Interrogating melodic similarity: a definitive phenomenon or the product of interpretation?

This is a pre-print of an article to be published in *Journal of New Music Research*, vol.41, no.4. Details of the definitive version may be found at www.informaworld.com/jnmr.

Alan Marsden

Lancaster Institute for the Contemporary Arts
Lancaster University, UK

A.Marsden@lancaster.ac.uk

Abstract

The nature of melodic similarity is interrogated through a survey of the different means by which the phenomenon has been studied, examination of methods for measuring melodic similarity, a Monte Carlo analysis of data from the experiment which formed the basis for the 'ground truth' used in the MIREX 2005 contest on melodic similarity, and examples of interest in the music of Mozart. Melodic similarity has been studied by a number of means, sometimes quite contrasting, which lead to important differences in the light of the finding that similarity is dependent on context. Models of melodic similarity based on reduction show that the existence of multiple possible reductions forms a natural basis for similarity to depend on interpretation. Examination of the MIREX 2005 data shows wide variations in subjects' judgements of melodic similarity and some evidence that the perceived similarity between two melodies can be influenced by the presence of a third melody. Examples from Mozart suggest that he deliberately exploited the possibilities inherent in recognising similarity through different interpretations. It is therefore proposed that similarity be thought of not as a distinct and definite function of two melodies but as something *created* in the minds of those who hear the melodies.

1. What is melodic similarity?

A common theme of music-computing research in the last couple of decades has been measurement of melodic similarity. Much of this research has been in the context of query systems, with the aim of finding a way of organising and searching a database of music so as to retrieve melodies similar to a given query (Hu, Dannenberg & Lewis, 2002; Pardo, Shifrin & Birmingham, 2004). The idea has been used also as a basis for segmentation (Ahlbäck, 2007) and for music analysis (Adiloglu, Noll & Obermayer, 2006). The objective of this article is not to seek better measurement of melodic similarity, but rather to interrogate its essential nature and ask whether seeking a definitive measure of similarity is a reasonable research goal. One strand of this interrogation is to ask whether or not 'melodic similarity' is a single phenomenon. If it is not, there can be no single measure of similarity. Another strand questions whether the perception of similarity is dependent on interpretation and therefore, to some extent, is a *creative* act on the part of the perceiver. If it is, then it will be impossible to discover an accurate measure of similarity which is a function of two melodies alone.

Section 2 of the article surveys the many different empirical bases used in studies of similarity and questions whether these all show evidence of the same phenomenon. Section 3 discusses some mechanisms used in the measurement of similarity and again questions whether these can really measure the same thing. Section 4 examines the rich set of data used to form the 'ground truth' for the MIREX 2005 'symbolic melodic similarity' competition with the objective of querying whether the similarity judgements made by subjects in that experiment can be modelled as functions of pairs of melodies. Section 5 examines some cases of similarity in music by Mozart, pointing out how Mozart exploits the possibilities inherent in alternative interpretations. In Section 6 I attempt to draw some conclusions from the equivocal evidence arising from earlier sections. It is safe to conclude that melodic similarity is a complex phenomenon which is affected by context, but it is not possible to reach a definitive conclusion on the degree to which similarity can be measured. A conclusion consistent with the data is that the perception of similarity is indeed creative, but humans are sufficiently creative to invent a similarity metric when the context requires it, and this can be subject to modelling.

1.1 Similarity and measurement

It is useful first to clarify some fundamental issues about measurement. The simple observation that some melodies are similar while others are different, and that the similarity can be greater or lesser, suggests that similarity can be represented by a number. Similarity is limited at one extreme (when similarity becomes sameness) but not obviously limited at the other, so it is natural to think of similarity like distance, representable by a non-negative number. The fundamental question is whether or not we can hope to define a function δ(a, b) which yields this number representing the dissimilarity between two melodies **a** and **b**.

The most useful measurements have the four properties of a metric: non-negativity, self-identity, symmetry, and triangle inequality (Armstrong, 1983, p.38). The first two properties seem self-evident for melodic similarity. With respect to symmetry, however, factors which have been demonstrated to cause asymmetry in other domains, such as salience and prototypicality (Tversky, 1977), are likely to be important in melodic similarity also. It has been demonstrated that manipulating subjects' familiarity with colours induces asymmetries in their judgements of similarities between those colours (Polk et al., 2002). The familiarity of melodies varies enormously, so we should assume that it will also lead to asymmetry in melodic similarity: an unfamiliar melody is likely to be judged as more similar to a familiar one than is the familiar melody to the unfamiliar one. The literature on melodic similarity does not include discussion of such asymmetry, though, and the published models do not account for it.

The property most commonly questioned is triangle inequality, and the common grounds for this are that melody **a** might be similar to melody **b** by virtue of property or component *x*, while melody **b** might be similar to melody **c** by virtue of a different property or component *y*. In such a situation there is no reason to expect the dissimilarity between **a** and **c** to be limited. Despite such easily imagined counter-examples, those who use systems of measurement with the property of triangle inequality have not reported failure to match human judgements of melodic similarity on the grounds that those judgements do not exhibit triangle inequality. Indeed it is not uncommon to adapt a measure precisely so that it has the property of triangle inequality (for example the development of Proportional Transportation Distance (Giannopoulos & Veltkamp, 2002) from Earth Mover's Distance) with the objective of facilitating the organisation and searching of a database.

(Meanwhile, others have taken the alternative path of investigating means for organising and searching databases without the need of triangle inequality (Typke & Walczak-Typke, 2008).)

2. Empirical bases

Most studies have grounded their work on some kind of empirical basis, some raw 'truth' that certain melodies are similar and others are not. When we look at the detail, however, we find that very different paradigms have been used, firstly in the source of that 'truth' and secondly in the kind of relationship tested between melodies.

2.1 Experimental paradigms

Many studies ask experimental subjects, often experts, to judge the similarity between pairs of melodies or extracts of melodies on a rating scale (Eerola et al., 2001; Eerola & Bregman, 2007; Müllensiefen & Frieler, 2004; Müllensiefen & Frieler, 2007; Schmuckler, 2010). This has the advantage of directly generating measures of difference which will almost certainly have the first three properties of a metric. A rating of 0 or below is not an option; subjects are not asked to compare a melody to itself; and the set-up usually discourages asymmetric judgements. There is no guarantee, however, of triangle inequality. One objection to experimental procedures like this is that they are not realistic: musicians are rarely (if ever) in a situation when they have to match the similarity between melodies to a number. Such direct rating was avoided in another study which also used expert judgement but subjects were asked to rank a set of melodies by their similarity to a reference melody rather than to simply compare pairs of melodies (Typke et al., 2005; Typke, Wiering & Veltkamp, 2007). (This study is examined in more detail below.) A measure of difference can be derived from the relative positions of melodies in the rankings, but this measure can only be relative, unlike the potentially absolute measure derived from direct rating of similarity. I say only 'potentially' because in practice the ratings will depend also on the set of melodies presented to the rating subjects. In other domains, Tversky (1977) demonstrated how the similarity between a pair of objects can be influenced by the presence of other objects for comparison. We should expect this effect to apply in the case of melodies also, and it is likely to be stronger in the case of the ranking paradigm because a larger set of melodies is continuously present. Another paradigm which avoids an artificial direct rating of similarity is to present subjects with three melodies and ask them to indicate the pair which are most alike the pair which are least alike (Allan, Müllensiefen & Wiggins, 2007; Novello, McKinney & Kohlrausch, 2011). This approach is the one which places the least burden on experimental subjects, and it appears to have been successful for non-expert subjects, unlike the paradigms mentioned above. On the other hand, deriving measurements from these observations requires a method such as multi-dimensional scaling, and a large quantity of observations.

