Using Transportation Distances for Measuring Melodic Similarity

Rainer Typke, Panos Giannopoulos, Remco C. Veltkamp, Frans Wiering, René van Oostrum
Institute of Information and Computing Sciences
University of Utrecht
3584CH Utrecht, The Netherlands
+31-(0)30-2531172
{rainer.typke,panos,remco.veltkamp,frans.wiering,rene}@cs.uu.nl

Abstract

Most of the existing methods for measuring melodic similarity use one-dimensional textual representations of music notation, so that melodic similarity can be measured by calculating editing distances. We view notes as weighted points in a two-dimensional space, with the coordinates of the points reflecting the pitch and onset time of notes and the weights of points depending on the corresponding notes' duration and importance. This enables us to measure similarity by using the Earth Mover's Distance (EMD) and the Proportional Transportation Distance (PTD), a pseudometric for weighted point sets which is based on the EMD. A comparison of our experiment results with earlier work shows that by using weighted point sets and the EMD/PTD instead of Howard's method (1998) using the DARMS encoding for determining melodic similarity, it is possible to group together about twice as many known occurrences of a melody within the RISM A/II collection. Also, the percentage of successfully identified authors of anonymous incipits can almost be doubled by comparing weighted point sets instead of looking for identical representations in Plaine & Easie encoding as Schlichte did in 1990.

1 Introduction

Representing music as a weighted point set in a two-dimensional space has a tradition of many centuries. Ever since the 13th century, music has been written as a set of notes (points) in a two-dimensional space, with time and pitch as coordinates. Varying characteristics are associated with the notes by, for example, using different symbols for different note durations. The look of written music has changed somewhat over the last 8 centuries, but the basic idea of representing music as a weighted point set has been followed for almost a millenium, and it has served composers and performers well. Since weighted point sets seem to be so well suited to representing music, it feels natural to measure melodic similarity directly by comparing weighted point sets instead of first transforming the music into one-dimensional abstract representations.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. ©2003 Johns Hopkins University.

We studied the use of the Earth Mover's Distance (EMD), which measures a minimum flow for transforming one weighted point set into another, for the purpose of measuring melodic similarity. Because the triangle inequality does not hold for the EMD, we also used a modified version of it, the Proportional Transportation Distance (PTD), which was proposed by Giannopoulos and Veltkamp (2002). A distance measure for which the triangle inequality holds can be used to make database searches more efficient by using indices.

The music database we used for evaluating the EMD and PTD as similarity measures contains about half a million musical incipits[1] from the RISM A/II collection (Répertoire International des Sources Musicales, 1995-2002).

We evaluated the appropriateness of the EMD and PTD for measuring melodic similarity by constructing groups of similar melodies within the RISM A/II collection and comparing our results to the "Frankfurt Experience" and "Harvard Experience" of sorting RISM incipits described by John B. Howard (1998). We were able to identify about twice the percentage of melodies by anonymous composers and group together 76 % instead of 46 % of the known occurrences of a tune called "Roslin Castle".

2 Melodies as weighted point sets

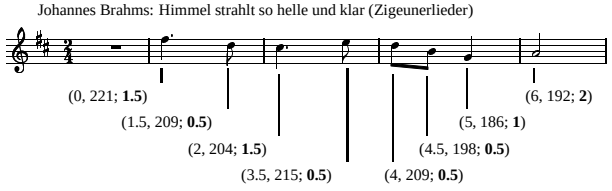

Figure 1: An example of music represented with a weighted point set. Format: (Time, Pitch; **Weight**). In this example, the weights only reflect the note durations. Because of this, the time coordinate here equals the sum of the weights of preceding notes. Pitches are specified using Hewlett's (1992) base-40 system.

In order to be able to apply a transportation distance measure, we must transform the melodies we want to compare into sig-

[1] Incipits are the beginnings of pieces, typically about 20 notes long.

natures. By signature, we mean a set of points in the two-dimensional Euclidean space where each point has a weight associated with it. The two dimensions are time and pitch.

When transforming melodies into signatures, we create one point for each note. Rests are encoded implicitly as the time spans that are not covered by points. As a consequence, we do not distinguish between two subsequent quarter rests and one half rest, but we do distinguish between two subsequent quarter notes and a half note; only the latter sounds differently.

2.1 The time coordinate

In our database, durations of notes and their positions within measures are specified using divisions of a quarter note, in a way similar to the MIDI format. With every melody, the number of divisions per quarter note is stored. This number is chosen such that the duration of every note in the melody can be specified as a whole number. For example, if there are 96 divisions per quarter note, a quarter note has duration 96, a half note has duration 192, and a sixteenth 24.

We want time coordinates in signatures to be independent of the number of divisions chosen for a particular melody. Therefore, we calculate the time coordinate of a note as the sum of the lengths of measures preceding the note plus the note's position within its measure, divided by the number of divisions for a quarter note. Measure lengths are calculated as follows: for each note or rest in a measure, the duration is added to the position within the measure. The maximum of all of these end points of notes and rests is then taken as the measure length.

