The perception of similarity in court cases of melodic plagiarism and a review of measures of melodic similarity

Anna Wolf
HMTM Hannover
Emmichplatz 1
30175 Hannover
+49 (0)176 631 45 309
mail@annamwolf.de

Daniel Müllensiefen
Department of Psychology
Goldsmiths, University of London
New Cross Road, New Cross,
SE14 6NW, London
+44 (0)20 7919 7895
d.mullensiefen@gold.ac.uk

ABSTRACT

This study examines the perception of melodic similarity applied to cases of melodic plagiarism under a review of similarity measures. An implicit memory task was designed to test the extent of the participants' confusability of two similar melodies. The participants were able to distinguish between such melodies involved in cases with and without actual copyright infringement. Many of the applied measures of similarity relate very well to the results of the implicit memory task and the court case decisions, such as a Tversky feature-based measure ($r = 0.514$) and a weighted Edit Distance ($r = 0.515$) for the psychological data and certain Earth mover's distance measures for the court decisions (AUC of .84).

Keywords

Melodic similarity, melodic plagiarism, Tversky feature-based similarity

1. INTRODUCTION

Finding an empirical approach towards melodic similarity is of great importance: Search engines focus on the retrieval of melodies based on a hummed melody, a tapped rhythm or a basic contour representation. To successfully find the queried melody they need to understand the human perception of melody and how people group, reproduce and confuse similar musical information (i. e. Musipedia, [1]).

Improving the measurement of melodic similarity, which will be of use for such databases, will above all also progress our understanding of the perception of similarity in melodies. We already know about some factors that give rise to a perception of similarity in melodies, such as a similar melodic contour [2] or a similar rhythm or harmonic content [3]. Hence the construct of a contour or a harmonic content are both simplifications of the actual stimulus and indicate that our perception of similarity relies not only the stimulus itself, but on our percept of it.

There has been an application for melodic similarity for many decades: In legal plagiarism cases a judge has to yield a decision whether two melodies are similar enough to convict a musician of having plagiarised a piece of music. These decisions are reached by including assessments of musical experts in this musical genre, and for the future it conceivable to employ computational measures of similarity to enclose a more impartial and all-knowing assessment.

In many cases, the legal decisions about melodic plagiarism can be simplified to two categories of outcome: The decision can either be pro plaintiff, therefore the defendant's melody is a theft of intellectual property, or it is contra plaintiff, and the defendant has not infringed the law in the process of writing the respective melody. The law for the copyright of music and musical plagiarism varies in different countries to some extent; the copyright law of the United Kingdom is still based on the principle of "sweat of the brow" which does not require originality in the work to be protected by the law, but only hard work and adequate skill. A similar law has been in place in the United States until 1991, but has been dropped at that time [4]. This study will include melodies from court cases from the U.S. and the U.K. as well as one case from both Canada and Australia, whose origins in the copyright law and whose changes over time are shared within the Commonwealth and with the U.S.

The decision in a case of melodic plagiarism does not only rely on the similarity of the melody, the judge will also include a possibly similar or same title of both songs, similar lyrics and the likeliness of the defendant's knowledge about the plaintiff's song. These and other factors that influence the final ruling of a court case are not included in this study which focuses on the perception of melodic similarity.

To substantiate the approach towards the psychological aspect of similarity, Gärdenfors [5] suggests three positions. First he argues that similarity "is something that exists in the world, independent of any perceptual or other cognitive processes". Second, similarity is an empirically examinable entity that can be approached by investigating the perception of similarity. Third, its cognitive domain can be used to shape models of i. e. categorisation. The first position will not be approached in this thesis as it is of a more philosophical nature. The third position has already been laid out by presenting several algorithmic models. The second position, addressing the perception of similarity and different psychological methodologies to measure it, will be laid out in the following and lead to the research question for this study.

To define a similarity measure it is necessary to first define a melody space M, which is a subset of the Cartesian product of a time coordinate representing the notes' onsets and a real-valued coordinate representing pitch.

Early assumptions of the concept of similarity perception focused on the properties extracted in the process of perception and the distance within. These distances satisfy the definitions of a metric [5].