Other studies have avoided direct judgment of similarity, whether by experts or naive listeners. Some have depended on categorisation of melodies either from existing musicological studies (Müllensiefen & Frieler, 2007; Volk et al., 2008) or on the basis of geographical origin (Juhasz, 2006). In these cases a useful measurement cannot be derived from the empirical data, since distances between melodies are all either 0 or 1 according to whether or not the melodies belong to the same category. However, the data can still be used to verify a computational model on the grounds that the computed distance for melodies within a category should be less than the distance between melodies from different categories.

Yet other studies have attempted to judge similarity on the basis of some real musical activity. Studies aimed at producing measurements for use in query-by-humming systems have been based on asking subjects to sing a known melody (Hu, Dannenberg & Lewis, 2002; Pardo, Shifrin & Birmingham, 2004). The subjects make mistakes, so the resulting melody is not the same as the original, but it is assumed to be more similar to the original than to other melodies. Subjects can also be asked to deliberately vary a melody (Bernabeu et al., 2011), and once again the variations are assumed to be more similar to the original than to other melodies.

2.2 Similarity and cognition

Do all these paradigms study the same thing? There are other musical phenomena whose underlying models are robust under different experimental paradigms (models of tonal perception via pitch-frequency profiles are one example (Krumhansl, 1990)), and these suggest stable underlying cognitive functions. I am not aware of evidence that judgements of melodic similarity are consistent across different paradigms. Indeed, there is clear evidence for what one might expect from other aspects of human behaviour: that judgements of melodic similarity are dependent on context. Müllensiefen and Frieler have demonstrated that a different model is required to account for similarity judgements which use the same paradigm but in which the set of melodies to be compared is different (2007).

In fact, the contexts in these various experiments have been very different. The nature of melodic materials has varied widely, and crucially the instructions and information given to the subjects have also varied. Sometimes subjects have been given no further instruction than to rate the similarity between two melodies. On other occasions they have been given guidance such as to imagine that the comparison melody is a student's attempt to reproduce a teacher's melody and to think of the similarity rating as a mark (Müllensiefen & Frieler, 2007). (Note that in this case the similarity judgement can no longer be assumed to be symmetric.) Sometimes subjects' attention has been drawn to particular aspects of the melody, for example by being told in advance that the experiment was concerned with contour (Schmuckler, 2010).

The differences in paradigm also introduce significant issues. If data is derived from real musical behaviours which do not involve explicit similarity judgements, we can only assume that similarity is a governing factor; if data is not derived from real musical behaviours we cannot be certain that it has any real musical relevance. Even in the cases based on explicit expert judgements of similarity, there are important differences. As stated above, we cannot be certain that judgement of melodic similarity has the property of triangle inequality. Even if it does not, subjects can give answers with confidence when asked to rate the similarity between two melodies, or even to judge the most similar and least similar pairs in a triple. However, in a ranking task such as used in (Typke et al., 2005) the subjects might be in a position of having to balance competing similarity judgements, depending on how they interpret the instructions. If they consider their task to be simply to ensure that the melody ranked x is no less similar to the reference than the melody ranked $x + 1$, no competing rankings can arise. If, however, they also believe that a ranking implies that the melody ranked $x + 2$ is less similar to the one ranked x than the one ranked $x + 1$, then in the absence of triangle inequality, a subject might find it impossible to find a ranking which meets both criteria: melodies **a**, **b** and **c** might have decreasing similarity to the reference, and so be ranked x, $x + 1$ and $x + 2$, but **c** might be more similar to **a** than **b**, implying instead the ranking x, $x + 2$ and $x + 1$.

It is not safe, therefore, to assume that these studies investigate the same phenomenon of melodic similarity. Until there is evidence that data produced under these various paradigms is compatible it is probably safer to consider melodic similarity to be a family of possibly related phenomena.

3. Bases for measuring similarity

Writing about theories of similarity in general, Tversky noted that 'theoretical analysis of similarity relations have been dominated by geometric models [which] represent objects as points in some coordinate space such that the observed dissimilarities between objects correspond to the metric distances between the respective points' (1977, p.327). He proposed instead an analysis based on the representation of objects as sets of features, and gave the following formula for the similarity of objects based on their shared and distinctive features, where *A* and *B* are the feature sets of object *a* and *b* respectively (p.333):

$$S(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A-B) + \beta f(B-A)} \tag{1}$$

He was able to explain various similarity phenomena on the basis of assumptions about the nature of the function *f*, the values of the coefficients α and β, and factors such as focus, salience and prototypicality. With a suitable representation of quantities in terms of possibly overlapping feature sets, the model can encompass 'geometric' analyses of similarity also. It also depends crucially, as Tversky acknowledged himself (p.331), on the interpretation of objects in terms of sets of features.

In the following sections, I discuss how the same factors can be seen in methods of modelling melodic similarity: models based on quantitative difference or distinctive features; and a crucial role for interpretation.

3.1 Models of similarity

Many models of musical similarity start from the assumption that a melody is a sequence of pitches, and so similarity relationships between melodies can be expressed in terms of relations in pitch and time. Some are thoroughly geometrical, representing a melody as a set of points in a pitch-time space and then measuring the distance between points (see, for example, Hofmann-Engl, 2003), or some other mechanism for measuring difference (e.g., difference between curves fitted to the notes, Urbano et al., 2011). Sometimes these measures of difference are mediated also by an alignment of the notes of one melody with another through a mechanism such as Dynamic Time Warping. Alternatively the shifts in time which this implies are 'measured' together with changes in pitch through the editing operations required to transform one melodic sequence into another, such as in Levenshtein distance, or Earth Movers' Distance (Typke, Wiering, Veltkamp, 2007). Alignment in time has its analogue in the pitch domain whereby some models use pitch class (chroma) instead of pitch to neutralise differences of octave, while others use intervals to concentrate ignore absolute pitch differences altogether. When dealing with audio rather than symbolic data, equivalent 'alignment' in the pitch domain can be achieved by extracting 'shift invariant' features such as the power spectrum of the chromagram (Marolt, 2008).

Other models start from the assumption that behind the sequence of pitches which makes up a melody is a musical structure, and melodic similarity is best modelled by similarities in these structures rather than by comparing melodies note-by-note. A melody is represented in a tree structure, constructed through a process of reduction which progressively removes decorative notes until only the main outline of the melody is left. These models are worth describing here in more detail because they illustrate clearly the importance of interpretation, to be discussed below.

