Chapter 16

Examining Fixed and Relative Similarity Metrics through Jazz Melodies

David J. Baker and Daniel Shanahan

Louisiana State University,
Baton Rouge, LA 70803, USA

16.1. Introduction

Similarity, defined here as the degree to which multiple items might share overlapping properties, inherently requires a comparison to be made between the two objects [5]. How the comparison of two or more musical ideas can be carried out is, at times, a perplexing question. A musical idea might be considered similar by specific parameters such as its rhythmic, melodic, harmonic features, or broader, more difficult to define aspects, such as contour and overall shape. Because of this, we might argue that the question of musical similarity occupies the space where music theory and music psychology intersect. Loosely defined, musical similarity is the extent to which two musical objects are thought to share similar characteristics. In this definition there are two parameters to consider. Firstly, which two musical objects are to be compared to one another? Secondly, which criteria should be used in order to make that comparison?

A special issue of *Musicae Scientiae* published in 2007 focused specifically on questions of musical similarity, and provides a many resources on the topic. For example, Typke, Weiring, and Veltcamp suggested that using the earth mover's distance can be used as a way to classify music that is in agreement with expert human judgments [22], Müllensiefen and Frieler modeled expert notions of melodic similarity and concluded that experts are, in fact, able to consistently make similarity judgments across different musical stimuli. They additionally matched algorithms that mirrored the expert judgments and argued that these methods could be used for real world problems such as folk song categorization and musical analysis [15]. Additional work has also shown that listeners are unconsciously sensitive to various statistical properties in melodies such as pitch distribution and the number of notes, and that these can be used to predict perceived similarity, although listeners may be unconsciously aware of it [9].

16.1.1. *Fixed and relative similarity*

At this point, it might be worth distinguishing between two modes of similarity measurements, one we might call a fixed, and a second we could refer to as relative. The first is an fixed, stable, and somewhat objective view of similarity, in which a mathematical measurement provides some theoretical understanding of a relationship between X and Y, and is simply asking "how similar are these two objects given a fixed parameter?" A second conceptualization of similarity, which we will refer to as "relative similarity", asks "when accounting for a humans response to this questions what external information helps guide a choice of similarity between objects X, Y, and maybe even Z, in a given context?" Although objective (fixed) measurements have been used a great deal for mathematical and computational expediency, psychologists argue that perceived similarity is, by its nature, perceived in relation to a network of the objects that inform each contextual decision. For example, Cambouropoulos [4, p. 10] writes that "[s]imilarity is by definition relational and, therefore, cannot be objective".

There are, however, benefits to both approaches. On the one hand, two objects are able to be seen as similar regardless of whether or not a third object is there. For example, if one is drinking out of the same type of paper coffee cup as a colleague, they are able to understand that those paper coffee cups are quite similar, regardless of whether a third friend is present with an ornate ceramic mug. This is because an individual brings their own memories and beliefs to any situation where they are made to make a choice. On the other hand, it could be argued that the only reason the two paper cups are seen as similar is due to the internalized understanding of what dissimilar coffee cups would look like. These ideas are very much at the core of prototype theory, which claims judgments of similarity are informed by the distance one member of a category is to its central, defining member [8, 19].

It would follow that depending on which set of musical objects one chooses, as well as which characteristics one chooses to focus on in a relative context, the answer to the question of "how similar is X to Y?" is subject to change. Contrastingly, if modeling this question mathematically as an objective measurement of similarity, the answer will be the same each time the computation is run. While tackling this question from a strictly mathematical vantage point has its advantages (such as classifying folksongs, or searching for repeated structures in a large corpus of music [12] or building computational models to formalize arguments), questions of psychological similarity will involve some sort of human judgment, variance, and error, which are missing in this sort of computation.

