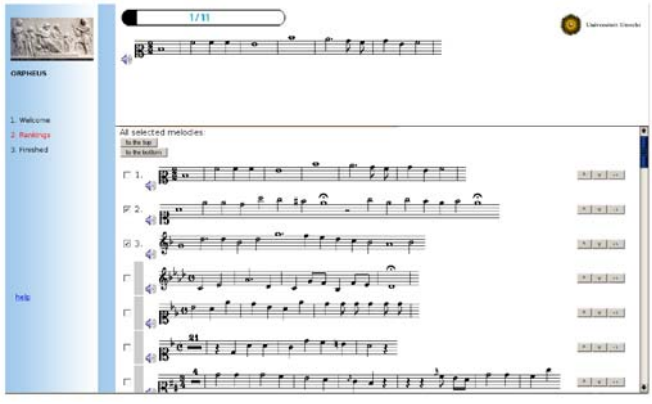

Figure 1: The user interface for the experiment. MIDI fiels are provided for listening to incipits. In the bottom half of the screen, the subjects can change the order of the candidate incipits, while the query always remains visible at the top.

3.2 Experts

Miillensiefen et al. point out [6] that music experts tend to have stable similarity judgements, in other words, do not change their mind on what is melodically similar when asked to perform the same judgements a few weeks apart. Subjects with stable similarity judgements, in turn, seem to have the same notion of melodic similarity. In order to establish a meaningful ground truth, we therefore tried to recruit music experts as our experimental subjects. We asked people who either have completed a degree in a music-related field such as musicology or performance, who were still studying music theory, or who attended the International Conference on Music Information Retrieval Graduate School in Barcelona 2004 to participate in our experiment.

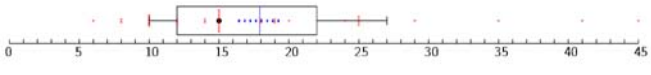

Figure 2 The experience of our experts, in years. The box extends from the first to the third quartile. The whiskers mark the bottom and top 10 percent. Every data point is shown as a little dot. The me-dian is marked with a fat dot, the mean is shown as a vertical line. The dashed horizontal line around the mean marks one standard deviation below and above the mean.

Lilypond (see http://lilypond.org) is an open source music typesetter.

All of our experts play at least one instrument or sing, most play several instruments. See Figure 2 for a box-and-whisker plot showing their musical experience in years.

3.3 Instructions, tasks

We asked the subjects to rank all candidates that resemble the query by their melodic similarity to the query. Candidates that seemed completely different from the query could be left unranked. The ranking was to be done by reordering the given candidates such that the candidate most similar to the query was at the top, followed by less and less similar candidates, and finally a number of candidates without any assigned ranks that did not resemble the query at all. By asking people to reorder a list instead of picking a rank from a scale, we avoided suggesting how long the ranked list should be, and we also made it easy for the experts to judge whether they ranked all candidates correctly by looking at a local ordering only. It was sufficient to ensure that for each pair of consecutive candidates in the ranked part of their reordered list, the incipit that was ranked higher was more similar to the query than the other incipit of the pair. We asked the experts to regard transpositions of a melody as identical, as well as melodies that are notated slightly differently, but in a way that does not affect the way they sound. For example, melodies that are notated with different clefs, but are otherwise the same, should not be viewed as different. In cases where two incipits were taken from similar pieces, but covered different amounts of musical material, we asked the subjects to only consider the common parts of the two incipits for the comparison.

We asked every subject for about 2 hours of his time. If somebody did not manage to work on all queries within that time, we ended the experiment anyway. Therefore, not all queries were judged by all experts. For example, only 25 out of all 35 experts ranked the top candidate shown in Table 3. We asked the experts to work carefully, even if that meant that they could not finish all 11 queries within two hours.

3.4 Threats to the validity of results

- **Filtering errors.** It is possible that we filtered out some incipits although they are similar to the query. Our ground truth, therefore, could be incomplete. However, this does not threaten the validity of the ranking of those candidates that we did include.