Rizo (2010) and Bernabeu et al. (2009) derive the reduction of a melody by selecting one of the notes occurring in each span based on a small number of rules. The spans are determined by the metre, so that, in 4/4 for example, there is a span for each bar, at the next level down two spans for the minims (half notes), then four spans for the crotchets (quarter notes), etc., halving each span at the level above. There are also higher-level spans which group bars into pairs, etc. The result is a tree structure in which each node corresponds to a specific time span, and the rhythm of the melody is completely defined by the tree structure. The reduction is built bottom-up by

(a) always selecting a note in preference to a rest,
(b) selecting a harmonic note in preference to a non-harmonic one, and
(c) selecting the note at the head of the span if both are harmonic.

A harmonic analysis of the melody must be generated before reduction, and this is currently done by hand. A measure of similarity based on the tree edit distance between the reductions of melodies was compared with edit distance on the melodic surfaces alone. The reduction-based similarity measure proved to perform better at distinguishing variations of a melody from unrelated melodies (Bernebau et al., 2009).

The approach of Orio & Rodà (2009) is similar, in that it generates a tree based on the metrical structure, and notes are selected within each span partly on the basis of a harmonic analysis. The selection, however, is based on a more complex set of weights using the relation of the note to the underlying harmony (fifth, third or root), the function of that harmony, and the position in the metre. Furthermore, similarity between melodies is not based on the edit distance between trees. Melodies are segmented (using pre-existing segmentation schemes) and the segmentation propagated to higher levels of the tree. The resulting melodic segments, expressed as interval patterns, are placed in a directed acyclic graph (DAG) in which parent-child relations between segments copy those relations in the reductions. The difference between two segments is then measured by the minimum path length between the segments in the DAG, and the difference between two melodies is the average difference between their component segments. This method was not tested against other measures of melodic similarity.

The reductions produced by my own system (Marsden, 2010a, 2010b) are intended to more closely mimic the reductions of Schenkerian analysis (Schenker, 1935). Furthermore, they are based not just on melodies but on a full musical texture (generally extracts from piano pieces). The reduction process is therefore considerably more complex than those outlined above. In particular, the reduction tree does not necessarily follow the metrical structure (as indeed it does not in many Schenkerian analyses), and no prior harmonic analysis is necessary (though specification of the key and metre is). While early results matched actual analyses to a promising degree (2010b), an attempt to use the same system of reduction for demonstrating the similarity underlying themes and variations produced less promising results (2010a). Matching themes and variations via reductions proved no better than matching on the basis of the surfaces alone.

A final class of model is based on neither sequences of pitches nor on structures, but on more abstract features of a melody. A common paradigm is to extract features from a melody through some analytical process, and then on the basis of a set of empirical data to determine, through a technique such as machine learning or statistical analysis, which features in which combination provide the best model of similarity. Bohak & Marolt (2009), for example, found the following five features to be useful in distinguishing folk-song variants: melodic expectancy, entropy, phenomenal accent synchrony, 'melodiousness' (based on prime factorisation of frequency ratios), and melodic originality. This is a surprising result, because the features are quite different from the measures of distances in pitch and time used in geometric models or the comparisons of structure used in reduction-based models. The features do not need to be necessarily melodic ones ('entropy' for example, is not a specifically melodic feature), so it is easy for this type of model to move away from melodic similarity *per se* to more general musical similarity. Novello, McKinney & Kohlrausch (2011), for example, in a statistical analysis based on similarity judgements between audio clips of popular music, found three significant dimensions which they described as vocal-non vocal, slow-fast, and synthetic-acoustic.

We find, therefore, a continuum of melodic-similarity models, from those based on distances in time and pitch, through those based on progressively abstracted features such as pitch class, the power spectrum of the chromagram, and structural reductions, to models based on sets of derived features. As with empirical studies of melodic similarity, models show a multiplicity of different bases and different kinds. If these even deal with related phenomena, let alone the same phenomenon, a significant task of meta-modelling will be required to reach a unified understanding of melodic similarity.

3.2 The importance of interpretation

The role of interpretation in similarity is particularly clear in the case of models based on reduction. The reduction systems of Rizo and colleagues and of Orio & Rodà each produce a single definitive reduction of a melody, but they will not always produce the same reduction for a given melody. While my system can produce a single reduction, one important finding is that a very large number of reductions is possible on the basis of the 'rules' inferred from writings on Schenkerian analysis alone (Marsden, 2010b). Indeed, music analysts commonly recognise that alternative analyses of the same piece of music are possible and valid. If multiple reductions are possible, how should a similarity-measurement procedure based on reduction select which reduction to use? From the perspective of Tversky's feature-set model, a melody will be represented by a different set of features according to different ways of reducing it. We should therefore expect the similarity between two melodies to vary according to different reductions. The structure of one melody is likely to prime certain structural interpretations of another, so we should expect the presence of one melody to influence the reduction of another, and a lack of triangle inequality to follow as a natural consequence. There are likely to be sets of melodies *a*, *b*, *c* such that *a* and *b* can be reduced to appear similar, and *b* and *c* also be reduced to appear similar, reducing *b* in a different way, but no reduction of *a* and *c* makes them appear similar. This is likely to be true of any model of similarity which depends on interpretation.

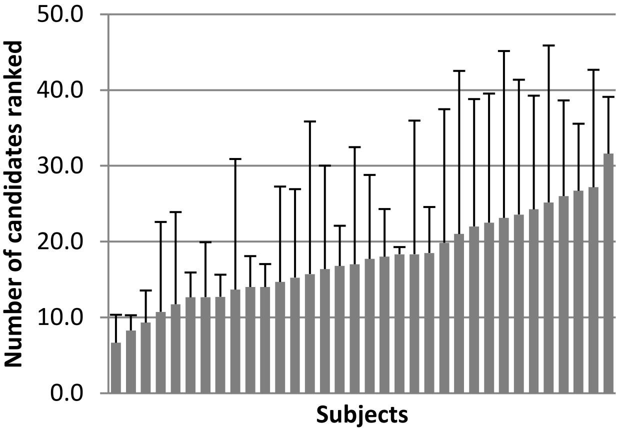

Figure 1. Average number of candidates ranked by each subject, with error bars showing standard deviation.

4. Modelling the MIREX similarity data

The experiment by Typke et al. (2005, 2007) is one of the most extensive and thorough in the literature. Several aspects of the experiment's design reflect the objective of establishing a set of 'ground truth' data for the MIREX 2005 melodic similarity information-retrieval software competition (Downie, 2008). All the musical materials were drawn from RISM A/II (a catalogue giving the initial melodies of pieces in a large number of music manuscripts dating from 1600 onwards) and so were extracts from genuine pieces of music. Eleven items were selected to be 'queries' against which other melodies ('candidates') were to be compared for similarity. For each query, between 45 and 70 candidates were selected and presented to subjects, in music notation, who were asked to rank the candidates according to their similarity to the query. Subjects were not required to rank all candidates. There were a total of 34 subjects, all with some degree of musical training, but not all subjects ranked candidates for all queries. Each query was ranked by at least 25 subjects. The essential raw data, therefore, was a distribution of ranks for each candidate from 1 to a maximum (varying from 45 to 70), representing its similarity to the query. For the MIREX contest, the median rank was taken as an indication of the similarity of the candidate to the query.