Definition 1. A distance is defined as a map d : M × M → R such that:

1. Identity: d(a, a) = 0 and non-negativity: d(a, b) ≥ 0

2. Symmetry: d(a, b) = d(b, a)

3. Triangle inequality: d(a, b) + d(b, c) ≥ d(a, c)

The third item is substituted for some similarity measures by the assumption of Transposition-, Time translation- and time dilation invariance. This means invariance concerning changes of the melody in pitch and in time, namely both in time shift and time stretch so that tempo changes and changes of the point in time when the melody begins will not change the result of the measure [6]. Most of the considered algorithms will fulfil these properties, however, a few exceptions will for instance be the Earth Mover's distance which does include tempo information (time dilation). Another group of measures based on Tversky's [7] concept of similarity perceptions explicitly argues that the perception of similarity is not symmetrical as many of our statements about resembling structures have a direction which neglects the property of symmetry as well.

Such properties collected from different objects are compared and lead to a perceived distance which is the determining factor in similarity: Similarity is here considered to follow an exponentially decaying law, which means a linear increase of distance leads to an exponential decrease in similarity [13, 14].

The measurement of similarity compliant to this theory is approached by similarity rating of pairs of melodies, multi-dimensional scaling for stimuli with more dimensions [15] and ranking of comparison stimuli depending on their similarity with a target stimulus.

The similarity measurements used in this study are from diverse areas of computational similarity research. Either using a MIDI or a textual music representation (CSV file) which contains a transcription of the respective melody. All of the following algorithms use monophonic melodies given the information of pitch, onset and duration.

The simplest algorithm used is the Levenshtein distance, which is the standard **Edit distance** measure. It compares two strings and computes the minimally required number of changes necessary to convert either string into the other one. For each insertion, deletion or substitution of an element a cost of one is added to the final result; this will then be divided by the length of the longer string to achieve a normalised measure with a range of [0,1] [3].

Based on the raw Edit distance there is a slightly more sophisticated measure, the **weighted Edit distance**, which is weighted by the duration of the notes. All notes are split into as many notes of the smallest rhythmical unit (the tatum) present in that melody, which are necessary to still represent their original duration. This will lead to a melody representation consisting of usually only quavers or semiquavers with the effect that longer notes will be "more expensive" to be edited as they are indicated by a high number of tatums which have to be deleted, inserted or replaced. On the other hand, originally shorter notes are "less expensive" as they are represented by a smaller number of tatums. Taking account of the note's rhythm, the duration and therefore the note's importance on a very basic level is expressed in this similarity measure. The actual Edit distance is applied to this string and is divided by the count of the notes, which now is the number of the existing tatums [3].

The **opti3** measure has been developed by Müllensiefen and Frieler [3] as a hybrid measure including the rhythfuzz measure described previously, as well as a measure of harmonic content and an Ukkonen n-gram pitch measure.

Algorithms such as the **Earth Mover's Distance** belong to a rather new approach developed to grasp similarity. The representational distortion theory takes the relation between the items of a stimulus into account so that similarity "is a function of the 'complexity' required to 'distort' or 'transform' the representation of one into the representation of the other" [11]. This regards similarity perception as the endeavour made by the human mind to transform one percept into another one.

The application of the Earth mover's distance to music is accomplished by a mapping of (usually) onset time and pitch onto a subset of a two-dimensional space, the note's duration is visualised by the weight of each point. The similarity of two melodies is calculated by the amount of work necessary to change the location and weight of the point set representing one melody in order to convert it into the other melody [8].

To correct for the key all melodies are mapped by either centering the notes around middle C or by using intervals instead of the particular pitches. Additionally, the distance of the transformation of one melody mapping into the other can be either calculated as a Euclidean distance (the length of a straight line between two points, intuitive measurement) or city block distance, which is conceivable by imagining the route through a city with a regular grid layout.

Tversky's theory about similarity is based on the perception of features that two objects share and the salience of these features. People tend to compare the less salient stimulus with a more salient one ("an ellipse is like a circle" and not "a circle is like an ellipse") and by establishing this order our judgments of similarity are shaped alike [7].

