MIREX 2010 SYMBOLIC MELODIC SIMILARITY: LOCAL 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 four submissions to the Symbolic Melodic Similarity task of the MIREX 2010 edition. All four submissions rely on a local-alignment approach between sequences of n-grams, and they differ mainly on the substitution score between two n-grams. This score is based on a geometric representation that shapes musical pieces as curves in the pitch-time plane. One of the systems described ranked first for all ten effectiveness measures used and the other three ranked from second to fifth, depending on the measure.

1. INTRODUCTION

The problem of Symbolic Melodic Similarity, where a retrieval system is expected to retrieve a ranked list of musical pieces deemed similar to another one (i.e. the query), has been approached from very different points of view [1]. Some techniques are based on geometric representations of music, others rely on classic n-gram representations to calculate similarities, and others use editing distances and alignment algorithms.

In a previous work we mixed these three major approaches [2]. We modeled melodies as sequences of overlapping n-grams of 3consecutive notes, and then they were compared using a modified version of the Smith-Waterman local-alignment algorithm [3]. The substitution score between two n-grams was calculated based on a geometric interpretation of the notes within the n-grams, which considers musical pieces as curves in the pitch-time plane. We have improved this approach and submitted four variations to the current 2010 edition of MIREX: *Domain*, *PitchDeriv*, *ParamDeriv* and *Shape*.

In the next section we describe the local-alignment approach we followed, discussing the insertion, deletion and match scores common to all four submissions. Section 3 describes how the substitution score is calculated in each case and Section 4 shows the re-ranking phase. Section 5 discusses the results and the paper then finishes with conclusions and discussion.

2. LOCAL-ALIGNMENT

We implemented a heuristic very similar to the classical TF-IDF (Term Frequency-Inverse Document Frequency) in Text Information Retrieval: the more frequent an n-gram is in the document collection, the less important it is for the comparison of two documents. Thus, the insertion, deletion and match scores between two n-grams are adapted as follows:

- Insertion: $s(-, n) = -(1 - f(n))$. An infrequent added n-gram penalizes more than a frequent one.
- Deletion: $s(n, -) = -(1 - f(n))$. An infrequent missed n-gram penalizes more than a frequent one.
- Match: $s(n, n) = 1 - f(n)$. An infrequent matched n-gram rewards more than a frequent one.

where $f(n)$ indicates the frequency of the n-gram n in the document collection. The representation schema used for the n-grams at this point is directed-interval.

3. SUBSTITUTION SCORES

The four systems submitted differ on the substitution function $s(n, m)$ used by the local-alignment algorithm. Next, we describe how they are calculated in each case.

3.1 JU1: Domain

The substitution score $s(n, m)$ is calculated as the average of the absolute values of the directed interval differences between the corresponding notes of the two n-grams. For example, $s(\langle 71, 70, 71\rangle, \langle 78, 73, 74\rangle)$ would be:

$$\frac{|(70-71)-(73-78)|+|(71-70)-(74-73)|}{2}=2$$

This system ignores completely the time dimension of music, but presents the advantage of being transposition invariant.

3.2 JU2: PitchDeriv

In this case, the n-grams are represented as curves in the pitch-time plane. Each note is arranged in the plane according to its pitch height and its onset time, and then we calculate the interpolating curve passing through the notes (see Figure 1). From that point on, only the curve is used to compare the n-gram to another one.

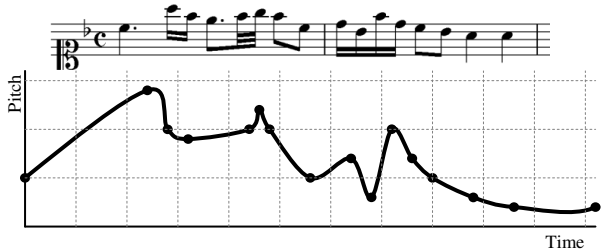

Figure 1. Melody represented as a curve in the pitch-time plane.

We used Uniform B-Splines to interpolate through the notes [4]. This gives us a parametric function for the spline: one function for the pitch dimension and another one for the time dimension. The first derivative of these functions measures how much the melody is changing at any point. This representation is also transposition invariant, as the curve of a transposed melody is the same, just shifted up or down. Therefore, the first derivatives are equal (see Figure 2).