In order to skip leading rests – we do not want to distinguish between melodies that differ only in the duration of leading rests –, we then subtract the very first note's time coordinate from all time coordinates, thereby shifting all notes so that the first note starts at time 0.

For a complete example, see Figure 1 and Table 1.

Note Number	Measure Number	Pitch40	Duration	Position in bar
1	1	0	1920	0
2	2	221	1440	0
3	2	209	480	1440
4	3	204	1440	0
5	3	215	480	1440
6	4	209	480	0
7	4	198	480	480
8	4	186	960	960
9	5	192	1920	0

Table 1: The database contents for the melody shown in Figure 1. There are 960 divisions per quarter note, and rests are coded as notes with pitch 0. To arrive at Figure 1, we first normalize the time coordinates (i. e., divide them by 960). The durations are then: 1920/960=2, 1.5, 0.5, 1.5, 0.5, 0.5, 0.5, 1, 2. All measure lengths are 1920/960=2. Therefore, the note onset times are: 0, 2, 3.5, 4, 5.5, 6, 6.5, 7, and 8. This still includes the leading rest, which we want to ignore, so finally, we skip the leading rest and subtract its duration from all subseqent notes: 0, 1.5, 2, 3.5, 4, 4.5, 5, 6. These are the time coordinates in Figure 1.

Our method of determining the length of each measure without relying on the time signature ensures that we get sensible coordinates even in cases where the notes in a measure do not match the time signature. This actually happens with the RISM data. See, for example, the bottom right incipit in Figure 5, where not only the octaves are encoded incorrectly for some notes, but there is also a mismatch of the time signature and the contents of measures.

2.2 The pitch coordinate

Unlike MIDI files, our database contains the pitch in Walter Hewlett's (1992) Base-40 notation. This notation distinguishes between notes with the same pitch, but different notations. Like MIDI pitches, it is a number-line representation of musical pitch notation, but with the added advantage of being interval-invariant. I. e., the difference between any two base-40 pitch numbers will correctly determine the notated interval name between those pitches.

2.3 Weights

Increasing a note's weight increases the importance of it having a counterpart of similar weight at the same position in the compared melody. A natural method of using weights is to make them reflect note durations. That way, differing note durations at corresponding positions lead to an increase in the resulting distance. For instance, in Figure 1 the note weights reflect only the durations. All results in this paper were obtained with weights that only depend on note durations. By adding more components, however, additional desirable effects could be achieved. Two promising weight components are stress weight and note number weight.

2.3.1 Stress Weight

There are cases where melodies clearly differ, but a distance measure which ignores the positions of notes within measures fails to distinguish between them. For example, the two melodies in Figure 2 would not be distinguished by the simple distance measures used for Figure 5. By adding more weight to notes at positions in measures which are usually emphasized, e. g. the first beat, the measure structure can be taken into account as well.

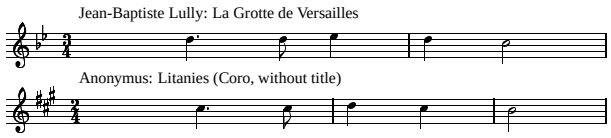

Figure 2: By adding a stress-based weight component, the distance measure can be made to reflect different measure structures. Without that, the distance would be zero for these clearly different melodies, provided that transpositions are allowed.

2.3.2 Note Number Weight

In the RISM database, there are no clear rules about how many notes are included in the incipits. Therefore, it happens that very similar or identical melodies differ mainly in the number of notes that are included in the incipit. As we shall see later, for example in the right column of Figure 5, there are instances where the distance between melodies becomes very large because one of them is cut off after fewer notes, not because they contain very different musical material. One possible way of addressing this problem is to add an extra weight component to each note that depends on how many notes precede it. That way,

notes close to the beginning are made more important than extra notes at the end which might not be present in all occurrences of a melody in the database.

In Section 4.2, we will describe some adjustments of the signatures which we do before applying a distance measure.

3 Similarity Measures for Weighted Point Sets

For a similarity measure (formally speaking, a function on a set S, $d : S \times S \rightarrow \mathbb{R}^+ \cup \{0\}$), the following properties are usually desirable:

i. *Self-identity*: For all $x \in S, d(x,x) = 0$.

ii. *Positivity*: For all $x \neq y$ in $S, d(x,y) > 0$.

iii. *Symmetry*: For all $x,y \in S, d(x,y) = d(y,x)$.

iv. *Triangle inequality*: For all $x,y,z \in S$, $d(x,z) \leq d(x,y) + d(y,z)$.

A measure with all of these properties is called a metric, while a measure with only properties i, iii, and iv is called a pseudo-metric. Depending on the application, different properties are relevant. For measuring melodic similarity, we need self-identity and symmetry. The triangle inequality is useful for efficiently searching the database (Barros et al., 1996). Positivity is not necessarily always desired. The EMD's partial matching property, which is closely related to its lack of positivity (see Section 3.1.2), can be useful.