- **Sequence effects.** The initial order of candidates as well as the order in which queries are presented to the experts could have an impact on the results. Experts could be tempted to leave the order similar to the initial order, and they get more tired and at the same time more skilled at using our interface over the course of the experiment. We addressed these problems by randomizing the order of queries for every participant, and we also put the candidates in a new random order whenever a new query appeared on the screen.
- **Carelessness of experts.** For some queries, such as the "White Cockade" shown in Table 3, we included the query itself among the candidates. Careful experts should put it at the very top of the ranked list. Not everybody did, but enough of the experts were careful. This query was recognized as most similar to itself with high statistical significance: the Wilcoxon rank sum test, which we used as described in Section 4.1, shows that for every candidate that was not identical to the query, the probability of the null hypothesis is < 0.0001123.

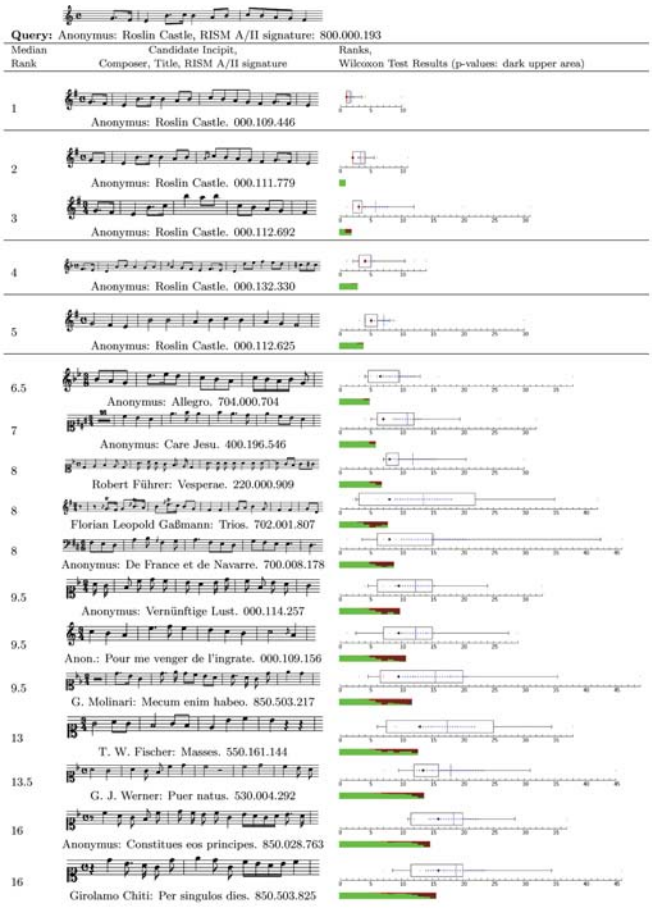

Table 1: Ground truth for "Roslin Castle". Table contents: median rank, incipit with title and RISM A/II signature, box-and-whisker plot showing the ranks assigned by our subjects, and a bar composed of squares visualizing the Wilcoxon rank sum test results for every preceding incipit. For details see Section 4.2.

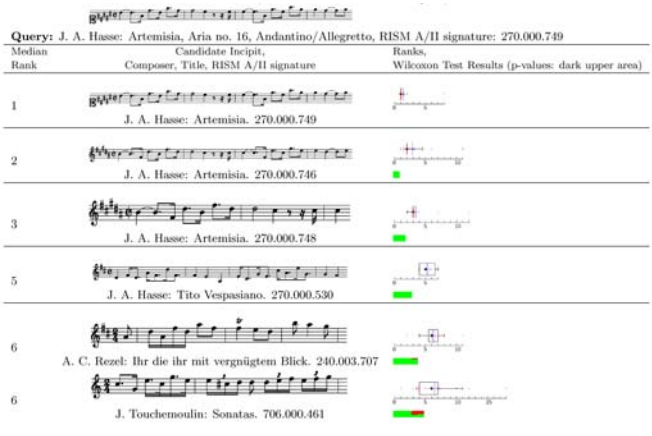

Table 2 Ground truth for an Aria by Johann Adolf Hasse as query. For details see Section 4.3.