16.1.2. *Considering human judgment*

A large amount of studies exploring musical similarity have employed similarity measurements in their research that juxtapose two melodies and ask participants to rate the degree to which they are similar (for example, see [9, 18]). While this approach has its advantages, one drawback is that these measures can only capture the sets of perceptual qualities of the two items in question. One point of interest that has been raised is that the context of a similarity judgment can influence ratings of similarity even by the same participant [21]. For example, if given the context two separate natural harmonic minor melodies and a major key melody, an individual might select the two minor melodies as most similar in one context. Though if those two same minor melodies are present in a subsequent trial with the addition of a different, yet novel harmonic minor melody, the two harmonic minor melodies might be then rated as more similar — perhaps even to a larger extent. We chose to investigate how the context of melodies might in fact impact judgments of similarity in both humans and computers in two ways:

(1) An experiment asks participants to gauge how similar two objects are when given *three* choices of melody.
(2) A computational model is given the same three choices and by using an objective measurement of similarity, also makes decisions about which two objects are the most similar.

We then compare results between the listeners behavior and the computational model, examining the degree to which a distance metric can be used to predict human choices when given a paradigm based on *relative* similarity. The process of modeling with computational tools how humans perform when carrying out behavioral tasks is increasing, but there is still a great deal of word to be done. Müllensiefen and Pendzich [17] examined the accuracy of similarity metrics in predicting the jury responses in musical plagiarism cases and found that while certain measures such as edit distance performed somewhat well, they were surpassed by Tversky's [21] model of similarity which takes into account the number of shared categorical features between two sets to be compared. Given the success of edit distance in this and other contexts, we choose to use this as a stepping off point for investigating the effects of context on musical similarity.

16.2. Distance as Transformation and Similarity

Looking at musical distances can be operationalized a handful of ways depending on how the question being asked. In this chapter we have chosen to view musical distance as a transformation, conceiving the distance between two objects as a result of the number of operations that are needed to move from one object to another. This method has been recently successful in modeling musical similarity [18] and here we take a related approach.

One of the most commonly-used transformation metrics is the Damerau–Levenshtein distance measure (also commonly referred to as edit distance). In some ways, the Damerau–Levenshtein distance could be thought of as a proxy for measuring human behavior similarity judgments in that it makes a calculation of the number changes needed to transform one object into another. Although the Damerau–Levenshtein distance measure is a rather simple transformative model, it has been applied successfully in music information retrieval tasks [5, 13] and has predicted melodic similarity well in previous research by Müllensiefen *et al.* who suggested that the distance that best matches human judgments of melodic similarity can be modeled with combination of the edit distance and an n-gram distance. We therefore chose to implement the edit distance as an entry point to implement and reasoned that once a paradigm was designed that could map human judgments on to those of the computer, it would then be possible in future research to explore other options of comparison.

Unlike prior research that often relies on participants making continuous ratings between two separate musical stimuli against one another [18], we opted for using a forced choice task that would provide an insight into how humans behave when confronted with having to pick between stimuli. Using the three way alternative forced choice (3AFC) paradigm listed above not only provided the opportunity to do analyses like those found in the aforementioned papers, but also created data that counts the number of times a participant selected a certain stimuli when presented alternative options. This not only allowed for our analyses to look at trial by trial inter-rater agreement, but also provided a paradigm where an algorithm could hypothetically partake in the experiment as if it were a participant. While a computation almost by definition will not display the same sort of variability as a human responses, this paradigm affords analyses that are easily interpretable when comparing human judgments of similarity to those of the computer.

16.3. Experiment

We designed an experiment with the goal of exploring how different computational and algorithmic measures can be used to mirror human behavior using a contextual paradigm. This goal led to employing a design that was not dependent on any sort of explicit understanding of Western musical notation. While these aspects of the music were necessary in order to carry out some of the above studies, it could be argued that some of the results depend heavily on an individual's musical background as well as their training with the context of Western music notation which could cause the listeners to consciously or unconsciously focus on aspects of the music more linked to the notation than the actual aural events.

Work in the field of music psychology has suggested that many factors involved in music perception do not even require any sort of formalized training [3] and that factors outside of musical training can be used to model performance on perceptual tasks [2]. In light of these considerations, we set out to investigate if listeners, regardless of any sort of domain specific expertise, were able to perform consistent and reliable similarity judgments in a context that is not dependent on anything but auditory information.

