Measuring Musical Rhythm Similarity: Edit Distance versus Minimum-Weight Many-to-Many Matchings

Godfried T. Toussaint¹, and Seung Man Oh²
¹Computer Science Program, New York University Abu Dhabi, UAE
²Freelance IT Developer, Seoul, Korea

Abstract—Musical rhythms are represented as binary symbol sequences of sounded and silent pulses of unit-duration. A measure of distance (dissimilarity) between a pair of rhythms commonly used in music information retrieval, music perception, and musicology is the edit (Levenshtein) distance, defined as the minimum number of symbol insertions, deletions, and substitutions needed to transform one rhythm into the other. A measure of distance often used in object recognition is the minimum-weight many-to-many matching distance between the object’s features. These two approaches are compared empirically, in terms of how well they predict human judgments of musical rhythm similarity, using a real-world family of Middle-Eastern rhythms.

Keywords: musical rhythm, perception, edit distance, many-to-many matchings, similarity measures, Hungarian algorithm

1. Introduction

For a variety of applications such as Music Information Retrieval (MIR) [22], automated music generation and composition [23], phylogenetic analysis of rhythms [21], [16], playlist generation [2], and music perception [4], it is desirable to have a mathematical measure of rhythm similarity (or distance) that predicts, or correlates well with, human judgments. A plethora of distance measures exists for computing the dissimilarity of two pieces of music that depend on how music is modeled. The present study is concerned with symbolically notated musical rhythms (rather than acoustic inputs). Furthermore, rhythms are represented in their most skeletal form as sequences of binary symbols denoting unit duration pulses that are either sounded (called onsets) or silent (called rests). For example, the rhythm used by Steve Reich in his piece titled Clapping Music [24] is represented by [x x x . x x . x x .], where ‘x’ and ‘.’ denote, respectively, a sounded and a silent pulse.

1.1 The Edit Distance

A distance measure between a pair of rhythms commonly used in computational music is the edit (also Levenshtein [5]) distance, defined as the minimum number of symbol insertions, deletions, and substitutions needed to transform one rhythm into the other [18], [4]. An insertion of a sounded or silent pulse lengthens the duration of the rhythm, a deletion shortens its duration, and a substitution (reversal) of a sounded pulse by a silent pulse (and vice-versa) leaves the rhythm duration unaltered. Consider the two binary sequences A and B in Figure 1, expressed in box-notation, in which an empty box denotes a silent pulse and a filled box a sounded pulse. There are many possible series of edits that can transform A into B. One possibility is to delete silent pulses 8, 9, 10, and 11, followed by reversals of pulses 0, 1, 2, 5, 6, and 7, yielding a total of 10 edit operations. However, the minimum possible number of edits is 5, and may be obtained as illustrated in the figure. Sequence A* is obtained by deleting sounded pulse 0 in sequence A. Sequence A** is obtained by a reversal of pulse 7 in sequence A*, from a silent pulse to a sounded one. Finally, sequence B is obtained by deleting the last three silent pulses (8, 9, and 10) in sequence A**.

Pulse No. 0 1 2 3 4 5 6 7 8 9 10 11
Sequence-A

Sequence-B

Sequence-A*

Sequence-A**

Sequence-B

Fig. 1: The edit distance between sequences A and B is 5.

1.2 The Many-to-Many Matching Distance

A measure of distance frequently used in object recognition is the weight of the minimum-weight many-to-many matching between the features of one object and the features of the other [7]. Distances based on many-to-many matchings have also been applied to measure the similarity of two origami crease patterns corresponding to two different folded paper objects [10], to measure the difference between musical chords [11], and to compare relational data [6]. The many-to-many matching is a generalization of the simpler one-to-one and many-to-one matchings [12]. If two rhythms A and B have the same number of onsets, then the one-to-one matching assigns the i-th onset of A to the i-th onset of B, and the overall distance between A = (a₁, a₂, ..., aₙ) and B = (b₁, b₂, ..., bₙ), where the aᵢ and bᵢ are the x-coordinate
values of the onsets, is the sum over all i of the absolute values of the differences between the pulse locations of the corresponding i-th onsets, given by the equation:

