MELODIC SIMILARITY ALGORITHMS – USING SIMILARITY RATINGS FOR DEVELOPMENT AND EARLY EVALUATION

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 focuses on gathering similarity ratings for use in the construction, optimization and evaluation of melodic similarity algorithms. The approach involves conducting listening experiments to gather these ratings for a piece in Theme and Variation form.

Keywords: melodic similarity, score, algorithm, perception, similarity ratings, listening experiments

1 OVERVIEW OF PROBLEM

The MIR research community draws its members from many research disciplines, including musicology and music analysis. It is of benefit for musicologists to be able to search digitized score databases (corpora) for exact and similar melodies. Melodic similarity algorithms play an important role in automating this process. Such algorithms calculate a measure that reflects the degree of similarity/dissimilarity between a pair of melodies or melodic segments.

Many algorithms used to measure melodic similarity are text-based string-matching algorithms that have either been adopted directly or somewhat altered to suit this new role [1-4]. One of the most commonly used of these is the edit-distance family of algorithms (along with variations), which essentially calculates the "cost" of taking one string/melody and converting it into the other [5-7]. However, the issue of identifying melodies that are perceptually similar means that there is not a clear analogy between comparing words/sentences in text to comparing musical melodies which are multi-dimensional, and for which operations such as simple addition or deletion of notes is problematic.

This research is concerned with identifying successful algorithms for determining melodic similarity using music perception principles as a guide and employing a relevant testbed in the development stage to aid the process. We are currently focussing on monophonic music.

2 WHICH MUSICAL FEATURES TO USE AND HOW TO COMBINE THEM.

The most basic features that can be used to describe a note are pitch and duration. In its most basic form, a melody could be described as a sequence of such pitches and durations. Further features that could also be used to describe a notated melody in more detail include rests, phrasing, dynamics, tempo, articulation and other expression indications. Metrical accents (the pattern of strong and weak beats relating to the time signature of a piece) are implicit in scores and are instinctively performed by musicians. These accents are also included in the list of possible features. In the past researchers have mainly focussed on using pitch sequences alone [8-9] or both pitch and duration sequences [1, 5, 10-11]. Some researchers have used other melodic features include metrical stress [12-13] and dynamics [14].

Music perception and cognition research provides a useful starting point for deciding which features might be most relevant. Quite a bit of work has been done to discover the ways in which we remember, identify and recall melodies. Much of this work has focussed on pitch aspects of music only [15-19], although rhythm is considered to a lesser extent [20]), as well as the effects of melodic and rhythmic accents [21-24], among other features.

If more than one musical feature is used, it becomes important to consider the relative importance of each and to explore fruitful ways of combining the associated measures. If only pitch and duration are used, a decision to weight them equally would require some justifications. There is also the issue of how to combine such features (e.g. by addition, or by multiplication, or by some other means). When a decision is taken on ways in which the various feature measures are to be combined, there is still the remaining task of selecting weights to apply to each measure, Finding the proportionate weights for internal parameter values forms part of the development and "tweaking" stage of an algorithm. Often a researcher may simply pick arbitrary values, combination schemes, and weightings that instinctively make sense in order to get some satisfactory results. Others have "tweaked" settings as a direct result of the output of the algorithm and thus tried to improve and optimise the algorithm in some way [7]. Again, research in the music perception area may be useful here to indicate the relative importance certain musical features play.

The approach adopted here is to use relevant research as a guide to inform choices made in the design and development stage of an algorithm, rather than to attempt to create a perceptual model of melodic similarity. Schulkind et. al., in a paper that deals with how people identify melodies, urges caution as "features that are easily perceived may not necessarily be those used to distinguish melodies"[25]. In this research we propose to examine the extent to which such features may be useful in melodic similarity algorithms.

3 EXPERIMENTAL METHOD

We propose to use a small test collection based on a real piece of music during the development stage of this project. As Pampalk indicates "..even a tiny music collection can be used to identify weaknesses of a measure and compare measures to each other"[26]. Theme and Variation form are very good candidates for this early assessment of algorithm performance because they consist of musical material that is of varying degrees of similarity, where some variations may be very similar to the theme and others greatly alter the theme. In order to assess the performance of algorithms and various combinations of weighting of features we need to have an objective idea of the similarity of each variation. We propose gathering similarity ratings by conducting listening experiments that ask subjects to rate the similarity of the theme to each variation.

