Melodic Similarity: A Re-examination of the MIREX2005 Data

Alan Marsden

Lancaster Institute for the Contemporary Arts, Lancaster University, UK
a.marsden@lancaster.ac.uk

ABSTRACT

Despite a considerable body of research, there is no clarity about the basic properties of melodic similarity, such as whether or not it constitutes a metric space, or whether it is a more complex phenomenon. An experiment conducted by Typke et al., used as a basis for the MIREX2005 melodic-similarity modelling contest, represents a particularly rich source of data. In the experiment, for each of eleven queries (melodies taken from RISM A/II), about 25 experts ranked some of about 50 candidates for similarity with the query. A 'Monte Carlo' approach has been taken in re-examining this data, simulating data in the same form on the basis of simple assumptions about the nature of melodic similarity. Statistical properties of the actual data were compared with the same properties for 10,000 sets of simulated data, allowing estimation of the significance of differences found. In terms of overall measures such as the ranking profile for each candidate, quite good simulations (i.e., sets of simulated data in which the original falls within the second and third quartiles in the measured property) arose from stochastic ranking based only on the mean and variance of the actual ranking for each candidate and on the likelihood of the candidate being selected for ranking. However, the simulations did show evidence, in a substantial minority of cases, of an effect for some candidates to be ranked higher or lower dependent on the presence of another candidate, and of the influence of similarity between candidates.

I. INTRODUCTION

Previous research has demonstrated that naive listeners and experts are able to state the similarity of two melodies. Psychologists have been interested in studying the basis for the perception of similarity between two melodies as a means of exposing underlying mechanisms of cognitive representation of melody (e.g., Eerola & Bregman, 2007; Schmuckler, 2010). Researchers in the field of Music Information Retrieval have also been interested in the concept, but more as a means of organizing search engines and other software tools for processing and organizing large quantities of musical data (e.g., Pardo, Schifrin & Birmingham, 2004; Novello, McKinney & Kohlrausch, 2011). Research in the first field is characterized by controlled and precise experimentation with small quantities of music, whereas the second field has typically used large quantities of music with less fine-grained examination of musical details. In neither field is there any clarity about the fundamental nature of melodic similarity. Does it constitute a metric space in which, if we only knew the dimensions and how to measure them, we could place any set of melodies? Does the similarity between two melodies depend only on the properties of those melodies, or does it depend on emergent properties of the juxtaposition of melodies or even on other context- dependent factors? Is it a unitary phenomenon at all? (Further discussion can be found in (Marsden, forthcoming).)

This paper aims to explore approaches to answers to these questions through analysis of data from a previous experiment (Typke et al., 2005; Typke, Wiering & Veltkamp, 2007) which is an important exception to the characterization of the two fields above. The quantity of material used in this experiment is as large as found in experiments in Music Information Retrieval, but in other respects it exhibits the quality of control and analysis found in experiments in Music Psychology. This experiment formed the basis for the 'ground truth' used in the MIREX melodic-similarity contest in 2005 (Downie, 2008).

A. The MIREX 2005 experiment

Details of the original experiment are given in (Typke et al., 2005). Eleven melodic incipits were selected from the RISM A/II database to be 'queries' against which other melodies ('candidates', also drawn from RISM A/II) 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 candidates ranked candidates for all queries. Each query was ranked by at least 25 subjects.

For the MIREX ground-truth data, the median ranking of a candidate was taken as an indication of the similarity of the candidate to the query accompanied by analysis to determine the significance of difference in ranking. This data was then used to evaluate software designed to measure the similarity between melodies. However, the data also provides a rich source of other information, which is subject to further analysis here.

II. DISTRIBUTIONAL ANALYSIS

Candidates varied markedly in two measures: the number of subjects who included the candidate in their ranking, and the mean rank across the rankings by all subjects. Among the candidate melodies was often included an exact copy of the query, and, as expected, this was almost always included in the ranking, and usually ranked first.