$$\text{Dist}(A, B) = \sum_{i=1}^{n} |a_i - b_i|$$

In this case the one-to-one minimum-weight matching distance is equivalent to the swap distance [16]. A swap is an interchange of a sounded and silent pulse that are adjacent to each other. For instance, in Figure 2 in order to move the sounded pulse in location 2 to location 13, a total of 11 swaps are necessary, which is equivalent to $|13 - 2|$. When A and B have an unequal number of onsets then the one-to-one matching distance finds the minimum-weight perfect matching, thus ignoring the extra onsets of the denser rhythm [3]. If it is desired to assign all the onsets of the denser rhythm to the onsets of the sparser rhythm, with the constraint that every onset of the sparser rhythm must be matched to at least one onset of the denser rhythm, then the matching is a many-to-one matching, and the minimum-weight many-to-one matching is called the directed swap distance in computational music [8], and the restriction scaffold assignment distance in computational biology [9]. These two measures have achieved limited success in their applicability to accurately measure musical rhythm similarity because there are cases in which they give counter-intuitive and unsatisfactory results. One example that illustrates their weakness is shown at the top of Figure 2. Both the one-to-one and the many-to-one distance measures assign onset No. 2 of rhythm A to onset No. 13 of rhythm B, yielding a large distance of $1 + 11 + 1 = 13$, violating the Gestalt principle of proximity. Indeed, experimental results have provided evidence that the edit distance yields higher correlations with human judgments than these two matching distances [19]. In fact, the edit distance also has been shown to perform better than statistical features of the inter-onset-histograms of the rhythms [1], and better than global structural music-theoretical features of the rhythms [25]. Thus, the evidence garnered to date suggests the hypothesis that the edit distance is superior to all other measures of rhythm similarity with respect to correlation with human judgments.

In this paper the edit distance is compared empirically with the minimum-weight many-to-many matching distance, in order to test the above hypothesis, using a family of Middle-Eastern rhythms. A minimum-weight many-to-many matching is a matching of minimum cardinality that assigns each onset of one rhythm to at least one onset of the other rhythm, and vice-versa. The many-to-many matching between rhythms A and B seeks the minimum weight matching in a complete bipartite graph in which the nodes correspond to the onsets of the rhythms, and an edge is inserted between an onset of rhythm A and an onset of rhythm B such that its weight is equal to the absolute value of the distance (duration) between the two onsets. As the example at the bottom of Figure 2 illustrates, the many-to-many matching reflects more faithfully the Gestalt principle of temporal proximity, and makes musical sense to a greater extent, than the one-to-one or many-to-one matchings: pulse No. 2 in rhythm A is now matched to pulse No. 1 in rhythm B instead of pulse No. 13, and pulse No. 13 in rhythm B is matched with pulse No. 14 in rhythm A, yielding a distance of $2 + 1 + 1 = 4$.

Fig. 2: The one-to-one (top) and many-to-many (bottom) minimum-weight matchings between sequences A and B.

2. Experiments and Results

The one-to-one and many-to-many minimum-weight matching distance measures were compared with the edit distance using a dataset consisting of nine Middle-Eastern musical rhythms, for which human similarity judgments were available from previous listening tests performed at the Radcliffe Institute for Advanced Study at Harvard University in the spring of 2010. For details about the human subjects, stimulus materials used, and the experimental procedures applied, the reader is referred to the paper by Toussaint, Campbell and Brown [19].

2.1 Computing the Edit Distance

The edit distance between two sequences A and B of lengths n and m may be computed in $O(nm)$ time using dynamic programming. Edit distance online calculators are readily available on the Internet, and open source code in several programming languages is also obtainable. In the present study the Wagner-Fisher algorithm implemented by Malcolm Campbell [19] was used, in which the cost (weight) of all three operations (deletions, insertions, and reversals) was set equal to 1. Note that a reversal may be considered to be a deletion followed by an insertion, and therefore if the cost of a reversal is set to 2, it becomes superfluous.
2.2 Computing the Many-to-Many Matching

The minimum-weight many-to-many matching distance for rhythms with unequal numbers of pulses and onsets was computed by combining two existing algorithms. The well-known Hungarian algorithm, also known as the Kuhn-Munkres algorithm [13], [14], computes a minimum-weight one-to-one (perfect) matching in a bipartite graph, and code is available in the Python library [26]. However, this algorithm does not solve the many-to-many matching problem. Fortunately, Eiter and Mannila [12] provide an algorithm for transforming one bipartite graph \( G \) into another bipartite graph \( G^* \) that contains twice as many nodes, such that the minimum-weight one-to-one perfect matching in \( G^* \) is an optimal solution to the many-to-many minimum-weight matching in \( G \). This algorithm was programmed in the Python language to complement the code for the Kuhn-Munkres algorithm [26]. This approach to computing the many-to-many matching has the advantage that it applies to graphs in general, but the drawback that the computational complexity is \( O(n^3) \) for sequences of length \( n \). This is not a problem for short rhythms such as those tested here. Furthermore, versions of the Hungarian algorithm with improved expected running times are also available [15]. For long sequences faster \( O(n \log n) \) algorithms are known for the special case of one-dimensional sequences [20].