Various kinds of rating scales, including similarity rating scales, have been commonly used in music perception research but often as a mechanism to derive information other than melodic similarity [18]. Also, some recent research in the area of melodic similarity has also used similarity ratings gathered from subjects so there is a body of research with which to compare methods and results with [14, 28-31]. Another recent project involved asking subjects to rank melodies in order of similarity [32]. The current focus of our research is on carrying out listening tests to gather the similarity ratings we need for future work. The methods used for the test are briefly described in section 4. The above mentioned papers present a number of mechanisms for ensuring the reliability and consistency of user ratings and some guidance is taken from this research.

Once a set of reliable similarity ratings has been identified we will focus on creating and evaluating algorithms that produce comparable results with these ratings.

4 GATHERING SIMILARITY RATINGS FOR A SMALL TESTBED

For the initial phases of this research we are using a set of variations on Twinkle, Twinkle, Little Star composed for recorder by Duschenes [33]. This piece was used by Mongeau and Sankoff [7] for evaluating the success of their distance measure algorithm. Initially this collection was considered because the Mongeau and Sankoff paper provided some discussion and comment on the musical material and research findings. On closer examination this test collection demonstrated very good examples of the various kinds of problems that a melodic similarity algorithm is faced with. Different time signatures, different keys, augmentation of theme (1 bar stretched to 2 bars), notes replaced by shorter repeated notes, triplets, elaborations of theme by stepwise motions and by leaps, notes occurring an octave higher and hiding of theme notes are all included in the 10 short variations (see Figure 1 for an example). This set of Theme with nine variations consists mostly of 12 bars in $\frac{4}{4}$ time. One variation is in $\frac{3}{4}$ time and is 24 bars long and a further variation is in $\frac{6}{8}$ time. The test material tends to segment quite naturally, and this allows us to ignore an issue that would be more problematic with other music forms and pieces. The form of the piece is very obviously ABA/ABA', with each section lasting for four bars.

One of the key issues when using similarity ratings is the reliability or consistency of the data. As we intend to develop improved versions of existing algorithms based

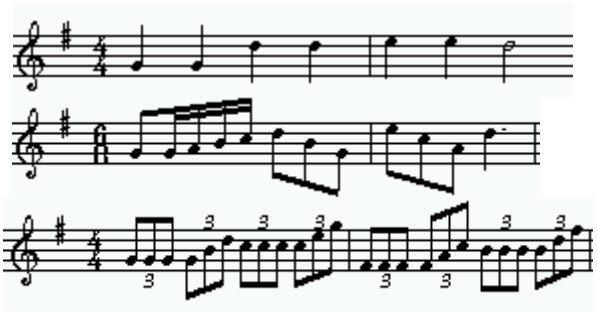

Figure 1: The first two bars of the Theme, Variation 4 and Variation 7, illustrating some of the variations of the theme found in this piece.

on these gathered ratings it is essential to ensure that the data is as "true" as possible. As previously mentioned, the form of this set of Theme and Variations easily allows for dividing each into short four bar phrases, which makes it easier for subjects to compare than the entire 12 bars and therefore provides more reliable data. The melody is also very well known which means that the subjects should not have difficulty remembering the reference melody. Many perception experiments use unknown melodies but there are examples of using known melodies [34-35] and we believe it increases the usefulness of the data we are gathering.

5 THE LISTENING EXPERIMENT

The main part of the listening experiment is structured as shown in Table 1 below and is run on computer using Roger Kendall's MEDS (Music Experiment Development System [36]). Subjects hear a series of pairs of melodies and are asked to rate the similarity of the melodies in each pair. Each segment is 8 seconds long, although one variation is 12 seconds long. There is a .5 second pause between the first and second melody of

each pair then a rating scale is shown. As soon as the subject inputs their rating (there is no time limit), the next pair of melodies is played. Re-testing, split-testing or repeated random trials are often used in these kinds of experiments and later checked for consistent results from subjects. We decided in this case to repeat the basic test in random order for later comparison. There is a one minute pause between Part A and B.