In the following subsections, we will describe the two transportation distances which we used.

3.1 The Earth Mover's Distance (EMD)

The Earth Mover's Distance between two weighted point sets measures the minimum amount of work needed to transform one into the other by moving weight. Intuitively speaking, a weighted point can be seen as an amount of earth or mass; alternatively it can be taken as an empty hole with a certain capacity. We can arbitrarily assign the role of the supplier to one set and that of the receiver/demander to the other one, setting, in that way, the direction of weight movement. The EMD then measures the minimum amount of work needed to fill the holes with earth (measured in weight units multiplied with the covered ground distance). See Cohen's Ph.D. thesis (1999) for a more detailed description of the EMD.

3.1.1 Definition

Let $A = \{a_1, a_2, .., a_m\}$ be a weighted point set such that $a_i = \{(x_i, w_i)\}, i = 1, .., m$, where $x_i \in \mathbb{R}^k$ with $w_i \in \mathbb{R}^+ \cup \{0\}$ being its corresponding weight. Let W=$\sum_{j=1}^{n} w_i$ be the total weight of set A.

The EMD can be formulated as a linear programming problem. Given two weighted point sets A, B and a ground distance d, we denote as f_{ij} the elementary flow of weight from x_i to y_j over the distance d_{ij}. If W, U are the total weights of A, B respectively, the set of all possible flows $\mathcal{F} = [f_{ij}]$ is defined by the following constraints:

1. $f_{ij} \geq 0, i = 1, ..., m, j = 1, ..., n$
2. $\sum_{j=1}^{n} f_{ij} \leq w_i, i = 1, ..., m$
3. $\sum_{i=1}^{m} f_{ij} \leq u_j, j = 1, ..., n$
4. $\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = min(W, U)$

These constraints say that each particular flow is non-negative, no point from the "supplier" set emits more weight than it has, and no point from the "receiver" receives more weight than it needs. Finally, the total transported weight is the minimum of the total weights of the two sets.

The flow of weight f_{ij} over a distance d_{ij} is penalized by its product with this distance. The sum of all these individual products is the total cost for transforming A into B. The EMD(A, B) is defined as the minimum total cost over $\mathcal{F}$, normalized by the weight of the lighter set; a unit of cost or work corresponds to transporting one unit of weight over one unit of ground distance. That is:

$$\mathrm{EMD}(A,B) = \frac{\min_{F \in \mathcal{F}} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{min(W,U)}$$

3.1.2 Properties and Computation

The most important properties of the EMD can be summarized as follows:

1. The EMD is a metric if the ground distance is a metric and if the EMD is applied on the space of equal total weight sets.

2. It is continuous, in other words, infinitesimal small changes in position and/or weight of existing points cause only infinitesimal change in the its value. Moreover, the addition of a point with an arbitrarily small weight, i. e. noise (which can be seen as increasing its weight from zero to a positive value) leads to an arbitrarily small change in the EMD's value.

3. It does not obey the positivity property if the sums of the weights of the two sets are not equal. In that case, some of the weight of the heavier distribution remains unmatched. Therefore, the EMD allows for partial matching. As a result, there are cases where it does not distinguish between two non-identical sets. Sometimes this can be useful, for example when two incipits contain identical melodies which are cut off after different numbers of notes. On the other hand, this also leads to effects like the one we see with incipit number 12 in the left column of Figure 5, where the EMD yields a relatively low distance. Here the surplus of weight is not all concentrated at the end of the melody, but distributed over several rests and other places, which leads to a false positive.

4. In the case of unequal total weights, the EMD does not obey the triangle inequality. A simple counterexample would be three melodies called A, B, and AB. Let us assume that AB is the concatenation of A and B, and let us assume that A and B are chosen so that the EMD yields a distance of 1 between them. If A and B are positioned accordingly, both the distance between A and AB and the distance between B and AB can be zero (because both A and B are parts of AB). Then, d(A, B) $>$ d(A, AB) + d(AB, B).

As a result, methods that rely on the triangle inequality for speeding up database retrieval cannot be used in conjunction with the EMD.

The EMD can be computed efficiently by solving the corresponding linear programming problem, for example by using a streamlined version of the simplex algorithm for the transportation problem (Hillier and Lieberman 1990). We used Rubner's (1998) EMD function, which implements Hillier's and Lieberman's algorithm. It is possible that the simplex algorithm performs an exponential number of steps. One could use polynomial algorithms like an interior point algorithm, but in practice that would outperform the simplex algorithm only for very large problem sizes. Since the transportation problem is a special case of the minimum cost flow problem in networks, a polynomial time algorithm for that could be used as well.

3.2 The Proportional Transportation Distance (PTD)

Giannopoulos and Veltkamp (2002) proposed a modification of the EMD in order to get a similarity measure based on weight transportation such that the surplus of weight between two point sets is taken into account and the triangle inequality still holds. They call this modified EMD the "Proportional Transportation Distance" (PTD) because any surplus or shortage of weight is removed in a way that the proportions are preserved before the EMD is calculated. The PTD is calculated by first dividing, for both point sets, every point's weight by its point set's total weight, and then calculating the EMD for the resulting point sets.