The results of this experiment (kindly made available by Rainer Typke) constitute an extremely rich source of data with respect to judgements of melodic similarity. To what degree does this data support the idea that similarity can be modelled as a function of two melodies? What is there about the data which supports instead the idea that similarity depends also on other factors, such as the other melodies presented to the subjects at the time? (Further detail on my analysis of this data set can be found in Marsden, 2012.)

4.1 Distributional analysis of the MIREX data

It is important at the outset to get clear the essential features of the data. The difference in the number of candidates ranked by each subject is striking (Figure 1). Clearly they had different ideas about how similar candidates needed to be in order to be included in the ranking. The deviation in the number of candidates ranked also varied enormously from one subject to another, though the figures do indicate that in the case of three or four subjects it is possible that they started by ranking almost all the candidates but then changed their strategy to rank only a few.

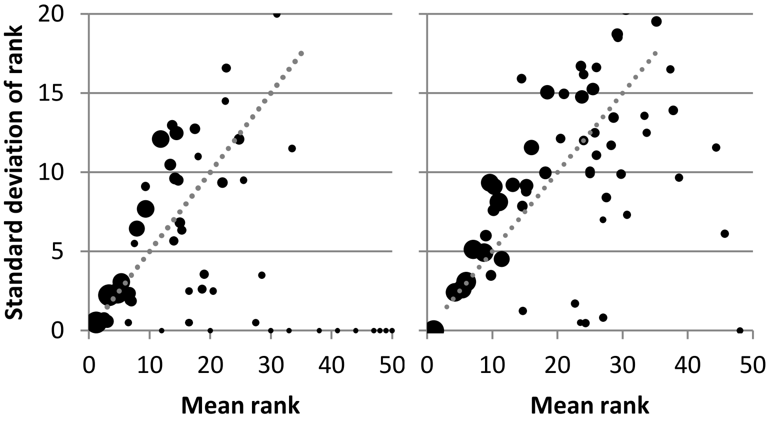

Figure 2. Mean ranking of candidates against standard deviation for three queries. The size of dots indicates the number of times the candidate was included in the ranking. The diagonal dots indicate the relation between mean and standard deviation expected from an even (rectangular) random distribution.

While the rankings of candidates do allow clear differentiation in similarity to the query to be inferred (so fulfilling the objective of the experiment in providing ground-truth data for MIREX) the rankings also showed wide variation. Figure 2 shows graphs of the mean ranking and standard deviation for each candidate with respect to three of the eleven queries. (The patterns for the other queries are similar.) The candidates ranked most often (indicated by the larger sized dots) are generally ranked most similar (i.e., closer to 1) but the variation in ranking for each candidate is generally rather large. As shown on the graphs, the deviation is usually *greater* than would be expected from a random selection of rank from any value up to a maximum of 2 times the mean ranking for each candidate. In other words, for most candidates there is a long 'tail' of some subjects placing this candidate much later in the ranking than other subjects. The relationship between the number of times a candidate is included by a subject in a ranking and the mean rank of that candidate is confirmed in the graphs in Figure 3, which also suggest, by the pronounced 'elbow' in the first and third graph, the employment of two strategies by subjects: either to rank candidates up to a limit of judged similarity, or to rank almost all of the candidates (adopted by one subject in the case of the first query and by four in the third).

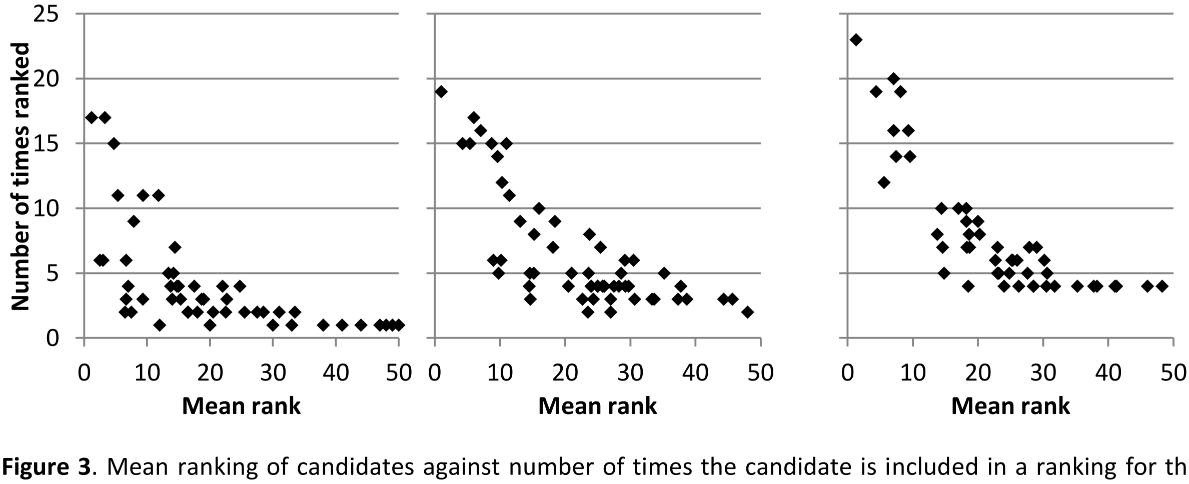

Figure 3. Mean ranking of candidates against number of times the candidate is included in a ranking for the same three queries as in Figure 2.

4.2 Monte Carlo modelling of the MIREX data

The data were further analysed through a method of 'Monte Carlo' simulation. This is a mechanism for testing assumptions about the mechanisms underlying a set of data. A method is designed for simulating data which embodies those assumptions but (ideally) is otherwise random. After many simulations the random effects should disappear, and the distribution of the simulated data should arise solely from the assumptions. If the observed data are 'typical' within that distribution, then the assumptions can be considered to model the mechanisms underlying the observed data.

In this analysis new sets of rankings were generated stochastically on the basis of the following three assumptions:

1. For each query, each subject determined *a priori* how many candidates should be included in the ranking.
2. For each candidate and each query, there is a fixed likelihood of being included in the ranking, relative to other candidates.
3. For each candidate and each query, there is a fixed probability function for the position it will take in the ranking.

The following procedure was used to simulate data. For each query, a number of rankings were simulated equal to the number of subjects who ranked candidates against that query. Each ranking was generated as follows:

A. In order to select appropriate candidates for ranking, an initial ranking was made by assigning a number to each candidate and placing the candidates in ascending order according to the number assigned. The first *n* candidates were selected from this initial ranking, where *n* is the number selected by the subject whose ranking was to be simulated. The numbers used in this initial ranking were selected at random from one of two distributions:

   proportion: a uniform distribution between 1 and *t* / *s*, where *s* is the number of subjects who selected that candidate and *t* is the total number of subjects who were presented with the query, or

   gamma: a gamma distribution based on the mean and variance of the actual ranks for that candidate, augmented to ensure that all candidates had an equal-area distribution. This augmentation was made on the basis of the assumption that, if every subject had ranked all the candidates, the ranks for unranked candidates would have been distributed evenly throughout the unused ranks.

