Evaluating the Earth Mover's Distance for measuring symbolic melodic similarity

Rainer Typke, Frans Wiering, Remco C. Veltkamp
Utrecht University
Department of Information and Computing Sciences
Padualaan 14, 3584 CH Utrecht, The Netherlands
rainer.typke,frans.wiering,remco.veltkamp@cs.uu.nl

ABSTRACT

We present two variants of an algorithm for measuring melodic similarity. The algorithm is based on the Earth Mover's Distance (EMD), which measures the amount of work one needs to transform one weighted point set into another. We describe how to represent melodies as weighted point sets and how to apply the EMD to compare them. The simpler algorithm variant first uses an evolutionary algorithm for finding a good alignment of two weighted point sets and then applies the EMD. We also present a second, more complicated algorithm which segments the query, thereby improving partial matching and making the method more robust against fluctuations of tempo or pitch within the query. The first algorithm is then used for the segments, and the results for individual segments are combined into one overall result. The more complicated algorithm was submitted to MIREX. Out of the other algorithms submitted to MIREX, three performed better than ours and three performed worse. We believe that although our result looks mediocre at first glance, our similarity measure still deserves to be developed further because of its built-in tolerance against distortions of the query and because of its continuity.

Keywords: MIREX, symbolic melodic similarity, Earth Mover's Distance.

1 INTRODUCTION

One task at MIREX 2005 was to retrieve the most similar incipits from a subset of the RISM A/II collection, given one of the incipits as a query. For each algorithm, the result lists for 11 queries were compared to a ground truth that was established as described in Typke et al. (2005a) (for a set of queries that was different from the set in this paper). The results were evaluated with four measures: Average dynamic recall, normalized recall at group boundaries, average precision, and precision at N documents (where N is the number of relevant documents).

We submitted an algorithm that compares melodies by transforming notes into a two-dimensional weighted point set. For each note, the coordinates are the onset time and pitch values, and the weight is the duration. The weighted point sets are then compared with the Earth Mover's Distance (EMD). The EMD is continuous and provides partial matching. By changing the weighting scheme and ground distance, one can tune it for different purposes. Its continuity makes it suitable for matching queries that are generated by humans (sung or played on a MIDI piano) with entries of a database of symbolic music (Typke et al., 2005b), without the need for quantizing, time warping, or any other form of tempo or pitch tracking. This strength does not matter in the MIREX 2005 task of matching notated music against other notated music. However, our method still ranks in the middle of the other methods, which do not work as well with distorted queries.

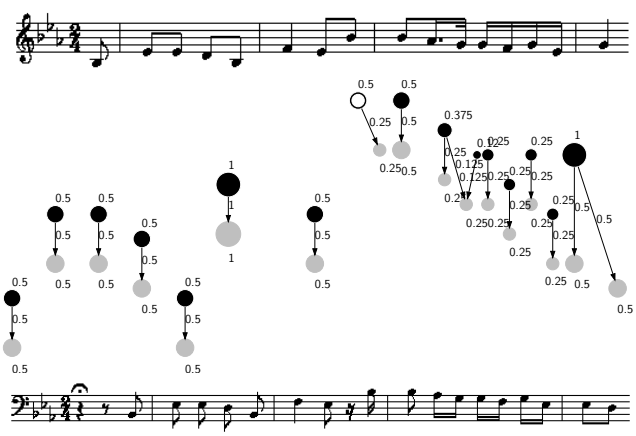


Figure 1: Two pieces of music and their representations as weighted point sets, along with the optimum flow that turns one of the point sets into the other. To make the two point sets easier to distinguish, they are shifted apart. For the actual distance calculation, they are positioned on top of one another.

The simpler algorithm variant has two main steps: finding a good combination of scaling and translation and then applying the EMD to compute the dissimilarity. It works well only if the compared melodies are not too different in length, as is the case in this task.

Our second, more complicated algorithm first segments the query before matching every query segment in a way similar to the first algorithm. This leads to result lists for the query segments which have to be recombined into an overall result. The segmentation makes the algorithm more robust against tempo or pitch fluctuations.

