MelodyShape at MIREX 2014 Symbolic Melodic Similarity

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

ABSTRACT

This short paper describes our three submissions to the 2014 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, 2012 and 2013 editions, our systems ranked first for all effectiveness measures. The main difference with last year is that we submitted a re-implementation of all algorithms, contained in the new open source library MelodyShape.

1. INTRODUCTION

For the 2014 edition of the MIREX Symbolic Melodic Similarity task we submitted the same three algorithms as last year. JU1-ShapeH implements the same algorithm that has consistently obtained the best or second-best results in MIREX 2010–2013. The second submission is called JU2-ShapteTime, and it contains the same algorithm as in MIREX 2013 and 2012. It works like ShapeH, except that the top-k retrieved results are further re-ranked using the third system, called JU3-Time (also submitted in MIREX 2013 and 2012). This system was shown to be especially good at ranking results, so it is used to complement ShapeH for rank-aware measures.

We submitted these algorithms again to evaluate them with a different set of queries and assessors, and to serve as strong and cross-year baselines to measure possible improvements in other submissions. In addition, we recently published MelodyShape [1] [8], a Java tool an library implementing all our algorithms since 2010 from their original code in C#. The three submissions to MIREX 2014 are the exact implementations found in MelodyShape v1.1.

In MIREX 2010, 2011, 2012 and 2013 all our systems ranked first [2–5]. In this MIREX 2014 edition the three systems again ranked at the very top [6].

[1] https://github.com/julian-urbano/MelodyShape

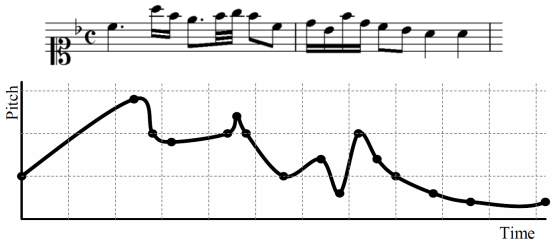

Figure 1. Melody as a curve in the pitch-time plane.

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 [10].

We use Uniform B-Splines to interpolate through the notes [1], which give 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 because two transposed melodies have the same first derivative (see Figure 2). It is also time-scale invariant because 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

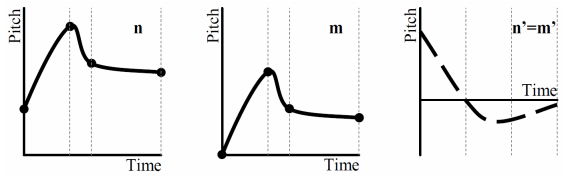

Figure 2. Transposition invariance with the derivatives.

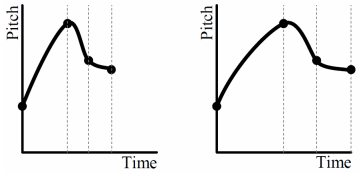

Figure 3. Time normalization in system Time. The span in the left side is transformed into the span in the right side.

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

A hybrid sequence alignment algorithm is used to compare splines [12]. 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:
  $s(-, n) = -diff_p(n, \phi(n)) - \lambda k_t \cdot diff_t(n, \phi(n))$.
- Deletion:
  $s(n, -) = -diff_p(n, \phi(n)) - \lambda k_t \cdot diff_t(n, \phi(n))$.
- Substitution:
  $s(n, m) = -diff_p(n, m) - \lambda k_t \cdot diff_t(n, m)$.
- Match:
  $s(n, n) = 2\mu_p + 2\lambda k_t \mu_t = 2\mu_p(1 + k_t)$.

where $diff_p(n, m)$ and $diff_t(n, m)$ 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 $-diff_p(n, \phi(n))$ 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 $\mu_t = 0.4772$) [10]; $k_t = 0.5$ 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 [10].

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. $Fine$) [11]. In 2012 we decided to submit the ShapeTime variant, which basically runs ShapeH and then re-ranks the top-k documents according to Time [12]; the

	ShapeH		ShapeTime		Time	
NRGB	0.679	(3)	0.749	(2)	0.760	(1)
AP	0.734	(3)	0.753	(2)	0.799	(1)
PND	0.736	(3)	0.744	(2)	0.761	(1)
Fine	0.538	(2)	0.546	(1)	0.512	(3)
PSum	0.558	(3)	0.565	(1)	0.563	(2)
WCSum	0.501	(3)	0.504	(2)	0.514	(1)
SDSum	0.473	(3)	0.474	(2)	0.490	(1)
Greater0	0.730	(2)	0.747	(1)	0.710	(3)
Greater1	0.387	(2)	0.383	(3)	0.417	(1)
Median rank	3		2		1	

Table 1. MIREX 2014 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.

average improvement in rank-aware measures was 4.7%. In 2013 we repeated this submission to confirm this observation, and found an average improvement in rank-aware measures of 4.4% [9]. This year we repeated this submission again to obtain more data.