A. Mean ranking and deviation

The variation in the rank to which candidates were assigned by different subjects was quite large. Figure 1 shows graphs for each of the eleven queries showing the mean rank for candidates in the horizontal dimension and their standard deviation in ranking in the vertical dimension. The size of each dot corresponds to the number of subjects who ranked the candidate. For candidates ranked more than a handful of times, the deviation in ranking is quite large. The diagonal line shown in grey dots on each graph shows the relationship between mean and standard deviation which would arise from a random distribution of ranks for a candidate selected with equal probability in a range between 1 and twice the mean rank. For most candidates ranked a significant number of times, the deviation is greater than this, indicating a relatively long 'tail'

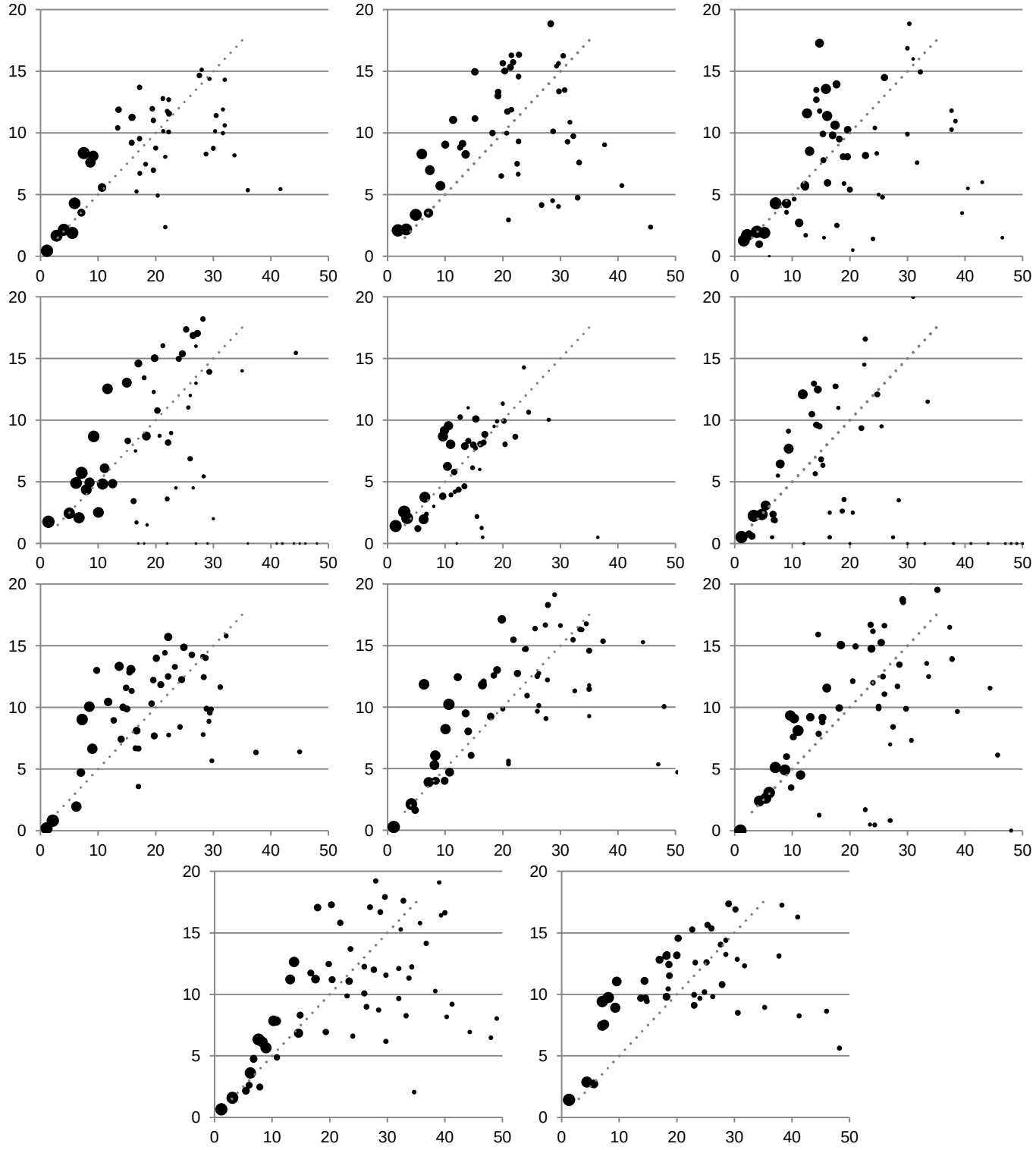

Figure 1. Mean rank of candidates (horizontal) to standard deviation (vertical) for each of the eleven queries. The size of dots corresponds to the number of subjects who included the candidate in their ranking.

in which a candidate can sometimes be placed quite low in the ranking.