The PTD is defined as follows:

Let A, B be two weighted point sets, W, U the total weights of A and B, and d a ground distance. The set of all feasible flows $\mathcal{F} = [f_{ij}]$ from A to B is defined by the following constraints:

1. $f_{ij} \geq 0, i = 1,...,m, j = 1,...,n$
2. $\sum_{j=1}^{n} f_{ij} = w_i, i = 1,...,m$
3. $\sum_{i=1}^{m} f_{ij} = \frac{u_i W}{U}, j = 1,...,n$
4. $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = W$

The PTD(A, B) is given by:

$$\mathrm{PTD}(A,B) = \frac{\min_{F\in\mathcal{F}} \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} d_{ij}}{W}$$

Constraints 2 and 4 force all of A's weight to move to the positions of points in B. Constraint 3 ensures that this is done in a way that preserves the old percentages of weight in B.

The PTD is a pseudo-metric. In particular, it obeys the *triangle inequality*. It still does not have the *positivity* property since the distance between positionally coinciding sets with the same percentages of weights at the same positions is zero. However, this is the only case in which the distance between two non-identical point sets is zero. The PTD will distinguish between two sets B and B' which differ in only one point. It has all other properties of the EMD for equal total weight sets.

4 Adjustments of coordinates and weights, ground distance

4.1 The ground distance

For all results in this paper, we used the Euclidean distance as ground distance. I. e., the distance between two notes with the coordinates (t_1, p_1) and (t_2, p_2) is $\sqrt{(t_1 - t_2)^2 + (p_1 - p_2)^2}$.

An interesting variation, especially for polyphonic music, would be to make the distance in the pitch dimension depend on harmony instead of just calculating the difference of pitches.

4.2 Adjustments of coordinates and weights

Before applying one of the similarity measures for weighted point sets described above, we adjust the signatures of the two melodies we want to compare in several ways:

- In order to be able to recognize augmented or diminuted versions of a melody as similar (like for example in the right column of Figure 5, second group, melody 4), it can be necessary to normalize the range of time coordinates. We chose to stretch the melody with the smaller maximum time coordinate over a longer time such that after the adjustment, both melodies' maximum time coordinates equal the larger maximum time coordinate before the adjustment. Note that with a less careful normalization, e. g. the adjustment of a randomly chosen melody to the other melody's length, one can easily lose the symmetry property. We did this alignment of durations when we used the PTD, where there is no partial matching; when we used the EMD, we compared the distances with and without duration alignment and took the minimum.

- It is desirable to make the distance measure independent of transpositions. This could be done by moving one of the two melodies up or down in pitch to a position where the distance is minimal (Ó Maidín, 1998). Since finding the optimum transposition would require the repeated application of the similarity measure, which would take a lot of time, we chose to transpose one of the melodies so that the weighted average pitch is equal. This way, the similarity measure for weighted point sets needs to be applied only once, but this is not always the optimum solution. However, this approximation usually works well enough for transposed versions of the same melody to appear closer than other melodies from the database, see Figure 5.

- When the transportation distance is calculated, the transportation of weight from one note to another can happen in the time dimension, the pitch dimension, or a combination of the two. Therefore, the range of numbers in both dimensions affects the results. For all results shown in this paper, we multiplied the time coordinates with 3 in order to avoid all points to be placed in a very narrow, long strip like in Figure 1, where the pitch ranges from 186 to 221 (a range of 35), while the time only ranges from 0 to 6. An arrangement of the points like in Figure 1 would make it too cheap to move weight in the time dimension in comparison to the pitch dimension, which would lead to notes being matched with notes that do not really correspond with them.

4.3 An example

Figure 3 shows the weight flow for signatures of two melodies after adjusting them as described above. Unlike the signature shown in Figure 1, the time coordinates are now multiplied with 3 so that weight transportation in the time and pitch dimensions are similarly expensive. In Figure 1, the range of pitches is much larger than the range of time coordinates so that transportation distance measures would match notes which do not correspond with one another. Also, the top melody in Figure 3 was stretched so that the maximum time coordinates are both 28.5, and the top melody was also slightly shifted in the pitch dimension so that the weighted average pitches are the same.Without the pitch and duration alignments, the distance between these two melodies would be 5.41825 instead of 0.739529. For the sake of simplicity, we treated the grace note in the bottom melody like any other eighth note, thereby overemphasizing it and influencing the time coordinates of subsequent notes. A special treatment of grace notes would probably lead to better results.

In Figure 3, an arrow indicates the flow and the transported weight for each pair of weighted points between which any weight is transported. Consider, for example, the first two notes of both melodies. Since the second melody starts with a dotted eigth note and a sixteenth note, while the first one starts with two eigth notes, half of the weight of the second note of the first melody is transported to the first note of the second melody, while the other half goes to the second note. The quarter note which is represented as a hollow circle is only partially matched. It has a capacity of 1, but only 0.5 weight units are transported into it.

