MIREX 2012 Symbolic Melodic Similarity: Hybrid Sequence Alignment with Geometric Representations

Julián Urbano, Juan Lloréns, Jorge Morato and Sonia Sánchez-Cuadrado
University Carlos III of Madrid
Department of Computer Science
jurbano@inf.uc3m.es llorens@inf.uc3m.es jmorato@inf.uc3m.es sscuadra@bib.uc3m.es

ABSTRACT

This short paper describes our five submissions to the 2012 edition of the MIREX Symbolic Melodic Similarity task. All five submissions rely on a geometric model that represents melodies as spline curves in the pitch-time plane. The similarity between two melodies is then computed with a sequence alignment algorithm between sequences of spline spans: the more similar the shape of the curves, the more similar the melodies they represent.
As in MIREX 2010 and 2011, our systems ranked first for all effectiveness measures used. However, this year there was only one competing system, so we employ this report mainly to describe and compare results within our systems.

1. INTRODUCTION

For the 2012 edition of the MIREX Symbolic Similarity task we submitted five systems. ULMS1-ShapeH is the exact same system that obtained the best results in the MIREX 2010 [6] and 2011 editions [8] (JU4-Shape and UL1-Shape back then, respectively). We submitted it again to evaluate it with a different set of queries and to serve as a baseline to measure possible improvements in our other algorithms.

Systems ULMS2-ShapeL and ULMS3-ShapeG are modified versions of ShapeH that use a different sequence alignment algorithm. In particular, they use a local and a global alignment algorithm, respectively. These are very common choices in the literature, so we submitted these versions to compare these alignment options with the hybrid approach we have followed so far in ShapeH.

ULMS4-ShapeTime is the same system as ShapeH, except that the top-k retrieved results are re-ranked using ULMS5-Time, which is the same as the UL3-Time system submitted last year and that was shown to be especially good at ranking results.

In MIREX 2010 and 2011 all our systems ranked first [2, 3]. In this MIREX 2012 edition the five systems again ranked at the very top, though we note that this year there was only one competing submission [4].

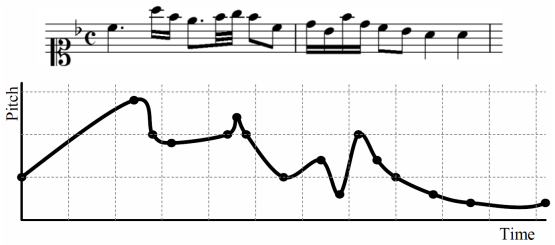

Figure 1. Melody as a curve in the pitch-time plane.

2. GEOMETRIC MELODY REPRESENTATION

Melodies are represented as curves in the pitch-time plane, arranging notes according to their pitch height and onset time. For the pitch dimension we use a directed interval representation, while for the time dimension we use the onset ratio between successive notes. We then calculate the interpolating curve passing through the notes (see Figure 1). From that point on, only the curves are used to compute the similarity between melodies [7].

We use Uniform B-Splines to interpolate through the notes [1], which gives us a parametric polynomial piecewise function for the spline: one function for the pitch dimension and another one for the time dimension. Their first derivatives measure how much the melodies change at any point. This representation is transposition invariant, as two transposed melodies have the same first derivative (see Figure 2). It is also time-scale invariant, as we use duration ratios within spline spans instead of actual durations.

A melody is thus represented as a sequence of spline spans, each of which can be considered the same as an n-gram. Given two arbitrary melodies, we compare them with a sequence alignment algorithm, which computes the differences between two spans based on their geometry.

3. SYSTEM DESCRIPTIONS

3.1 ShapeH, ShapeL and ShapeG

In these systems we completely ignore the time dimension and use spans 3-notes long, which result in splines defined by polynomials of degree 2. These are then differentiated, so we actually use polynomials of degree 1 to represent melodies. In addition, we implemented a heuristic very similar to the classical idf (Inverse Document Frequency)

System	Penalizes the beginning	Penalizes the end
ShapeH	yes	no
ShapeL	no	no
ShapeG	yes	yes

Table 1. Rough differences between the hybrid, local and global alignment approaches.

in Text Information Retrieval: the more frequent a spline span is in the document collection, the less important it is for the comparison of two melodies. Thus, the similarity between two spline spans is computed as follows:

