

An evaluation of melodic similarity models

Ludger Hofmann-Engl
The Link
Beddington
London
hofmann-engl@chameleongroup.org.uk

chameleongroup online publication 2005

Ludger Hofmann-Engl
Flat 2
59 Warminster Road
London SE25 4DQ
UK

Abstract

The advance of music information retrieval (MIR) has brought about a strong interest in melodic similarity models. In fact, the majority of these models derive their rationale from the implementation within a MIR context. This is, so authors argue in general, a model is suitable if it retrieves the desired melody or a close variation of it. This post hoc method does not permit the evaluation and comparison of these models (although such a “comparison” competition has been proposed for ISMIR 2005). The author of this paper approaches the issue from a different angle; rather than testing models in an unsystematic fashion, the author will discuss the underlying cognitive principles of the similarity models. Here, it will be shown that four principle strategies exist: The contrast models, the distance models, dynamic programming and transition matrices. The author will then demonstrate that a variety of distance measures based upon a specific representation of melodies appears most promising while insisting that no single measurement can be seen as the answer to everything.

1. Introduction

The interest in melodic similarity has mushroomed over the last few years. Although there exists some interest in the issue at hand from within disciplines such as ethnomusicology (e.g. Toivianen & Eerola, 2002) and the cognitive sciences (e.g. Clarke & Dibben, 1997), the strongest interest exists within the league of MIR researchers. For the fifth year running, the International symposium on music information retrieval (ISMIR, 2000, 2001, 2002, 2003 & 2004) included several papers on the issue in the year 2004 just as it did in previous years: Melodic similarity lies at the heart of MIR. This is, in order to retrieve a melody from a database, a query has to be placed. However, in many cases such a query is partly corrupted (e.g. when the user of a database does not remember parts of the wanted melody). In case that such a query is partially corrupted, an algorithm is needed to search the database for melodies which are similar to the query. Hence a working model of similarity is required.

Now, as we will see, although not all models are derived from the background of MIR, most models are rooted within the MIR approach and hence they are tested within MIR applications only. Typically, a researcher has access to a musical database which contains a number of melodies (e.g. the Essen database; compare Smith, McNab & Witten, 1998). In order to search such a database, a query in form of a notated or hummed melodic fragment is being placed. The researcher then looks at the melodies, which the similarity algorithm retrieves from the database, and expresses satisfaction that the retrieved melodies are close matches to the query indeed, based upon her or his subjective judgement. Examples of this kind of studies are: Doraisamy & Rüger (2002), Pienimäki, (2002) and Typke, Giannopoulos, Velkamp, Wiering, Ren Oostrum, (2003). Although some researchers make references to psychological data and similarity theories (e.g. Hofmann-Engl, 2001; Pauws, 2002) and Müllensiefen & Frieler, 2004b), such knowledge is generally ignored rendering these melodic similarity algorithm speculative at its best and wrong as its worst. Additionally, only very recently have there been made attempts to compare melodic similarity models such as the studies by Grachten, Arcos & Mántaras (2002) and Müllensiefen & Frieler (2004a & 2004b). However, none of these studies are comprehensive and in the case of Grachten, Arcos & Mántaras (2002) only models belonging to the class of dynamic programming have been

considered. Furthermore, in this case no reference to psychological similarity models is given and the evaluation is conducted in the typical MIR fashion. Finally, a paper which in its title “A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures” (Berenzweig, Logan, Ellis & Whitman, 2003) suggested to address the issue, refers to listeners’ judgments about how similar two artists are when compared to each other and does not address the issue at hand.

As almost every researcher has her/his own melodic similarity algorithm, it would be impractical to examine each one of them. Here, it seems more efficient to classify existing approaches into categories and evaluate the underlying principles, and then to consider examples of melodic similarity models illustrating their advantages or their disadvantages. This is, what will be done in chapter 2. However, before we start this discussion, the author wants to stress, that, as we are dealing with similarity algorithms applied to a cognitive context (a model ought to be correlated to the cognition of melodic similarity), the consideration of psychological and cognitive aspects will be of paramount importance.

2. The four approaches to melodic similarity

It is a rather difficult task to classify existing approaches to the issue of melodic similarity in a systematic fashion. This is, some of the existing approaches are motivated by cognitive and psychological research, while others are simply algorithms which are constructed for specific purposes (such as music information retrieval). However, interestingly, it seems possible to group all existing approaches into four classes. Two of these classes are rooted within the cognitive science. These are the models based upon the contrast model as introduced by Tversky (1977) and the distance model as developed by Shepard (1987). A third approach known as dynamic programming, which was first proposed for the purpose of measuring similarity by Goad & Kanehisa (1982), has been used in many contexts, such as bio-informatics, chemistry, music and multimedia information retrieval. More recently, transition matrices have been employed to measure melodic similarity by Hoos, Renz & Görg (2001). The author decided it would be best to discuss each of the different approaches in general terms (except transition matrices as it appears

that this is an approach used only in music information retrieval), offer some critical remarks and look at their implementation in the context of melodic similarity. Chapter 3 and 4 are dedicated to some general observations and an outlook into what melodic similarity model ought to deliver, suggesting that a melodic representation as introduced by Hofmann-Engl (2001, 2002, 2003, 2004) appears to be promising.

2.1. Contrast models

Central to Tversky's contrast model is the assumption that similarity is related to the weighted difference of measures of their common and distinct features. Thus, two objects A and B will be the more similar the more features they have in common and the less similar the more features the objects do not have in common. The model is usually presented in the following form:

$$S = \vartheta c + \alpha a + \beta b \quad (1)$$

where S is the similarity, ϑ , α , β empirical constants, c the count of common features, a the count of features present in object A but not in B and b the count of features not present in object A but in B .

Applying this model, we find, that a white door with 4 wooden panels (a), will be more similar when compared to a white door with two wooden panels (b), than it will be when compared to a white door with no wooden panels (c). This situation might change if additional features are considered. For instance if the white door with 4 wooden panels (a) is the same standard size as the door without wooden panels (b) and the white door with 2 wooden panels (c) is in fact part of a doll house. Still, whether this additional feature will effect a reversal in similarity, so that (a) is more similar to (c) than (a) to (b) will depend on the empirical constants ϑ , α , β .

Apparently, Tversky did not, as pointed out by Bradshaw (1997), take any features into account which are not present in both of the compared objects. Although this critique might seem formalistic (i.e. to say: "we count features both objects have, features the first object has but not the second one and features the first object does not have but the second one has, hence we also have to count the features neither object have"), but it is substantial in as much as it is related to a critique

by Goodman (1972). According to Goodman the question, whether an object X is similar to an object Y is meaningless if not stated in respect to a property Z . For instance, if the comparative property Y is colour, then all three doors of our previous example have the same similarity status. However, setting the comparative property Y to be functionality, door (a) will be more similar to door (b) than to doll house door (c). Tversky justifies his approach by stating that, "when faced with a comparison or identification problem we extract a limited number of relevant features on the basis of which we perform the required task." Although this might be the case, we still face the problem of identifying how and which are these features and how they are extracted. This is an issue raised by Barsalou (1982), who found that raccoon and snake are more similar when compared without further context specification than when compared in the context of the category pets. We might argue that Barsalou's similarity experiment investigates conceptual similarity, rather than cognitive similarity and that Tversky model is applicable to cognitive similarity and hence not adversely affected by Barsalou's findings. Here, we understand conceptual similarity as the similarity between two concepts such as Tel-Aviv and New York. Cognitive similarity on the other hand is the similarity between items which we actually perceive, such as the similarity between two snakes we see (or two melodies we hear). However, Medin, Goldstone & Gennter (1993) demonstrated in an experiment that cognitive similarity is likely to be affected by context as well. These researchers found that when an object which has ambiguous features (a drawing which can be interpreted as either a three dimensional or a two dimensional representation) is compared with an unambiguous object (a drawing which can only be interpreted as two dimensional) participants of the experiments adopted the unambiguous feature to interpret both objects (both objects will be seen to be two dimensional). If such cross-influences occur, it seems an unlikely assumption that similarity should be independent of context. It appears that context might pose a more serious threat to Tversky's model than he seemed to confess.

This deficiency is heightened by Tversky's own discovery of asymmetrical similarity judgements. This is, an object A compared with an object B will not necessarily produce the same similarity measure when the comparison order is reversed to comparing object B with object A . Thus for instance, Tversky found that the similarity between Tel Aviv and New York is greater than the similarity between New York and Tel Aviv. A reason for such asymmetrical judgements is

given by Nosofsky (1991). He argues that an object with high frequency presence (e.g., an object which we see often) is likely to be stored in the memory more strongly than an object with low frequency (e.g., an object we rarely see). Further, he maintains that an object with higher memory strength (in this case New York) will be activated more by an object of low memory strength (in this case Tel Aviv) than vice versa. Although this seems to be a reasonable explanation, we also can explain asymmetries by referring to the previous paragraph: Comparing Tel Aviv to New York (both are multi cultural and have a beach) will produce the selection of different relevant features than comparing New York with Tel Aviv (New York is a metropolis, but Tel Aviv is not). The advantage of this explanation is that it does not involve vague theoretical concepts. However, the point is that such asymmetries are seemingly in conflict with Tversky's model. This conflict can only be resolved when we introduce the set \mathcal{A} consisting of all relevant features for a specific comparison task comparing object A and B , where we set \mathcal{A} to be a function of the comparison order with $\mathcal{A}(A,B) \neq \mathcal{A}(B,A)$. Admitted, this is no elegant solution. Still, if we consider the context where asymmetries have been reported (e.g., Medin, Goldstone & Gentner, 1993), we find that such asymmetries seem only to occur under conceptual similarity tasks and not under cognitive similarity tasks. (It does not matter whether we look at a snake 1 first and then at snake 2 or vice versa in order to determine whether they are similar). Interpreting Tversky's model as an exclusively cognitive model, we might safely ignore asymmetries.

