MIREX 2011 SYMBOLIC MELODIC SIMILARITY: 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 jorge@ie.inf.uc3m.es ssanchec@ie.inf.uc3m.es

ABSTRACT

This short paper describes our three submissions to the MIREX 2011 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 MIREX 2010, our systems ranked first for all effectiveness measures used.

1. INTRODUCTION

For the MIREX 2011 edition of the Symbolic Similarity task, we submitted three systems. *UL1-Shape* is the exact same system that obtained the best results in the MIREX 2010 edition (*JU4-Shape* back then) [6]. We decided to submit it again to evaluate it with a different collection and to serve as a baseline to measure possible improvements in our other algorithms. The second system, *UL2-Pitch* is a modified version of the *JU2-PitchDeriv* system we submitted last year, which obtained the second best results overall. The third system, *UL3-Time* is a modified version of the *JU3-ParamDeriv* system we submitted last year, which obtained the third best results overall. In this MIREX 2011 edition, the three systems again ranked first, second and third respectively [3].

In the next section we briefly describe our geometric model, and in Section 3 we detail our three systems. Section 4 shows the re-ranking phase, and Section 5 discusses the results. The paper then finishes with the conclusions in Section 6.

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].

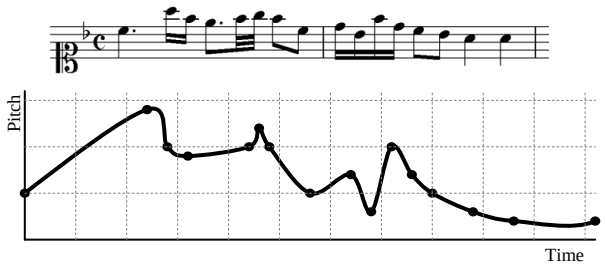

Figure 1. Melody represented as a curve in the pitch-time plane.

We used Uniform B-Splines to interpolate through the notes [1]. This 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 are changing 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, because we use duration ratios within spline spans instead of actual durations.

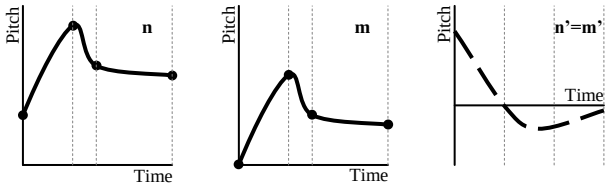

Figure 2. Transposition invariance with the first derivatives.

3. SYSTEM DESCRIPTIONS

3.1 UL2-Pitch

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 is defined as follows:

- Insertion: $s(-, n) = -diff_p(n, \phi(n))$.
- Deletion: $s(n, -) = -diff_p(n, \phi(n))$.
- Substitution: $s(n, m) = -diff_p(n, m)$.
- Match: $s(n, n) = 2\mu_p$.

where $diff_p(n, m)$ measures the area between the first derivatives of the two spans' pitch 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; and μ_p is the mean score returned by the $diff_p$ function over a random sample of 100,000 pairs of spline spans drawn from the Essen Collection ($\mu_p = 2.1838$) [7]. This system thus ignores the time dimension altogether, but it remains transposition invariant.

3.2 UL3-Time

This system employs the same representation and rationale as UL2-Pitch, but it does take the time dimension into account. The similarity function is:

- Insertion: $s(-, n) = -diff_p(n, \phi(n)) - \lambda k_t diff_t(n, \phi(n))$.
- Deletion: $s(n, -) = -diff_p(n, \phi(n)) - \lambda k_t diff_t(n, \phi(n))$.
- Substitution: $s(n, m) = -diff_p(n, m) - \lambda k_t diff_t(n, m)$.
- Match: $s(n, n) = 2\mu_p(1+k_t)$.

where $diff_t(n, m)$ measures the area between the first derivatives of the two spans' time functions; $k_t = 0.5$ 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 ($\mu_t = 0.4772$ for the Essen Collection). 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].

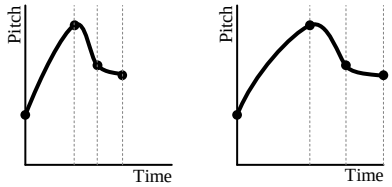

Figure 3. Time normalization in UL3. The span in the left side is transformed into the span in the right side.

Span durations are normalized to length 1, so the system 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 UL1-Shape

In this system we also ignore the time dimension, but in this case we use spans 3-notes long, which result in splines defined by polynomials of degree 2. 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:

- 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. In this case a more naive rationale is followed for the substitution score: 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. This system is also transposition invariant.

4. RE-RANKING

The sequence alignment algorithm may return the same similarity score for different documents, so a ranking process is performed to solve ties. For every document in a tie, the sequence alignment algorithm is run again, but with an absolute representation instead. Therefore, transposition-equivalent documents that ranked equally would be re-arranged with this process, ranking first those less transposed from the query.

5. RESULTS

Table 1 shows and excerpt of the official MIREX 2011 results [3], with the overall figures for the systems described. Notably, our three systems ranked in the top 3 for all 10 effectiveness measures, except for *UL1* with ADR and NRGB. Most importantly, the best score in all measures was obtained by one of our systems.