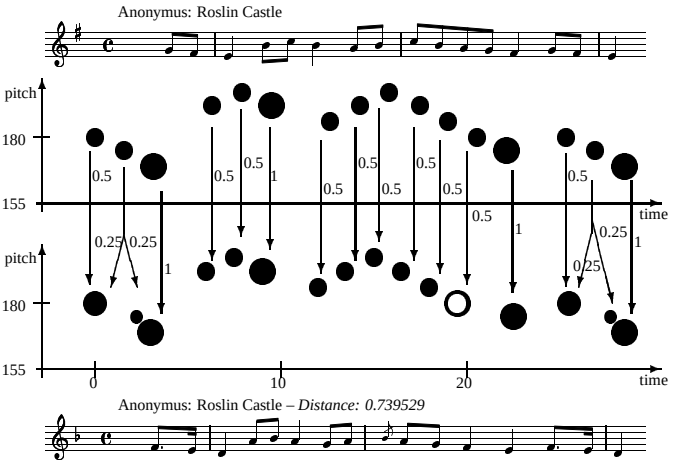

Figure 3: An illustration of a weight flow with the EMD; the coordinates are adjusted as described in Section 4.2. The signatures of melodies 1 and 11 from Figure 5 (left) after the adjustments, shown in the format (Time, Pitch; **Weight**):
Top: (0, 180.138; **0.5**), (1.58333, 175.138; **0.5**), (3.16667, 169.138; **1**), (6.33333, 192.138; **0.5**), (7.91667, 197.138; **0.5**), (9.5, 192.138; **1**), (12.6667, 186.138; **0.5**), (14.25, 192.138; **0.5**), (15.8333, 197.138; **0.5**), (17.4167, 192.138; **0.5**), (19, 186.138; **0.5**), (20.5833, 180.138; **0.5**), (22.1667, 175.138; **1**), (25.3333, 180.138; **0.5**), (26.9167, 175.138; **0.5**), (28.5, 169.138; **1**)
Bottom: (0, 180; **0.75**), (2.25, 175; **0.25**), (3, 169; **1**), (6, 192; **0.5**), (7.5, 197; **0.5**), (9, 192; **1**), (12, 186; **0.5**), (13.5, 192; **0.5**), (15, 197; **0.5**), (16.5, 192; **0.5**), (18, 186; **0.5**), (19.5, 180; **1**), (22.5, 175; **1**), (25.5, 180; **0.75**), (27.75, 175; **0.25**), (28.5, 169; **1**)

5 Results and comparison with earlier results

5.1 Comparison with Schlichte's "Frankfurt Experience"

Joachim Schlichte (1990) describes an attempt of grouping similar incipits together. This work was based on 83,243 incipits from the RISM A/II collection. Schlichte shows that omitting "insignificant" musical phenomena immediately leads to useless results, even among the small subset of 83,243 out of the 476,000 incipits that are currently available to us. The methods he classified as useless are:

- Converting the incipits into strings of intervals and comparing those, thus ignoring rhythm, absolute pitch, and rests, leads to distances of zero between very different pieces. Schlichte gives some examples.
- Comparing strings of pitches, ignoring only rhythm and rests, still leads to too many false positives. Transposing all pieces into the same key before applying this method makes matters worse.

Schlichte therefore only looked for identical incipits, which already give music scholars valuable pointers to interesting facts. One of the interesting applications is the identification of anonymous pieces. Schlichte writes that among the 14,000 anonymous works in his data collection, 292, i. e. about 2 %, can be automatically associated with a composer by looking for identical incipits. This comparison is based on the Plaine & Easie encoding (Selfridge-Field, 1997) in which all RISM A/II incipits are stored.

We compared approximately 80,000 incipits by unidentified composers in the RISM A/II collection to all other incipits; the result can be seen at `http://give-lab.cs.uu.nl/MIR/anon/idx.html`. About 13 % of these incipits lie within a distance of less than 1 (PTD; weights: duration only; base-40 pitches; time coordinates multiplied with 3) from other incipits. This includes trivial cases where the incipits are identical, but also more interesting cases like the one shown in Figure 4, where added notes, augmentation, transposition, and differences in rhyhtm, contour and the sequence of intervals make it more difficult to recognize the similarity.

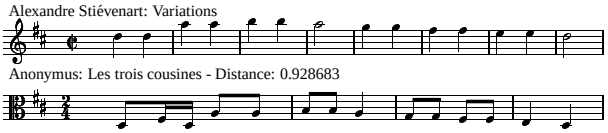

Figure 4: These two versions of the "Ah! vous dirai-je Maman" theme are recognized as similar (with PTD, weights: duration only). Note the extra notes in the second to last measure of Stiévenart and the first measure of Anonymus, which lead to differences both in the sequence of intervals and the contours, and the fact that the Stiévenart version is an augmented and transposed version.