Tversky also demonstrated, in reference to his model, that the so called triangle inequality does not hold scrutiny in the context of similarity judgement tasks: The triangle inequality, until Tversky's seen as a psychological fact, states that the psychological distance between two points a and c is lower than or equal to the sum of a to b plus b to c . This is $D(a,b) + D(b,c) \geq D(a,c)$. Quite clearly, this law does not hold when we consider the following example: Given an object A (red triangle), an object B (blue triangle) and an object C (blue square), we find, that although A is close to B because of shape and B is close to C because of colour, A and C are not close. This discovery is of crucial importance when similarity ratings are made on larger sets of objects necessitating the comparison of each object with each object.

There have been several proposals on how to modify Tversky's model, with the model by

Markman & Gentner (1993, 1996, 1997) being possibly the most interesting one. They proposed a structure-based model. Here, feature commonalities and feature differences are replaced by alignable commonalities, alignable differences and non-alignable differences. When comparing two objects, an alignable commonality is a shared feature which does not only exist in both objects but is also structurally at the same position (isomorphic) in both objects. For instance the wheel on a bicycle is isomorphic to the wheel of a motorcycle, but not to the wheel on a sewing machine. Shared features which are not alignable are called non-alignable. Alignable differences are deviations in features at the same position (e.g., the bicycle has pedals instead of the engine on the motorcycle). Non-alignable differences are features at a position in one object while there are no features at all at the other object (e.g., the tank on a motorcycle). The authors have been able to produce some evidence that alignable differences influence similarity judgements more than non-alignable differences do. This seems to confirm the validity of their approach. However, a major logical problem underlies their understanding of isomorphism. Even if Markman and Gentner understand the isomorphism in a more colloquial sense, it might be useful to consider a more formal definition of what an isomorphism is. Mathematically speaking, two objects A and B are isomorphic, if all positions of A can be mapped unequivocally onto a corresponding position in B by a function (generally written as: $F(a_1 + a_2) = F(a_1) + F(a_2)$ with $a_1, a_2 \in A$ and $F(a_1), F(a_2), F(a_1 + a_2) \in B$). Quite clearly, a bicycle and a motorcycle do not fulfil this criterion, or generally speaking, isomorphism is not a requirement for us to consider two objects A and B to be similar. The only way for Markman and Gentner to save the idea of isomorphism, requires the consideration of a local isomorphism by segmenting the objects into sections. Thus, we might segment A into $A_1, A_2 \dots A_n$ and $A'_1, A'_2 \dots A'_n$ while B might be segmented into $B_1, B_2 \dots B_n$ and $B'_1, B'_2 \dots B'_n$. An isomorphism might be established between A_1 and B_1 , between A_2 and B_2 and so on (alignable segments), while some segments $A'_1, A'_2 \dots A'_n$ and $B'_1, B'_2 \dots B'_n$ might remain without such an isomorphism (non-alignable segments). However, such a segmentation is already ambiguous and so is the question which segment to map onto which other segment (e.g. A_1 onto B_1 or A_1 onto B_2). For instance, the question, which metal bar on the bicycle should be mapped onto which metal bar on the motorcycle, appears to be a rather difficult one. Another, maybe more obvious, example would be given by the comparison of a chair with four differently shaped legs with a chair with three differently shaped legs. The question which legs are to be aligned or are isomorphic and which leg is to remain non-

aligned, might turn into an artful task. Moreover, such a segmentation will threaten the overall meaningfulness, as two features might show local isomorphism, and yet this local isomorphism might be accidental when considering the objects as a whole. Thus, the spinning wheel on a car might be aligned with the spinning disk of hard-drive in a PC, but whether there lies any meaning in doing so is another question. In fact, an appropriate aligning seems to imply that we understand the functionality of the compared objects. However, such understanding implies underlying theoretical constructs which themselves will have implications on similarity. Thus, the author concludes that, although Markman and Gentner's model might shed light onto some simple examples (e.g. comparing Motel with Hotel and Hotel with Motorcycle), their approach seems to produce more problems than it sets out to solve.

2.1.1. Contrast models in music

There have also been several applications of Tversky's model in the context of melodic similarity (e.g. Kluge, 1996; Uitdenbogerd & Zobel, 1998). However, it seems that the most far reaching attempt was undertaken by Cambouropoulos (1998).

Cambouropoulos considered similarity in the context of categorization in an effort to offer a computational model of melodic segmentation. The underlying principle is the idea to vary a threshold h so as to allow the similarity of motives to generate categories so that similar motives will be found in the same category. Cambouropoulos defines the following relations:

$$d(x, y) = \sum_v w_{x_i} w_{y_i} (1 - \delta_{x_i y_i}) \quad (2)$$

and

$$S_h(x, y) = 1 \text{ if } d(x, y) \leq h \text{ and } S_h(x, y) = 0 \text{ if } d(x, y) > h \quad (3)$$

where $d(x, y)$ is the distance between the entities (motives) x and y , w_{x_i} and w_{y_i} are weighting factors (which are under-defined in Cambouropoulos's work), x_i and y_i the i th feature of the entities x and y respectively, v the number of features, $\delta_{x_i y_i}$ the Kronecker delta (with $\delta_{x_i y_i} = 1$ for $x_i = y_i$ and $\delta_{x_i y_i} = 0$ for $x_i \neq y_i$), $S_h(x, y)$ the similarity between the entities x and y (either similar or dissimilar) and h is the threshold (variable number).

Once two entities x and y reach a level $d(x,y)$ above the threshold h , they are considered dissimilar and if $d(x,y)$ lies below h , they are considered similar, which then serves in his unscramble algorithm as the criterion to draw up categories. Unfortunately, the question which are the relevant entities (features) x_i and y_i is assumed to be answered without ever being asked. Thus, Cambouropoulos uses the features: exact pitch intervals, contour and durations. As demonstrate by Hofmann-Engl (2003), these features alone are insufficient to describe melodic similarity and this is particularly true for the counter. In an experiment conducted by Hofmann-Engl (2003), it was found that exact interval difference is a similarity predictor rather than contour. This is, a melody ascending 3 steps (e.g. 3 semi-tones) compared to a melody ascending 1 step only, produces an interval difference of $3 - 1 = 2$. The same is true for a melody ascending one step +1 compared with a melody descending one step -1. The interval difference here is: $1 - -1 = 2$. Hofmann-Engl found that the measured similarity ratings for both cases were the same although the contour differs in the second case but not in the first. Using multiple correlation, it was shown that contour is an insignificant predictor. Moreover, Cambouropoulos's model seems to imply complexity, but de facto it is a reduced form of Tversky's model, only considering commonalties and not taken into account differences and it appears to be an inferior version. Clearly, a model of melodic similarity will have to refer to pitch (or some correlate of pitch), duration and dynamics. So, for instance, a model could count how many pitches two melodies A and B have in common, how many pitches are in A , in case A has different length than B , which are not in B and how many pitches are in B which are not in A . Additionally, higher level features such as tone repetitions or symmetries (e.g. sequences) are features which will allow for counting differences. Additionally, the findings by Egmond, Povel, & Maris (1996) and Hofmann-Engl (2003) are in contrast to Cambouropoulos's model. These researchers found that similarity judgements decrease with increasing transposition interval. According to the above described model, however, we find that all transpositions are treated as equivalent. Finally, considering the discussion above about Tversky similarity, it is questionable whether contrast models are of meaning in the context of melodic similarity.

2.2.The distance models

The second approach to a similarity measures was put forward by Shepard (1987). Here,

similarity is ultimately related to the distance between all the points of the objects' attributes. Thus, if the attributes of two objects A and B fall into five dimensions (for instance: weight, colour, volume, shape, sound characteristics), we will obtain a 5-dimensional attribute vector for each object. The similarity then is a function of the distance between the attribute vector of object A and object B . We give a physical example: Object A is a cube (12 sides), 5 kg, red (let us say wavelength is 660 nm) and produces a low frequency of 200 Hz. Object B is a pyramid with square base (8 sides), 4 kg, blue (let us say wavelength is 460 nm) and produces a high frequency of 2000 Hz. Now, difference in sides is $12 - 8 = 4$, in weight $5 \text{ kg} - 4 \text{ kg} = 1 \text{ kg}$, in colour $660 \text{ nm} - 460 \text{ nm} = 200 \text{ nm}$ and in frequency $2000 \text{ Hz} - 200 \text{ Hz} = 1800 \text{ Hz}$. Thus, the similarity will be a function of the differences. Although the author used a physical example for the purpose of clearness, Shepard constructs an "abstract psychological space" for the similarity measure. However, he discerns various dimensions of this abstract space as being approximated by physical dimensions (e.g., psychological space distance as measured by a Euclidean metric, or in the case of pitch by the frequency ratios). Referring to this model, we find that the above mentioned experiment by van Egmond, Povel & Maris (1996) can be easily explained in form of a 1-dimensional distance similarity measure. If we form the distance between the first pitch p_a of melody A and the first pitch p_b of melody B , we obtain the transposition interval $I = p_a - p_b$. Thus, the similarity S will be proportional to I :

$$S \propto I \quad (4)$$

where S is the similarity and I the transposition interval between two melodies.