Table 1: layout of listening test

Part A	Theme & Variations 1-9 Bars 1-4	sequential
	Theme & Variations 1-9 Bars 1-4	random
Part B	Theme & Variations 1-9 Bars 5-9	sequential
	Theme & Variations 1-9 Bars 5-9	random

In a pilot experiment we used a 7-point scale that ranged from "very dissimilar" to "very similar" and did not tell the subjects that all the pairs they heard would be similar in some way. This did include an introductory description of the test with played examples and a practice test with three pairs. We found that subjects found the use of the words dissimilar and similar confusing and often were reluctant to use the extremes of the scale until they had heard all melodies played through once.

We are currently in the process of running these listening experiments and due to findings in the pilot test have changed the rating scale to a 7-point scale using the descriptors "hardly similar at all" and "very similar" at the opposite poles of the scale (1 and 7 respectively). Additionally we are now spending c.10 minutes of preparation time with the subject describing and discussing the test. In the introduction they are told that all the melodies they will be comparing to the theme will be similar in some way to it and a demonstration using 7 variations on another well known tune are played and discussed with the subject. These variations are based on the sort of modifications made to the "Twinkle, Twinkle" melody and three pairs are used in a practice run before the test proper begins. Subjects can ask questions and discuss the issue of similarity during this period but no comments were made about which musical features subjects should use to make their judgements. Subjects are encouraged to use the full range of the scale. Subjects are asked to complete a short questionnaire on musical experience and comment sheet the end of the experiment so that we can determine if significant effects occur due to musical background.

6 INITIAL RESULTS

At the time of writing 13 subjects have participated in the experiment, all of whom are musicians but not all are from a classical music background. Initial results show that there is general agreement between subjects on the ratings given to each variation. As might have been expected, subjects were most in agreement regarding variations that were least similar and most similar (i.e. ratings of 1 and 7), while the results for variations that were only somewhat similar were less clear. Many subjects showed very high correlation between their ratings for the sequential and random playings (only one subject was less than .7) and there was high inter-subject correlation for the most part. Most subjects did consistently use the full range of the scale. We have yet to fully analyse the results gathered so far but are confident that these initial results show that reliable data can be obtained from such an experiment and intend to continue these listening tests with further subjects.

REFERENCES

[1] Crawford, T., and Iiliopoulos, C. Strong-Matching Techniques for Musical Similarity and Melodic Recognition. Computing in Musicology II, Melodic Similarity - Concepts, Procedures and Applications, MIT Press, 1998.

[2] Downie, J.S. Music Retrieval as Text Retrieval: Simple yet Effective. SIGIR 1999, Berkeley, California.

[3] Droettboom, M., Fujinaga, I., MacMillan, K., Patton, M., Warner, J., Choudhury, S., and DiLauro, T. Expressive and Efficient Retrieval of Symbolic Musical Data. ISMIR 2001, Indiana.

[4] Lemstrom, K. String Matching Techniques for Music Retrieval. PhD Thesis, Series of Publications A, Report A-2000-04, Department of Computer Science, University of Finland.

[5] Smith, L., McNab, R., and Witten, I. Sequence – Based Comparison: A Dynamic-Programming Approach. Computing in Musicology II, Melodic Similarity - Concepts, Procedures and Applications, MIT Press, 1998.

[6] Uitdenbogard, A., and Zobel, J. Melodic Matching techniques for Large Music Databases. ACM Multimedia 1999, Orlando, Florida.

[7] Mongeau., M, and Sankoff, D. Comparison of Musical Sequences. Computers and the Humanities, 24, (1990), 161-175.

[8] Blackburn, S., and DeRoure, D. A Tool for Content Based Navigation of Music. ACM Multimedia 1998, Bristol, UK.

[9] Downie, S., and Nelson, M. Evaluation of a Simple and Effective Music Information Retrieval Method. SIGIR 2000, Athens, Greece.

[10] Hoos, H., Renz, K., and Gorg, M. GUIDO/MIR – an Experimental Musical Information Retrieval System based on GUIDO Music Notation. ISMIR 2001, Indiana.

[11] McNab, R., Smith, L., Bainbridge, D., and Witten, I. www.dlib.org/dlib/may97/meldex/05witten.html

[12] Typke, R., Giannopoulos, P., Veltkamp, R., Wiering, F., and Oostrum, R. Using Transportation Distances for Measuring Melodic Similarity. Technical Report, UU-CS-2003-024, Institute of Information and Computing Sciences, Utrecht University.