B. The final ranking was made by once again assigning a number to the selected candidates and placing the candidates in ascending order according to the number assigned. Where more than one candidate was assigned the same number, they were ordered in the final ranking at random. The numbers used in this ranking were again selected at random from one of two distributions:

   distribution: the actual distribution of ranks for that candidate in the original data, or

   gamma: a gamma distribution based on the mean and variance of the actual ranks for that candidate (not augmented as described above).

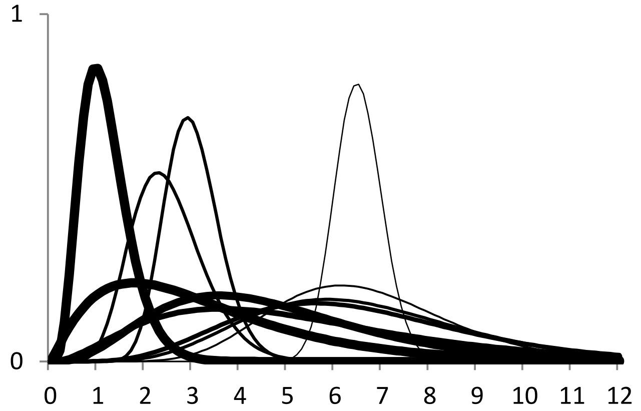

Figure 4. Gamma distributions for simulating rankings of the ten highest ranked for the first query illustrated in Figures 2 and 3. The weight of lines corresponds to the number of times the candidate was included in a ranking.

A gamma distribution is the distribution for a random variable with range from 0 to infinity for a given mean and variance with maximum entropy. Any other distribution would have introduced other non-random factors into the model. Figure 4 illustrates the gamma distributions used in step B for the ten highest ranked candidates for the first query illustrated in Figures 2 and 3. (All these candidates occupy the bottom left corner of the first graph in Figure 2.) Note that the gamma distributions used in steps A and B were not identical. As explained above, those used in A took into account the non-inclusion of candidates in some rankings so as to form a better basis for selecting which candidates to rank. Simulation was also tested using the same gamma distributions for step A as in step B, but this produced rather poor results in terms of the test of fit described below. This suggests that ranking involved two kinds of decision for the subjects: whether or not to include a candidate in the ranking, and where to place it in the ranking.

The two points of choice in the procedure outlined above combine to make four methods of simulation constituting four models. The 'proportion-distribution' model used the first distribution in each of steps A and B, the 'gamma-distribution' used the first distribution in step A and the second in step B, and so on. For each model, the fit with the data was tested in a manner which can be related to tests of statistical significance. In a standard test of significance, one determines the expected distribution of values under the assumption of the null hypothesis. If the distribution has only a small area 'beyond' the observed value, i.e. the likelihood of values at least as extreme as the observed value—the *p*-value—is small, then we can conclude that the null hypothesis is likely to be false. In Monte Carlo simulation one determines the distribution of values expected under the test assumptions (not the null hypothesis) by generating a large number values stochastically. The assumptions are supported if the observed value falls in the middle of the distribution, i.e. values at least as extreme as the observed value have a high likelihood. 'At least as extreme' means greater than or equal to the observed value if that value is above the mean or less than or equal to if it is below. A perfect fit between model and data would result in a *p*-value of at least 0.5.

It is not immediately obvious what the test statistic should be for fit with the data in this case, so a number of test statistics were used, as follows:

a. **mean rank:** the overall mean rank,

Table 1. Fit of stochastic models to actual data according to six criteria. A value of at least 0.5 would indicate a perfect fit.

Model	a. mean ranking	b. mean variance	c. mean rank distribution	d. times ranked mean *p*	e. ranking mean *p*	f. rank distribution mean *p*
proportion-distribution	0.191	0.202	0.020	0.487	0.345	0.322
proportion-gamma	0.196	0.173	0.052	0.486	0.336	0.295
gamma-distribution	0.165	0.130	0.079	0.402	0.337	0.286
gamma-gamma	0.174	0.113	0.092	0.402	0.332	0.276

b. **mean variance:** the overall mean variance of rank for each candidate,

c. **mean rank distribution:** the overall mean of the difference in distribution of ranks for each candidate and the average simulated distribution for that candidate, as measured by the sum of squares of difference,

d. **times ranked mean *p*:** the mean *p* of the number of times each candidate is ranked,

e. **ranking mean *p*:** the mean *p* for the mean rank of each candidate, and

f. **rank distribution mean *p*:** the mean *p* for the difference in distribution of rankings for each candidate and the average simulated distribution for that candidate.

For each of the four models, each of these test statistics was determined for each of the eleven queries on the basis of 10,000 sets of simulated data. The average for all eleven queries for each model is reported in Table 1. As can be seen from the table, the fit was far from perfect. For the first of the eleven queries, for example, the mean rank for all candidates was 18.724. In the corresponding data generated stochastically by the 'proportion-distribution' model, the mean rank was 18.894, and only 34% of the sets of generated rankings had an average ranking of 18.724 or less, giving a *p*-value of 0.34. The *p*-values for other queries were mostly lower than this, resulting in the overall average of 0.191 reported in the table.

The fit with respect to the test statistics **a-c**, which measure the fit with respect to overall averages, is particularly poor, indicating that factors other than the three assumptions outlined above had an influence on ranking. These measures of fit with respect to overall averages are more likely to show the cumulative effect of small but systematic deviations between the model and the data, and indeed the other three values, which instead measure the mean fit for each candidate and query, show a moderate degree of fitting to the actual data. While the best fit is achieved by the first model, which makes most use of the detail of the actual data (using the number of times each candidate is ranked as a basis for selection, and the actual ranking profiles as a basis for ranking) the fit for the models which make use of gamma distributions is not markedly worse. The 'proportion-gamma' model, for example, fits the data moderately well on the basis of just three values for each candidate compared to about 30-70 values for the 'proportion-distribution' model. These three values are the likelihood of the candidate being included in the ranking, the mean rank, and the variance in rank. These constitute two kinds of measures of similarity and a measure of uncertainty.

4.3 Evidence for similarity between candidates to be ranked

In placing candidates in a ranking rather than simply judging the similarity between two or three melodies, as required in most of the experimental paradigms referred to in Section 2.1 above, it is possible that subjects might have been influenced by perceived similarity between the candidates to be ranked as well as by their perceived similarity to the query. Three kinds of ways in which one candidate might influence the ranking of another were tested using the same Monte Carlo method as above, but in this case a *lack* of fit between simulated data and the actual data was evidence in support of the hypothesis. One is seeking evidence for an effect *not* captured by the assumptions embodied in the models. The three values tested for lack of fit were:

a. the number of times two candidates were both included in a ranking, testing whether the inclusion of one candidate would influence the inclusion of another,
b. the difference in rank between two candidates compared to the difference in mean rank for those candidates, testing for an influence which placed candidates closer together or further apart in the ranking than implied by their respective similarity to the query, and
c. the mean rank for a candidate when a second candidate is included in the ranking compared to the mean rank when the second candidate was not included, testing for an influence which causes a candidate to appear more or less similar to the query depending on the presence of another candidate.