In order to investigate these aspects of musical similarity, we employed a paradigm that that did not rely on any sort of musical formal training expertise [1]. This approach has been used successfully with naive listeners to investigate questions of genre perception where the authors concluded that based on similarity judgments alone, listeners were able to pick out the hypothesized salient features [10].

16.3.1. *Methodology*

Participants ($N = 44$, 13 women, $\bar{X} = 201.16$, SD = 1.45, R = 18–24) for this experiment were selected from the subject pool at Louisiana State University's School of Music. All participants were given partial course credit for taking part in the experiment. Stimuli from this experiment were randomly selected from a subset of the database of encoded bebop transcriptions. It was decided *a priori* that seven phrases, or licks, of 24 notes would be chosen as this number of notes would create three bar phrases. Each excerpt was converted into eighth notes, transposed to the key of C major, set to quarter note = 120, and presented as MIDI recordings with a uniform timbre in order to control for any musical features outside of melody. Stimuli from the experiment are shown in Figs. 16.1–16.7. Software

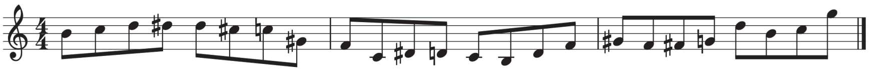

Fig. 16.1. From Charlie Parker's "Warming Up on a Riff".

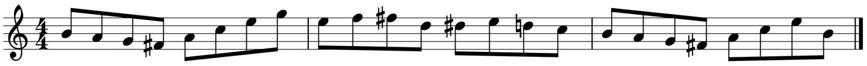

Fig. 16.2. From Charlie Parker's "Passport".

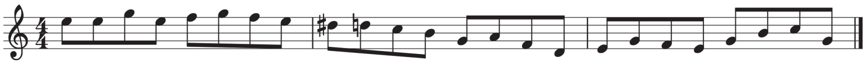

Fig. 16.3. From Charlie Parker's "Chasing the Bird".

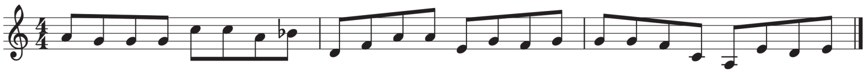

Fig. 16.4. From Charlie Parker's "Cardboard".

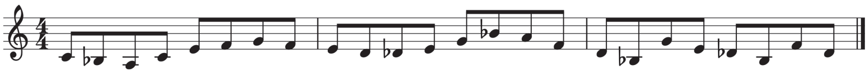

Fig. 16.5. From Charlie Parker's "Bird Gets the Worm".

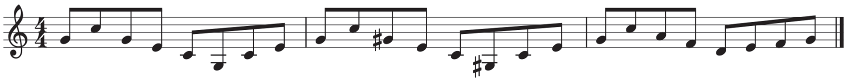

Fig. 16.6. From Charlie Parker's "Anthropology".

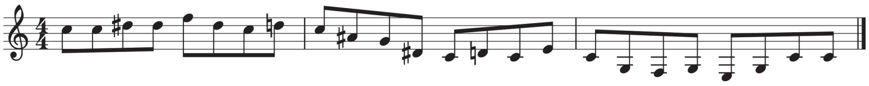

Fig. 16.7. From Charlie Parker's "Another Hairdo".

for the experiment was provided by the authors of Farrugia *et al.* [10].

16.3.1.1. *Procedure*

Participants first completed forms relating to their general demographic information, as well as their musical background as measured by the Goldsmiths Musical Sophistication Index [16]. After completing these forms, the experimenter directed their attention to an interface in Max/MSP where participants were told they needed to make a subjective choice about which two musical excerpts they found to be most similar. Participants were not prompted to focus on any sort of musical parameters and if they inquired about what to pay attention to, the experimenter would tell to focus on whatever aspects they found to be most salient. Participants were allowed to listen to any stimuli in any order, as many times as they would like to. The amount of times a participant listened to a track was recorded, unknown to them.

Additionally before running the experiment, we ran an *a priori* effect size calculation using the *pwr* package in [6] to identify approximately how many subjects we needed in order to generate a large effect size (0.5) [7]. Calculations indicated that with $df = 2$, $\alpha = 0.05$, $\beta = 0.8$ we would need approximately 39 participants.