Shepard's model is usually written in the form:

$$d(x, y) = \left(\sum_{k=1}^D |x_k - y_k|^p \right)^{1/p} \quad (5)$$

where $d(x,y)$ is the generalized distance of the objects x and y within the psychological space of dimension D , x_k and y_k are the psychological quantities of object x and y along the k th dimension, p is an empirical constant.

Applying this model to the findings in Egmond, Povel & Maris's study concerning the transposition interval, we get $p = 1$ and $D = 1$. This metric ($p = 1$) is called city-block metric in contrast to a metric with $p = 2$ which is called Euclidean metric.

This approach is not only supported by the study above and the experiments by Hofmann-Engl (2003), but by several studies conducted by Shepard (for instance that the length of time it takes participants to make same/different judgements about pairs of shapes, one in standard position and the other rotated, is proportional to the degree of rotation). However, it seems that there are two major problems with Shepard's model as it stands. Firstly, as observed by some researchers (e.g. Cardie & Howe 1997) the model does not incorporate the weighting of specific dimensions, although it seems highly unlikely that all psychological dimensions will weigh the same (for instance the loudness dimension versus the pitch dimension). However, this is easily fixed by introducing a weighting factor w_k for each attribute (as we will see, this is exactly what O'Maidín (1998) proposed). More serious might seem the second issue: The asymmetries as found in Tversky's experiments are not built into the model. The minimal expense required to solve this problem will call for a weighting factor which depends on the comparison order, possibly in the form w_{kx_y} for comparison of the object x with object y and w_{ky_x} for comparison of the object y with the object x . However, if we understand the distance model as a cognitive and not as a conceptual similarity model such weighting might not be necessary. It appears that a suitable melodic similarity model might belong to the class of distance models.

2.2.1 Distance models in music

A modification of Shepard's model has been put forward by Kluge (1996), who is apparently unaware of Shepard's model, for the application to music analysis. Hereby, Kluge proposes a city-block matrix. However, he sees that the similarity distance should be weighted by the amount of attributes. Thus, we get:

$$d(x, y) = \frac{\sum_{k=1}^n |x_k - y_k|}{n} \quad (6)$$

where $d(x,y)$ is the generalized distance of the objects x and y within the psychological space, x_k and y_k are the psychological quantities of object x and y along the k th dimension and n the amount of attributes.

However, Kluge does not specify the attributes of a melody which will have to be taken into account. Thus, his model remains abstract and how to apply the model to music remains unclear. The omission of p as found in Shepard's model also seems to weaken this model, as there is no means of adapting this model to empirical data.

A more elaborate model was put forward by O'Maidín (1998). He proposed the following model:

$$difference = \sum_k^n |p_{1k} - p_{2k}| w_k w_{s_k} \quad (7)$$

where p_{1k} is the pitch of the note from the first segment at the k th window, p_{2k} is the pitch of the note from the second segment at the k th window, w_k is the width of the k th window, w_{s_k} is the weight derived from metrical stress for the window k and n is the amount of windows.

Before we will interpret this formula, we can see some improvements and some impoverishment when compared to both Kluge's and Shepard's model: Firstly, this model does not contain the empirical constant p as in Shepard's model, which will imply a reduced empirical adaptability. It also does not divide the sum by the amount of summands n in contrast to Kluge's model. This again seems problematic as it implies that the longer two melodies, the less similar they are regardless of any other features. However, his model shows some strength by introducing two weighting factors. O'Maidín suggested to use a weighting factor w_k , which basically gives more weight to notes of longer durations (duration of a "window"). Thus a crotchet might fetch the value $w_k = 1$, while a minim might fetch the value $w_k = 2$. The second factor w_{s_k} gives weight according to metrical stress. Thus, an upbeat note might fetch the value $w_{s_k} = 4$, while a down beat might fetch the value $w_{s_k} = 2$. However, the choice of the weights is, according to O'Maidín, arbitrary. This seems to be an unsatisfactory point of view as the choice of the weights can affect the order of similarity of three motives. For instance, let us assume we have three motives M_a , M_b and M_c . All motives consist of four crotchet notes, written in a 3/4 times and starting on the first beat of bar 1

and lasting to the first beat of bar 2. With motive $M_a = [c, d, e, d]$, motive $M_b = [c\#, d, f, d]$ and $M_c = [c, d\#, c, d]$, we obtain the difference for $\Delta(M_a, M_b) = 2 + 0 + 1 + 0 = 3$ and $\Delta(M_a, M_c) = 0 + 1 + 3 + 0 = 4$ for $w_{S_k} = 2$ for the first beat of a bar and $w_{S_k} = 1$ for any other beat and with 1 semitone = 1 as pitch unit. Thus we find that motive M_a is more similar to M_b than to M_c . However, if we change $w_{S_k} = 4$ for the first beat of a bar and $w_{S_k} = 1$ for any other beat, we obtain: $\Delta(M_a, M_b) = 4 + 0 + 1 + 0 = 5$ and $\Delta(M_a, M_c) = 0 + 1 + 3 + 0 = 4$. Hence, motive M_a is now more similar to motive M_c than to motive M_b . This renders the proposed algorithm an arbitrary tool of low reliability. Surely, weights will have to be adjusted empirically. O'Maidín also suggested to integrate the variable m into the model, where we obtain:

$$difference = \sum_{k=1}^n |p_{1k} - p_{2k} - m| w_k w_{S_k} \quad (8)$$

where p_{1k} is the pitch of the note from the first segment at the k th window, p_{2k} is the pitch of the note from the second segment at the k th window, w_k is the width of the k th window, w_{S_k} is the weight derived from metrical stress for the window k and n is the amount of windows and m an integer.

He suggests to vary the value m , until the difference takes a minimum value. The purpose of this is clear: Should the second fragment be a transposition of the first fragment, we obtain for all $p_{1k} - p_{2k} = a$ with a as a constant. Setting $m = a$, we obtain a difference of 0 between the two fragments (maximum similarity). Thus, the introduction of m renders the model transpositional invariant. However, the meaning of m becomes more obscure when the two fragments are not identical; an issue surely to be investigated. True, by varying m , we might obtain a minimal difference, but whether this minimal difference implies maximum similarity is questionable. For instance, the model regards all differences according to a city-block metric without considering whether other metrics might be more appropriate (e.g., Euclidean metric, which might produce different minimal values). A second objection against this model is based on its computational implications. Assuming we are comparing just five motives with each other (all in all 15 comparisons) and assuming these motives are not more than two octaves apart from each other, we will have to compute 15 times 24 (= 360) differences, which will have to be compared and evaluated. This is surely no elegant solution. However, the main objection arises when considering

the findings by Egmond & Povel & Maris (1996) and Hofmann-Engl (2003), who demonstrated that transposition is a melodic similarity predictor. This is, the larger the transposition interval the smaller are the similarity ratings.

A third, more recent application of a distance model, was introduced by Typke, Giannopoulos, Veltkamp, Wiering & Oostrum (2003) utilizing the Earth Mover's distance. The Earth Mover's is the distance measure between discrete, finite distributions such as: $X = \{ (x_1, w_1), (x_2, w_2), \dots, (x_m, w_m) \}$ and $Y = \{ (y_1, u_1), (y_2, u_2), \dots, (y_n, u_n) \}$. Hereby, $x_i - y_j$ represents the distance d_{ij} between the weights w_i and u_j and the difference $w_i - u_j$ represents the flow f_{ij} from X_j to Y_j . The entire weight of X is: $W_x = \sum_{i=1}^m w_i$ and the weight of Y is: $W_y = \sum_{j=1}^n u_j$. We arbitrarily set: $W_x \geq W_y$

The following rules apply:

1. The flow goes from the heavier set X to the lighter set Y and it has to be positive. This is: $x_i - y_j > 0$
2. The weight of x_j can flow to several points in Y , but can not exceed the weight of x_j : This is: $\sum_{j=1}^{m_i} f_{ij} \leq w_i$ Additionally, a point in Y has to absorb exactly the amount of it own weight: This is: $\sum_{i=1}^{n_j} f_{ij} = u_j$.
3. The total transported weight is the minimum sum of all flows, which is identical with the sum of all weights of the lighter set and equals W_y .

Finally, the Earth Mover's distance is calculated as:

$$EMD(X, Y) = \min \frac{\sum_{i=1}^t \sum_{j=1}^{m_j} f_{ij} d_{ij}}{W_y} \quad (9)$$

where $EMD(X, Y)$ is the Earth Mover's distance between the sets X and Y , f_{ij} is the flow from x_i to y_j , d_{ij} is the distance between x_i and y_j , t is the number of elements in X where flow occurs, m_i is the number of elements in Y the element x_i flows to and the sums are to minimal.