The procedure was as follows. As before 10,000 sets of simulated data were generated using each of the four models. For each pair of candidates, the values of the first, fifth, ninety-fifth and ninety-ninth percentile for each of the test statistics in the distribution of the simulated data were determined, in other words, the values which would be matched or exceeded with only 1% or 5% likelihood. Then the values of test statistics **a-c** in the actual data were compared to these for the same candidates and the percentage of pairs for which the value was lower than the fifth or higher than the ninety-fifth percentile entered in the column of Table 2 headed '$p < 0.05$' for that statistic, and the percentage of pairs for which the value was lower than the first percentile or higher than the ninety-ninth entered in the column headed '$p < 0.01$'. For the null hypothesis that the assumptions model the data to be rejected, and for us to conclude that there is an effect of one candidate on the ranking of another, the values under '$p < 0.05$' should be above 5% and those under '$p < 0.01$' above 1%. In other words, extreme values which are rare in the simulated data should be more common in the actual data.

The results shown in Table 2 are somewhat equivocal. The values for statistic **a**, the percentage of extreme values in the actual data for candidates being included together in the ranking, is lower than would be expected rather than higher, meaning that the null hypothesis cannot be rejected. Probably this is a result of the fact that this value is expressed in small integers, and so a large number of simulated cases will be equal to the actual values rather than greater or less. Extreme values for the other two statistics, however, are about two to three times as common as expected from the model. In other words, candidates are ranked significantly closer or further apart than would be expected from their difference to the query alone more often in the actual ranking data than in the simulated data, and candidates are ranked significantly higher or lower when another candidate is included in the ranking compared to when it is not more often also. This suggests that in

Table 2. Proportion of pairs of candidates with extreme values for the number of times both candidates are included in a ranking, for the difference between the difference in mean ranks of the candidates and the mean difference in the actual ranks, and for the difference between the mean rank for the first candidate when the second is included in the ranking and the mean rank when the second is not included in the ranking. A value is considered 'extreme' if it is greater (or less) than 5% of the corresponding simulated data ($p < 0.05$), and greater (or less) than 1% ($p < 0.01$).

Model	a. times pairs ranked		b. difference in mean ranking – difference in actual ranking		c. mean difference in rank dependent on inclusion of other candidate	
	$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$	$p < 0.05$	$p < 0.01$
proportion-distribution	1.06%	0.36%	12.88%	2.98%	10.31%	3.57%
proportion-gamma	1.09%	0.30%	13.09%	2.75%	10.91%	4.03%
gamma-distribution	1.75%	0.18%	12.33%	2.59%	9.64%	3.61%
gamma-gamma	1.74%	0.18%	12.47%	2.58%	10.54%	4.29%

the actual experiment there was an effect between candidates as well as between the candidate and the query, and that similarity therefore cannot be modelled simply as a function of two melodies.

The data reported in Table 2 should themselves be subject to a test of significance. Is the difference between, for example, the observed 12.88% and the expected 5% significant in the context of the simulated data? The significance of these figures was tested by counting for each set of simulated rankings for a query, the number of pairs of candidates with an 'extreme' value for the test statistics **b** and **c**, where 'extreme' is defined as before as below the fifth or above the ninety-fifth percentile at the $p < 0.05$ level and below the first or above the ninety-ninth at the $p < 0.01$ level. The proportion of simulated sets of data with a number of such pairs greater than or equal to the number of such pairs in the actual data could then be determined, to yield a measure of the likelihood of observing the actual data under the null hypothesis. (It was necessary to run the simulation twice, i.e., 20,000 times, to achieve this: once to determine the percentile values and once to count the number of pairs exceeding these. Because of the high computational load, this was only done for the 'proportion-gamma' model.) The results (reported in Marsden, 2012) showed that, when averaged across all eleven queries, an effect could be observed only at significance levels ranging from $p < 0.09$ to $p < 0.18$, which would not normally be regarded as a sufficiently low level to reject the null hypothesis. However, the significance level varied widely from one query to another, ranging from $p < 0.005$ to $p < 0.49$. For only one query was the p-value less than 0.05 for both test statistics at both levels for the definition of 'extreme'.

4.4 Conclusions from the MIREX data

The fact that the simulation fits the data moderately well and the weak evidence for inter-candidate effects suggests that melodic similarity, at least as understood by the subjects in this experiment, is indeed a function of two melodies. The high degree of variance in ranking, however, shows that either this function differs widely from one individual to another, that it is highly non-deterministic,

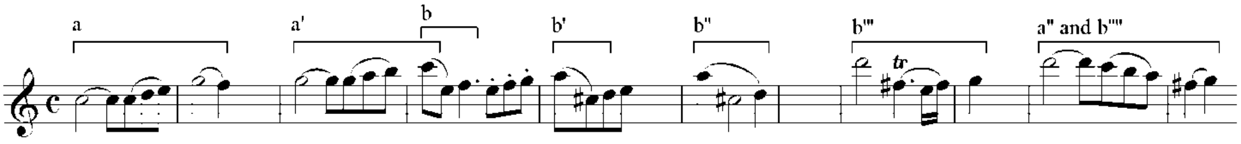

Figure 5. Extracts from Mozart's string quartet in C major, K. 465, first movement.

or that it is influenced by other as yet unidentified factors. The analysis above suggests that the influence of one candidate on the similarity of another to the query might be one of these factors, but this cannot be concluded with any confidence from this data. It seems that even a data set as rich as this one is not sufficiently large for drawing complex conclusions.

5. Examples of similarity and variation

The conclusions of Tversky (1977) and my argument in Section 3.2 emphasise the importance of interpretation in the recognition of similarity. Some styles of composition, in particular the writing of variations and the use of recurrent motives, depend on the listener recognising similarity, and so it seems entirely plausible that a listener will *seek* similarity in what he or she hears, and so interpret what is heard precisely in such a way as to maximise the perception of similarity. Here I outline two interesting examples from the music of Mozart where similarity is crucially related to alternative interpretations.

5.1 Example 1: Mozart K. 465

Figure 5 shows extracts from the first violin part of Mozart's string quartet in C major, K. 465 ("Dissonance"). (For more detailed discussion see (Marsden, 1989).) The allegro begins with the theme shown as **a**. This is immediately repeated a tone higher (not shown) and then, with a slight modification, as **a'**. The last note of **a'** begins a new motive **b** which appears to contrast with **a** (descending instead of rising; made up largely of shorter notes; containing a large leap instead of mostly steps). This is repeated at **b'** (reinforcing the identity of the motive) and then in rhythmic transformation some bars later at **b''** (where the recognition of similarity is aided by using exactly the same pitches). Several bars later the figure identified as **b'''** is heard, whose similarity to **b''** is aided by the equivalent durations of the second note (though in the case of **b'''** it is decorated with a trill). Finally, beginning on the same pitch as **b'''** and ending with the same pair of pitches, a figure is heard which is also clearly similar to **a** by inversion. (Indeed, to help make this clear, the intervening music has presented several other versions of **a** without inversion.) This figure is easily recognised as similar to *both* **a** and (with the aid of the intermediate transformations) **b**.