In order to see how many of these matches are actually useful and would not have been found by just looking for identical

pieces, we manually checked 100 randomly chosen search results with distances below one. 55 % of these works only match with other anonymous works. For 19 %, a composer could be found because the compared incipits are identical. This is similar to Schlichte's result – 19 % of 13 % are 2.47 %, while Schlichte's figure is 2.08 %. We expect a slightly higher percentage because we do not compare the Plaine & Easie encodings, but our database contents as described in Section 2, which means that we view more melodies as identical than Schlichte did. For example, we ignore beaming. For another 11 %, we could determine the composer although the incipits are not identical. Therefore, our method leads to the identification of about 3.9 % of all anonymous pieces instead of Schlichte's 2.08 %.

5.2 Comparison with Howard's "Harvard Experience"

Howard (1998) describes a later attempt of grouping together similar incipits from the RISM A/II collection. This work was based on a subset of our collection with at most half as many[2] incipits. The U. S. RISM officials did not, like their Frankfurt colleagues, compare Plaine & Easie encodings, but converted the incipits into the DARMS (Selfridge-field, 1997) encoding language. They compared sorting results of five encoding types:

1. the complete encoding with all parameters,
2. the complete encoding transposed to a common pitch register,
3. the encoding stripped of such features as beaming, bar lines, and fermatas,
4. the encoding stripped of the items given in (3) plus grace notes,
5. the encoding stripped of the items given in (3) and (4) plus rhythmic values, rests, and ties, with the transposition to a common register (2) but with preservation of repeated notes.

None of the five encoding types lead to more than 6 out of 13 known occurrences of a song called "Roslin Castle" being sorted together among not more than 230,000 incipits.

Figure 5 shows the results of some queries for the same song with the EMD (left column) and PTD (right column). The EMD query groups together 11 out of 16 known occurrences among 476,000 incipits, i. e. a larger percentage among more than twice as many potential false positives in comparison with the "Harvard experience". If one does not count the 16th occurrence, shown at the bottom right, because it is not encoded correctly, our method compares even more favourably (73 % versus 46 %). The PTD result shown in Figure 5 does not group together more occurrences, but at least the false positives are musically very similar.

Figure 5 also illustrates the different properties of the EMD and PTD:

- The EMD groups more occurrences together in just one query result. Among the 16 known occurrences of "Roslin Castle", there are 3 groups with similar numbers of notes. The fact that the EMD allows partial matching, while the PTD matches all notes, leads to a clear distinction of these groups by the PTD, but not the EMD.
- Because of the weight normalization, the PTD recognizes augmented or diminuted versions of the same melody as similar. In the second group of melodies in the right column of Figure 5, melody number 4 is recognized as similar to the other melodies in the group, while the use of the EMD leads to a rather large distance since there is a large difference in weight between corresponding notes.
- The false positives in the right column of Figure 5 (numbers 6 and 8 in the first group) are more similar to the musical material in the rest of the query result than the false positives in the left column (numbers 12 to 16). The reason is that the EMD allows an unmatched weight surplus to be spread over the whole melody. In other words, this distance measure does not distinguish between a few extra or missing blocks of notes on the one side and differences between many individual notes or rests on the other side. When the PTD is used, blocks of extra or missing notes lead to the wrong notes being compared to one another, which usually dramatically increases the distance. Any differences between individual notes are also penalized.

6 Indexing

We eliminated the need for calculating the expensive transportation distance to a query for every melody in the database by exploiting the triangle inequality and using vantage objects (Vleugels and Veltkamp, 2002). As a preparation, k vantage objects are randomly chosen from the database containing n point sets, and the distance of each of the n point sets to each of the k vantage objects is calculated in the feature space by determining a transportation distance. These transportation distances can be viewed as the coordinates of the point sets in a k-dimensional Euclidean space.

Thanks to the triangle inequality, the Euclidean distance between two point sets in the k-dimensional space is a lower bound for the transportation distance in the feature space (Barros et al., 1996).

The search for point sets which are closer than r to the query point set can now be limited to those point sets whose Euclidean distance from the query in the k-dimensional Euclidean space of transportation distances is less than r. Only for those, the transportation distance needs to be calculated. If the query object is not yet in the database, its distances to the k vantage objects need to be calculated as a first step. Then, by performing an approximate nearest-neighbour search in the Euclidean space, one can answer a query by performing $O(m \log n)$ Euclidean distance calculations (Arya, Mount, Netanyahu, Silverman, and Wu, 1994) plus m expensive transportation distance calculations, where m, the number of reported point sets, depends on how the weighted point sets are distributed in the Euclidean space. If one prefers an exact nearest neighbour search, one can query a k-dimensional kd-tree using $O(n^{1-\frac{1}{k}} + m)$ Euclidean distance calculations. In practice, with our database of 476,000 point sets and a maximum distance r of 5, we need

[2] Howard does not clearly say how many, but from the introduction to his paper it can be inferred that the number was probably below 230,000.