It is clear, therefore, that, for most candidates, either subjects had very different ideas about how to determine their similarity to the query, or subjects' judgements of similarity are rather non-deterministic. It would be interesting to conduct an experiment similar to that by Typke et al. which retested subjects after an appropriate interval to determine whether the source of the large deviation in ranking is differences between subjects or inherent variability in judgements of similarity.

B. Mean ranking and probability of selection

The graphs in Figure 1 suggest that candidates which are selected more often for ranking are likely to be ranked higher. This is confirmed by a correlation of mean ranking with the number of times a candidate is selected for ranking, but graphs of this relation (a sample of which is shown in Figure 2) exhibit

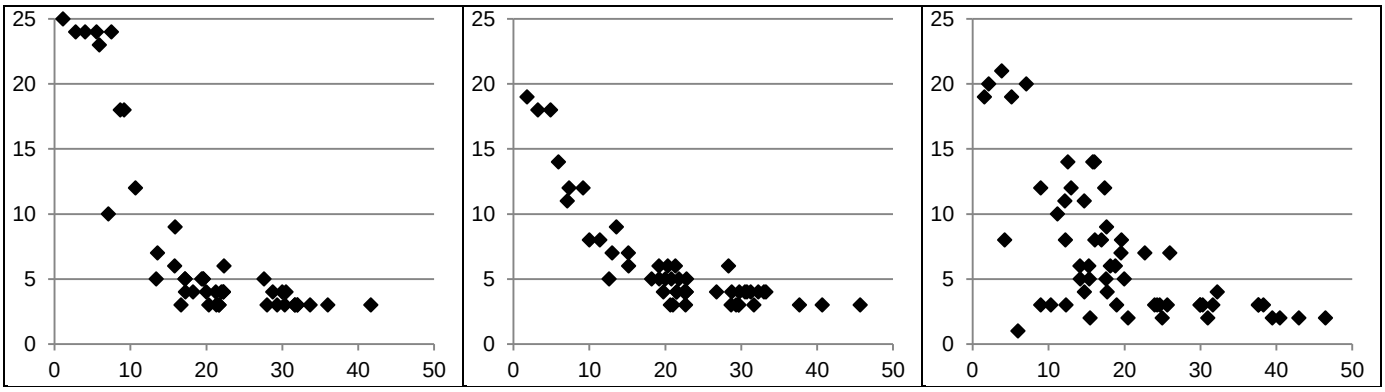

Figure 2. Mean rank of candidates (horizontal) to number of subjects selecting the candidate for ranking (vertical) for the first three queries shown in Figure 1.

a shape characterized by a negative slope on the left and a horizontal line on the right, suggesting that subjects operated one of two strategies: either they ranked only those candidates which they regarded as similar to the query or they ranked all or most of the candidates. Subjects varied greatly in the number of candidates they ranked (see Figure 3), and obviously the maximum rank assigned by a subject who ranked only a few candidates was smaller than the maximum ranked by a subject who ranked many, resulting in a tendency for the frequently selected candidates to be ranked more highly than those selected less frequently and a negative slope in the graph. Rankings by subjects who ranked most or all of the candidates, on the other hand, gave rise to the horizontal trend observed on the right of the graphs.

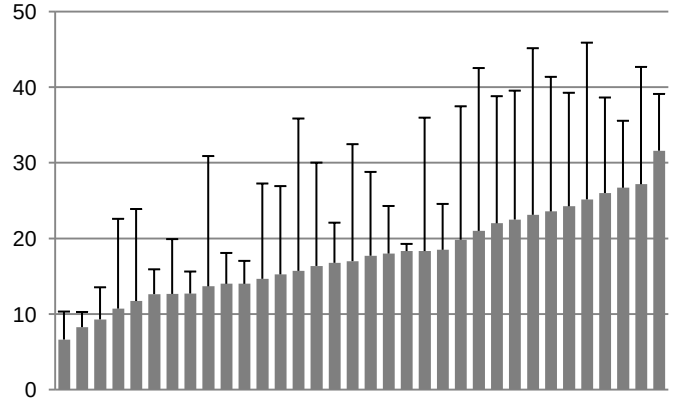

Figure 3. Mean number of candidates selected for ranking by each subject, with error bars indicating standard deviation.