- Insertion:
  $s(-, n) = -(1 - f(n))$.
- Deletion:
  $s(n, -) = -(1 - f(n))$.
- Match:
  $s(n, n) = 1 - f(n)$.

where $f(n)$ indicates the frequency of the spline span n in the document collection. For the substitution score we follow a naive rationale: if two spans have roughly the same shape they are considered the same, no matter how similar they actually are. For example, the polynomials $t^2 + 4$ and $0.5t^2 + 3t - 1$ are considered equal because they are both monotonically increasing. To this end, we only look at the direction of the splines at the beginning and at the end of the spans:

- If the two curves have the same derivative signs at the end and at the beginning of the span, the penalization is the smallest.
- If the two curves have opposite derivative signs at the end and at the beginning of the span, the penalization is the largest.
- If the two curves have the same derivative sign at one end of the span but not at the other, the penalization is averaged.

Because these splines are defined by polynomials of degree 2, they can change their direction just once within the span, so looking at the end points is enough.

3.1.1 Sequence Alignment

The only difference between these systems is in the sequence alignment algorithm they use. Let H be the dynamic programming table filled by the algorithm to compare sequences a and b. ShapeL employs a local alignment approach, where the score corresponding to an arbitrary cell is computed as:

$$H(i,j) = \max \left\{ \begin{array}{c} 0 \\ H(i-1, j-1) + s(a_i, b_j) \\ H(i-1, j) + s(a_i, -) \\ H(i, j-1) + s(-, b_j) \end{array} \right\}$$

and the bottom-right cell corresponds to the similarity between the two sequences. On the other hand, ShapeG employs a global alignment approach, where the score of an arbitrary cell is computed as:

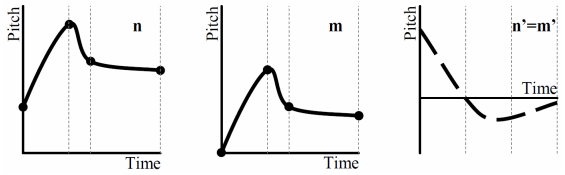

Figure 2. Transposition invariance with the derivatives.

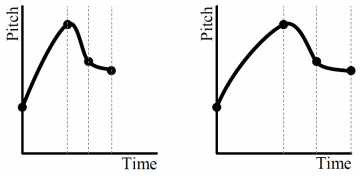

Figure 3. Time normalization in system Time. The span in the left side is transformed into the span in the right side.

$$H(i,j) = \max \left\{ \begin{array}{c} H(i-1, j-1) + s(a_i, b_j) \\ H(i-1, j) + s(a_i, -) \\ H(i, j-1) + s(-, b_j) \end{array} \right\}$$

In the ShapeH system we employ a variant of the global alignment approach, where the similarity between the two sequences corresponds to the maximum score in the table, regardless of its position. Very roughly speaking, the main difference between the three systems is that ShapeL does not penalize differences at the beginning of the sequences, while ShapeG and ShapeH do; and ShapeL and ShapeH allow differences at the end of the sequences, but ShapeG does not (see Table 1). With this hybrid approach we therefore assume that human listeners pay attention to the beginning of the melodies, but not to the end.

3.2 Time

This system uses spans 4-notes long, which result in spline spans defined with polynomials of degree 3. These are then differentiated, so we actually use polynomials of degree 2 to represent melodies. The similarity function between two spline spans does take the time dimension into account:

- Insertion:
  $s(-, n) = -\mathit{diff}_p(n, \phi(n)) - \lambda k_t \cdot \mathit{diff}_t(n, \phi(n))$.
- Deletion:
  $s(n, -) = -\mathit{diff}_p(n, \phi(n)) - \lambda k_t \cdot \mathit{diff}_t(n, \phi(n))$.
- Substitution:
  $s(n, m) = -\mathit{diff}_p(n, m) - \lambda k_t \cdot \mathit{diff}_t(n, m)$.
- Match:
  $s(n, n) = 2\mu_p + 2\lambda k_t \mu_t = 2\mu_p(1 + k_t)$.