As abstract as these definitions appear, an example will help illustrating that the underlying

mechanism displays some simplicity. The following example is taken from Cohen (1999):

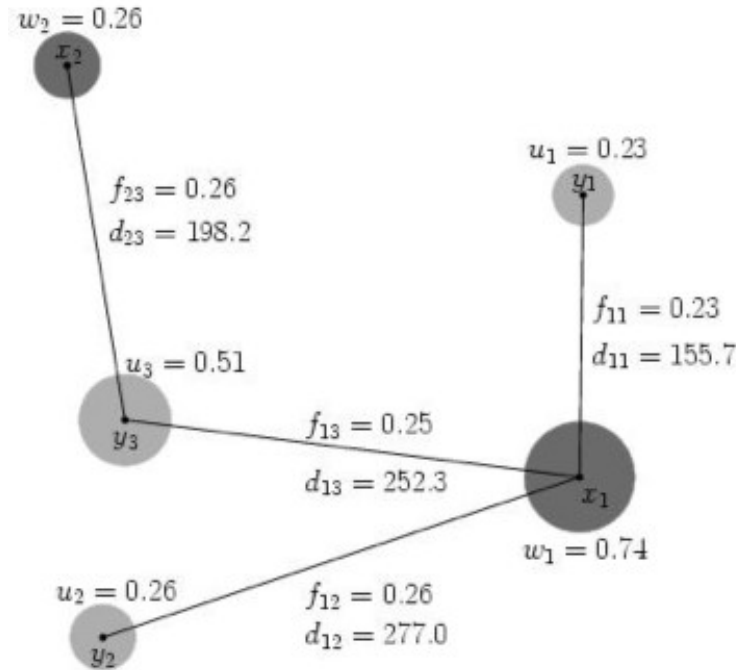


Figure 1: The application of the Earth Mover's distance to two sets of points in a two-dimensional space as taken from Cohen (1999).

The algorithm works in the following manner: Firstly, the distances between all elements of X and all elements of Y have to be calculated. Subsequently, the weight of the element of X with the smallest distance to an element of Y transports its weight to this element. The same happens then to the element which is second closest, followed by the third closest and so on, until all elements of Y are filled. This means, for the example above: x_1 is closest to y_1 ($d_{11} = 155.7$) and transports 0.23 of its weight. The second smallest distance is $d_{23} = 198.2$. Thus, the entire weight of x_2 ($= 0.26$) flows into y_3 . Next is the distance $d_{13} = 252.3$ which means that another 0.26 have to flow from x_1 into y_3 . The remaining 0.26 of weight flow into y_2 . We obtain the $EMD(X, Y)$:

$$EMD(X, Y) = \frac{155.7 * 0.23 + 198.2 * 0.26 + 252.3 * 0.25 + 277.0 * 0.26}{1.0} = 222.4$$

The proof that weights always have to flow over the smallest distance in order to generate minimal EMD is given here:

Let x_1 and x_2 be points in X and the point y_1 a point in Y . Let further d_{11} be the distance between x_1 and y_1 and d_{21} the distance between x_2 and y_1 . Let further y_1 have the weight u_1 .

Now, allowing both points of X to flow into y_1 filling it, we obtain the distance:

$$D = d_{11} f_{11} + d_{21} f_{21}$$

As f_{11} and f_{21} fill y_1 , we obtain: $f_{11} + f_{21} = u_1$. We further assume that f_{11} fills u_1 partially and we set: $f_{11} = u_1 - c$. Consequently, we obtain for $f_{21} = c$. Finally, we assume that d_{21} is larger than d_{11} by the value e . Thus, we obtain:

$$\begin{aligned} d_{11}(u_1 - c) + (d_{11} + e)c &= d_{11}u_1 - d_{11}c + d_{11}c + ec \\ &= ec + d_{11}u_1 \end{aligned}$$

This equation is either minimal if $e = 0$ (when both x_1 and x_2 have the same distance to y_1), where it does not matter how the flow occurs, or if $c = 0$, which means that no flow occurs from the point further away from y_1 . Hence, flow always occurs along the smallest distance.

Clearly, at this point we ought to comment on the critical issues connected with this approach, but this will be easier and more stringent when it is done in the context of its application to music. Still, one point might be mentioned here; having to compute all the differences between the points of X and Y , having to order the differences and then to compute the distances results in expensive computational times. As reported by Typke, Giannopoulos, Veltkamp, Wiering & Oostrum (2003), this might result in a running time of ca. 70 min when searching a larger database.

This specific type of distance model has been implemented into a melodic similarity model by Typke, Giannopoulos, Veltkamp, Wiering & Oostrum (2003). An example of it is given below.

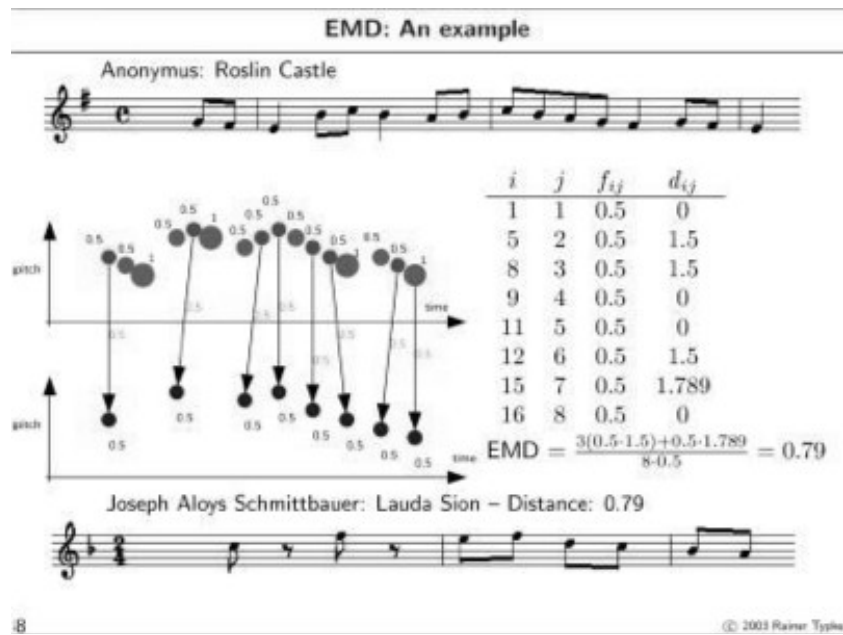


Figure 2: The application of the Earth Mover’s distance comparing to melodies as taken from Typke, Giannopoulos, Veltkamp, Wiering & Oostrum (2003).

Here, the distance d_{ij} is calculated as the Euclidean distance consisting of the pitch component and the onset time of a note, such as: $d_{ij} = \sqrt{(p_i - p_j)^2 + (t_i - t_j)^2}$. In the example above, the top melody has been transposed up by a perfect fourth. Hence the distance between the first note of the top melody and the first note of the bottom melody is 0. The next match is $i = 5$ and $j = 3$, there is no pitch difference, but time difference of 1.5 (the units are set arbitrarily). The same holds true for the third and sixth match. There is no difference between match 4, 5 and 8. Only match 7 displays both a time difference of 1.5 and pitch difference of 0.95 (again units are set arbitrarily). The durations are represented in form of weights. In the above example one quaver fetches the value 0.5 and a crotchet the value 1. Thus, we obtain the EMD:

$$EMD = \frac{0 + 0.75 + 0 + 0 + 0.75 + 0.8945}{8 * 0.5} = 0.79$$

As interesting as this application is, it does not come without serious problems. Considering the above figure, we find that the way the EMD is defined, that not all weights flow from the heavier to the lighter set. This means that several notes of the top melody are deleted without being accounted for. This is in contradiction with Tversky’s model as much as it is in contradiction with

dynamic programming (see below). According to Hofmann-Engl (2003), melodic intervals are a significant melodic similarity predictor. Hence, a note which is deleted between two others (let us say a note between the two notes c and e) will effect similarity (e.g. whether we delete the note d between c and e or the note g). This means, not only that deletion ought to be accounted for but additionally, the exact note value too. A second issue is similar to a critique expressed earlier when discussing O'Maidín's model: the researcher applying the EMD to melodic similarity treats similarity as a phenomenon which is transposition invariant. As mentioned, not only that this conflicts with experimental data (Egmond & Povel & Maris ,1996 and Hofmann-Engl, 2003), but aligning melodies so as to find the best transposition in order to obtain minimal EDM is consuming computational time. However, possibly the most serious problem is concerned with the calculation of the distances; not only that pitch and time are treated as if they were of the same perceptual dimension (as they are added in order to obtain the Euclidean distance) although they are not, what units ought to be applied to the measurement of time and which one to the measurement of pitch is likely to turn into a question which cannot be answered. This is, for instance in our example above, had we changed the pitch units of 0.95 per semitone to 0.5 per semitone, some distances d_{ij} would have been affected (the ones where the pitch differences were unequal zero) while others would have remained the same. This would change not only the distances but ultimately the flow and which notes would have been deleted as well. Other questions such as whether to use a city block matrix, an Euclidean or other non linear matrices, or the issue of how to set the units of the weights, would require extensive investigation and experimentation. This not to say that the model might be useful in certain contexts but due to the fact that it contains systematical errors, its validity and reliability will remain limited.

2.3. Dynamic Programming

Another recent approach to similarity seems to have emerged from the biological sciences, where scientist endeavour to analyse the similarity between different DNA proteins (Goad & Kanehisa, 1982). An example might be given by the comparison of the first 11 amino acids of human Alpha haemoglobin to the first 11 amino acids of the human beta haemoglobin:

Alpha *Hb* human:

g s a q v k g h g k k ...