2.3 The Rhythm Dataset

The dataset consisting of the nine Middle-Eastern rhythms is shown in box-notation in Figure 3. These rhythms vary in their number of pulses and onsets, and thus allow for the differences between the distance measures to be fleshed out. The filled boxes come in two varieties: a solid black disk indicates a low-pitched “dum” sound, whereas a white-filled disk indicates a high-pitched “tak” sound. For this reason they are referred to as dum-tak rhythms [17]. Although the two types of onsets (“dum” and “tak”) differed in the acoustic samples heard by the participants in the listening tests, the distance measures did not differentiate between them, and treated them both simply as sounded pulses.

2.4 Results and Discussion

For every pair of rhythms in the dataset the distance measure between the pair was computed, yielding a distance matrix. The various distance measures were then compared with the matrix of human similarity judgments, by calculating the correlations between the corresponding matrices, using the Mantel test [27] available in the \( R \) software package [28]. The Mantel test is a statistical test expressly designed to test the correlation of structures that contain dependencies between their elements (such as distance matrices), which invalidate the assumption of independence. The resulting Mantel correlations and their \( p \)-values are listed in Table 1. The edit and many-to-many matching distances correlated significantly and almost equally with human judgments.

By contrast, the one-to-one matching distance showed no significant correlation with human judgments. Although this distance fared well in previous studies that compared rhythms that had equal numbers of pulses and onsets [19], it is not surprising that for the dum-tak rhythms in Figure 3 the measure is inadequate, since in the comparison of rhythms such as the 3-onset laz with the remaining 5-onset rhythms the calculation ignores two onsets that may play a significant role in the human perception of these rhythms.

Correlations were also calculated between the edit distance and the many-to-many and one-to-one matching distance measures, and the results are shown in Table 2. In spite of the disparity between the correlation coefficients for the many-to-many and one-to-one distances when compared with human judgments in Table 1 and with the edit distance in Table 2, the many-to-many and one-to-one matching distances are significantly correlated with each other (correlation coefficient = 0.758 with \( p = 0.0001 \)). This suggests that the seemingly small structural difference between the two matching distances is nevertheless an important factor for modeling the Gestalt principle of proximity in human rhythm perception.

<table>
<thead>
<tr>
<th>Distance Measure</th>
<th>Corr</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edit Distance</td>
<td>0.677</td>
<td>0.001</td>
</tr>
<tr>
<td>Many-to-Many Matching</td>
<td>0.661</td>
<td>0.0006</td>
</tr>
<tr>
<td>One-to-One Matching</td>
<td>0.195</td>
<td>0.175</td>
</tr>
</tbody>
</table>

Table 1: Correlation with Human Judgments.

Matching Distances with Edit Distance

<table>
<thead>
<tr>
<th>Matching Distance</th>
<th>Corr</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Many-to-Many Matching</td>
<td>0.566</td>
<td>0.007</td>
</tr>
<tr>
<td>One-to-One Matching</td>
<td>0.205</td>
<td>0.168</td>
</tr>
</tbody>
</table>

Table 2: Correlation with Edit Distance.
3. Conclusion and Ongoing Research

The goal of this ongoing research project is to determine if the minimum-weight many-to-many matching distance as defined in the Introduction can yield better correlations with human judgements than the edit distance. The results obtained so far, with the dataset consisting of the *dum-tak* rhythms in Figure 3, provides no evidence contrary to the hypothesis that the edit distance is superior to all other distance measures. Similar studies are under way using several other datasets consisting of synthetic rhythms as well as rhythms from different genres of music, to test this hypothesis further. In the present study no information was used that distinguishes between the two different sounding onsets ("dums" and "taks"), because previous studies that used a 3-symbol edit distance with different symbols for "dums," "taks," and silent pulses, revealed no differences in performance compared to the 2-symbol edit distance [19]. However, the many-to-many distance measure suggests an alternative method for incorporating information that distinguishes the "dums" from the "taks." The results reported here were obtained by computing the minimum-weight many-to-many matching on the complete bipartite graph in which all onsets of rhythm *A* are connected to all onsets of rhythm *B*. However, it may be that using incomplete bipartite graphs in which all "dums" in *A* are connected only to all the "dums" in *B*, and all "taks" in *A* are connected only to all the "taks" in *B*, may yield better results. There is also evidence that distance measures that take into account perceptually relevant musical factors can improve performance. The weights of the edges in the bipartite graph of the many-to-many matching distance, as well as the costs of the edit operations in the edit distance, may be chosen to reflect such perceptual information. These modifications of the distance measures will also be compared using all the datasets.

4. Acknowledgments

This research was supported by a research grant from the Provost Fund at New York University Abu Dhabi, and a Research Enhancement Fund grant from the NYUAD Institute, for the project titled: *Cross-Disciplinary and Multi-Cultural Perspectives on Musical Rhythm*. The research was done while the second author was a Global Academic Fellow at New York University Abu Dhabi, United Arab Emirates.

References