Correspondingly, the initial rationale of similarity as a function of perceived distance has been disputed, especially the assumption of symmetry (see also [15]). But also the common assumption of triangle inequality (d(a, b) + d(b, c) ≥ d(a, c)) has been confuted as they have shown in an experiment that participants systematically violate this property [16]. Tversky claimed that human perception was less influenced by continuous characteristics but rather by well-defined features and also the salience of these features. The salience of a stimulus is influenced by intensity, frequency, familiarity, good form and informational content [7] and therefore salience is a result of the basic principles of perception which everybody shares and also of habituation (familiarity), knowledge and interest in that domain (informational content).

The measures based on this concept of similarity also include a weighting scheme developed from a large corpus of pop melodies [9]. Interval n-grams, subsequences of n items from a given object, with of a length of n=3 are quantified in the corpus. The rationale of this step is that two melodies sharing more frequent features are less similar than two melodies sharing rather infrequent features. This assumption is based on a perception of similarity, which only grounds on the number of shared features and not their order in the piece. This weighting scheme is used as the function to determine the salience of a structure so that not only the features' salience in both melodies is used, but actually the frequency of features for a representative corpus of melodies; this generalisation leads to a broadened evaluation of frequency weighting. Four variations of weighing have been conceptualised. The Tversky.equal measure is the only symmetrical variant evaluating the salience of the shared features with respect to the salience of all features. Tversky.plaintiff.only and Tversky.defendant.only evaluate the

non-mutual n-grams' originality for either the defendant's or the plaintiff's melody. Tversky.plaintiff.only could therefore lead to a wording such as "the defendant's melody resembles the plaintiff's melody"; with the consequence that the melodic features presented by the plaintiff mainly define the defendant's melody. Tversky.weighted defines the amount of weighting by the extent to which the mutual n-grams span either melody.

These algorithms use various approaches towards the perception of similarity, but they can be generalised to either a spatial account [10] or a feature-based account [7]. All the Edit distances and the Earth mover's distance belong to the spatial representations, the Tversky-based algorithms are obviously part of the feature-based representations. Both simplifications are very elementary in their representation of similarity, they either refer to points in space and measure distances to compare two objects, or they count feature sets [11].

The psychological experiment is a follow-up study to Müllensiefen and Pendzich [9]. They have mainly investigated the legal background and applied similarity measures on melodies pairs as they are dealt with in plagiarism cases. The implementation of feature-based algorithms was a major objective of which Tversky.plaintiff.only achieved the highest correct classification of the American court cases. The present study includes court cases from English and other courts as suggested to extend the database, and the psychological data will presumably be consistent with both the court rulings and calculated classification accuracy of the similarity measures.

The first test examines the confusability of two similar stimuli by priming the participant with one melody and later presenting a similar melody. Then the participant is asked about the level of certainty of "remembering" this melody. Applying this test to investigate melodic similarity we are hypothesising that participants will be able to distinguish between those paired melodies in between which there is a case of plagiarism and therefore an assumed higher level of similarity, and those melodies for which a court has decided there was no copyright infringement and therefore a lower level of similarity. However, such a confusability test has not been conducted with melody pairs of different grades of similarity yet so this data will have to be absorbed carefully. This thesis will also compare and contrast similarity measures associated to the different theories of melodic similarity regarding both the psychological data and the court decisions as ground truth.

2. METHOD

2.1 Design

The experiment examines the perception of melodic similarity implicitly by presenting melodies several times during the study phase. In the testing phase participants were played melodies that were categorised as "same", "similar" and "neutral" and they were asked how sure they were about having heard this melody during the study phase. The "similar" melodies were compiled from court cases of melodic plagiarism with a dichotomous outcome (either an existing case of plagiarism or not). The "same" melodies were played in both the study and testing phase and they were also chosen from court cases. The "neutral" melodies were only played in the testing phase. Assuming that similar melodies sometimes tend to be confused with their counterpart from the court case it is expected that the listener will state that he has heard this melody although he has only heard a similar melody. The aim is to find the level of confusability for these similar songs and if this number is correlated to the court decisions in the plagiarism cases. The same and neutral melodies are used as control melodies and to detect participants who are only performing at chance level.