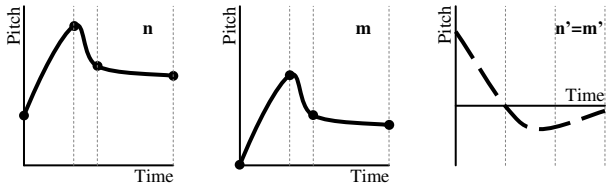

Figure 2. Transposition invariance with the first derivatives.

This system ignores the time dimension and only considers pitches. To this end, every note lasts the same. For example, the third note in the left-most n-gram in Figure 3 is actually moved to the right up to position 2/3. The substitution score in this system is calculated as the area between the first derivatives:

$$s(n, m)=\int |N'(t)-M'(t)|dt$$

where $N(t)$ and $M(t)$ are the pitch-wise interpolating functions of n and m respectively (see Figure 3).

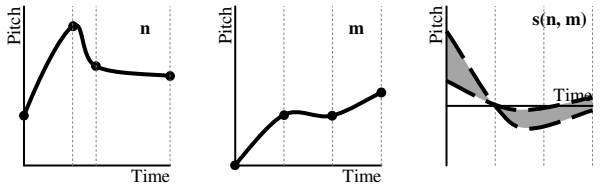

Figure 3. Comparison in JU2 with the area between derivatives.

3.3 JU3: ParamDeriv

In this system the time dimension is considered. The n-gram duration is normalized to 1, so the system is also time-scale invariant. For example, the second note in the left-most n-gram in Figure 4 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.

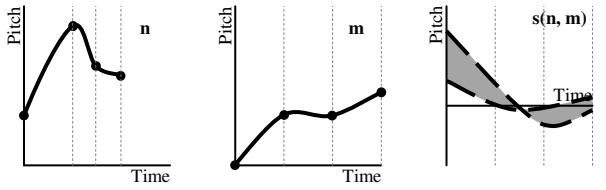

Figure 4. Comparison in JU3 with the area between derivatives.

The substitution score in this system is calculated as the area between the first derivatives of the parametric curves (see Figure 4).

3.4 JU4: Shape

In this case a more naive score is calculated. If two n-grams have the same shape they are considered the same, no matter how sharp they are. For example, the curves defined by the polynomials t^2 and $4t^2$ are considered the same, though the second one is sharper. To this end, we consider only the value of the first derivative at the beginning and the end of the n-gram:

- 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 use polynomials with low degrees, the curves cannot wiggle in the middle, so that considering the values at the beginning and at the end is sufficient. This system is also transposition invariant.

4. RE-RANKING

The local-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, a regular local-alignment algorithm is run also with n-grams, but with an absolute representation instead. Thus, transposition-equivalent documents that ranked equally will 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 results [5], with the overall figures for the systems described. Notably, all our four systems ranked in the top 5 for all 10 effectiveness measures (5th only in 4 of the 40 cases).

	JU1	JU2	JU3	JU4
ADR	0.307 (5)	0.309 (3)	0.317 (2)	0.371 (1)
NRGB	0.297 (3)	0.294 (4)	0.288 (5)	0.328 (1)
AP	0.300 (3)	0.299 (4)	0.301 (2)	0.349 (1)
PND	0.373 (2*)	0.373 (2*)	0.368 (4)	0.399 (1)
Fine	0.579 (5)	0.583 (2)	0.581 (3)	0.606 (1)
Psum	0.613 (4)	0.620 (2)	0.615 (3)	0.642 (1)
WCsum	0.559 (3*)	0.563 (2)	0.559 (3*)	0.580 (1)
SDsum	0.532 (3)	0.535 (2)	0.531 (4)	0.549 (1)
Greater0	0.777 (5*)	0.790 (3)	0.783 (4)	0.827 (1)
Greater1	0.450 (2*)	0.450 (2*)	0.447 (4)	0.457 (1)
Median Rank	3	2	3.5	1

Table 1. MIREX overall results for our four systems. Ranks per effectiveness measure appear in parentheses. * for tied ranks.