16.3.2. *Algorithm implementation*

Mirroring the experimental design, we used the Damerau–Levenshtein distance measure from the Humdrum Toolkit and the Humdrum Extras package [11, 20] to simulate what the Damerau–Levenshtein distance would "select" in each permutation of the stimuli in our experiment. To do this, we computed Damerau–Levenshtein distance measures between each three stimuli for every unique trial and then used the smallest distance between the three melodies to select the two most similar stimuli. Computing a three way distance measure between each triad of stimuli and selecting the two which were most similar enabled use to somewhat anthropomorphize the algorithm. By doing this, we were able essentially have an algorithmic participant that would be able to always select the same melody in each context.

16.3.3. *Analysis*

Our initial analyses sought out to investigate two hypotheses:

(1) H_1: Participants will identify at each three alternative forced choice (3AFC) in accordance with edit distance.
(2) H_2: Participant choices will most often match an algorithm when given identical choices, based on relative similarity.

Before comparing the human responses to any sort of algorithm, data from participants was examined separately. Since participants not only rated each permutation of seven different stimuli, but did so three times, first data was checked for reliability in order to investigate the aforementioned question about if non-expert listeners are able to provide an acceptable degree of reliability between the rating blocks. Reliability between blocks was checked at the block level by running a Spearman rank order correlation between the trials selected most often between participants. A high degree of correlation was found between blocks 1 and 2 ($r = 0.88$), 2 and 3($r = 0.89$), and 1 and 3($r = 0.90$).

Because of the high degree of consistency between participants, deemed it was appropriate to pool participant responses and examine the degree to which human responses were in agreement with edit distance.

After computing all edit distance measures, we then tallied the amount of times that each participant was either in agreement or disagreement with what Damerau–Levenshtein distance selected and performed a chi-square test on each of the unique trials to determine if humans agreed with the Damerau–Levenshtein distance measures a significant amount of times to investigate H_1. Type I error inflation rates were controlled by using a Bonferroni correction. Results of H_1 suggest that, within a set of trials, there was a fair degree of agreement within some of them, but after correcting for multiple tests any sort significant effects disappeared due to the extremely conservative alpha ($p < 0.001$).

Table 16.1. Block comparison.

H_2 analysis
Block 1: $\chi^2(2,35) = 0.25, p > 0.05$
Block 2: $\chi^2(2,35) = 0.03, p > 0.05$
Block 3: $\chi^2(2,35) = 0.71, p > 0.05$

After this step, we then looked at unique trial sets on the whole and computed a second, experiment level chi-square test on unique sets of trials to determine if Damerau–Levenshtein distance on the whole aligned with human choices a significant amount of times. While Damerau–Levenshtein distance did correctly predict human choices in some sets of trials, they did not predict human decisions on a global level as demonstrated in Table 16.1.

Although edit distance was not able to mimic human behavior in our experimental design, we used this as a stepping off point to inspect melody sets where humans did and did not agree with edit distance. The truncated Table 16.1 shows a listing of the highest and lowest agreements between our participant data and that of the edit distance.[a]

Upon inspecting the data, it appeared that similarities in contour variation (see stimuli 3 and 7) as well as the outlier status of stimuli 4 lead to a fair degree of consistency in ratings between edit distance and the participant judgments of similarity. This outlier status was also apparent when looking at the amount of times that a stimuli at the global level of the experiment was chosen as the oddball.

16.3.4. *Discussion*

In this study we sought out to investigate the extent to which Damerau–Levenshtein is able to mirror how humans would behave when presented with forced choice tasks. Before evaluating how well both algorithms and humans faired when compared to one another, it is first worth noting that within this dataset most

[a] The term "oddball" refers to the stimuli not selected in the group of three, thus selecting the other two as most similar.

participants exhibited a high degree of reliability in their ratings. This is in contrast to earlier research that screened participants for the reliability [14], but our findings may be due to the fact that our participant pool consisted of trained musicians. Alone this finding is important in that it suggests that, when prompted, participants are depending on a consistent and reliable process in their decision making. It is additionally worth noting that this degree of consistency was achieved via listening, and was not dependent on any sort of symbolic notation which carries many preconceived notions of Western music education.