Beta *Hb* human:

g n p k v k a h g k k ...

where the letters *a, g, h, k, n, p, q, s, v* represent specific amino acids

Both sequences match at places 1, 5, 6, 8, 9, 10 and 11. Additionally, they show similar amino acids at place 2 and 4. Thus, the alpha is supposed to be similar to the beta haemoglobin. In order to measure the degree of similarity, dynamic programming is used. That is, one sequence (like the alpha sequence) is transformed into another sequence (like the beta sequence). Then the editing steps are counted (edit distance). The longer the edit distance the less similar the sequences. Three different edit operations are used: Insertion, deletion and substitution. In our example above, we might substitute *s* for *n* (2nd place), *a* for *p* (3rd place), *q* for *k* (4th place) and *g* for *a* (7th place) in the beta sequence in order to transform the beta sequence into the alpha sequence. Thus, we performed four edit operations (edit distance 4). We might have chosen to delete *s, a, q* and *g* in the alpha sequence and *n, p, k, and a* in the beta sequence, resulting in an overall 8 edit operation (edit distance 8). This shows that the edit distance depends on which edit operations we choose to perform. This also implies that we can at best obtain a minimal edit distance only through a trial and error procedure (hence the name dynamic programming). Generally, the similarity will be rated according to an algorithm of the form:

$$S = am - bi - cs \tag{10}$$

where *S* is the similarity rating, *a, b,* and *c* are weighting factors, *m* is the amount of matching places, *i* the amount indels (delete or addition) and *s* amount of substitutions.

The resemblance of this model with Tversky's model is striking. Equating the amount of

matches with the count of common features, we find that indels (amount of operations) corresponds to features present in one object but not the other, whereas substitutions are a cross between common features (both sequences have an item at the place of substitution) and features present in one but not the other (the items at the place of substitution differ). This means that the critique brought forth against Tversky, applies to this model.

2.3.1. Dynamic Programming in music

The dynamic programming approach has been utilized by a variety of researchers (e.g. McNab, Smith, Witten, Henderson & Cunningham, 1996 and Ning Hu, Dannenberg, Lewis, 2002). However, so it seems to the author, the most far reaching application was introduced by Mongeau & Sankoff (1990). In order to apply dynamic programming the authors regarded a melody as a sequence of tones t_1, t_2, \dots, t_n , where a tone is seen as possessing the two features pitch (p) and duration (d). Thus, we might compare a sequence given as $S_1 = t_{11}, t_{12}, \dots, t_{1n} = (p_{11}, d_{11}), (p_{12}, d_{12}) \dots (p_{1n}, d_{1n})$ with a sequence $S_2 = t_{21}, t_{22}, \dots, t_{2n} = (p_{21}, d_{21}), (p_{22}, d_{22}) \dots (p_{2n}, d_{2n})$. The authors then produced a matrix calculating the distance I_{ij} between each tone from S_1 with each tone from S_2 , where the distance I_{ij} between two tones t_{1i} and t_{2j} with $t_{1i} \in S_1$ and $t_{2j} \in S_2$ is given as: $I_{ij} = (|p_{1i} - p_{2j}| + |d_{1i} - d_{2j}|) / 2$. Hereby, the pitch p is measured in semitones and the durations as multiples of a basic beat (e.g., in case the basic beat is measured in semi-quavers a quaver receives the value 2, a crotchet the value 4 etc.). Thus, the authors produce a matrix of the following format:

$$\begin{array}{cccc}
 I_{11} & I_{12} & \dots & I_{1n} \\
 I_{21} & I_{22} & \dots & I_{2n} \\
 \dots & \dots & \dots & \dots \\
 I_{m1} & I_{m2} & \dots & I_{mn}
 \end{array}$$

Starting with a distance I_{i1} in the first column a sequence of distance is constructed and summated to an overall distance $D = I_{i1} + I_{j2} + \dots + I_{kp} + I_{l(p+1)} \dots + \dots I_{mn}$, with $i \leq j \leq k \leq p \leq l \leq p+1 \leq m \leq n$. The starting point I_{i1} and all possible combinations for the subsequent summands will be varied until a minimal value for D_{\min} is found. We will give an example:



Sequence 1

Sequence 2



Sequence 1 is: $S_1 = (d, 3/8), (b, 1/8), (c, 1/4)$ and sequence 2 is: $S_2 = (e, 1/4), (d, 1/4), (c, 1/4)$. We obtain the matrix comprising the following elements (with one semitone and one quaver fetching the value 1): $I_{11} = 1.5, I_{12} = 0.5, I_{13} = 1.5, I_{21} = 3, I_{22} = 2, I_{23} = 1, I_{31} = 2, I_{32} = 1, I_{33} = 0$, written in matrix form:

$$\begin{matrix} 1.5 & 0.5 & 1.5 \\ 3 & 2 & 1 \\ 2 & 1 & 0 \end{matrix}$$

As we can see, the minimal distance is given by $D_{\min} = I_{11} + I_{12} + I_{33} = 2$. Seemingly, this approach has little to do with dynamic programming except the need for variation. However, as we will see, there exists a strong link: Assuming, that a tone t_{1i} of S_1 has the same value as a tone t_{2j} of S_2 , we have a match. In case there is a difference between these two tones, one tone will have to be substituted where the weight of this edit operation will depend on the distance between these two tones (this is reminiscent of Shepard's model). In case S_1 contains one more tones than S_2 , dynamic programming requires that either a tone of S_1 will have to be deleted or another tone will have to be added to S_2 . However, this is not exactly what Sankoff and Kruskal do. If, for instance, we "delete" the last tone t_{1n} of S_1 , without any further deletion or addition, this will mean that tone t_{1n-1} will be edited to equal t_{2m} as well as tone t_{1n} ; the two last tones of S_1 will be mapped onto the last tone of S_2 . In our example above, we find that tone $t_{11} = (d, 3/8)$ of S_1 was mapped onto tone $t_{21} = (e, 1/4)$ as well as onto the first quaver duration of the tone $t_{22} = (d, 1/4)$ of S_2 , while the tone $t_{12} = (b, 1/8)$ of S_1 was mapped onto the second half of the tone $t_{22} = (d, 1/4)$. Tones t_{13} and t_{23} proved to be a match. Although this is, strictly speaking no deletion, it can be interpreted as such, where the distance between t_{1n} and t_{2m} will be interpreted as the weight of deletion.

No doubt, this is an interesting approach combining elements from other models (it is Tverskian in as much as dynamic programming is Tverskian and it is Shepardian in as much as its

weighting factors are determined). However, there are a number of serious problems with the model. The main issue was addressed by Smith, McNab & Witten (1998): The model produces a number of possible sequences of edit operation, all producing minimal edit distance. Further, some of these edit sequences might, in the words of Smith, McNab & Witten, “not make sense”. This implies, in case none of the edit sequences which make sense produce minimal edit distance, that the similarity rating is overrated. The fact that results will have to be evaluated on the basis of musical judgements decreases its applicability and value significantly. However, the main problem with this model appears to be its missing support through empirical data. For instance, the implementation which gives more weight to pitch than to duration ($I_{ij} = (|p_{1i} - p_{2j}| + |d_{1i} - d_{2j}|) / 2$), is entirely arbitrary, and yet similarity ratings will crucially depend on the adjustment of these weights. Moreover, just as the issue was raised when discussing the Earth Mover’s distance, mixing the pitch dimension with the time dimension seems extremely problematic. Clearly, the validity of this model is questionable.

2.4 Transition matrices

We finally consider transition matrices; a model which is based upon a concept firstly introduced by Fucks (1965). Fucks intended to find a measurement in order to describe historical musical development. He produced transition matrices for melodies and found that while for instance 17th century music displayed low entropy, 20th century music displayed high entropy. Hoos, Renz & Görg (2001) presented a melodic similarity model where transition probabilities are obtained for melodies, and where two melodies are rated similar if they produce the same transition matrices. We will give an example: The melody: *e, d, c, d, e, e, e, d, d, d, e, g, g, e, d, c, d, e, e, e, e, d, d, e, d, c* (Marry had a little lamb) produces the following transition matrix:

	c	d	e	f	g
c	0.33	0.67	0	0	0
d	0.3	0.4	0.4	0	0
e	0.1	0.35	0.45	0	0.1
f	0	0	0	0	0
g	0	0	0.5	0	0.5

Table 1: Transition matrix for the song *Marry had a little lamb*.

If we now changed, let us say the first *d* to a *c*, the transition matrices would largely remain the same.

	c	d	e	f	g
c	0	1	0	0	0
d	0.3	0.4	0.4	0	0
e	0	0.45	0.45	0	0.1
f	0	0	0	0	0
g	0	0	0.5	0	0.5

Table 2: Transition matrix for the song *Mary had a little lamb* with the note *d* changed to *c*.

Although it might appear that this model captures changes sufficiently, we find that this is not the case, for several reasons: The main problem might be that melodies which are different can produce the same matrices (e.g., *c, d, d, e, d, e, c, e* and *c, d, d, e, c, e, d, e*). This means that ratings according to this model will produce erratic material because it is based on a misconception. Further, we find that changing a tone within the melody will effect changes in the transition matrix in three places, while changing the last note will effect two changes only. Considering the recency/primacy effect, we would expect the last note to be of greater importance rather than smaller importance. Thus, the model seems to disregard cognitive principles. Finally, it is entirely unclear how to rate changes within the transition matrix, as there are no empirical data available to determine the values of possible parameters. The author concludes that this model has no future.