2.1.1 Excursus to Receiver operating characteristics

The previously mentioned term "area under curve" is an index for the performance of a classifier. It can range from 0.5, chance performance, to 1.0, perfect performance, (see Figure 1), and is a simple way to display the participants' performance according to the classifier's assignation of this piece of data. Such an area under curve is the result of a Receiver operating characteristic (ROC) curve, which is defined by the true positive rate (TPR or hit rate) and the false positive rate (FPR or false alarm rate) of a data set. The higher the TPR and the lower the FPR, the larger the AUC will be. This means belongs to signal detection theory and it can include representations of sensitivity (= TPR) and specificity (= 1 - FPR).

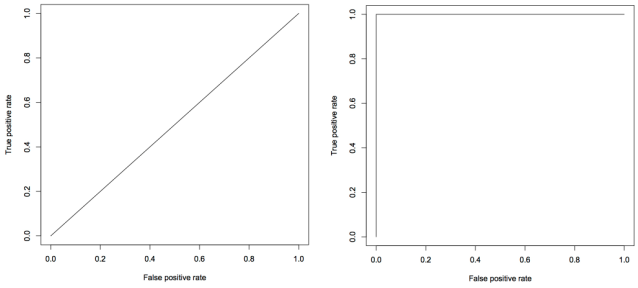

Figure 1: Comparison of different AUC (left: AUC = 0.5, right: AUC = 1.0)

Signal detection theory, in particular receiver operation characteristic (ROC), was used to identify participants only performing a chance level (with an area under curve of about 0.5). It is also applicable to compare the dichotomous court decisions with the experimental data and the algorithmic measures of similarity. The relationship between the results of the psychological experiments and each of the algorithms (Edit distances, opti3, Earth Mover's distances, Tversky's feature based similarity) was calculated by a Pearson correlation comparing the experimental data with one algorithm at a time.

2.2 Participants

The participants in this study are a group of 32 people (18 female, 14 male). Their mean age is 25.4 years (SD = 3.91, range from 19 to 39). Participants were recruited with only a few restrictions; professional musicians were not allowed to take part and people would not be recruited if they had a hearing deficit.

2.3 Materials/Stimuli

The core of the melodies used in the experiment was collected from melody plagiarism cases from the United States of America (17), United Kingdom (6), Australia (1) and Canada (1). All melodies were presented as midi files with the timbre of a piano.

These melodies were compiled from a bigger collection of melodies which was first divided into two groups of cases including rather well-known and less-known pieces: Well-known melodies were likely to be a confound in the implicit similarity test as people would not remember the melody but rather verbal information such as the artist or the title.

The melodies of twenty plagiarism cases with all less-known titles were cut to similar length and presented in the implicit similarity test. For all the melodies the keys were changed so that two melodies from the same court case were presented in fairly distant keys on the circle of fifths. The neutral melodies were selected from a list of less known melodies of the Goldsmiths College's Earworm project (http://www.gold.ac.uk/music-mind-brain/earworm-project/) and also adjusted in length.

The questionnaire on musicality used in this study is the Goldsmiths Musical Sophistication Index (GoldMSI) ([17] and http://www.gold.ac.uk/music-mind-brain/gold-msi/). It contains seven music-related subscales called Importance, Perception and production, Musical training, Emotions, Body, Creativity and Openness and events and therefore covers a broad spectrum of musicality. Demographic data, such as age, gender, education and current occupation was enquired at the end of the questionnaire. This questionnaire was included to form a distracting task between the study and the test phase of the implicit memory test with the aim that the memory for the melodies would stabilise before the test phase.

