Assessing the Performance of Melodic Similarity Algorithms Using Human Judgments of Similarity

Margaret Cahill **Donncha Ó Maidín**
Centre for Computational Musicology and Computer Music
Department of Computer Science and Information Systems
University of Limerick
Ireland
margaret.cahill, donncha.omaidin@ul.ie

Abstract

This paper outlines a project to identify reliable algorithms for measuring melodic similarity by using melodies extracted from a piece of music in Theme and Variations form, for which human judgements of similarity have been gathered.

Keywords: melodic similarity, human similarity judgments, scores

1. Background Work and Context

This research is concerned with identifying reliable algorithms for measuring melodic similarity in scores. Often, pitch alone or pitch with duration, are the only musical features used to measure similarity. A wide range of further musical information is available from a score. The approach taken here is to study music perception research, especially melody identification and melodic memory research for indications of which musical features of a score would be most useful for measuring similarity. Along with the question of which musical features to use in an algorithm, there is the issue of how to mathematically and proportionally combine these features and what values to assign to internal algorithm parameters and weightings.

The aim of this research is to use a test-bed of musical material for which we have gathered reliable human judgments of similarity to evaluate the performance of algorithms, use of particular features, and for improving and tweaking internal weights and parameter values. This paper gives an overview of a listening experiment, an initial algorithm with variations, and some discussion of issues in comparing the human and algorithmic measures of similarity.

2. Overview of the Experiment and Results

A listening experiment was carried out to gather human judgments of similarity for the algorithmic test-bed. Real melodies from a piece of monophonic music in Theme and Variations style were chosen as the Variations demonstrate varying degrees of similarity to the Theme. The piece of music chosen was a set of nine variations on "Twinkle, Twinkle, Little Star" for recorder [1]. The piece contains among other things, rhythmic and elaborate melodic variations, different time and key signatures, augmentation and diminution in the time domain, triplets and octave replacements of notes. The Theme and Variations easily segments into distinct four bar sections (ABA) and two separate melodies for the experiment were created using the first and second four bars of the Theme and each Variation. It is important for the success of this research that the human judgments of similarity are as accurate as possible for comparison with the algorithm results. The use of similar melodies helps the user make a judgment on the relative similarity of each. The fact that the melodies are short and well-known means the subjects do not have problems remembering the main melody and can concentrate on assessing the similarity of the Variation melody. The short melodies, comprehensive introduction and demonstration phase that makes use of a similar melody with variations, and the short listening time (c. 15 mins) also contribute to the collection of accurate similarity judgments. The experiment involved playing subjects pairs of the Theme and a Variation melody, and asking them to give a rating on a scale of 1 to 7 indicating the degree of similarity between the two melodies. Each melody extracted from the piece generated 9 pairs of melodies for comparison. Each pair was repeated in random order so that subject consistency could be checked.

The correlation coefficient (Spearman's non-parametric) was calculated for each individual subject using their ratings for the first and repeated playing of each pair of melodies. Only the data of the most highly consistent subjects was kept (those showing significance at the .01 level). There was quite a high level of agreement between subjects with a mean and median inter-subject correlation of .78 and .84 respectively for the first melody and .71 and .75 for the second. The ratings are pooled for future use, by using the median ratings as a measure of central tendency.

3. The Algorithms

The first algorithm used to compare to these human similarity judgments is listed as Equation 1. [2]

$$difference = \frac{\sum_{k=1}^{n} | p_{1k} - p_{2k} | w_k}{totaldur} \quad (1)$$

k = the time windows of the score
p_1, p_2 = pitch values of the first and second melodies in a window
w_k = the weight associated with that time window
totaldur = the duration processed

This algorithm processes the score in time windows. A time window unit in this case is the duration of the shortest full note at that particular point in the score. The basic form of the algorithm multiplies a weight value by the pitch difference in each time window. Various ways of incorporating duration (possibly separately to window width) so that longer notes are weighted as being more important than short notes, using pitch difference values based on tonality as Mongeau and Sankoff did [3], using metrical stress weights, so that notes on stronger beats of the bar are weighted more than notes on weak beats, and using weights according to the melodic accents in a melody [4], are currently being explored. If using metrical accents, there is the question of how to choose appropriate values for different beats of a bar. Further treatments for comparing melodies in different time signatures and identifying transpositions are also implemented. The performance of a number of variations of this algorithm are being evaluated, with implementations of other algorithms planned so that comparison of performance can be made.

4. Issues in Comparing the Algorithm Output to Human Ratings

One of the considerations when comparing the human ratings to the output of the algorithms is the difference in scale and range of both. In this case, the human ratings ranged from 1 (least similar) to 7 (most similar), while the algorithm ranges from 0 (the same) to some higher value that represents a degree of difference. This is a similar case in many such algorithms, where each difference found between melodies cumulatively contributes to an overall measure of similarity.

Correlation can be a useful metric here, as the direction and difference in scale do not affect the calculation. The correlation between the human ratings and each algorithmic similarity measure for all nine variations was used as a rough metric to identify the most successful algorithms. Correlation can only be used to assess the performance of the algorithm across all nine variations. It may also be useful to be able to assess the performance of particular algorithms for particular variations. Each variation varies the original theme in a different way, and an algorithm that performs well with a particular variation, may not perform well with others. It would also be useful to isolate the results for those variations with different time signatures to the theme to assess the way these have been handled.

In order to compare the algorithmic and human judgments for individual variations, the scales of both need to be normalized in some way first. The human rating results can be inverted by subtracting each from the maximum rating. This results in a range of values from 0 (most similar) to 6 (least similar), and both sets of results use the same direction of scale. There are a number of ways in which the sets of values can be normalized. Standard score, such as z-score, can be used to express the values in terms of the mean and standard deviation. These can be problematic, however, when outliers occur in the data set, as was found in these algorithmic measures. The min-max normalization was chosen, as this allows both sets of values to be represented in a chosen range, such as 1-10, or 1-100. The human ratings and algorithm output can then be compared for individual variations. A metric such as the sum of the absolute differences between both for each variation, or the Euclidean distance for the variations, can then be calculated and used as a measure of how that algorithm performed overall. Although it is important to thoroughly examine the closeness of results between the algorithmic and human judgments, it is also important to be wary of over-fitting algorithms to the particular melodies chosen here.

Further work involves detailed analysis of the performance of algorithms, along with the implementation and comparison of further algorithms, and the introduction of further musical test-beds for this early evaluation phase.

References

[1] M. Duschenes, “Variations on Twinkle, Twinkle, Little Star,” in *Method for the Recorder – Tunes and Exercises*. Berandol Music Limited, Ontario, Canada, 1962.

[2] D. Ó Maidín. “A Geometrical Algorithm for Melodic Difference,” Melodic Similarity - Concepts, Procedures and Applications, Computing in Musicology II, MIT Press. 1998.

[3] M. Mongeau and D. Sankoff. “Comparison of Musical Sequences,” *Computers and the Humanities*, 24, pp. 161-175, 1998.

[4] M. Jones. “Dynamic Pattern Structure in Music: Recent Theory and Research,” *Perception & Psychophysics*, 41, 6, pp. 621-634, 1987.