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 2014 results [6], with the overall scores for the systems described here [2]. The bottom row shows the median rank for each system. Although results are quite similar across systems, Time does generally outperform the others, and ShapeH returns this time the least relevant material. These results directly contradict those in 2012 [12] and 2013 [9], where ShapeH retrieved more relevant material but then failed to rank it properly; ShapeTime did then provide the best ranking. This year, Time retrieved slightly more highly-relevant documents than the others to begin with, and then ranked them correctly. 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 at the bottom.

[2] 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.

Compared to the system by the other participant, Time obtained an average improvement of 29% in rank-aware measures and 60% in rank-unaware measures.

6. DISCUSSION

We have submitted three systems to the 2014 edition of the MIREX Symbolic Melodic Similarity task. Our systems again ranked at the top for all measures [6]. 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, 2010, 2011, 2012, 2013 and 2014 collections [2–6, 10].

However, the results obtained this year contradict the conclusions from the last two years. We observed better performance 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 year though, the best results were obtained when retrieving *and* ranking using both pitch and time; re-ranking with time the top-k retrieved with pitch alone did improve ranking too.

These contradictions confirm our comments from last year that the collections used in this task are unreliable [9]. Our ShapeH system has been evaluated in all five editions since 2010, and it has obtained average performance scores that differ in over 200% from year to year. This year the general results contradict the past two editions. We can not calculate confidence intervals on the average scores or test statistical significance because neither the raw system outputs nor the per-query scores are available. Notwithstanding, such large differences across years clearly show one of two problems. First, that the query selection method is not valid (probably not random). Since the musical content of the queries is hidden as well, we cannot verify this point. Second, that 30 queries are just too few to have reliable estimates of true performance in this task. 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.

We employed the GT4IReval [3] [7] tool to run a quick analysis of reliability with Generalizability Theory, using the available *Fine* scores from 2012, 2013 and 2014. The results indicate that, in the particular case of our three algorithms, we would need over 500 queries to reliably detect differences. This means that either a) the algorithms are indeed *very* similar to each other in terms of *Fine* scores (rank-unaware), or b) the query selection method is indeed very biased, leading to invalid results. The evidence suggests again that the Symbolic Melodic Similarity task is using too few queries.

7. ACKNOWLEDGMENTS

This work was supported by an A4U postdoctoral grant and Sergio Ramos' goal in the 93rd minute that lead us to win La Décima.

[3] http://github.com/julian-urbano/GT4IREval

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. Online: http://www.music-ir.org/mirex/wiki/2010:Symbolic_Melodic_Similarity_Results, 2010.

[3] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2011 Symbolic Melodic Similarity Results. Online: http://www.music-ir.org/mirex/wiki/2011:Symbolic_Melodic_Similarity_Results, 2011.

[4] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2012 Symbolic Melodic Similarity Results. Online: http://www.music-ir.org/mirex/wiki/2012:Symbolic_Melodic_Similarity_Results, 2012.

[5] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2013 Symbolic Melodic Similarity Results. Online: http://www.music-ir.org/mirex/wiki/2013:Symbolic_Melodic_Similarity_Results, 2013.

[6] International Music Information Retrieval Systems Evaluation Laboratory. MIREX 2014 Symbolic Melodic Similarity Results. Online: http://www.music-ir.org/mirex/wiki/2014:Symbolic_Melodic_Similarity_Results, 2014.

[7] J. Urbano. GT4IREval: An R package to Measure the Reliability of an Information Retrieval Test Collection with Generalizability Theory. Online: http://github.com/julian-urbano/GT4IREval, 2013.

[8] J. Urbano. MelodyShape: a Library and Tool for Symbolic Melodic Similarity based on Shape Similarity. Online: https://github.com/julian-urbano/MelodyShape, 2013.

[9] J. Urbano. MIREX 2013 Symbolic Melodic Similarity: A Geometric Model supported with Hybrid Sequence Alignment. Technical report, Music Information Retrieval Evaluation eXchange, 2013.

[10] 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.

[11] 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.

[12] 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.