It is clear therefore that not only did subjects vary greatly in their ranking of candidates, but they varied also in their interpretation of the task required in the experiment.

III. MONTE-CARLO ANALYSIS

A 'Monte Carlo' analysis was carried out to investigate the data further and test models of the subjects' ranking process. Rankings were generated in a stochastic fashion on the basis of sets of assumptions, which together constituted models of the ranking process. If the generated data matched the original, according to certain statistical criteria, then the model could be regarded as being confirmed. On the other hand, differences between the original and simulated data indicated phenomena in the actual ranking process not accounted for in the model being tested. Just as a classic p-value gives a measure of the significance of a statistic by estimating the proportion of the population expected to have values as extreme as that statistic under the null hypothesis, the modelling process similarly allowed the estimation of the significance of a difference found between the original and simulated data: a large number of sets of data could be simulated, and the proportion of sets which have a value as extreme as the corresponding value in the original data counted.

A. Modelling assumptions and process

The modelling process was based on the following basic 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.

In initial testing, the first assumption was replaced by an assumption that candidates were selected for ranking on the basis only of their perceived similarity to the query, determined by either of the two methods outlined under (2) below, but this failed to produce data which came close to matching the original, once again providing evidence for the conclusion drawn at the end of section II above, that subjects had quite different interpretations of the task required.

The following procedure was used in simulation of data:

1. For each query, a number of rankings were simulated equal to the number of subjects who ranked candidates against that query.
2. For each ranking (simulating the ranking of candidates made by a subject), the first n candidates were selected from an initial ranking of candidates, where n is the number selected by the subject whose ranking was to be simulated. The initial rank for each candidate was determined at random either from

   proportion: 0 to 1/s – 1, where s is the proportion of subjects who selected that candidate, or

   gamma: a gamma distribution based on the mean and variance of the actual rank for that candidate.
3. The selected candidates were then ranked by once again generating an initial rank for each candidate and placing the candidates in the corresponding order. Where more

than one candidate shared the same initial rank, they were ordered at random. The initial rank was determined at random either from

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 rank for that candidate.

A gamma distribution is the maximum entropy distribution for a random variable with range from 0 to infinity for a given mean and variance. This is therefore an appropriate distribution for stochastic modelling of ranking since ranks vary from 1 to an unbounded value, making no further assumptions beyond a mean value and variance. Figure 4 illustrates the gamma distributions used in step (3) for the ten highest ranked candidates for the sixth query illustrated in Figure 1. (All these candidates occupy the bottom left corner of the corresponding graph in Figure 1.)

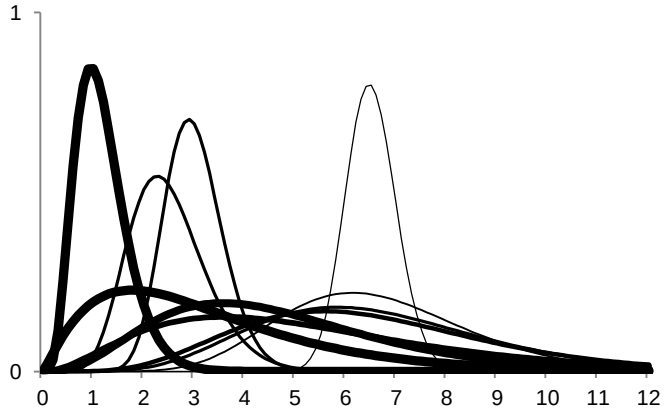

Figure 4. Illustration of gamma distributions for simulating rankings of the ten highest ranked candidates for the sixth query illustrated in Figure 1. The weight of lines corresponds to the number of times the candidate was included in a ranking.

Note that the gamma distributions used in steps (2) and (3) were not identical: those used in (2) took into account the non-inclusion of a candidate in some rankings so as to form a better basis for selecting which candidates to rank. For the purpose of determining the mean and variance for the distribution, the rank for such non-ranked candidates was taken to be evenly distributed between the rank one beyond the last candidate ranked by that subject and the last rank if all the candidates had been ranked. The rationale here was that if the subjects had been required to rank all the candidates, non-ranked candidates would have been ranked somewhere within this range, but we cannot predict where. Simulation was also tested using (a) the same gamma distributions for step (2) as in step (3), and (b) for a gamma distribution based on an assumed distribution where all non-ranked candidates were assigned one greater than the maximum rank for a query. Both of these produced results with a consistently poorer fit to the original data than the procedure outlined above, with the exception of the measure of the number of times each candidate was ranked, for which the procedure (b) above produced a marginally better fit.