2.4 Procedure

The participants had to listen to the particular phrases from the less-known melodies involved in melodic plagiarism court cases three times. Each of the three repetitions was accompanied by a different cover task and the order of the melodies was different for every condition. Each participant was randomly allocated to one of two sequences of melodies so that one sequence contained ten plaintiff's and ten defendant's melodies all from strictly different court cases. Their paired melodies were used in the other sequence of melodies, respectively. The forty melodies from the twenty court cases were split with the result that only one melody of all the chosen court cases was present in either sequence. There was a break of about four seconds between all the melodies, and the melodies were presented from a CD. First, the participants had to state their familiarity with the specific melody ("familiar" to "non-familiar" on a Likert-type scale from 1 to 7) and they were offered a box to write down the title in case they remembered it. Second, they had to note how much they liked the presented melody ("like very much" to "like not at all" on a scale from 1 to 7). Third, they had to attribute a mood to each of the melodies ("happy", "sad" or "neutral").

After the study phase the participants had to fill in the complete questionnaire of the Goldsmiths Musical Sophistication Index.

In the testing phase the participants listened to five melodies from the exposure phase again ("same"), fifteen melodies which were the paired melodies from the remaining fifteen melodies from the exposure phase ("similar", possible case of plagiarism) and ten melodies from the earworm project ("neutral") in a randomised order. Again, all melodies have been of approximately the same length, they were presented from CD and there was a break of approximately four seconds between all the melodies. The participants had to state if they have heard the melody before on a 7-point scale from "very sure heard" (1) via "not sure" (4) to "very sure not heard" (7).

The whole testing time lasted approximately 35-45 minutes depending on the time needed by the participant to fill in the questionnaire. The duration of the initial exposure phase of the test was 15-16 minutes, the testing phase of the test was 7-8 minutes. In between participants have filled in the GoldMSI, which has taken approximately 12-20 minutes.

All answers were given on an paper questionnaire. Before the experiment started participants were asked for their consent and they were told that they could withdraw from the experiment at any time, no participant did make use of this possibility. After the experiment participants were handed a debriefing sheet and they were invited to ask questions about the purpose of this study.

3. RESULTS

There has been two cases of which all the collected data had to be excluded which were the court cases AS01 and US09.

US09 consists of two very particular melody phrases: One of them is a repetition of one pitch for several bars, the other melody is almost the same, however, it changes the pitch once, but except for that change, the phrases are identical. In this case the confusability of two melodies did not work, because this one single change in pitch or a melody without any changes in pitch are both very remarkable and therefore participants were very sure that they did not remember one of the two melodies from the study phase when hearing the paired one. Second, it is questionable if this "pitch construct" is a melody in a musical way at all, or if this musical "background" only serves the lyrics, which are presumably the most important feature of this song. On that account the experimental data based on these songs has been excluded from the data set.

Concerning AS01, Men at work have integrated a flute melody into their hit song "Down under" which is an exact copy of the first phrase of the tune "Kookaburra sits in the old gum tree", a well-known Australian children's rhyme. First, this is the only case in which the defendants have used the identical melody from the plaintiff's work. Second, this was the song most people have known, 8 have written down its title. Therefore the participants could not have confused two songs with one another because they are the same and also, they have not necessarily remembered the melody itself, but rather the title and as this melody was played again without any changes in the test phase, they have reliably remembered it ($m = 1.16$ on a scale from 1 to 7, 1 is "very sure heard"). For these reasons the song has also been excluded from further analysis.

3.1 Implicit memory test

None of the 32 participants was considered to be an outlier in the implicit memory test. All of them did distinguish well between the same and neutral melodies with a large area under curve (AUC) of 0.86 and higher, 23 participants did ideally segregate these two conditions and attained a perfect AUC of 1. The internal consistency of the participants was very high, for the rating of the 30 melodies they have listened to in the testing phase, Cronbach's alpha was as high as $\alpha = 0.955$ which is an excellent level of internal consistency.

The application of the receiver operating characteristic in this psychological data set will classify the participants' responses to the similar condition into either cases of actual plagiarism or cases of no infringement. The empirical data will be included as arithmetic means normalised to a scale from [0,1] of the participants' answers. The more the participants have gotten confused about the decisively plagiarised melodies, the higher the means of that empirical data, the higher the rank of that mean, and the bigger the AUC. The less the participants have gotten confused about the decisively non-plagiarised melodies, the lower the means of that empirical data, the lower the rank of that mean, and the bigger the AUC. This comparison between the dichotomous court decisions and the participants' confusability of plagiarised and non-plagiarised songs confirms

the hypothesis: The participants have successfully detected higher and lower levels of similarity which are reflected in the court's ruling for these melodies. The AUC for this classifier's performance is 0.69 (Figure 2), and since melodic plagiarism often involves more factors than just the melodies' similarity (such as the title or the lyrics and the likeliness of the defendant having known the plaintiff's melody), this result can be seen as a confirmation of the initial hypothesis.