3 Critiques on the conception of similarity

Summarizing the features of the models discussed above, we find one particular critique recurring throughout: It seems none of the authors writing on melodic similarity are considering that an appropriate model, should such a model be available, will at least have to include some empirical constants. Instead, these authors seem to imply that we already know the relevant features and the empirical parameters of melodic similarity which they then input into a model. Although the author is aware that there are some empirical studies available (e.g., Cuddy, Cohen & Miller

1979; Dowling & Harwood, 1986; van Egmond & Povel & Maris 1996; Francès, 1988 Gabriellson, 1973; White, 1960; Hofmann-Engl, 2003) there is still not enough empirical information available in order to identify the relevant features. Thus, even if the models did not show deficiencies and inconsistencies, they still would remain purely speculative. For instance, none of the models incorporated dynamics, although it seems extremely unlikely that dynamic variations will not influence similarity judgements. In fact, the difficulties with all these models are so substantial that we might ask the question whether a model of melodic similarity is attainable or at least desirable.

This question, whether similarity measures are attainable and desirable, was asked by Clarke & Dibben (1997), who posed the question: “Does it really make sense to ask whether musical event X is more similar to Y than Z. Is this a judgement anyone often (or ever) makes?”. Their argument, so they claim, is supported by the fact that nobody so far - referring to Nattiez (who, in the opinion of these authors, should have been able by now to deliver a more formalized method of identifying and classifying motivic material) - has yet been able to deliver an operational model. Although the exposition of several existing models above renders such a claim an over-generalization, the non-existence of a tool does neither mean that such a tool is unattainable nor that the development of such a tool is not desirable. However, Clarke and Dibben are right to bring to our attention the question of what we actually are seeking. True, if these authors are correct with their opinion that similarity judgements are hardly ever made, then there is no need for the development of such a model, indeed. However, this seems not to be the case. The author will give eight examples which will involve some form of similarity judgement: (a) The comparison of different interpretations of a specific composition, (b) a student trying to reproduce just the sound the teacher produces on her/his instrument, (c) a musician working out a specific interpretation of a composition for performance based on melodic comparisons, (d) an analyst performing a motivic analysis, (e) an ethnomusicologist tracing the origin of melodic material, (d) a judge deciding whether a copyright infringement suit over a motive should be granted or not, (e) a composer producing a variation of a theme, and most importantly perhaps (f) the classification and retrieval of melodic material in a data base (MIR). True, a composer might produce a variation according to some abstract algorithm without considering cognitive implications, the judge just wants to find the liar and similarity is one means to this end, the student is probably not even aware of any similarity judgement, while the



ethnomusicologist is or should be interested in cognitive processes. Admittedly, the strategy and result of similarity judgements might depend on context, but the author hopes that the examples given are sufficient to disprove Clarke's & Dibben's claim as unsubstantiated. Surely, a model is desirable, but whether it is attainable and if so in what shape and form seems to be the question. We will commence with the investigation of the first part of the question of whether a model might be attainable.

A major issue raised in the context of similarity is the issue of categorization. A seemingly popular theory (e.g., Posner & Keele, 1968; Reed, 1972; Rosch & Mervis, 1975) understands the relationship between similarity and categorization in the following manner: an object *a* is more likely to be classified as belonging to the category *A* than belonging to the category *B*, if object *a* is more similar to all the objects in category *A* than it is to all the objects in category *B*. This link between categorization and similarity can be expected to hold true in a reversed relationship: Once two objects are assigned to two different categories, they will also be seen as less similar than two objects from the same category, even if they should share more relevant common features to a stronger degree. However, Goodman (1972) remarks that this approach implies a philosophical weakness. He argues that, for instance, assigning the letter *A* to the category of *As*, because of its similarity to this category, requires the existence of the category of *As* and thus similarity does not explain categorization. It seems that Goodman is referring to existing categories, and no doubt similarity is insufficient in explaining the assignment of elements to existing categories. However, the question whether similarity is a factor in the ontogenesis of categories is unaffected by his argument. Moreover, the argument formulated above will also imply that an appropriate similarity model would ideally consider categorization. Without going into a lengthily discussion, it seems that existing approaches to melodic categorization have been found unsatisfactory. For instance, Adams (1976) suggested to classify material according to contour features, but his approach was subsequently refuted by Marvin & Laprade (1987). Much of Nattiez's (1975) analytical technique is based on melodic comparison and melodic classification. However, Clarke & Dibben (1997) expressed their concern that Nattiez has failed to bring his method into a cohesive system. Lerdahl's and Jackendoff's hierarchical structuring (1983) produces a segmentation of melodic material, which implies categorization (as utilized by Cambouropoulos, 1998). However, their methods have

been so widely criticized (e.g., Rosner, 1984; Clarke, 1986; Cross, 1998) that even a summary of these critiques exceeds the framework of this article. It seems apparent that there does not exist a sound understanding of melodic categorization. Hence, a model of melodic similarity cannot be built upon a theory of melodic categorization. However, building a melodic similarity model (admitted of limited validity), we might enhance the investigation into issues of categorization. For instance, assuming we developed a reliable model which is operational in various contexts and assuming further that we then find that the predictions of the model do not coincide with empirical data in a new context, we might be able to explain this deviation as a result of categorization. Such new knowledge itself then would lead to a modification of a similarity model and so on. Thus, the author concludes that the absence of a theory of melodic categorization at the present time increases the need to develop a melodic similarity model independent of categorization. However, in order to establish which features this model should incorporate, we will consider a second critique by Goodman.

We considered earlier Goodman's critique - in the context of Tversky's similarity model - where he stated that a comparison of two items requires a frame of reference (i.e., similar according to a specific measure). It seems that researchers consider six main factors which influence or determine the frame of reference. These factors are: context, culture/language, expertise, age and experimental method.

There seems to be strong empirical evidence, that similarity judgements are context dependent (Goldstone, Medin & Halberstadt, 1997), and we referred earlier to Barsalou's (1982) experiment, who found that raccoon and snake are less similar if no context is given than when compared in the context of pets. Barsalou (1983) also showed that seemingly highly dissimilar objects receive a high similarity rating when put into specific context (e.g., jewellery and children in context of "things to retrieve from a burning house"). However, this argument seems to indicate that context changes the absolute scale but not necessarily the relative scale. To give an example 50 cents is half of \$1, which can seem a lot more. However, if we see this in the context of \$1,000,000 there seems to be not much of a difference between 50 cents and \$1. Nevertheless, 50 cents is still just half of \$1. Thus, it seems we are dealing here with a measurement issue (influence of

experimental method) rather than an issue concerning context. There is however, a second way of how to interpret Barsalou's findings. While our example above involves a one-dimensional quantity comparison, the comparison of children and jewellery involves features which are far less well defined. In fact, it seems that an almost infinite amount of quantitative and qualitative features can be assigned to both children and jewellery, so that context is desperately needed in order to select the relevant features. If no context is given, we might speculate that participants of an experiment will create their own context. Still, how relevant this observation are in the context of melodic similarity remains an unsolved issue. Comparing two melodies is supposed to be a cognitive, perceptual and almost automatic process, while the comparison of children and jewellery is a judgement of conceptual similarity, and as mentioned above, cognitive similarity seems to be less dependent on context than conceptual similarity. Finally, the comparison of two melodies seems to involve a limited number of features (such as pitch, dynamics, tone-colour and rhythm), hence we would expect that context will be far less important in the selection of the relevant features. Still, we might argue that it seems unlikely that a transformation of a transitional passage of a composition will be as significant as the transformation of the main theme of a composition. Similarly, we might expect that the change of a specific rhythm such as  to be more significant if the same rhythm surrounded by the same rhythm such as . However, it seems that, at this point, a contextual melodic similarity model is unattainable. True, that this will put constraints onto the model, but a context free model will at least produce some testable hypotheses. Still, even if we now developed a context free model (disregarding features such as harmonic implications), Goodman's argument holds true in as much as a frame of reference will be required. Thus, a model will have to be developed in such a way that there is scope for adaptability, not just in form of empirical constants but on a more fundamental level.

In a classic text by Whorf (1941), we find as Goldstone (1994) reports, a rather intriguing Wittgensteinian offshoot on language and similarity. During his studies of the Shawnee Native American language, Whorf seemed to have confirmed that language and culture are strongly interlinked, not just the vocabulary but even the syntactic organisation of language (for instance the temporal structure). Consequently, we would also expect that such interdependency of language and culture will affect similarity judgements. Indeed, Whorf gives us an example. He argues that for a

Shawnee Native American the two sentences: "I pull the branch aside" and: "I have an extra toe on my foot" are highly similar sentences. This is, more literally translated, the first sentence takes the form: "I pull it (something like the branch of a tree) more open or apart where it forks", while the second sentence becomes, "I have an extra toe forking out like a branch from the normal toe". However interesting this example might be, it does not demonstrate as Goldstone (1994) seems to imply, that syntactic similarities will induce semantic similarities which then evoke cognitive similarities. However, we might argue in a Shepardian fashion, that cultural difference will lead to difference in categorization, which will then somehow be reflected in the user's language and hence similarity judgements and language are expected, if not to be causally related, so still to be correlated. Consequently, we would want a melodic similarity model to be sensitive towards culture and possibly sensitive towards language. However, the development of such a model would appear as rather overambitious at this point. This is not to say that a more abstract model will be of no value in a cross-cultural setting. Should, for instance, such a model lead to predictions which will deviate from measured data in a given culture, we might be able to generate a better understanding of this given culture in reference to these deviations.