That two different gamma distributions produced a better result than one 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.

Some subjects sometimes left gaps in their ranking, though this was not common (an average of 5.5 subjects left an average of 3.7 gaps). It is possible that (a) this was a deliberate strategy to indicate greater dissimilarity between the query and some candidates, or (b) that it was simply a mistake and the subjects would have moved candidates up to fill gaps if they had realised, of (c) that the mistake was to leave candidates in the ranking which the subject had decided no longer warranted ranking. Each model was tested under each of these assumptions, (a) by leaving the rankings unchanged from the original data, (b) by moving candidates up in to fill gaps, and (c) by deleting candidates in rankings after any gap. It proved difficult to model the data with gaps in the rankings as models under this assumption consistently had a poor fit with the data. The fit for assumptions (b) and (c) did not differ significantly, and all results reported below are for modelling under the assumption (c) that candidates after gaps should be deleted from the ranking.

For each model, 10,000 sets of data were simulated to test the degree of fit with the original data.

B. Measuring the fit of a model

The following measures were used to determine the degree of fit between a model and the original data. The term 'rank distribution' here is used to mean the number of times a candidate is ranked at each possible rank from 1 to the maximum ranking.

- **mean ranking:** the overall mean ranking of all candidates,
- **mean variance:** the overall mean variance of ranking for each candidate,
- **mean distribution difference:** the overall mean of the difference in distribution of rankings for each candidate and the average simulated distribution for that candidate, as measured by the sum of squares of difference,
- **fit of times ranked:** the mean *p* for the number of times each candidate is selected for ranking,
- **fit of mean rank:** the mean *p* for the mean ranking of each candidate, and
- **fit of rank distribution:** the mean *p* for the difference in distribution of rankings for each candidate and the average simulated distribution for that candidate.

In each case, the degree of fit was measured by counting the proportion of sets of simulated data which had a value as extreme as the value for the original data (i.e., a value equal to or greater than the original if the original was above the mean value for all the sets of simulated data, and equal to or less if the original value was below the mean). This value is referred to above as *p*. A perfect fit would be indicated by a value for *p* of at least 0.5. (Values of greater than 0.5 are possible in the case of discrete data because an appreciable number of the sets of simulated data will have a value equal to the original data.)

C. Results

Tables 1 and 2 show the results for fit with the data for each of the queries for the model which used the proportion of subjects who ranked a candidate as the basis for selecting candidates for ranking and a gamma distribution to determine where to place it in the ranking. As can be seen the degree of fit of the model with the data varied considerably from query to

query. Furthermore, despite the generally small differences between the mean rank and standard deviation in mean ranking for the actual data and the model, the degree of fit between the data and the model is generally poor (average 0.19 and 0.17). The same is true for the fit for the rank distributions (average just 0.05), but the average fit for the measures for individual candidates (Table 2, right three columns) was generally good (average 0.49, 0.34 and 0.30). This was typical for all of the different models, for which overall averages of fit are given in Table 3.

Table 1. Fit of model to data according to general ranking measures. Each row gives the figures for each of the eleven queries, and the bottom row the average for all eleven. 'Mean rank' is the mean rank for all candidates, and 'St. dev. rank' the standard deviation in mean ranking.

Mean rank			St. dev. rank		
data	model	fit	data	model	fit
18.7	18.8	0.40	8.91	8.21	0.04
21.5	21.1	0.10	10.67	10.08	0.07
20.1	21.3	0.01	9.32	9.53	0.33
25.2	27.2	0.01	10.76	10.13	0.19
13.7	15.3	0.01	6.90	6.92	0.48
19.7	19.9	0.41	7.27	7.71	0.29
19.3	19.4	0.39	10.40	9.94	0.11
22.0	22.2	0.34	11.71	10.89	0.06
14.1	15.0	0.06	6.63	6.21	0.20
24.5	24.0	0.16	11.30	10.23	0.02
18.7	19.1	0.26	9.73	9.07	0.11
19.8	**20.3**	**0.19**	**9.42**	**8.99**	**0.17**

Table 2. Fit of model to data according to other measures. Each row gives the figures for each of the eleven queries, and the bottom row the average for all eleven. 'Mean rank distribution fit' is the degree of fit for the mean sum of squares of difference between the distribution of ranks for each candidate and the average distribution for all sets of data. The right three columns give the average fit for each candidate (i.e., the proportion of simulated data which is as extreme as the original) for the number of times the candidate is selected for ranking, for the mean rank of that candidate, and for the rank distribution of that candidate.