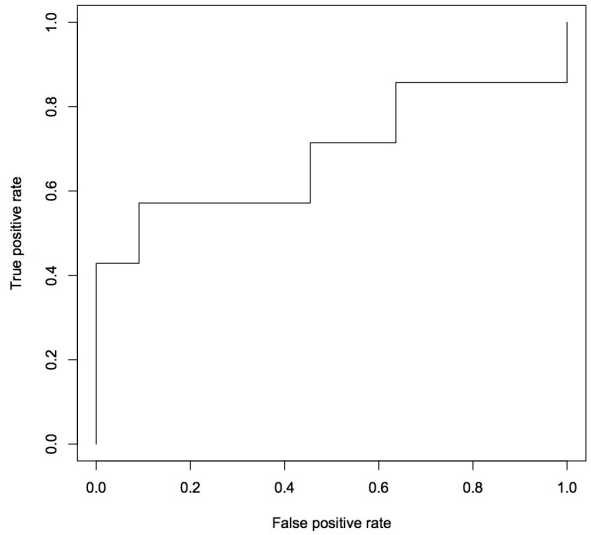

Figure 2: ROC curve for the court decisions and the empirical data (AUC is 0.69)

3.2 Comparison of psychological results, court decisions and similarity measures

The comparison between the court decisions and the different similarity measures is conducted by measuring the AUC for each combination; to compare the psychological data with the similarity measures a Pearson's correlation has been calculated.

All variations of the Earth mover's distance algorithm include a dimension of the midi onset in addition to the mentioned dimensions and methods of normalisation. This piece of information and the information about the pitch or the interval is taken to produce a map of onset and pitch of which the Earth mover's distance is calculated. In general, Edit distance measures with a focus on pitch perform well in the correlation with the psychological data (No. 1 and No. 3), surpassed only by the much more complicated opti3 measure (No. 2) and the feature-based measures (No. 12, 13, 15, 16). Amongst the plain Earth mover's distances the one using the city block distance and complete matching of pitch interval sequences performs best (No. 8) to recreate the court decisions. Of the four Tversky measures the focus on the features presented in the plaintiff's melody is the most promising approach for both the psychological data and the court decisions (No. 13 and 16). A recent implementation in Fantastic (http://www.doc.gold.ac.uk/isms/mmm/) produces an even better correlation coefficient for the psychological data; however, it performs less well in the comparison with the court decisions.

Table 1: Comparison of implicit memory test data, court decisions and similarity measures

No		Name and short explanation of measure	Corr.: psych. data	AUC: court decision
1	ED	Edit distance with pitch	**0.374**	0.591
2	hybrid	opti3, hybrid measure	0.185	0.571
3	ED	rawEdw, Edit distance with pitch weighted by duration	**0.515**	0.571
4	EMD	Pitch centralised, city block distance, complete matching	0.278	**0.779**
5	EMD	Pitch centralised, city block distance, partial matching	**0.284**	0.623
6	EMD	Pitch centralised, Euclidean distance, complete matching	0.264	0.766
7	EMD	Pitch centralised, Euclidean distance, partial matching	0.273	0.597
8	EMD	Pitch intervals, city block distance, complete matching	0.315	**0.844**
9	EMD	Pitch intervals, city block distance, partial matching	**0.326**	0.805
10	EMD	Pitch intervals, Euclidean distance, complete matching	0.302	**0.844**
11	EMD	Pitch intervals, Euclidean distance, partial matching	0.299	0.805
12	Tver	T.Tversky.equal	0.414	0.675
13	Tver	T.Tversky.plaintiff.only	**0.499**	**0.766**
14	Tver	T.Tversky.defendant.only	0.119	0.455 (sic)
15	Tver	T.Tversky.weighted	0.414	0.675
16	Tver	Tversky.plaintiff.only, new implementation in Fantastic	**0.514**	0.688

Aside from simple EMD measures a fitted model has also been created to recreate the psychological data using city block distance and partial matching. The model's weighting is $1 * \text{interval} + 0.03767 * \text{onset}$ and leads to a Spearman (rank) correlation coefficient of $r = 0.710$, but also an AUC of 0.597 which is considerably less than for the simple EMDs. These weightings have not been cross-validated yet so certainly overfit this data.