2 BASIC ALGORITHM

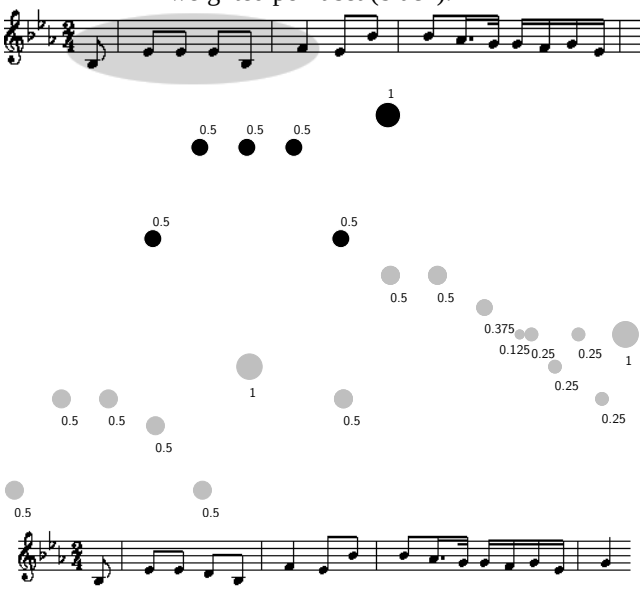
The steps of the algorithm are:

1. Represent melodies as weighted point sets. Every note is represented by a point in a two-dimensional space of onset time and pitch. The weight of the note represents its duration and possibly its position within a measure. This is illustrated by Figure 1.

2. Find a good alignment of the two point sets to be compared. Scaling in the time dimension and translation of a point set in both time and pitch dimensions are allowed since neither tempo changes nor transposition or the position of a melody within a piece fundamentally change the character of a melody. For aligning two point sets such that the EMD is minimized, we use the evolutionary optimization function evofunc (Min) of the library "Reusable Evolutionary Algorithms in Shape Matching" (REALISM), part of the Shape Matching Environment (SHAME). See Figures 2 and 3 for an illustration.

3. Use the EMD between the optimally aligned point sets as distance measure.

Top: Query. Only the shaded segment is converted into a weighted point set (black).



Bottom: A different piece to which the query is compared.

Figure 2: Problem: Before calculating the EMD, we need to somehow find out that the black point set should be moved to the beginning of the grey one for minimizing the EMD.

3 MORE COMPLEX ALGORITHM

We submitted the following, more complex algorithm to MIREX.

In order to improve robustness against tempo and pitch fluctuations in the query, this algorithm segments the

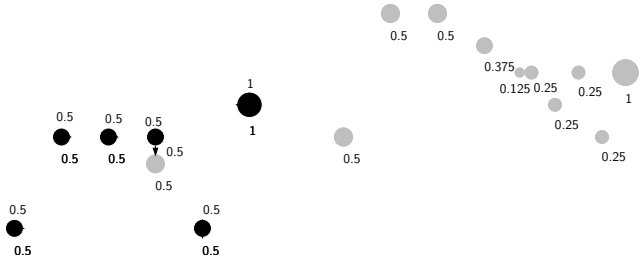


Figure 3: The optimum alignment of the two point sets from Figure 2 so that the EMD is minimized. The black points without arrows hide grey points.

query and uses the basic algorithm for finding matches of the segments. The steps are:

1. Segment the query into possibly overlapping segments with a given number of consecutive notes. The length of these segments is largely independent on the number of voices since we do not count all notes in the segment, but the number of notes that follow one another. For details on this segmenting method, see (Typke et al., 2004).

2. For every query segment, apply the basic algorithm described in the previous section.

3. Now we have one result list of matching documents and their distances for every query segment. Combine them as follows:

   - Calculate normalized distances as follows: Let e be the EMD distance between a query segment and a document. Let c be a large cutoff distance that is larger than a distance observed when there is some meaningful melodic similarity. Pick ε with $0 < \varepsilon < 1$. The normalized distance is 1 for EMD distances $\geq c$, and it is $\varepsilon + \frac{(1-\varepsilon)e}{c}$ otherwise. This ensures that the normalized distance is in the interval $[\varepsilon, 1]$.
   - For every document that occurs in any list of matches for a query segment, construct a list that contains: Query segment number, normalized distance to the document, and the onset times of the first and last query note within the matched document. This list might contain the same query number multiple times (if it matches at more than one place within the document), and it might contain different query segment numbers (if more than one query segment matches the document).
   - For every document, calculate an overall distance score by finding the minimum product of normalized distances for a legal combination of segments that match with this document. By "legal combination", we mean that:
     - No segment number occurs twice
     - Segments with a higher number are matched at a later position within the document
     - For segments with consecutive numbers: the overlap of the matched areas corresponds to

Table 1: Result quality for all MIREX symbolic melodic similarity submissions. The average precision is non-interpolated, and for the precision at N documents, N is the number of relevant documents.