Mean rank distribution fit	Average fit for each candidate		
	times ranked	mean rank	rank dist.
0.00	0.49	0.35	0.20
0.25	0.46	0.37	0.29
0.06	0.53	0.29	0.27
0.01	0.51	0.30	0.36
0.17	0.50	0.27	0.30
0.07	0.53	0.32	0.34
0.00	0.44	0.36	0.28
0.00	0.47	0.33	0.27
0.01	0.54	0.42	0.38
0.00	0.44	0.35	0.30
0.01	0.45	0.33	0.28
0.05	**0.49**	**0.34**	**0.30**

To fit with the general measures (the first three in the measures described above and the first three columns in Table 3) is quite a hard test of a model because these are average measures for all candidates. The degrees of fit with the per-candidate measures (the second three measures) is, by contrast, quite good. This difference indicates that there are small but systematic effects in the actual ranking of candidates which are not captured in the models.

Table 3. Degrees of fit for four different models.

Model	fit of mean rank	fit of std.dev. rank	fit of mean rank dist.	mean fit of times ranked	mean fit of mean rank	mean fit of rank dist.
proportion-distribution	0.19	0.20	0.02	0.49	0.35	0.32
proportion-gamma	0.19	0.17	0.05	0.49	0.34	0.30
gamma-distribution	0.17	0.13	0.08	0.40	0.34	0.29
gamma-gamma	0.17	0.11	0.09	0.40	0.33	0.28

The data in Table 3 show that the model which selects candidates on the basis of the number of subjects who selected the candidate for ranking, and then ranks the candidates on the basis of their actual ranking distributions is (not surprisingly) consistently the model which best fits the data (disregarding the very low values of fit for the mean rank distribution). However, the differences in fit between this model and the other three are not large. The second model, which similarly selects candidates on the basis of the number of subjects who selected the candidate for ranking, but then ranks the candidates on the basis of a gamma distribution, requires only three pieces of data for each candidate—the likelihood of its selection for ranking, the mean rank, and the standard deviation of rank—in contrast to up to 70 pieces of data per candidate for the first model. Ranking, therefore, can be modelled on the basis of two kinds of measure of similarity (since, as shown in Figure 2 the likelihood of a candidate being ranked is correlated with its mean rank), and a measure of uncertainty in ranking. Furthermore, the fact that the two models based on selection using a gamma distribution produce degrees of fit almost as good suggests that the decision to select a candidate to be ranked is made on a similar basis to the decision of where to place it in the ranking.

D. Between-candidate similarity

It is possible that in ranking candidates, subjects were influenced not simply by the perceived similarity between the query and each candidate, but also by perceived similarities and differences between candidates. The same Monte-Carlo approach was used to test for such an effect. The average difference in ranking between pairs of candidates was compared with the difference in their mean ranking. If subjects tended to rank candidates perceived to be similar close to each other, then the mean difference in rank would be smaller than the difference in mean rank. However, we cannot conclude that there is an effect simply from finding such differences in the actual data; some random differences are to be expected.

The approach taken here has been to measure this difference between mean rank differences and difference in mean rank for each pair of candidates in each set of simulated data, using the same models for simulation as described above. This makes it possible to estimate the degree to which the differences found in the actual data are a result of random effects. The models have no component which takes account of similarity between

candidates when ranking, and so can be regarded to generate data according to the null hypothesis in this respect.

The results showed that there was no significant difference in the average value for this measure between the actual data and the simulations. However, a different means of seeking significant difference did show some evidence of an effect. This was to count the number of pairs of candidates which had an extreme value for this difference. 'Extreme' meant pairs of candidates for which no more than 1% or 5% of the sets of simulated data had a value as large or as small. This count could then be tested for significance as before by determining the proportion of sets of simulated data which had a count greater than or equal to the count for the actual data. (This measure of significance should be interpreted in the same way as classic p-values.) The results, once again for the proportion-gamma model, are presented in Table 4.