4. DISCUSSION

4.1 Implicit memory task

The implicit memory task has emerged as a useful test to investigate melodic similarity. Some parameters will have to be adjusted, for example the rather high number of melodies (20) in relation to the few listening repetitions (3). This ratio could

be disputed as too big and that the participants' memory would benefit from more repetitions in addition to a longer break between the study phase and the test phase than the approximately 12-15 minutes.

Participants' responses in comparison to the court decisions attain an AUC of 0.69, which already corrects for the unequal distribution of the 18 melody pairs of which only 7 were an actual case of copyright infringement. However, three of these 7 melody pairs were rated as rather less similar; and two of these three are melodic phrases from the same songs, which was a case of plagiarism due to these two phrases. To comprehend this decision it would be expedient to consult the case file and it shows that the lyrics are of vital importance in this law suit, and the judge stating in the end that this was "an extremely difficult case" for him (Case file of "There's Nothing like a Dame" from the musical South Pacific vs. National Express: Parody of "There's Nothing Like a Dame": Williamson Music Ltd. and Others v. The Pearson Partnership Ltd. and Another, 1986). This shows the importance of including the information from the court case to reconstruct all factors leading to the decision.

An interesting follow-study of this experiment would be a strict separation of the plaintiff's and defendant's songs and instead of randomising the direction of the implicit comparison one should include this as a further parameter since the directed similarity measure Tversky.plaintiff.only has shown such good results. Also, if this setup played the plaintiff's melody first to let it be compared with the defendant's it would match the actual listening situation of a person noticing a striking similarity between two songs.

On a small scale this data could provide a first impressionof divising the listening orders into plaintiff's melodies in the study phase and defendant's melodies in the test phase and vice versa. For the confusability responses emerging from first listening the defendant's melody followed by the plaintiff's melody in the test phase one might expect a better performance of the Tversky.defendant.only measure. There the participants will rather conduct a comparison of the plaintiff's melody contingent upon the previously heard defendant's melody.

4.2 Performance of the similarity measures for the implicit memory task

The most promising similarity measures to model the court decision's classification were Earth mover's distance with pitch intervals (AUC of 0.844) and centralised pitch (AUC of 0.779). This study shares five measures with Müllensiefen and Pendzich [9] and all five measures are less successful for the present set of melodies in correctly classifying the melody pairs into cases of actual and missing copyright infringement.

Table 2: Comparison of performance of similarity measures for two overlapping data sets

AUC of measure	data 2009	data 2011
ED	0.74	0.591
Tversky.equal	0.85	0.675
Tversky.plaintiff.only	0.95	0.766
Tversky.defendant.only	0.64	0.455
Tversky.weighted	0.85	0.675

The present data set partially includes melodies from the same court cases, which were adjusted in length to suit the listening experiment. Another explanation for the decrease in classification is a difference in jurisdiction for UK and USA where most of the melodic material was involved in a court case. However, the main reason for the decline seems to be the constant incorrect classification of two melody pairs (UK05b and UK07), which contribute to the false positive rate of the ROC for all five measures because of their low melodic similarity coincidentally with a court deciding for plagiarism. In consideration of the psychological ground truth, UK05b was the melody pair with the lowest implicit similarity rating in general, but UK07 appeared as rather similar (0.322, rank 6).

Some of the applied similarity measures performed very well in modelling the participants' (implicit) perception of similarity. A new implementation of the Tversky.plaintiff.only measure has obtained a correlation coefficient of 0.514, the weighted Edit Distance is as high-performing with r = 0.515, and the (still over-)fitted EMD reaches a Spearman correlation of 0.71.