It has been argued that similarity judgements are dependent on expertise. For instance, Suzuki, Ohnishi & Shigemasa (1992) found that experts, when asked to compare various stages of the Hanoi puzzle with the completed puzzle, judged similarity by assessing how many steps were needed to complete the task, while novices rated similarity according to shared features between the various stages and the completed puzzle. However, this result seems to suggest that experts and novices interpreted the question of how similar two items are in a different fashion. While experts understood the question as, "How many steps are needed to complete the puzzle", novices interpreted the question as, "How similar are the looks of two visual images." From this point of view, we might argue that experts interpreted the question differently by rating similarity to how much work they would have to do. Thus, it appears to the author that this experiment cannot support the hypothesis that there is a significant difference in similarity judgements depending on expertise. However, it is such a common feature in music psychological experiments (compare Deutsch, 1999) to generate expertise dependent data, that we might suspect that the same will hold true for melodic similarity. Further, we might expect that musical experts will deliver more accurate

similarity judgements, meaning data with smaller variance, than musical novices do. Thus, the model will have to include some empirical constant which can be adjusted according to the level of expertise.

It also has been observed that similarity is age dependent (e.g., Shepp & Schwartz, 1976, Smith & Kemler, 1977, Smith 1989a). It seems to be the general point of view that young children judge similarity according to an overall similarity, while older children may choose a specific dimension in order to compare objects in reference to this one specific dimension ignoring other dimensions. Typically, a sample of preschool children is compared with a sample of school children. Similarity is rated on shapes which differ in size and color. While preschool children rate similarity along both dimensions (size and color), school children tend to regard two items as similar if they are identical along one dimension (e.g., size) without any consideration for the second dimension (e.g., color). Further, when asked, young children find it difficult to identify in what respect two objects are similar. This age dependency has been challenged by Smith (1989b) who found that not only preschool children use overall-similarity judgments but adults as well when the stimuli are more complex (varied over more than two dimensions). This is what Medin & Ortony (1989) seem to refer to as heuristic similarity. Putting similarity into an evolutionary context, they propose that information (such as 'this is a lion') will be analyzed according to overall similarity in order to derive competent decisions (such as running away, instead of petting). Quite clearly this also implies a certain amount of context independence. It seems a strong argument and has been juxtaposed by Goldstone (1994), who pointed out that we will behave cautiously if confronted with a snake which resembles a rattlesnake and the fact that snake rimes with snowflake will be of no significance. He further points out that where context dependency of similarity measures occur, it seems to be systematic rather than random. Thus, they can be integrated in a wider model.

4. A conceptual framework of melodic similarity

Summarizing, we conclude that a melodic similarity model is desirable, has to be context independent, has to identify the relevant features, has to include empirical constants and has to

incorporate some conceptual flexibility. The author further hopes to have demonstrated that all the reviewed models above are characterized by a series of deficiencies in one form or another. Moreover, all models do not provide conceptual flexibility. This inflexibility seems to stem from a distinct absence of a theoretical framework. The author feels, that asking the question, what are the relevant features and how these features are compared when rating the similarity between two melodies is far more promising.

This is indeed what Hofmann-Engl (2001, 2002, 2003, 2004) undertook. He identified melota (correlated to pitch), chronota (correlated to duration) and dynama (correlated to loudness) as the relevant parameters. He then introduced a representation of melodies in form of atomic beats. Note, that this representation was firstly used by Gustafson (1987), although he used a different terminology and applied this representation rhythms only. We will give an example referring to the musical example above (with added dynamics). The first sequence was: (*d*, dotted crotchet, forte), (*b*, quaver, mezzo forte), (*e*, crotchet, piano) and the second sequence was: (*e*, crotchet, piano), (*d*, crotchet, mezzo forte), (*c*, crotchet, forte). The smallest beat is a quaver (= 1/8), hence we rewrite the sequences as:

$$Ch_1 = [(d, 3, f), (d, 3, f), (d, 3, f), (b, 1, mf), (e, 2, p), (e, 2, p)](1/8) \text{ and}$$

$$Ch_2 = [(e, 2, p), (e, 2, p), (d, 2, mf), (d, 2, mf), (c, 2, f), (c, 2, f)](1/8)$$

This is to be read: Both chains (sequences) consist of 6 atomic beats (unit is 1/8). On the first atomic beat of ch_1 we find the meloton (pitch) *d*, the chronoton (duration) $3 * 1/8$ which makes a dotted crotchet (3/8) and the dynamon (loudness) *f* for forte. The second atomic beat contains the same values so does the third. The fourth atomic beat is *b*, $1 * 1/8$ (which is a quaver) and *mf* for mezzo forte. The other atomic beats have to be read in similar fashion.

Now, rather than inputting all three parameters straight into one model, Hofmann-Engl treats all three dimension separately (cross-interferences will need further investigation before they can be integrated into a melodic similarity model. More on this subject can be found below). Thus, we obtain the following graph for the melotonic similarity:

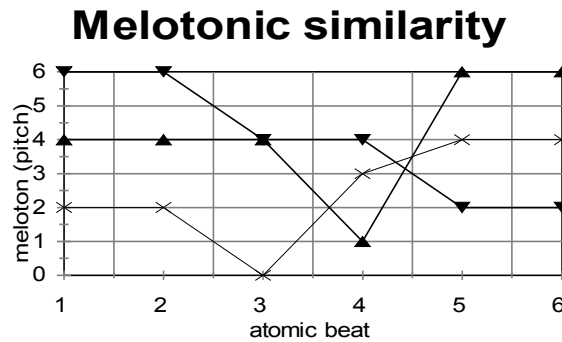


Figure 3: Melotonic (pitch) similarity between the two melodies: d, b, e and e, d, c . The dotted line represent the similarity: the less straight the line is and the further away from the x-axis the smaller the similarity. The values are $b = 1, c = 2, \dots$

Chain ch_1 is represented by the line with the standing triangles (d fetches the value 4, b the value 1 etc.), the second chain ch_2 is represented by the line with the upside down triangles and the third line (crosses) represents the difference between the two chains (first atomic beat is: $6 - 4$ (d against e), second beat: $6 - 4$, third beat: $4 - 4$ etc.). We will call this line the similarity line. Hofmann-Engl's claim is that the similarity line is the basis for the calculation of melotonic (pitch) similarity: Firstly, the melotonic difference of both chains is the larger the further this line is displaced from the x-axis and secondly, the interval difference of both chains is larger the less straight the line is. Both, the melotonic difference and the interval difference have been identified as melotonic similarity factors within two experiments (Hofmann-Engl, 2003). Thus, this approach has empirical support. Further, Hofmann-Engl maintained that not one single model will be correct, but depending on length and features such as stream segregation different models will have to applied. He also maintained that the differences should not be computed in al linear fashion but using the exponential function , $e^{-k\Delta}$ where k is an empirical constance and Delta the differences either between the relevant melota or the relevant intervals. The model also covers the situation when two melodies are of different length as well as glissandi. However, one simple form which can be considered to be approximative is the following:

$$S = \frac{\sum_{i=1}^n e^{-k*(f_1(x_i)-f_2(x_i))^2}}{n} \quad (11)$$

where S is the similarity, k an empirical constant, n the number of tones, $f_1(x_i)$ the melotonic chain ch_1 and $f_2(x_i)$ the melotonic chain ch_2 .

For more detail compare Hofmann-Engl (2001, 2003). However, in order to offer the reader a small insight, the author presents Hofmann-Engl's algorithm for glissandi both lasting the same length τ covering the aspect of interval difference similarity:

$$S_{interval} = \sqrt{\frac{\int_0^{\tau} (e^{-\frac{k_2(t)}{\tau^{c_2}} (\frac{dm(t)}{dt} - \frac{dm'(t)}{dt})^2)} dt}{\tau}} \quad (12)$$

where $S_{interval}$ is the melotonic interval difference similarity, $k_2(t)$ and c_2 empirical constants, t the time (length of the glissandi), $m(t)$ the melotonic glissando 1 and $m'(t)$ the melotonic glissando 2 and τ the length of the glissandi. Note. The time dependency of k_2 as due to the primacy/recency effect we can expect that the similarity rating of the beginning and ending of two glissandi will be more important than the parts in the middle.

Now, the chronotonic similarity is approached in a similar fashion.

Chronotonic Similarity

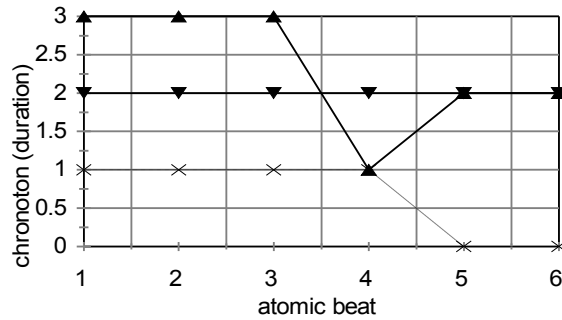


Figure 4: Chronotonic (duration) similarity between the two rhythms: dotted crotchet, quaver, crotchet and crotchet, crotchet, crotchet. The dotted line represent the similarity: the further away from the x-axis the smaller the similarity. The values are quaver = 1, crotchet = 2 ...