where $\mathit{diff}_p(n, m)$ and $\mathit{diff}_t(n, m)$ measure the area between the first derivatives of the two spans' pitch and time functions; $\phi(n)$ is a function returning a span like n but with no change in pitch, so that $-\mathit{diff}_p(n, \phi(n))$ actually compares n with the x axis. The constants μ_p and μ_t are the mean scores returned by the diff_p and diff_t functions over a random sample of 100,000 pairs of spline spans

drawn from the Essen Collection ($\mu_p = 2.1838$ and $\mu_t = 0.4772$) [7]; $k_t = 0.5$ is a constant that weights the time dissimilarity with respect to the pitch dissimilarity; and $\lambda = \mu_p/\mu_t$ is a constant that normalizes time dissimilarity scores with respect to the pitch dissimilarity scores. This normalization is used because time dissimilarity scores use to be between 5 and 7 times smaller than pitch dissimilarity scores, so that weighting by k_t alone can be deceiving [7].

This system is transposition invariant as well. Also, span durations are normalized to length 1, so it is also time-scale invariant. For example, the first note in the left-most span in Figure 3 is kept in position 0, the second note is actually moved to the right up to position 1/2, the third note is moved up to position 3/4, and the fourth note is moved to the end (position 1). This system is thus transposition and time-scale invariant.

3.3 ShapeTime

This system is an extension of ShapeH. Last year, we saw that the Time system performed very well for the rank-aware measures (e.g. ADR), while the Shape system performed better for the rank-unaware measures (e.g. $Fine$). Therefore, we decided to submit the ShapeTime variant this year, which basically runs ShapeH and then re-ranks the top-k documents according to Time.

4. RE-RANKING

The sequence alignment algorithms may return the same similarity score for different documents, so a re-ranking process is run to solve ties. For every document in a tie, the corresponding sequence alignment algorithm is run again, but with an absolute pitch representation instead. Therefore, all transposition-equivalent documents that ranked equally are re-arranged with this process, ranking first those less transposed from the query. Note that the re-ranking process in ShapeTime is different (see Section 3.3).

5. RESULTS

Table 2 shows and excerpt of the official MIREX 2012 results [4], with the overall scores for the systems described here [1]. The bottom row shows the median rank for each system. In general, the ShapeTime system does indeed outperform the others; and ShapeH does return again more relevant material than Time, but then fails at ranking it properly. In addition, we see that the hybrid alignment approach clearly outperforms the local and global versions.

5.1 Sequence Alignment

As shown in Table 2, the hybrid alignment approach clearly outperforms the local and global alternatives for all effectiveness measures. In fact, the relative performance is always the same: the global algorithm outperforms the local version, and both are significantly outperformed by the hybrid alternative. As Table 3 shows, the relative difference is about 15% with ShapeG and around 30% with ShapeL.

[1] The scores here do not exactly match the official scores in the MIREX site because we normalize between 0 and 1 to make discussion easier and comparable with previous years.

	ShapeH	ShapeL	ShapeG
ADR	0.609	0.483 (-21%)	0.542 (-11%)
$NRGB$	0.534	0.428 (-20%)	0.471 (-12%)
AP	0.532	0.273 (-49%)	0.418 (-21%)
PND	0.524	0.327 (-38%)	0.446 (-15%)
$Fine$	0.629	0.496 (-21%)	0.546 (-13%)
$PSum$	0.680	0.467 (-31%)	0.582 (-14%)
$WCSum$	0.629	0.391 (-38%)	0.532 (-15%)
$SDSum$	0.603	0.353 (-41%)	0.508 (-16%)
$Greater0$	0.833	0.693 (-17%)	0.730 (-12%)
$Greater1$	0.527	0.240 (-54%)	0.433 (-18%)

Table 3. Differences in performance between the hybrid and the local and global sequence alignment algorithms.

5.2 Time-based Re-Ranking

As shown in Table 2, ShapeH retrieves slightly more relevant material than Time and, as expected, pretty much the same as ShapeTime. We note that the rank-unaware scores are not exactly the same between ShapeH and ShapeTime because the latter also re-ranks those documents beyond the top-k that are tied with the k-th document, which can ultimately lead to a slight change in what documents are actually retrieved in the top-k. Most importantly, we see that re-ranking with the Time algorithm does indeed improve results acrross measures, especially when taking the ranking into account. For instance, there is an improvement of 8% in $NRGB$ and as much as 10% in ADR.