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

[14] Hofman-Engl, L. Melodic Similarity and Transformations: A Theoretical and Empirical Approach. PhD Thesis, Department of Psychology, Keele University, January 2003.

[15] Levitin, D. Absolute Memory for Musical Pitch: Evidence from the Production of Learned Melodies. Perception and Psychophysics, 56, 4 (1994), 414-423.

[16] Edworthy., J. Interval and Contour in Melodic Processing. Music Perception, 2 (1985), 375-388.

[17] Dowling, J. and Fujitani, D. Contour, Interval, and Pitch Recognition in Memory for Melodies. Journal of the Acoustical Society of America, 49, 2 (1971), 524-531.

[18] Dowling, J., Scale and Contour: Two Components of a Theory of Memory for Melodies. Psychological Review, 85, 4 (1978), 341-354.

[19] Dewitt, L. and Crowder, R., Recognition of novel melodies after brief delays., L., Bainbridge, D., and Witten, I. The New Zealand Digital Library MELody inDEX. D-Lib Magazine, May 1997, Music Perception, 3, 3 (1986), 259-274.

[20] Kidd, G., Boltz, M., & Jones, M. R. Some effects of rhythmic context on melody recognition. American Journal of Psychology, 97, (1984), 153-173.

[21] Jones, M. Dynamic Pattern Structure in Music: Recent Theory and Research. Perception & Psychophysics, 41, 6 (1987), 621-634.

[22] Thomassen, J. Melodic Accent: Experiments and a Tentative Model. Journal of the Acoustical Society of America., 71, 6 (1982), 1596-1605.

[23] Huron, D., and Royal, M. What is Melodic Accent? Converging Evidence from Musical Practice. Music Perception, 13, 4 (1996), 489-516

[24] Pfordresher, P. The Role of Melodic and Rhythmic Accents in Musical Structure. Music Perception, 20, 4 (2003), 431-464.

[25] Schulkind., M., Posner, R., and Rubin, D. Music Features that Facilitate Melody Identification: How do you Know it's "your" Song when they Finally Play It? Music Perception, 21, 2 (2003), 217-249.

[26] Pampalk, E., Dixon, S., and Widmer, G. Exploring Music Collections by Browsing Different Views. DAFX-03, London.

[27] Rosner, B., and Meyer, L. The Perceptual Roles of Melodic Process, Contour, and Form. Music Perception, 4, 1 (1986), 1-40.

[28] Eerola, T., Jarvinen, T., Louhivuori, J. and Toivainen, P. Statistical Features and Perceived Similarity of Folk Melodies. Music Perception, 18, 3 (2001), 275-296

[29] Mullensiefen, D., and Frieler, K. Measuring Melodic Similarity: Human vs. Algorithmic Judgements. Proceedings of the Conference on Interdisciplinary Musicology, Graz, Austria, April 2004.

[30] Schmuckler, M. Testing Models of Melodic Contour Similarity. Music Perception, 16, 3 (1999), 295-326.

[31] McAdams, S., and Matzkin, D. Similarity, Invariance, and Musical Variation. The Biological Foundations of Music, New York Academy of Sciences, 2001

[32] Typke, R., Hoed, M., de Nooijer, J., Wiering, F., and Veltkamp, R. A Ground Truth for Half a Million Incipits. Proceedings of DIR 05, 5th Dutch-Belgian Information Retrieval Workshop, Utrecht, The Netherlands, January 2005.

[33] Duschenes, M. Variations on Twinkle, Twinkle, Little Star from Method for the Recorder – Tunes and Exercises. Berandol Music Limited, Ontario, Canada, 1962.

[34] White., B. Recognition of Distorted Melodies. American Journal of Psychology, 73, (1960), 100-107.

[35] Dowling, W. J., and Hollombe, A. The Perception of Melodies Distorted by Splitting into Several Octaves: Effects of Increasing Proximity and Melodic Contour. Perception and Psychophysics, 21, 1 (1977), 60-64.

[36] Kendall, R. MEDS – Music Experiment Development System. www.ethnomusic.ucla.edu/systematic/Faculty/Kendall/meds.htm