4. RESULTS

4.1 Evaluation methodology

For every query, the subjects were asked to choose and rank as many of the candidates for matches as they thought had some similarity to the query. Those candidates without any similarity could be left unranked. This gives us a list of ranks for every candidate. These lists tend to be longer for the candidates that are more similar to the query. For example, 20 subjects ranked the first candidate in Table 1. Not all of them assigned rank 1 to this incipit, but the median is still 1. Only 6 people, however, ranked the incipit whose median rank is 13.5.

We did not exclude any expert's opinion from our final ground truth. Four data sets were left out of our evaluation because they had resulted from aborted experiments which were restarted from scratch by the same experts after encountering technical problems with browsers. If we had included those four data sets, we would have partially counted four experts' opinions twice.

To obtain a ground truth, we ordered the candidates by their median rank and then by their mean rank. In addition, for every ranked candidate, we applied the Wilcoxon rank sum test to the ranked candidate and every incipit that was ranked higher. The Wilcoxon rank sum test, given two samples, determines the probability of the null hypothesis (p-value), that is, the hypothesis that the median values are the same for the whole two populations from which the samples were taken. We used it to find out how likely it is that the differences in ranking observed by us are only caused by our choice of 35 people out of the whole population of music experts. A low p-value resulting from the Wilcoxon test means that the difference in medians is probably not a coincidence. A large p-value does not necessarily mean that the medians are the same, but just that we do not have compelling evidence for them being different.

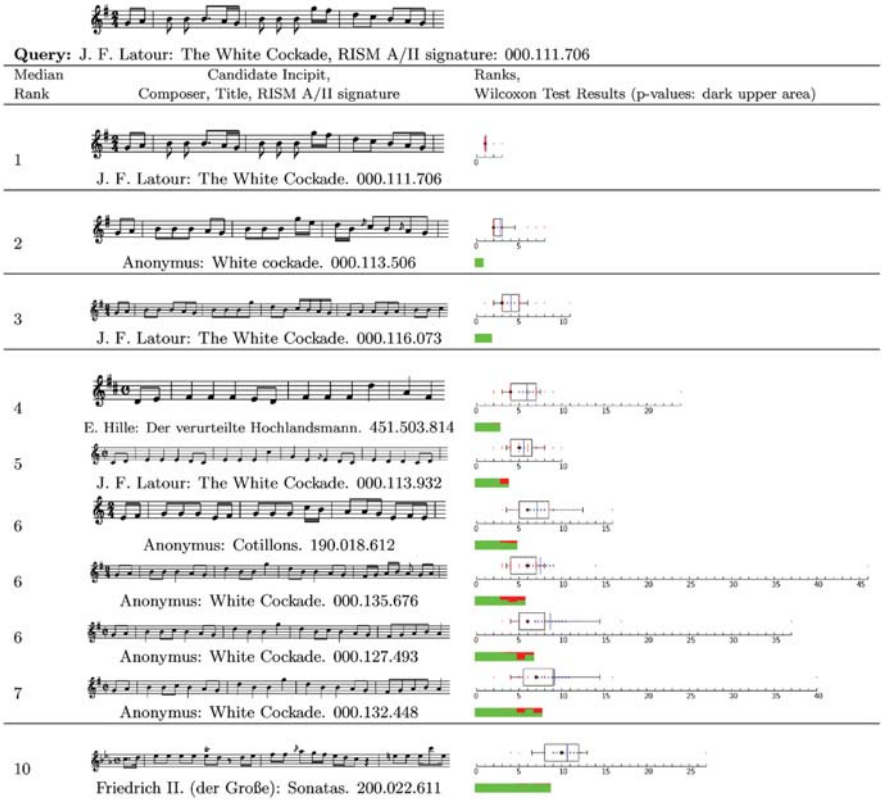

Table 3 Ground truth for "The White Cockade" by J. F. Latour as query. Only one out of the top nine pieces, "Cotillons", is not the same piece as the query. As one should expect, the Wilcoxon rank sum test results warrant a separator between the first nine incipits and the tenth, which is from a different piece and at the same time clearly different from the preceding incipits. For details see Section 4.3.