6. CONCLUSIONS

We have submitted five systems to the 2012 edition of the MIREX Symbolic Melodic Similarity task. Our systems again ranked at the top, but there was only one more competing team this year [4]. Nonetheless, we observed two improvements as expected. On the one hand, we obtained better performance when using a hybrid sequence alignment algorithm as opposed to the local and global versions traditionally employed. On the other hand, we obtained better performance when re-ranking the top-k results using the time dimension, as opposed to just the pitch dimension.

With the results of this new edition, our approach of melodic similarity through shape similarity seems to work very well across collections. In fact, these systems have obtained the best results reported to date for the MIREX 2005 [7], 2010 [2], 2011 [3] and 2012 [4] test collections.

After three editions evaluating the ShapeH algorithm (2010, 2011 and 2012), we make an observation regarding the evaluation framework. In terms of ADR and AP scores, the results obtained have been 0.371, 0.651 and 0.609; and 0.349, 0.626 and 0.532, respectively [2–4]. That is, there have been very large differences across years, showing a clear reliability problem in the current evaluation framework [9]. We can not calculate confidence intervals

	ShapeH	ShapeL	ShapeG	ShapeTime	Time
ADR	0.609 (3)	0.483 (5)	0.542 (4)	0.671 (1)	0.657 (2)
$NRGB$	0.534 (3)	0.428 (5)	0.471 (4)	0.579 (1)	0.567 (2)
AP	0.532 (2)	0.273 (5)	0.418 (4)	0.541 (1)	0.487 (3)
PND	0.524 (1)	0.327 (5)	0.446 (4)	0.516 (2)	0.487 (3)
$Fine$	0.629 (2)	0.496 (5)	0.546 (4)	0.635 (1)	0.626 (3)
$PSum$	0.680 (2)	0.467 (5)	0.582 (4)	0.685 (1)	0.663 (3)
$WCSum$	0.629 (2)	0.391 (5)	0.532 (4)	0.636 (1)	0.609 (3)
$SDSum$	0.603 (2)	0.353 (5)	0.508 (4)	0.611 (1)	0.582 (3)
$Greater0$	0.833* (1)	0.693 (5)	0.730 (4)	0.833* (1)	0.827 (3)
$Greater1$	0.527 (2)	0.240 (5)	0.433 (4)	0.537 (1)	0.500 (3)
Median rank	2	5	4	1	3

Table 2. MIREX 2012 overall results for our five systems, normalized to the range 0 to 1. Ranks per effectiveness measure are in parentheses. * for ties.

on those average scores because neither the raw system outputs nor the per-query scores are available [5], but such large differences across years (up to 75% in ADR and 80% in AP), clearly show that 30 queries are just too few to have reliable estimates of true performance. In fact, in the current framework only 6 queries are used, with four artificial changes that then count to 30 queries. Therefore, we can actually consider the evaluation as using only 6 queries. In previous work we showed that the number of queries used in the Audio Music Similarity task can be greatly reduced [9], and in fact it has dropped from 100 to 50 in the MIREX 2012 campaign. The evidence suggests that the Symbolic Melodic Similarity task is using too few queries, so we propose to use some of the leftover manpower from AMS to evaluate more queries in further editions of the SMS task.

7. REFERENCES

[1] C. de Boor. *A Practical guide to Splines*. Springer, 2001.

[2] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2010 Symbolic Melodic Similarity Results, 2010.

[3] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2011 Symbolic Melodic Similarity Results, 2011.

[4] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2012 Symbolic Melodic Similarity Results, 2012.

[5] J. Urbano. Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain. In *International Society for Music Information Retrieval Conference*, pages 609–614, 2011.

[6] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado. MIREX 2010 Symbolic Melodic Similarity: Local Alignment with Geometric Representations. Technical report, Music Information Retrieval Evaluation eXchange, 2010.

[7] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado. Melodic Similarity through Shape Similarity. In S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, editors, *Exploring Music Contents*, pages 338–355. Springer, 2011.

[8] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado. MIREX 2011 Symbolic Melodic Similarity: Sequence Alignment with Geometric Representations. Technical report, Music Information Retrieval Evaluation eXchange, 2011.

[9] J. Urbano, D. Martín, M. Marrero, and J. Morato. Audio Music Similarity and Retrieval: Evaluation Power and Stability. In *International Society for Music Information Retrieval Conference*, pages 597–602, 2011.