	UL1-Shape	UL2-Pitch	UL3-Time
ADR	0.651 (9)	0.675 (3)	0.726 (1)
NRGB	0.627 (10)	0.651 (3)	0.696 (1)
AP	0.626 (1)	0.624 (2)	0.612 (3)
PND	0.663 (1)	0.633 (2)	0.621 (3)
Fine	0.594 (1)	0.568 (2)	0.552 (3)
Psum	0.543 (1)	0.519 (2)	0.511 (3)
WCsum	0.615 (1)	0.575 (2)	0.572 (3)
SDsum	0.508 (1)	0.491 (2)	0.481 (3)
Greater0	0.83 (1)	0.743 (3)	0.753 (2)
Greater1	0.4 (2)	0.407 (1)	0.39 (3)
Median Rank	1	2	3

Table 1. MIREX 2011 overall results for our three systems. Ranks per effectiveness measure are in parentheses. * for ties.

The bottom row shows the median rank for each system. Like in MIREX 2010, *Shape* ranks first, followed by *Pitch* and then by *Time*. This pattern is held for most measures, with the glaring exception of ADR and NRGB.

All eleven systems submitted by ourselves and other participants achieved an ADR score between 0.646 and 0.676, and a NRGB score between 0.626 and 0.652. However, our *UL3-Time* system obtained scores well above those: 0.726 (+10.2%) for ADR and 0.696 (+9.2%)

for NRGB. We are unsure as to why *UL3-Time* worked that exceptionally well compared to the other systems. One explanation could be the test collections used to develop all these systems: both our three submissions [7] and the other participant's [4][8] where trained with the MIREX 2005 collections, which are evaluated with ADR. It could be possible that the highly similar scores achieved by all systems with ADR and NRGB were just a product of all systems being trained with the same measure. The choice of spans 4-notes long and the particular values for k_t, μ_p and μ_t used in *UL3-Time* was based on our experiments with that same collection and measure, so the scores could be that large due to some (unlikely) overfitting to those measures. On the other hand, it could just be that the weighted and normalized use of the time dimension proved to be informative to rank results. Not knowing the actual queries used, we would like to further explore the results and run other experiments before drawing any conclusion [5].

Most importantly, the *Shape* system (UL1) is ranked first for most effectiveness measures. Table 2 shows the scores of the best system compared with the second best system, excluding our two others, for each effectiveness measure. Unlike last year [2], the improvements are much larger with set-based measures than with rank-based measures. This could again be caused by all systems being trained with the rank-based measure ADR.

	Best UL	2nd best system	Improvement
ADR	0.726 (UL3)	0.676 (WK1)	7.3%
NRGB	0.696 (UL3)	0.652 (WK1)	6.7%
AP	0.626 (UL1)	0.610 (WK1)	2.6%
PND	0.663 (UL1)	0.605 (WK1)	9.5%
Fine	0.594 (UL1)	0.515 (WK1)	15.3%
Psum	0.543 (UL1)	0.457 (WK1)	18.8%
WCsum	0.615 (UL1)	0.497 (WK1)	23.7%
SDsum	0.508 (UL1)	0.437 (WK1)	16.2%
Greater0	0.83 (UL1)	0.657 (LJY2)	26.3%
Greater1	0.407 (UL2)	0.377 (WK1)	8%

Table 2. Scores of the best UL system per measure, compared with the second best scores, excluding our other two systems.

In any case, it strikes us that the *Shape* system, using the simplest technique, obtained the overall best results in MIREX 2010 and 2011. While the results in two different collections seem to support it as the system to go for, we would like to further explore these differences with other collections and more diverse data.

6. CONCLUSIONS

We have submitted three systems to the 2011 edition of the MIREX Symbolic Melodic Similarity task. Among the 11 systems evaluated this year, our three systems ranked always in the top 3 for all 10 effectiveness measures calculated, except for two cases. Repeating the MIREX 2010 results, our *Shape* system worked best overall, although this time the normalized and weighted use of the time dimension greatly improved the ranking of documents.

With the results of this 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 ever reported for the MIREX 2005 [7], MIREX 2010 [2] and MIREX 2011 [3] test collections.

REFERENCES

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

[2] IMIRSEL, "MIREX 2010 Symbolic Melodic Similarity Results," http://www.music-ir.org/mirex/wiki/2010:Symbolic_Melodic_Similarity_Results.

[3] IMIRSEL, "MIREX 2011 Symbolic Melodic Similarity Results," http://www.music-ir.org/mirex/wiki/2011:Symbolic_Melodic_Similarity_Results.

[4] J. Lee, S. Jo, and C.D. Yoo, "Coded Melodic Contour Model," *Music Information Retrieval Evaluation eXchange*, 2011.

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

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

[7] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado, "Melodic Similarity through Shape Similarity," in *Exploring Music Contents*, S. Ystad, M. Aramaki, R. Kronland-Martinet, and K. Jensen, (eds.), Springer, pp. 338-355, 2011.

[8] J. Wolkowicz and V. Kešelj, "Text Information Retrieval Approach to Music Information Retrieval," *Music Information Retrieval Evaluation eXchange*, 2011.