MIREX 2013 Symbolic Melodic Similarity: A Geometric Model supported with Hybrid Sequence Alignment

Julián Urbano
Universitat Pompeu Fabra
Barcelona, Spain
julian.urbano@upf.edu

ABSTRACT

This short paper describes our three submissions to the 2013 edition of the MIREX Symbolic Melodic Similarity task. All three 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 the previous MIREX 2010, 2011 and 2012 editions, our systems ranked first for all effectiveness measures.

1. INTRODUCTION

For the 2013 edition of the MIREX Symbolic Melodic Similarity task we submitted the same three systems as last year. JU1-ShapeH is the exact same system that obtained the best results in the MIREX 2010 [7] and 2011 editions [9] (JU4-Shape and UL1-Shape back then, respectively), and the second best results in 2012 [10]. We submitted it again to evaluate it with a different set of queries and to serve as a strong and cross-year baseline to measure possible improvements in other algorithms.

The second submission is called JU2-ShapteTime, and it contains the same system as ULMS4-ShapeTime in 2012. It works like ShapeH, except that the top-k retrieved results are further re-ranked using the third system, called JU3-Time (the same as ULMS5-Time in 2012 and UL3-Time in 2011). This system was shown to be especially good at ranking results, so it is used to complement ShapeH for rank-aware measures.

In MIREX 2010, 2011 and 2012 all our systems ranked first [2–4]. In this MIREX 2013 edition the three systems again ranked at the very top.

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 [8].

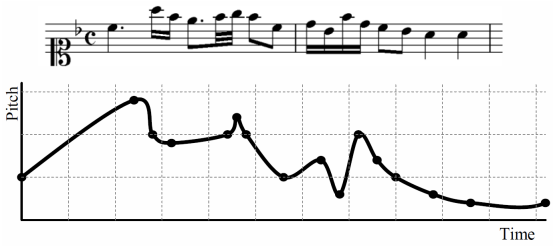

Figure 1. Melody as a curve in the pitch-time plane.

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

In this system 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) 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:

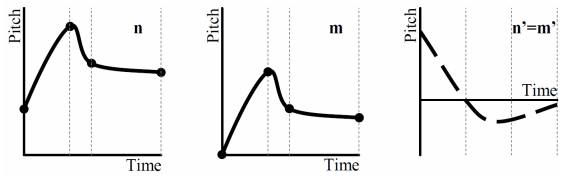

Figure 2. Transposition invariance with the derivatives.

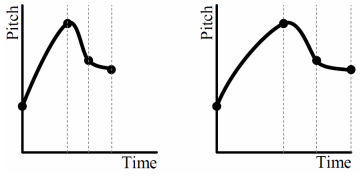

Figure 3. Time normalization in system Time. The span in the left side is transformed into the span in the right side.

- 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

A hybrid sequence alignment algorithm is used to compare splines [10]. This algorithm penalizes changes at the beginning of two melodies, but not at the end. Let H be the dynamic programming table filled by a global alignment algorithm to compare sequences a and b. The score of an arbitrary cell (i, j) is computed as:

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

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. 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 $\mu_t = 0.4772$) [8]; $k_t = 0.5$ 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 [8].

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. In MIREX 2011 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. $Fine$) [9]. In 2012 we decided to submit the ShapeTime variant, which basically runs ShapeH and then re-ranks the

	ShapeH		ShapeTime		Time	
ADR	0.734	(3)	0.794	(2)	0.798	(1)
$NRGB$	0.697	(3)	0.756	(1)	0.744	(2)
AP	0.690	(3)	0.708	(1)	0.694	(2)
PND	0.719	(1)	0.706	(2)	0.688	(3)
$Fine$	0.656	(1)	0.655	(2)	0.645	(3)
$PSum$	0.722	(1)	0.718	(2)	0.715	(3)
$WCSum$	0.673	(2)	0.676	(1)	0.668	(3)
$SDSum$	0.649	(2)	0.654	(1)	0.644	(3)
$Greater0$	0.867	(1)	0.847	(3)	0.857	(2)
$Greater1$	0.577	(2)	0.590	(1)	0.573	(3)
Median rank	2		1.5		3	

Table 1. MIREX 2013 overall results for our three systems, normalized between 0 and 1. Ranks per measure in parentheses. Measures at the top are rank-aware, measures at the bottom are not.

top-k documents according to Time [10]. This year we repeated this submission to confirm this observation.

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 1 shows an excerpt of the official MIREX 2013 results [5], with the overall scores for the systems described here [1]. The bottom row shows the median rank for each system. Although results are very similar across systems, ShapeTime does generally outperform the others; and ShapeH does return again more relevant material than Time, but then fails at ranking it properly. 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. In comparative terms, these are the same results observed last year [10].

Compared to the next best system by other participants, ShapeH obtained an average improvement of 389% in rank-aware measures and 232% in rank-unaware measures.

[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.

6. CONCLUSIONS

We have submitted three systems to the 2013 edition of the MIREX Symbolic Melodic Similarity task. Our systems again ranked at the top for all measures [5]. In general, the results obtained this year confirm the conclusions from last year: better performance is achieved when retrieving according to pitch alone and then re-ranking the top-k results using the time dimension, as opposed to using just one or another or both at the same time. This means that comparing pitch sequences performs best alone, but including time information further improves the ordering of documents.

With the results of this new edition, our approach of melodic similarity through shape similarity is confirmed to work very well across collections. In fact, these systems have obtained the best results reported to date for the MIREX 2005 [8], 2010 [2], 2011 [3], 2012 [4] and 2013 [5] test collections.

After four editions evaluating the ShapeH algorithm we again make an observation regarding the evaluation framework. In terms of ADR and AP scores, the results obtained have been 0.371, 0.651, 0.609 and 0.794; and 0.349, 0.626, 0.532 and 0.708, respectively [2–5]. That is, there have been extremely large differences across years *for the same system*, showing a clear problem in the current evaluation framework [13]. We can not calculate confidence intervals on those average scores because neither the raw system outputs nor the per-query scores are available, but such large differences across years (up to 214% in ADR and 203% in AP), clearly show that either a) 30 queries are just too few to have reliable estimates of true performance, or b) the query selection method is not valid (probably not random).

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 reduced [6, 11], and in fact it has dropped from 100 to 50 in the MIREX 2012 and 2013 editions. In addition, we showed that the annotation effort required to make judgments can also be reduced to less than 5% [6, 12]. The evidence thus 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. ACKNOWLEDGMENTS

This work is supported by an A4U postdoctoral grant.

8. 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] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2013 Symbolic Melodic Similarity Results, 2013.

[6] J. Urbano. *Evaluation in Audio Music Similarity*. PhD thesis, University Carlos III of Madrid, 2013.

[7] 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.

[8] 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.

[9] 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.

[10] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado. MIREX 2012 Symbolic Melodic Similarity: Hybrid Sequence Alignment with Geometric Representations. Technical report, Music Information Retrieval Evaluation eXchange, 2012.

[11] 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.

[12] J. Urbano and M. Schedl. Minimal Test Collections for Low-Cost Evaluation of Audio Music Similarity and Retrieval Systems. *International Journal of Multimedia Information Retrieval*, 2(1):59–70, 2013.

[13] J. Urbano, M. Schedl, and X. Serra. Evaluation in Music Information Retrieval. *Journal of Intelligent Information Systems*, 2013.