Despite the two cases for which the respective case file needs to be consulted to reconstruct the counter-intuitive court's decision, similarity seems to be the key factor in cases of melodic plagiarism and people are able to distinguish between successful and unsuccessful lawsuits. Furthermore, a number of similarity measures also differentiate between the dichotomous outcomes of lawsuits although there is still much room for improvement when compared to a recent study sharing the choice of measures and court cases [9]. Concluding from the current correlational data of the implicit data there are some similarity measures that simulate the participants' perception of similarity very well.

Ensuing a similar comment given by some participants after the experiment, it seems to be an easier task to decide that they do not recognise a song compared to recognising a song and then deciding how certain they were. This suggests that the process of recognising is more complex than the process of registering something unknown. An interesting question arising from this assumption would be the search for this perceptual "threshold" of actually confusing a stimulus with another one. This could be approached by using well-known melodies in the study phase of the task, which should be solidly represented in the participant's long-term memory and are only played to refresh their memory. The melodies should be deviated until the participant does not recognise them any longer, i. e. considering existing knowledge of contour and marked tempo changes. The implicit similarity scale would have to be changed to the certainty or quality of recognising this song, possibly under a misleading assignment. Such an assignment could consist of an instruction to evaluate the transcriptions of rehearsals of a band trying to cover popular songs.

In conclusion, the research area of melodic similarity benefits from investigations collecting psychological data and forging a link between these results and computational modellings such as the similarity measures from this study. A process of adapting known measures and including new ones can lead to an enhancement of the current knowledge in the field of music perception itself and the possibility to connect it to already existing applications such as music information retrieval, folk song research and copyright infringement of music.

5. ACKNOWLEDGMENTS

The authors are grateful for the contribution of Jamie Forth who provided all the calculations for the Earth Mover's distance.

6. REFERENCES

[1] Musipedia. Retrieved 31 August 2011, from http://www.musipedia.org/

[2] Dowling, W. J., & Fujitani, D. S. (1970). Contour, interval and pitch recognition in memory for melodies. The Journal of the Acoustical Society of America, 49(2), 524-531.

[3] Müllensiefen, D., & Frieler, K. (2004). Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. Computing in Musicology, 13, 147-176.

[4] Frieler, K., & Riedemann, F. (2011). Is Independent (Re)creation Likely to Happen in Pop Music? Musicae Scientiae, 15(1), 17-28.

[5] Gärdenfors, P. (2004). Conceptual spaces. The geometry of thought. Cambridge, London: The MIT Press.

[6] Müllensiefen, D., & Frieler, K. (2007). Modelling expert's notions of melodic similarity. Musicae Scientiae, Discussion Forum 4A, 183-210.

[7] Tversky, A. (1977). Features of similarity. Psychological Review, 84(4), 327–352.

[8] Typke, R., Wiering, F., & Veltkamp, R. C. (2007). Transportation distances and human perception of melodic similarity. Musicae scientiae, Discussion Forum 4A, 153-181.

[9] Müllensiefen, D., & Pendzich, M. (2009). Court decisions on music plagiarism and the predictive value of similarity algorithms. Musicae Scientiae, Discussion Forum 4B, 257-295.

[10] Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. Psychometrika, 22, 325-345.

[11] Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. Cognition, 87, 1-32.

[12] Hahn, U., & Chater, N. (1997). Concepts and similarity. In K. Lamberts & D. Shanks (Eds.), Knowledge, concepts and categories (pp. 43-92). Cambridge, London: The MIT Press.

[13] Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. Science, 237, 1317-1323.

[14] Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. Psychometrika, 27 (2), 125-140.

[15] Nosofsky, R. (1991). Stimulus bias, asymmetric similarity, and classification. Cognitive Psychology, 23, 94–140.

[16] Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. Psychological Review, 89, 123-154.

[17] Müllensiefen, D., Gingras, B., & Stewart, L. (in prep.). Piloting a new measure of musicality: The Goldsmiths Musical Sophistication Index. Technical report. Goldsmiths, University of London.