Is it true, then, that **a** is similar to **b**, despite the fact that at first the motives seemed to be contrasted? If it is, then we must reduce **a** in *different* ways to find maximum similarity in each case. To find maximum similarity between **a** and **a'**, we must reduce **a** by removing the appoggiatura on the last note, which implies that the remaining notes are passing notes from C to F. To find maximum similarity between **a** and **b**, on the other hand, the first step must be to reduce out the quavers in **a** and regard the appoggiatura (neighbour note) as prior. It was my contention in the original analysis (Marsden, 1987) that Mozart intended this play with our sense of the difference and similarity between these motives as a way of capturing the listener's interest.

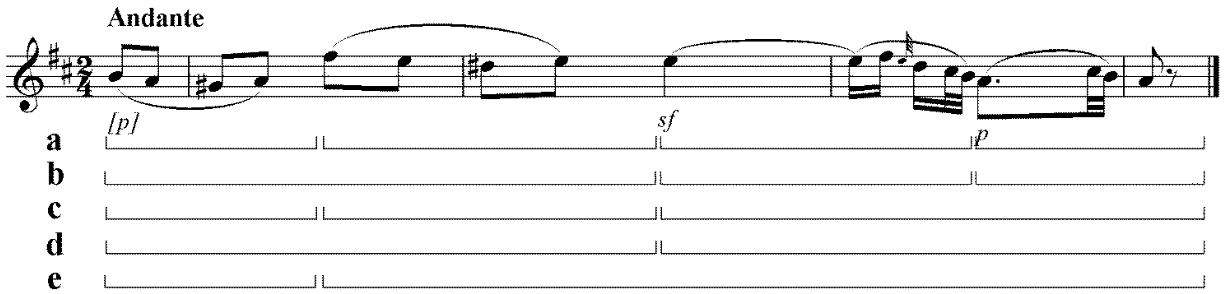

Figure 6. Alternative segmentations of the second phrase of the theme of the third movement of Mozart's string quartet in A major, K. 464.

5.2 Example 2: Mozart K. 464

Reduction is not the only aspect of music which is open to alternative interpretations. The same is true of segmentation. The second phrase from the opening of the theme from the third movement of Mozart's string quartet in A major, K. 464 is ambiguous in its segmentation, as illustrated in Figure 6. The articulation (indicated by the slurs) suggests a segmentation into four units each one bar long, as shown in segmentation **a**. Similarities and contrasts in the melodic material, on the other hand, suggests segmentation **b**, which puts the first two bars into a single unit on account of the repeated melodic pattern. Rhythmic proximity dictates that segments rarely end with short notes, and the rhythm of this phrase suggests the segmentation **c**, where the last two bars are grouped together because of the run of short notes in the penultimate bar. Segmentation **d** follows from the contrast in dynamics introduced by the *sforzando* on the crotchet E and is supported by the division of the phrase into two halves of equal length. Pitch proximity suggests segmentation **e**, which places a break between units at the only interval larger than a third.

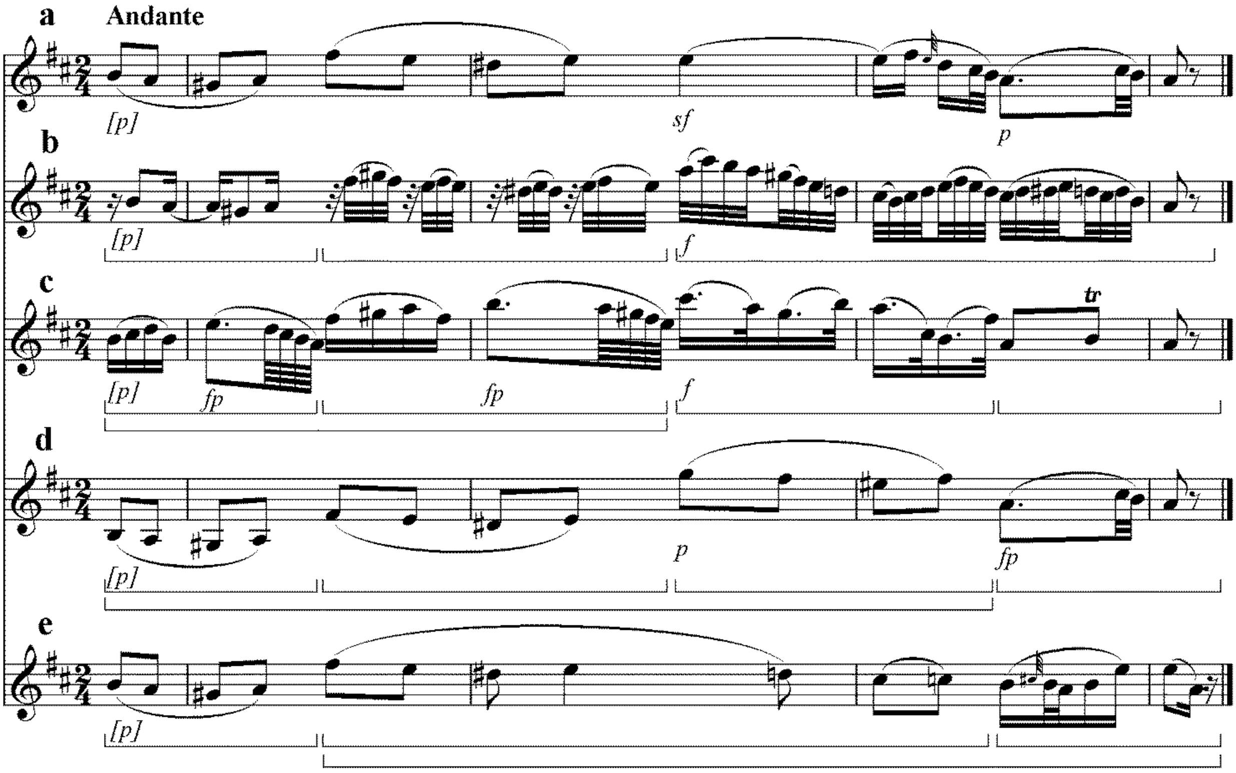

Figure 7. Different segmentations found in variations in Mozart's K. 464 of the theme from Figure 6.

Each one of these segmentations, and one other which divides the four bars into 3+1, can be found in the subsequent variations, as shown in Figure 7. Once again, Mozart seems to be deliberately exploiting multiple ways of interpreting the same music in order to provoke different perceptions of similarity.

6. Similarity and creativity

Each of the perspectives on melodic similarity in Sections 2-5 shows only questionable evidence for similarity as a distinct and objective function of a pair of melodies. The variety of ways in which the idea of melodic similarity has been used and tested in the literature suggests that it is perhaps not a single phenomenon. In at least one particular manner of measurement, using reduction, similarity depends crucially on an analysis of melodic structure which is susceptible to multiple interpretations. The analysis of the MIREX data gives the strongest support for similarity as a function of two melodies, but even here that is seen to be problematic. The examples from Mozart imply that for listeners his music is an invitation to interpret melodies in varied ways to find similarities. The subjects in the MIREX experiment were effectively explicitly invited to find similarity between the query and the candidate melodies. Similarity involves interpretation, and interpretation is always a *creative* act. When musicians say two melodies are similar, the arguments above suggest that the musicians have *created* that similarity as much as recognising it.