Table 4. Number of candidate pairs with extreme difference between mean difference in rank and difference in mean rank.

Extreme diff. (1% level)			Extreme diff.(5% level)		
data	**model**	**sig.**	**data**	**model**	**sig.**
27	20.7	0.23	146	103.1	0.03
47	24.6	0.03	152	122.4	0.13
41	33.2	0.24	197	165.5	0.14
94	48.7	0.02	336	240.3	0.02
32	18.3	0.08	138	92.7	0.03
27	25.9	0.43	128	127.7	0.49
34	23.6	0.14	189	117.3	0.00
43	32.5	0.20	208	160.3	0.06
55	18.9	0.00	173	94.4	0.00
57	34.3	0.06	223	171.6	0.06
45	22.6	0.03	179	112.2	0.00
45.6	**27.6**	**0.13**	**188.1**	**137.0**	**0.09**

These results suggest that there might be an effect on ranking from the perceived similarity between candidates, but an examination of the actual difference values showed that this effect is largely due to *larger* mean differences in rank compared to differences in mean rank, suggesting that the effect is more one of *dissimilarity* than similarity.

E. Dependence in similarity

Another possible effect in the ranking of candidates is for a candidate's perceived similarity to the query to be affected by the presence of another candidate in the ranking. The hypothetical mechanism here is that having perceived a similarity between the query and candidate **a**, the subject's attention is drawn to particular characteristics of the query which might then cause another candidate **b** to be perceived as more or less similar to the query than it would otherwise have been.

Evidence for this effect was sought by comparing the mean rank for each candidate when each other candidate was included in the ranking and when that other candidate was not included in the ranking. (Obviously, this measure depended on a candidate being ranked sometimes with the other candidate and sometimes without, so it could not be taken for every pair of candidates.) Once again, the significance of any difference in rank when the second candidate was present or absent was tested by counting the number of pairs of candidates with 'extreme' differences, and testing the significance of this count as before.

Table 5. Number of candidate pairs with extreme difference between the mean rank when the second candidate is present or absent in the ranking.

Extreme diff. (1% level)			Extreme diff.(5% level)		
data	**model**	**sig.**	**data**	**model**	**sig.**
59	35.6	0.17	181	153.0	0.28
54	38.8	0.28	180	163.0	0.37
195	58.8	0.00	509	261.5	0.00
134	74.5	0.13	420	356.6	0.26
117	35.3	0.01	251	161.0	0.05
140	40.9	0.01	245	183.3	0.18
96	38.6	0.03	243	167.9	0.10
156	50.9	0.02	365	217.2	0.04
164	44.0	0.00	373	186.9	0.00
122	57.5	0.06	303	252.2	0.25
43	34.2	0.35	139	148.4	0.47
116.4	**46.3**	**0.10**	**291.7**	**204.6**	**0.18**

The results are shown in Table 5. While some queries show evidence of a strong effect, the average significance is not great. We cannot therefore safely conclude that there is an effect of one candidate influencing the perceived similarity to the query of another.

IV. CONCLUSIONS

Overall, the results of this analysis are rather equivocal. Analysis of the distribution of rankings showed that judgements of similarity between melodies were extremely variable for all but the most similar melodies. On the one hand, modelling of the data using two measures of similarity and a measure of uncertainty for each candidate demonstrated a moderately good fit at the level of each candidate, though not at the global level. Evidence was found for an effect of similarity between candidates influencing ranking, rather than ranking being dependent simply between the query and candidates, but it is not strong. Similarly, while evidence was found for the influence of the presence of a third melody on the judgement of similarity between two melodies, this too is not consistently strongly present.

One definite conclusion is that the ranking paradigm used by Typke and colleagues has proven to be a rich source of data, and further research along similar lines, with queries and candidates specially selected to probe the issues in melodic similarity not clear from this study, would be likely to provide further insight.

ACKNOWLEDGMENT

I am very grateful to Rainer Typke for making available the original data from the experiment which has made this paper possible.

REFERENCES

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.

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

Marsden, A. (forthcoming). Interrogating Melodic Similarity, *Journal of New Music Research*, *41*(4).

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.

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.

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

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.