Given this degree of consistency, it follows that it should be possible to come up with some sort of mathematical or statistical model that could accurately mirror this behavioral process. We began to explore how measures of edit distance might compare to how human selections by looking for agreements between edit distance and our rating data. While the results from the H_2 analysis table did not yield any significant results, it may have been that the way in which we decided to compare edit distance to our human measures was too stringent of a test for edit distance to accomplish given this context. A significant p-value for this test, if found, would have suggested that edit distance follows our participant data to an extreme degree. Secondly, the statistics reported in the H_2 analysis table are based on a second-order level of statistical tests meaning that the chi square value reported here was derived from a family of other statistical tests in which the critical values at the participant level were transferred one higher in order to generalize to the entire experiment. This type of analysis is extremely conservative and protects for type I errors almost to a limiting degree. Upon examining the data *post-hoc*, the amount of individual trials where participants did agree with what edit distance would have chosen is impressive and may indicate that our initial parameters that we chose were too conservative and would not again be adopted if this type of study was run again (Table 16.2).

Inspecting the distribution of trials where a high degree of agreement was achieved, we additionally suggest that there may be

Table 16.2. Subject agreement.

	Set	Oddball stimuli	Trials agreed
1	437	4	96
2	714	4	86
3	675	5	85
4	134	4	77
5	674	4	76
6	367	6	75
7	136	6	73
.	.	.	.
29	642	2	24
30	265	2	24
31	573	7	20
32	352	2	19
33	563	6	18
34	531	1	14
35	245	5	11

different types of appraisals happening for different melody contexts. Given the non-uniform distribution of what sets of trials were agreed upon as displayed in Fig. 16.8, some salient features may have been used in order to make the each judgments. Of particular interest in this setting would be repeated opening notes from Melody 4, which may have caused significant attention from the participants as well as the global contours of pairs of melodies that were paired together frequently (Fig. 16.9).

The question of what salient features an individual attends to also brings up the question of if in the 3AFC task that we used, were participants making judgments of similarity, as they were prompted? Or were participants, much like the way that we describe our analysis, treating this as some sort of oddball task. Though the experimental instructions explicitly asked for participants to make a rating of similarity, it might have been that they were treating the task as picking out the melody that is not like the other. Returning to our previous example of the one ceramic and two paper coffee cups, the issue becomes more apparent. Is it that the two paper cups are seen as separate from the ceramic and share more features in common, or

AQ: As per the agreement there are no color pages. Please confirm whether we can print this figure in color.