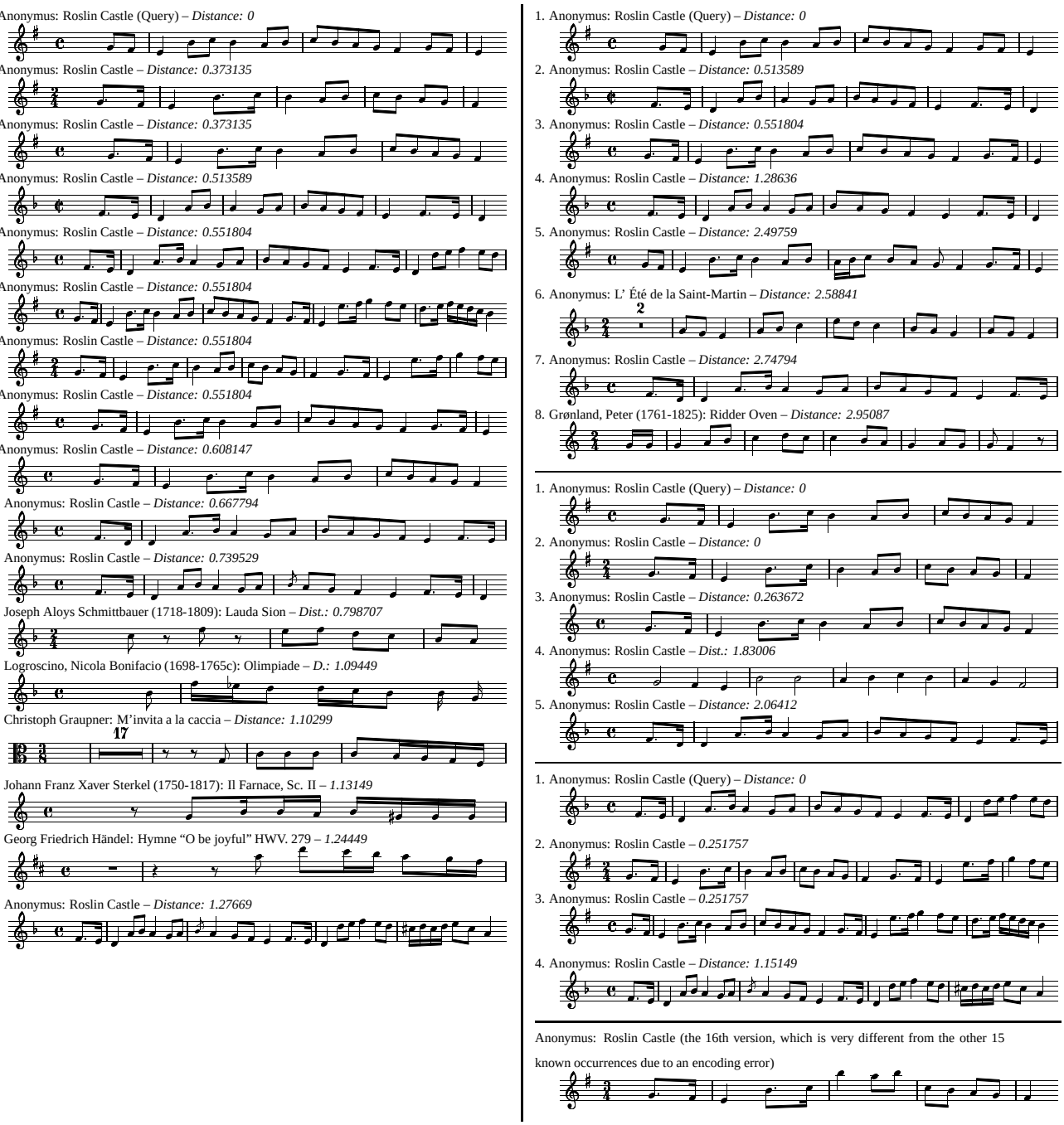

Figure 5: Query results for "Roslin Castle" among 476,000 melodies. Weights reflect only note durations. The left column shows the top 17 matches of an EMD-based query, containing 12 occurrences of "Roslin Castle". The right column contains all 16 known occurrences of "Roslin Castle", 15 of which are retrieved with 3 PTD-based queries whose results are separated with horizontal lines. There is an encoding error which prevents the 16th occurrence from being shown in other query results – in the Plaine & Easie format used for collecting the RISM data, it is easy to get the octaves wrong. For a discussion of the differences between EMD (left column) and PTD (right column), see Section 5.2.

less than 1000 expensive calculations instead of 476,000, which reduces the query running time on a 2-GHz Pentium 4 system with Windows XP from approximately 70 minutes to 9 seconds, without altering the result.

Although the triangle inequality holds only for the PTD and not generally for the EMD, we tried this indexing method for EMD distances as well. In most cases, the results are not distorted.