4.2 The resulting ground truth tables

We visualize the ranks assigned to each candidate with a box-and-whisker plot. The box extends from the first to the third quartile. The whiskers mark the bottom and top 10 percent. Every data point is shown as a little dot. The median is marked with a fat dot, the mean is shown as a vertical line. The dashed horizontal line around the mean marks one standard deviation below and above the mean. The numbers on the scales reflect ranks.

Below every box-and-whisker plot except for the first one, we visualize the Wilcoxon rank sum test results with a horizontal bar that is composed of one square for every incipit which is ranked higher than the current one. Each of these squares has a dark upper area with a fine pattern of horizontal stripes and a lower area with a lighter, solid colour. The size of the dark upper area reflects the p-value (see Section 4.1 for an explanation of what the p-value means).

For incipits where every square in the Wilcoxon visualization is almost entirely light-coloured, we can be reasonably sure that all preceding incipits should indeed be ranked higher. Wherever this is the case, we draw a horizontal line immediately above the incipit. For Table 1, we set the threshold for the maximum probability for the null hypothesis at 0.25. In other words, we draw a horizontal line above every incipit where the p-value is less than 0.25 for every single incipit that appears higher in the list. Most actual probabilities are much lower than that, as the visualization of the Wilcoxon tests in Table 1 shows.

For "Roslin Castle" (Table 1), we find five clearly distinguishable groups that way. The incipit with median rank 1 is generally considered the most similar incipit to the query. For the incipit with median rank 2, the Wilcoxon test shows that the probability for the null hypothesis is p=0.00006722. Therefore, we consider the difference in median values statistically significant and separate the second incipit from the first with a horizontal line. For the incipit with median rank 3, the difference in medians is statistically significant for the comparison with the first incipit (p=0.0002765), but not for the comparison with the second incipit (p=0.6363). This is reflected in the Wilcoxon visualization bar, which consists of one almost entirely light-coloured square on the left for the comparison of the third incipit with the first one, and one mostly dark square on the right for the comparison of the third incipit with the second one. Since there is no statistically significant difference between the second and third incipit, we group them together and do not separate them with a horizontal line. The third group consists of the incipit with median rank 4. The highest of its three p-values resulting from the Wilcoxon tests for its three

predecessors is 0.07633. The fourth group again consists of one single incipit, while for all other incipits, there are no statistically significant differences in median ranks. Either we did not have enough subjects who ranked these incipits, or people simply do not consider the dissimilarities between the remaining incipits and the query significantly different.

The tables shown in this paper are not complete. We cut them off a bit after the last detected border between clearly distinguishable groups because the ranking becomes less reliable and therefore less interesting towards the bottom of the tables. The complete data are available online at http://give-lab.cs. mi.nl/orpheus.

4.3 Musical properties of the identified groups

In Table 1 ("Roslin Castle"), the candidate with the highest rank looks as if it would begin with the query and therefore should, according to our instructions, be regarded as identical to the query since only the common part should be considered. If one looks more closely, however, one notices that the key signatures are different. The resulting differences in two notes, however, are not big enough for our experts to consider it very different from the query. The incipits with median ranks 2 and 3 constitute the second group. Both begin differently from the query - the incipit with median rank 2 has slight differences in rhythm at the beginning and two grace notes added in the second measure, while the incipit with median rank 3 has its second measure transposed by an octave. Otherwise their beginnings are the same as the query. Our experts agree that these incipits are both less similar than the incipit with median rank 1, but they disagree on whether the transposition of a measure by an octave or the modified rhythm and added grace notes should be seen as a greater dissimilarity. Because of this, these two incipits are combined into one group. The experts agree that the incipit with median rank 4 is significantly different from those preceding it. This is justified by a minor difference in rhythm in measure 1 and a major one in measure two - the first note is just a grace note, so there is no group of four descending eighth notes in that measure as in all preceding incipits. The incipit with median rank 5 is again significantly different. The rhythm is changed in several ways, leading to a very noticeable difference in measure 3. The third note in this measure corresponds to the first note in measure 2 of all preceding incipits. Because here this note is not at the beginning of the measure, it is much less emphasized, which changes the character of the melody. The last statistically significant border between groups is that between the incipits with median ranks 5 and 6.5. The latter is the first incipit of a different piece, and it also has a different time signature, so we would expect a border between groups here. Another border could be expected between the second and third incipit with median rank 9.5 because the interval sequence at the beginning changes noticeably here. However, at this point in the ranked list, we do not have enough votes per incipit for finding a statistically significant difference.