Others have argued that it is impossible to find a single measure of melodic similarity for all situations (e.g., Müllensiefen & Frieler, 2007). The arguments above suggest that it is impossible for *any* situation. The best one can hope for is a measure which will usefully approximate human judgements of similarity in a particular situation, and the analysis of the MIREX data suggests that we should expect quite a large degree of error. To me, this suggests that research to fine-tune models of similarity to a particular set of data is unlikely to be productive. Instead, we need to achieve a better understanding of the various factors which contribute to perceptions of melodic similarity, with the aim of modelling the variety of judgements of similarity in different situations and among different individuals.

Above all, I think it is important to recognise that melodic similarity is not a simple or even an objective phenomenon. It depends on musical culture, on the circumstances of comparison, and on the individual interpretation of the observer. Since music is a creative art, we will understand it better by acknowledging the creative aspects of its phenomena also.

References

Ahlbäck, S. (2007). Melodic similarity as a determinant of melodic structure. *Musicae Scientiae*, *11*(1 suppl.), 235–280.

Adiloglu, K., Noll, T., & Obermayer, K. (2006). A paradigmatic approach to extract the melodic structure of a musical piece. *Journal of New Music Research*, *35*(3), 221–236.

Allan, H., Müllensiefen, D. & Wiggins, G. (2007). Methodological considerations in studies of musical similarity. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, pp. 473–478.

Armstrong, M.A. (1983). *Basic Topology*. Berlin: Springer-Verlag.

Bernabeu, J.F., Calera-Rubio, J., Iñesta, J.M. & Rizo, D. (2009). Melodic identification using probabilistic tree automata, *Journal of New Music Research*, *40*(2), 93–103.

Bohak, C. & Marolt, M. (2009). Calculating similarity of folk song variants with melody-based features. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, pp. 597–601.

Downie, J.S. (2008). The music information retrieval evaluation exchange (2005–2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*(3) 247–255.

Eeerola, T., Järvinen, T., Louhivuori, J. & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies, *Music Perception*, *18*(3), 275–296.

Eerola, T. & Bregman, M. (2007). Melodic and contextual similarity of folk song phrases, *Musicae Scientiae*, *11*(1), 211–233.

Giannopoulos, P. & Veltkamp, R.C. (2002). A pseudo-metric for weighted point sets. *Proceedings of the European Conference on Computer Vision (ECCV 2002),* LNCS 2352, Berlin: Springer, pp. 715–730.

Hofmann-Engl, L. (2003). Melodic similarity and transformation: a theoretical and empirical approach. PhD thesis, University of Keele.

Hu, N., Dannenberg, R. & Lewis, A.L. (2002). A probabilistic model of melodic similarity. *Proceedings of the International Computer Music Conference (ICMC)*, San Francisco, pp. 509–515.

Juhasz, Z. (2006). A systematic comparison of different European folk music traditions using self-organising maps, *Journal of New Music Research*, 2006, *35*(2), 95–112.

Krumhansl, C. (1990). *Cognitive Foundations of Musical Pitch*. Oxford University Press.

Marolt, M. (2008). A mid-level representation for melody-based retrieval in audio collections. *IEEE Transactions on Multimedia*, *10*(8), 1617–1625.

Marsden, A. (1987). *Analysing music as listeners' cognitive activity, a study with reference to Mozart*, PhD thesis, Cambridge University.

Marsden, A. (1989). Listening as discovery learning, *Contemporary Music Review*, 1989, *4*(1), 327–340.

Marsden, A. (2010a). Recognition of variations using automatic Schenkerian reduction. *Proceedings International Conference on Music Information Retrieval (ISMIR)*, Utrecht, The Netherlands, pp. 501–506.

Marsden, A. (2010b). Schenkerian analysis by computer: a proof of concept. *Journal of New Music Research*, *39*(3), 269–289.

Marsden, A. (2012). Melodic similarity: a re-examination of the MIREX2005 data. *Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Conference of the European Society for the Cognitive Sciences of Music,* Thessaloniki, 653–659.

Müllensiefen, D. & Frieler, K. (2004). Cognitive adequacy in the measurement of melodic similarity: algorithmic vs. human judgments, *Computing in Musicology*, *13*, 147–176.

Müllensiefen, D. & Frieler, K. (2007). Modelling experts' notions of melodic similarity, *Musicae Scientiae*, *11*(1 suppl.), 183–210.

Novello, A., McKinney, M.M.F. & Kohlrausch, A. (2011). Perceptual evaluation of inter-song similarity in western popular music, *Journal of New Music Research*, *40*(1), 1–26.

Orio, N. & Rodà, A. (2009). A measure of melodic similarity based on a graph representation of the music structure. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Kobe, Japan, pp. 543–548.

Pardo, B., Shifrin, J. & Birmingham, W. (2004). Name that tune: a pilot study in finding a melody from a sung query, *Journal of the American Society for Information Science and Technology*, *55*(4), 283–300.

Polk, T.A., Behensky, C., Gonzalez, R., & Smith, E.E. (2002). Rating the similarity of simple perceptual stimuli: asymmetries induced by manipulating exposure frequency, *Cognition*, *82*(3), B75–88.

Rizo, D. (2010). *Symbolic music comparison with tree data structures*. PhD thesis, University of Alicante, Spain.

Schenker, H. (1935). *Der frei Satz*. Vienna: Universal Edition. Published in English as *Free Composition*, translated and edited by E. Oster, New York: Longman, 1979.

Schmuckler, M.A. (2010). Melodic contour similarity using folk melodies, *Music Perception*, *28*(2), 169–193.

Tversky, A. (1977). Features of Similarity, *Psychological Review*, *84*(4), 327–352.

Typke, R. & Walczak-Typke, A.C. (2008). A tunnelling-vantage indexing method for non-metrics. *Proceedings International Conference on Music Information Retrieval (ISMIR)*, Philadephia, pp. 683–688.

Typke, R., den Hoed, M., de Nooijer, J., Wiering, F. & Veltkamp, R.C. (2005). A Ground truth for half a million musical incipits, *Journal of Digital Information Management*, 3(1), 34–39.

Typke, R., Wiering, R. & Veltkamp, R.C. (2007). Transportation distances and human perception of melodic similarity, *Musicae Scientiae*, *11*(1 suppl.), 153–181.

Urbano, J., Lloréns, J., Morato, J., & Sánchez-Cuadrado, S. (2011). Melodic similarity through shape similarity. In S. Ystad, M. Aramaki, R. Kronland-Martinet & K. Jensen (eds.) *Exploring Music Contents* (LNCS 6684), Springer, 338–355.

Volk, A., van Kranenburg, P., Garbers, J., Wiering, F., Veltkamp, R.C. & Grijp, L.P. (2008). A manual annotation method for melodic similarity and the study of melody feature sets. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Philadephia, pp. 101–106.