7 Conclusions and future goals

In comparison to Schlichte's and Howard's experience with grouping similar melodies from the RISM A/II collection together, our transportation distance measures perform much better. It is possible to group together more occurrences of a melody among a larger total number of incipits. Also, with transportation distance measures, it is easy to recognize sim-

ilarities even if they are hidden by additional notes or different rhythm. Finally, there are transportation distance measures which obey the triangle inequality, e. g. the PTD, so that it is possible to efficiently search large databases. We used this fact for an interactive online search engine which searches all 476,000 melodies in our database.

One important strength of transportation distances which we have not exploited yet is the fact that they should allow us to compare polyphonic music in much the same way as monophonic music. For weighted point sets, it should not make much of a difference whether there are points which share the same time coordinate.

The partial matching provided by the EMD does not always make musical sense, as can be seen with the false positives in the left column of Figure 5. In order to be able to find motifs and themes within complete pieces, it might be necessary to split the pieces into small chunks and then use a transportation distance for comparing chunks. The EMD's partial matching, possibly in conjunction with a consonance ground distance, might then e. g. prove useful for searching full scores based on themes taken from a piano reduction.

Another promising application would be the use of transportation distance measures for building a Query-by-Humming system without explicit note onset detection. Note onset detection is a difficult problem (Pauws, 2002; Prechelt and Typke, 2001). This task could be delegated to the transportation distance measure, which would combine the two tasks of grouping FFT windows[3] into notes and comparing sets of notes.

8 Acknowledgements

We thank Han-Wen Nienhuys for the support he gave us when we used his Lilypond music typesetter for generating hundreds of thousands of music bitmaps.

References

Arya, S., Mount, D. M., Netanyahu, N. S., Silverman, R. & Wu, A. (1994). An optimum algorithm for approximate nearest neighbor searching. *Proceedings of the Fifth ACM-SIAM Symposium on Discrete Algorithms*, (pp. 573–582). Implementation retrieved April 1, 2003, from `http://www.cs.umd.edu/~mount/ANN/`

Barros, J., French, J., Martin, W., Kelly, P. & Cannon, M. (1996). Using the triangle inequality to reduce the number of comparisons required for similarity-based retrieval. *Proceedings of SPIE, Storage and Retrieval for Still Image and Video Databases IV*, 2670, (pp. 392–403).

Cohen, S. (1999). *Finding Color and Shape Patterns in Images.* Ph.D. thesis, Stanford University, Department of Computer Science.

Giannopoulos, P. & Veltkamp, R. C. (2002). A Pseudo-Metric for Weighted Point Sets. In Heyden, A., Sparr, G., Nielsen, M. & Johansen, P. (Ed.), *Proceedings of the 7th European Conference on Computer Vision (ECCV)* (pp. 715–730). Copenhagen, Denmark: Springer-Verlag.

[3]FFT windows: the results of Fast Fourier Transformations of short time windows of audio signal.

Hewlett, W. B. (1992). A Base-40 Number-line Representation of Musical Pitch Notation. *Musikometrika*, 4, 1–14. Retrieved April 1, 2003, from `http://www.ccarh.org/publications/reprints/base40/`

Hillier, S. & Lieberman, G. J. (1990). *Introduction to Mathematical Programming.* McGraw-Hill.

Howard, J. B. (1998). Strategies for Sorting Melodic Incipits. *Computing in Musicology, Melodic Similarities: Concepts, Procedures, and Applications*, 11, 119–128. MIT Press, Cambridge, Massachusetts.

Ó Maidín, D. (1998). A Geometrical Algorithm for Melodic Difference. *Computing in Musicology, Melodic Similarities: Concepts, Procedures, and Applications*, 11, 65–72. MIT Press, Cambridge, Massachusetts.

Pauws, S. (2002). CubyHum: A Fully Operational Query by Humming System. *ISMIR 2002 Conference Proceedings*, (pp. 187–196).

Prechelt, L. & Typke, R. (2001). An Interface for Melody Input. *ACM Transactions on Computer-Human Interaction* 8(2), 133–149.

Répertoire International des Sources Musicales (RISM). Serie A/II, manuscrits musicaux après 1600. (1996) K. G. Saur Verlag, München, Germany. `http://rism.stub.uni-frankfurt.de`

Rubner, Y. *Source code for the Earth Mover's Distance software*. (1998). Retrieved April 1, 2003, from `http://robotics.stanford.edu/~rubner/emd/default.htm`

Schlichte, J. (1990). Der automatische Vergleich von 83 243 Musikincipits aus der RISM-Datenbank: Ergebnisse - Nutzen - Perspektiven. *Fontes artis musicae*, 37, 35–46.

Selfridge-Field, E. (Ed.) (1997). *Beyond MIDI: the handbook of musical codes.* Cambridge: MIT Press.

Vleugels, J. & Veltkamp, R. C. (2002) Efficient Image Retrieval through Vantage Objects. *Pattern Recognition*, 35(1) (pp. 69–80)