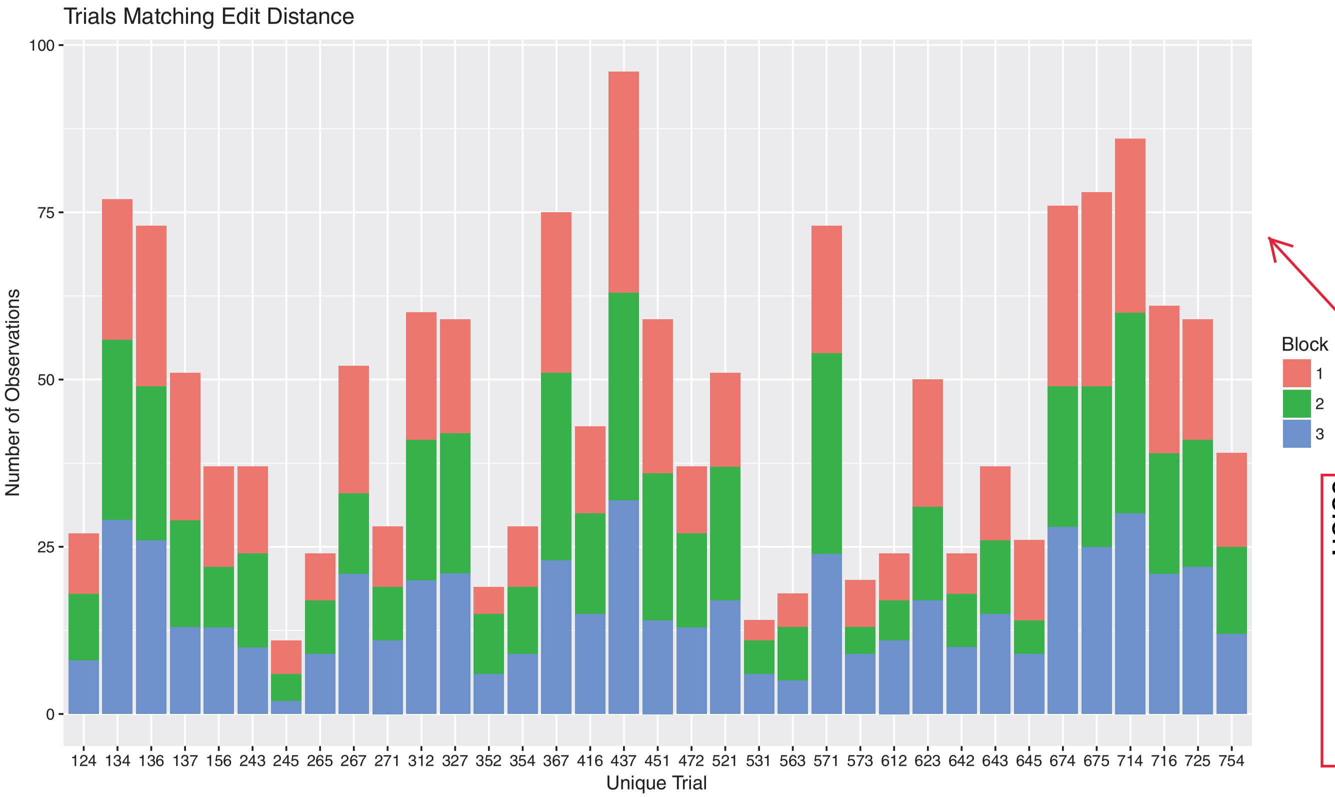

Fig. 16.8. The number of observations each participant was in alignment with the choices made using edit distance.

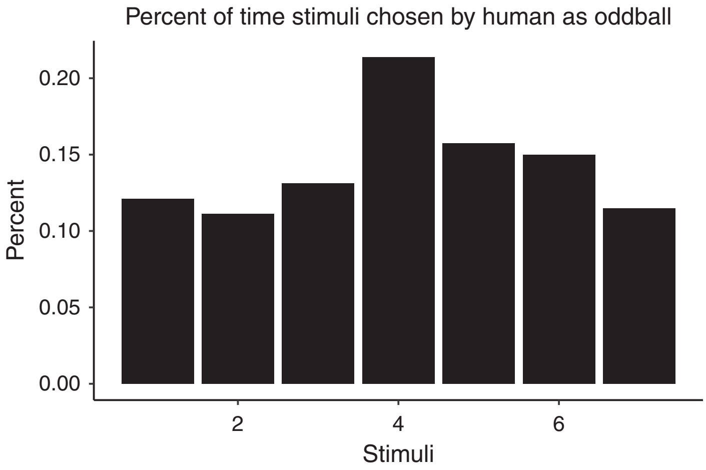

Fig. 16.9. The percentage of time each stimuli was chosen as an oddball.

is it that the ceramic cup shares less qualities with each of the two paper cups? While this pattern of results may appear to be the same end result, the means that an individual uses to reach this conclusion may be of interest. Future work might consider the directionality of paradigms like this. This type of discussion only bolsters claims made by Cambouropoulos *et al.* [4] mentioned earlier, who asserted that these questions are by nature relative, and in this case, directional.

If one were to accept our the logic of this experimental design, we could additionally argue that the reason that edit distance did not fair as well is because we are not mirroring any sort of cognitive decision process, only using a computational measure as a proxy for how someone might think.

Further, exploring other avenues of looking how computational measures might better reflect cognitive processes we can then turn to our *post-hoc* investigation on contour processing. While the idea was initially promising, the results of the analysis did not bring forth any sort of meaningful conclusions.