The bottom row shows the median rank for each system. Surprisingly, *ParamDeriv* (JU3) appears to perform the worst when considering the time dimension. While this could be caused by a inappropriate substitution score, it is interesting to see that, once again, the use of the time dimension does not improve the results.

On the other hand, it seems that the heuristic implement, mentioned in Section 2, does improve the local alignment approach.

Most importantly, the *Shape* system (JU4) is ranked first for all effectiveness measures. Table 2 shows the scores of JU4 compared with the second best system, excluding our three others, for each effectiveness measure.

	JU4	2nd best system	Improvement
ADR	0.371	0.308 (LL1)	21%
NRGB	0.328	0.299 (LL1)	10%
AP	0.349	0.256 (RI1)	36%
PND	0.399	0.353 (RI1)	13%
Fine	0.606	0.580 (LL1)	4%
Psum	0.642	0.600 (LL1)	7%
WCsum	0.580	0.548 (LL1)	6%
SDsum	0.549	0.522 (LL1)	5%
Greater0	0.827	0.797 (RI4)	4%
Greater1	0.457	0.443 (LL1)	3%

Table 2. Scores of JU4 compared with the second best scores, excluding our other three systems.

The improvement is much larger across the rank-based measures, which suggests that JU4 performs better not only in retrieving documents, but also in ranking them properly.

6. CONCLUSIONS AND DISCUSSION

We have submitted four systems to the 2010 edition of the MIREX Symbolic Melodic Similarity tasks. Among the 13 systems evaluated this year, our 4 systems ranked always in the top 5 for all 10 effectiveness measures calculated. One of them, *Shape* (JU4), always ranked the best of all 13 systems. These results support our approach of local-alignment with n-grams and the representation of melodies with curves, opening a new and promising line for further research.

Regarding the evaluation itself, we would like to propose further studies on two topics. First, the reconsideration of partially ordered lists [6] as the form of ground truth for the task. We have recently explored alternatives for their construction [7][8], which appear to mitigate some of their known problems [9]. Our current research is focused on more accurate and affordable ways to build partially ordered lists, which should lead to more robust and large-scale evaluations. Second, the consideration of using the results of rank-based measures such as ADR or AP instead of the set-based FINE score sum for finding significant differences among systems with the Friedman test. Rank-based measures are more suitable to model a real user, seeking the most relevant documents at the beginning of the results list. Being A a similar document and B a non-similar one, a system retrieving A after B is clearly better than one retrieving B after A. However, they would both have the same score with a set-based measure, whilst a rank-based measure would rank the former better than the later.

REFERENCES

[1] R. Typke, F. Wiering, and R.C. Veltkamp, "A Survey of Music Information Retrieval Systems," *International Conference on Music Information Retrieval*, 2005, pp. 153-160.

[2] J. Urbano, J. Lloréns, J. Morato, and S. Sánchez-Cuadrado, "Using the Shape of Music to Compute the Similarity between Symbolic Musical Pieces," *International Symposium on Computer Music Modeling and Retrieval*, 2010, pp. 385-396.

[3] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences," *Journal of Molecular Biology*, vol. 147, 1981, pp. 195-197.

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

[5] IMIRSEL, "MIREX 2010 Symbolic Melodic Similarity Results," 2010.

[6] R. Typke, M. den Hoed, J. de Nooijer, F. Wiering, and R.C. Veltkamp, "A Ground Truth for Half a Million Musical Incipits," *Journal of Digital Information Management*, vol. 3, 2005, pp. 34-39.

[7] J. Urbano, M. Marrero, D. Martín, and J. Lloréns, "Improving the Generation of Ground Truths based on Partially Ordered Lists," *International Society for Music Information Retrieval Conference*, 2010, pp. 285-290.

[8] J. Urbano, J. Morato, M. Marrero, and D. Martín, "Crowdsourcing Preference Judgments for Evaluation of Music Similarity Tasks," *SIGIR Workshop on Crowdsourcing for Search Evaluation*, 2010, pp. 9-16.

[9] J.S. Downie, A.F. Ehmann, M. Bay, and M.C. Jones, "The Music Information Retrieval Evaluation eXchange: Some Observations and Insights," *Advances in Music Information Retrieval*, W.R. Zbigniew and A.A. Wieczorkowska, Springer, 2010, pp. 93-115.