The top three candidates for J. A. Hasse's "Artemisia" (see Table 2) are very similar. The incipit with median rank 1 is identical to the query, that with median rank 2 is written with a different clef, but otherwise identical to the query, and the incipit with median rank 3 is identical to the first half of the query. Although they were instructed to disregard such differences, our experts still agreed that simply notating the query differently changes it less than omitting the second half, leading to statistically significant differences in the rankings. The incipit with median rank 5 is a somewhat similar melody by the same composer, but from a different work ("Tito Vespasiano"). It is similar to the query because it also begins with a motive built from notes from a triad (the tonic) and with a dotted rhythm, followed by a variation of the same motive that is based on another triad, the dominant. However, the rhythm is inverted in "Tito Vespasiano". All other candidates are ranked lower, but without further statistically significant differences in rank. The next candidates also begin with a triad that is split up in a similar way, sometimes also with a dotted rhythm, but not followed by a similar motive based on the dominant. Table 3 shows that our experts correctly recognized that the incipit that is most similar to the query is the query itself. The incipit with median rank 2 has some minor differences in rhythm, some added grace notes, and two different eighth notes instead of one quarter note in the last measure. Surprisingly enough, the incipit with median rank 3, about which we could say pretty much the same as about that with median rank 2, is ranked lower, and this difference is statistically significant. The remaining incipits of "The White Cockade" or a German version of the same song, "Der verurteilte Hochlandsmann", are all ranked lower, but without statistically significant differences in their median ranks. In that group, there is one incipit from a different piece (Anonymus: "Cotillons") that is melodically very similar. There is again a noticeable border between the incipit with median rank 7 and that with median rank 10. The latter is a sonata by Friedrich II, where only the first five notes are similar.

5. CONCLUSIONS

Our ground truth for 11 incipits from the RISM A/II collection can serve as a basis for a benchmark for evaluating music information retrieval systems. The complete ground truth, along with the sets of queries, candidates, and experimental results, can be found at http://give-lab.cs.uu.nl/orpheus. We encourage music retrieval researchers to apply their favourite methods to the RISM A/II collection and compare their results to our ground truth.

In order to use standard evaluation procedures using measures such as precision and recall, one could simply define a threshold and call every match in our ground truth that is ranked higher than the threshold "relevant" and the rest of the matches "irrelevant". Any report on the basis of this ground truth must then mention this threshold value. However, to take advantage of the ranking we established, one could also look at the precise order in which matches are returned. A good music information retrieval system should not only retrieve the matches that are ranked highly in our ground truth, but also return them in an order that does not violate the order of the groups of incipits we found by using the Wilcoxon rank sum test. The correct order of our incipits within groups, however, should be regarded as unknown. For example, for Table 1, we do not know whether the incipit with median rank 2 or that with median rank 3 should really be ranked higher, but we do know that both of them should be ranked lower than the incipit with median rank 1 and higher than any other incipit in our list of candidates. It is also important to take into consideration that we excluded candidates that are identical to other candidates or transpositions of other candidates, or candidates that are identical to the beginning of other candidates. Any additional such candidates should be regarded as belonging to the same group as their identical counterparts, transpositions, or the incipits with whose beginnings they are identical. Finally, for two reasons, there is the possibility that a good music retrieval system can find additional good matches: our filtering method is not guaranteed to be perfect, and later editions of the RISM A/II collection will always contain additional incipits. If such additional matches occur, one should check whether those matches were included in our lists of candidates.