Rank / Participant	Average Dynamic Recall	Normalized Recall at Group Boundaries	Average precision	Precision at N documents
1 Grachten, Arcos & Mantaras	65.98%	55.24%	51.72%	44.33%
2 Orio	64.96%	53.35%	42.96%	39.86%
3 Suyoto & Uitdenbogerd	64.18%	51.79%	40.42%	41.72%
4 Typke, Wiering & Veltkamp	57.09%	48.17%	35.64%	33.46%
5 Lemström, Mikkila, Mäkinen & Ukkonen (P3)	55.82%	46.56%	41.40%	39.18%
6 Lemström, Mikkila, Mäkinen & Ukkonen (DP)	54.27%	47.26%	39.91%	36.20%
7 Frieler & Müllensiefen	51.81%	45.10%	33.93%	33.71%

the overlap of the segments within the query (if there is any), and the matched areas are not too far apart.

4 DISCUSSION OF RESULTS

4.1 Result quality measures

Table 1 shows various measures of result quality for all algorithms that were submitted for the symbolic melodic similarity task at MIREX 2005. The measures can be split into two groups: those that work with a dynamic set of relevant documents and those that work with one fixed set of relevant documents. The former measures view some documents as relevant only from a certain position on. For example, a document that is somewhat similar to the query, but clearly less similar than two other documents, would be viewed as relevant only beginning with position 3 in the result list. The following measures work with a dynamic set of relevant documents:

- **Average Dynamic Recall:** This measure is described in Typke et al. (2005b). At any number of retrieved items, it gives the average recall among the documents that the user should have seen so far. For this comparison, it was measured at position N (where N is the number of relevant documents).
- **Normalized Recall at Group Boundaries:** The ground truth from Typke et al. (2005a) does not give one ideal ordering of results, but rather an ordering of groups where the ideal order within groups is not known. This measure is based on the recall at the boundaries of those groups.

The following measures view all relevant documents as equally relevant, even if they belong to different groups according to the ground truth constructed as described in Typke et al. (2005a).

- **Average Precision:** At every relevant document in the result list, the precision is measured. The average precision is the mean of the precisions at the positions of relevant documents in the result list.
- **Precision at N documents:** N is the number of relevant documents.

4.2 Result quality of our algorithm

Table 1 shows that our algorithm is ranked at the median position of all participants at MIREX 2005 for the measures which distinguish between different degrees of relevance. It would rank lower if the two measures which view all relevant items as equally relevant would be used.

We believe that even though our algorithm did not get top results at MIREX, it is still worth investigating further because it has some desirable properties that most of the other algorithms do not have:

- Our algorithm can handle cases where the query is distorted by tempo variations or pitch variations. This could make it suitable for query-by-humming applications without the need for an extra algorithm for pitch quantisation or tempo tracking.
- Our distance measure is continuous.

4.3 Future work

We see some possibilities for improving our algorithm:

- Our method of combining segment search results into one overall score could probably still be improved. There is a conflict between rewarding the occurrence of very few, very similar segments and, on the other hand, the occurrence of many segments with some similarity. In other words, it is not clear whether a document that matches very well with parts of the query, but not with the whole query should be ranked higher than a document that has some, but not very high similarity with the whole query.
- The calculation of the overall score could also be improved by taking into consideration how the transformations of individual query segments relate to one another. It can happen that different query segments undergo very different transformations (scaling and translation) for minimizing the EMD. Within one query, those transformations should not have to differ too much. If they do, this should lower the overall score.
- The evolutionary algorithm for finding an optimum alignment of two point sets sometimes gets caught

in a local optimum and fails to find the global optimum. If this happens, a segment is treated as less similar than it should. The evolutionary algorithm could be improved to make this less likely, or some other method for finding a good alignment of point sets could be used.

References

P. Min. Evofunc: evolutionary optimisation function of the library "Reusable Evolutionary Algorithms in Shape Matching" (REALISM). `http://www.cs.uu.nl/centers/give/multimedia/matching/shame.html`.

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*, 3(1):34–39, 2005a. Ground truth data available at http://give-lab.cs.uu.nl/orpheus/.

R. Typke, R. C. Veltkamp, and F. Wiering. Searching notated polyphonic music using transportation distances. In *Proceedings of the ACM Multimedia Conference*, pages 128–135, New York, 2004.

R. Typke, R. C. Veltkamp, and F. Wiering. A comparison of melody matching algorithms with a ground truth created by human experts. 2005b. Submitted for publication.