16.4. Conclusions

Overall the results of this study conclude that participants in a 3AFC reliably perform ratings of similarity across various conditions.

While this is not a new finding, it supports the idea that individuals use a consistent mental process when making similarity decisions about short melodies. In terms of modeling the degree to which humans make choices in accordance with the Damerau–Levenshtein edit distance, our data analysis did not yield any significant results. This could be in large due to the conservative preset alpha level decided *a priori*. Future work exploring this might consider how other algorithms might fare in this task as well as alternative ways of analyzing the dataset. The dataset used is publicly available at request of the authors.

Bibliography

[1] H. Allan, D. Müllensiefen and G. A. Wiggins, Methodological considerations in studies of musical similarity, in *ISMIR* (2007), pp. 473–478.

[2] D. J. Baker and D. Müllensiefen, Perception of leitmotives in richard wagner's der ring des nibelungen, *Front. Psychol.* **8** (2017).

[3] E. Bigand and B. Poulin-Charronnat, Are we experienced listeners? a review of the musical capacities that do not depend on formal musical training, *Cognition* **100**(1) (2006) 100–130.

[4] E. Cambouropoulos, How similar is similar? *Musicae Scientiae* **13**(1_suppl) (2009) 7–24.

[5] E. Cambouropoulos, T. Crawford and C. S. Iliopoulos, Pattern processing in melodic sequences: Challenges, caveats and prospects, *Comput. Humanities* **35**(1) (2001) 9–21.

[6] S. Champely, *Pwr: Basic Functions for Power Analysis. R Package Version 1.1. 1* (The R Foundation, Vienna, 2009).

[7] J. Cohen, Statistical power analysis, *Current Directions Psychol. Sci.* **1**(3) (1992) 98–101.

[8] I. Deliege, Grouping conditions in listening to music: An approach to lerdahl & jackendoff9s grouping preference rules, *Music Percept.* **4**(4) (1987) 325–359.

[9] T. Eerola, T. Jäärvinen, J. Louhivuori and P. Toiviainen, Statistical features and perceived similarity of folk melodies, *Music Percept.* **18**(3) (2001) 275–296.

[10] N. Farrugia, H. Allan, D. Müllensiefen and A. Avron, Does it sound like progressive rock? a perceptual approach to a complex genre, (2016), pp. 197–212.

[11] D. B. Huron, *The Humdrum Toolkit: Reference Manual* (Center for Computer Assisted Research in the Humanities, 1994).

[12] D. Meredith, Analysing music with point-set compression algorithms, in *Computational Music Analysis* (Springer, 2016), pp. 335–366.

[13] M. Mongeau and D. Sankoff, Comparison of musical sequences, *Comput. Humanities* **24**(3) (1990) 161–175.

[14] D. Müllensiefen and K. Frieler, Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments, *Comput. Musicology* **13** (2003) 147–176.

[15] D. Müllensiefen and K. Frieler, Modelling experts notions of melodic similarity, *Musicae Scientiae* **11**(1_suppl) (2007) 183–210.

[16] D. Müllensiefen, B. Gingras, J. Musil and L. Stewart, The musicality of non-musicians: An index for assessing musical sophistication in the general population, *PLoS One* **9**(2) (2014) e89642.

[17] D. Müllensiefen and M. Pendzich, Court decisions on music plagiarism and the predictive value of similarity algorithms, *Musicae Scientiae* **13**(1_suppl) (2009) 257–295.

[18] M. Pearce and D. Müllensiefen, Compression-based modelling of musical similarity perception, *J. New Music Res.* **46**(2) (2017) 135–155.

[19] Rosch, E. Principles of categorization, *Concepts: Core Readings* **189** (1999).

[20] C. Sapp, Humdrum extras, (2008), http://extras.humdrum.org/, http://extras.humdrum.org/ (accessed: 2017-11-1).

[21] A. Tversky, Features of similarity, *Psychol. Rev.* **84**(4) (1977) 327.

[22] R. Typke, F. Wiering and R. C. Veltkamp, Transportation distances and human perception of melodic similarity, *Musicae Scientiae* **11** (2007) 153–181.