Our ground truth is already applicable as it is, and the Wilcoxon test results give an indication of how much one can rely on the borders between the distinguishable groups we found. However, additional work could be beneficial, both for increasing the number of queries and for making more borders between groups emerge and for making the existing ones more reliable. Buckley's and Voorhees's work [1] indicates that having about 25 queries would be desirable for being able to apply statistical tests for comparing different retrieval systems. Also, more experiments could help if our candidate lists need to be extended either because a mistake in the filtering process gets noticed or because new incipits that are similar to a query were added to the RISM A/II collection after 2002.

Our ground truth can not only be useful as a benchmark for music retrieval systems, but also for finding out which musical features are relevant and how much they influence melodic similarity. For doing that, one needs to find incipits in the ground truth where very few features are different, but lead to significant differences in ranking. An example can be seen in Table 2, where the first three incipits show that notating music in a different clef leads to a smaller

difference than removing about two measures from the end. A somewhat less trivial observation that can be made in the same table is that repeated harmonic patterns matter, and that a dotted rhythm seems to be perceived as more similar to an inverted dotted rhythm than to a sequence of notes with the same durations.

6. ACKNOWLEDGEMENTS
We thank all participants of our experiments, both the ISMIR 2004 Graduate School attendees and the participants from Utrecht.

References

[1] C. Buckley and E. M. Voorhees (2000). Evaluating eval-uation measure stability. *Proceedings of the 23rd ACM Conference on Research and Development in Information Retrieval (SIGIR).* p. 33-40.

[2] J. S. Downie (2003). Toward the scientific evaluation of music information retrieval systems. *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003),* p. 25-32, Johns Hopkins University, Baltimore (MD), USA.

[3] D Hawking (2001). Overview of the TREC-9 Web Track. *Proceedings of the 9th Text Retrieval Conference TREC-9,* Gaithersburg, MD, 2001. http://trec.nist. gov/pubs/trec9/papers/web9.pdf

[4] J. Howard (1997). Plaine and Easie Code: A Code for Music Bibliography. *In*: *Beyond MIDI: The Handbook of Musical Codes,* edited by E. Selfridge-Field. p. 362-72

[5] K. Lemström, V. Mäkinen, A. Pienimaki, M. Turkia, E. Ukkonen (2003). The C-BRAHMS Project. *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR. 2003),* pp. 237-238, Johns Hopkins University, Baltimore (MD), USA.

[6] D. Müllensiefen and K. Frieler: Measuring melodic sim-ilarity: Human vs. algorithmic judgments. *R. Parn-cutt. A. Kessler & F. Zimmer (Eds.) Proceedings of the Conference on Interdisciplinary Musicology (CIM04) Graz/'Austria. 15-18 April. 2004.*

[7] L. Prechelt and R. Typke (2001). An Interface for Melody Input. *ACM Transactions on Computer-Human Interaction* 8(2), p. 133-149.

[8] *Repertoire International des Sources Musicales (RISM). Serie A/II. manuscrits musicaux apres 1600.* (2002) K. G. Saur Verlag, Miinchen, Germany, http://rism.stub.uni-frankfurt.de

[9] E. Selfridge-Field (1998). Conceptual and Representa-tional Issues in Melodic Comparison. *Computing in Musicology (11)* 3—64.

[10] R. Typke, P. Giannopoulos, R,. C. Veltkamp, F. Wiering, R,. van Oostrum (2003). Using Transportation Dis-tances for Measuring Melodic Similarity. *ISMIR 2003, Proceedings of the Fourth International Conference on Music Information Retrieval,* 107—114.

[11] R. Typke, R. C. Veltkamp, F. Wiering (2004). Searching notated polyphonic music using transportation distances. *Proceedings of the ACM Multimedia Conference. New York. October 2004,* 128-135.

[12] E. M. Voorhees, D. K. Harman (2000). Overview of the eighth Text REtrieval Conference (TREC-8). *Proceedings of the Eighth Text REtrieval Conference (TREC-8),* p. 1-24. NIST Special Publication 500-246. Electronic version available at http://trec.nist. gov/pubs.html

[13] F. Wilcoxon (1945). Individual comparisons by ranking methods. *Biometrics.* 1 (6):80-83.