The line with upright triangles represents the first chain and the line with the upside down triangles the second chain. The dotted line is the chronotonic similarity line. However, now we find that the distance of the similarity line to the x-axis is the only similarity factor. This is confirmed by one experiment conducted by Hofmann-Engl (2003). Still, just as before, he maintains that not one single model will be sufficient but that it will be necessary to consider several models based upon the similarity line and an exponential functions depending on factors such as length and stream segregation.

The dynamic similarity line is illustrated in the figure below.

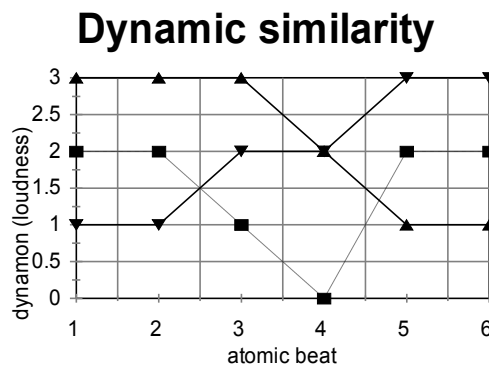


Figure 5: Dynamic (loudness) similarity between the two melodies: f, mf, p and p, mf, f . The dotted line represent the similarity: the less straight the line is and the further away from the x-axis the smaller the similarity. The values are: $p = 1, mf = 2 \dots$

Again, we obtain a line for the first and second chain and the similarity line which is marked by the crosses. Hofmann-Engl speculated that dynamic similarity will be correlated to the distance of this line from the x-axis and to the curviness of it (in analogy to melotonic similarity). However, this hypothesis has not been tested in an experiment.

The overall melodic similarity will be calculated as:

$$S = \mu S_m + \chi S_c + \delta S_d \quad (13)$$

where S is the overall similarity, μ , χ and δ are empirical constants, S_m is the melotonic similarity, S_c is the chronotonic similarity and S_d is the dynamic similarity.

Note, that all three dimensions are treated independently. As the literature indicates (e.g. Tekman, 1997), the three dimensions are somewhat interrelated and hence we might expect that formula (13) might have to be modified. However, there are no clear data available at present on how such a modification would have to be done and hence it would appear overambitious to do so at this point in time. Hofmann-Engl (2003) conducted three experiments which are in support of his approach. Two of these experiments were conducted in the context of melotonic similarity and one in the context of chronotonic similarity. Still, formula (13) remains speculative and so are Hofmann-Engl's assertions about the issue of dynamic similarity. Moreover, as pointed out by Toussaint (2004), there exists a large family of possible distance measure functions and the one proposed by Hofmann-Engl is just one amongst others, that it is to be seen which of those functions will deliver the best correlation between measured and predicted data. Hofmann-Engl (2003) maintained that mathematical and music theoretical aspects two should be considered. Following Toussaint's approach, we list the Kullback-Liebler divergence for discrete function given as:

$$KL = \sum_{i=1}^n f_1(x_i) \log \frac{f_1(x_i)}{f_2(x_i)} \quad (14)$$

where KL is the Kullback-Lieber distance, n the length of the chains, $f_1(x_i)$ the value of the chain ch_1 at the place i and $f_2(x_i)$ the value of the chain ch_2 at the place i .

and the Kolmogorov variational distance given as:

$$K = \sum_{i=1}^n |f_1(x_i) - f_2(x_i)| \quad (15)$$

where K the Kolmogorov variational distance is the distance, n the length of the chains, $f_1(x_i)$ the value of the chain ch_1 at the place i and $f_2(x_i)$ the value of the chain ch_2 at the place i .

Toussaint (2004) concludes that, in the context of rhythmic similarity measures, the chronotonic distance, as described above and computed along the formulae 14 and 15, is the overall best rhythmic similarity measure when compared with other rhythmic similarity measures. This adds strong support to Hofmann-Engl's conceptual framework. Finally, as found by Müllensiefen & Frieler (2004) melotonic interval similarity measures are the single best predictors.

Thus, although more experiments will have to be conducted in order to establish the reliability of Hofmann-Engl's conceptual framework, it appears that there exists already some strong support for this approach.

5. Conclusion

This paper set out to compare different approaches to melodic similarity classifying these into four groups: (a) Contrast models, (b) distance models, (c) dynamic programming and (d) transition matrices. There, it became clear that melodic similarity models belonging to class (a), (c) and (d) are weak in several aspects. Further, we found that existing melodic similarity models

belonging to class (b) too displayed deficiencies. The author then considered the existing literature on similarity from a psychological angle demonstrating that a great many issues are involved such as age and expertise dependencies as well as questions about categorization. In a final section the author introduced the reader to an approach to melodic similarity which belongs to class (b). Rather than presenting simply yet another algorithm, the author illustrated through an example the conceptual framework of melodic similarity in form of similarity lines. The advantage of this approach is firstly that it does not contain systematic errors, secondly that it does not conflict with existing empirical data and secondly and thirdly that it gives scope for implementation in a number of ways. The exact nature of this implementation is a matter of discussion at the forefront of the current research, but – so the author's claim – the exact mathematical form of a model will always depend on the context of the application.

Literature

Adams, C. (1976). Melodic Contour Typology. *Ethnomusicology*, vol 20.2

Barsalou, L. W. (1983). Ad hoc categories. *Memory and Cognition*, vol 11, 211-227

Barsalou, L.W. (1982). Context-independent and context-dependent information in concepts. *Memory & Cognition*, vol 10, 82-93

Berenzweig A., A., Logan, B., Ellis, D. & Whitman, B. (2003). A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures. In *Proceedings of ISMIR 2003*, Washington, D. C.

Bradshaw, J. (1997). Introduction to Tversky similarity measure. In *Proceedings of MUG'97*, Herts, UK

Clarke, E. F. & Dibben, N. (1997). An Ecological Approach to Similarity and Categorisation in Music. In *Proceedings of Simcat 97*, Edinburgh, 37-41

Clarke, E. F. (1986). Theory, Analysis and the psychology of Music: A Critical evaluation of Lerdahl, F. and Jackendoff, R., *A Generative Theory of Tonal Music*. *Psychology of Music* vol 14.1, 3-17

Cohen, S. (1999). Computing the Earth Mover's distance under Transformations. Online publication:

- Cross, I. (1998). Music Analysis and Music Perception. *Music Analysis*, vol 17. No.1
- Hu N., Dannenberg R., Lewis, A. L. (2002). A Probabilistic Model of Melodic Similarity. In *Proceedings of ICMC 2002*
- Doraisamy, S. & Rüger, S. (2002). A comparative and Fault-tolerance Study of the Use of N-grams with Polyphonic Music. In *Proceedings of ISMIR 2002, Paris*, 101-106
- Fucks, W. & Lauter, W. (1965). *Exaktwissenschaftliche Musikanalyse*. Westdeutscher Verlag, Köln
- Goad, W. B. & Kanehsia, M. I. (1982). Pattern Recognition in Nucleic Acid Sequences. *Nucleic Acids Research* vol 10.1, 247-263
- Goldstone, R. L, Medin, D. L. & Halberstadt J. (1997). Similarity in Context. *Memory & Cognition*, vol 25.2, 237 -255
- Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition* 52: 125–157
- Goodman, N. (1972). Seven strictures on Similarity. In: *Problems and Projects* (ed. Goodman), The Bobbs-Merrill Co, New York
- Grachten M., Arcos J. L., Mántaras de R. L. (2002). A comparison of two approaches to melodic similarity. In *Proceedings of ICMAI'02*, Edinburgh
- Gustafson, K. (1988) The graphical representation of rhythm. In (PROPH) *Progress Reports from Oxford Phonetics*, vol. 3, 6–26, Oxford
- Hofmann-Engl, L. J. (2004). Melodic similarity – providing a cognitive groundwork. Online publication. Chameleon Group of composers:
www.chameleongroup.org.uk/research/cognitive_similarity.pdf
- Hofmann-Engl, L. J. (2003). Melodic similarity - a theoretical and empirical approach. PhD thesis, Keele University (available online at: www.chameleongroup.org.uk/research/thesis.html)
- Hofmann-Engl, L. J. (2002). Melodic Similarity - a conceptual framework. In *Proceedings of Understanding and creating Music, 2nd International Conference, Naples 2002* (available online at: <http://www.chameleongroup.org.uk/research/A36.pdf>)
- Hofmann-Engl, L. J. (2001). Towards a model of melodic similarity. In *Proceedings of ISMIR 2001, Bloomington, Indiana* (available online at: <http://ismir2001.ismir.net/pdf/hofmann-engl.pdf>)
- Hoos, H., Renz K., Görg, M.(2001). Guido/Mir - an experimental Musical Information Retrieval System based on GUIDO Music Notation. In: *Proceedings of ISMIR 2001, Bloomington, Indiana*
- Kluge, R. (1996). Ähnlichkeitskriterien für Melodieanalyse. *Systematische Musikwissenschaft*,

4.1-2, 91-99

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press

Ó Muidín, D., (1998). A Geometrical Algorithm for Melodic Difference. In: *Melodic Similarity - Concepts, Procedures, and Applications* (ed. Hewlett & Selfridge-Field), MIT Press

Markman, A. B. & Gentner, D., (1997). The Effects of Alignability on Memory. *Psychological Science* vol 8.4, 363-367

Markman, A. B. & Gentner, D. (1996). Commonalities and differences in similarity comparisons. *Memory and Cognition*, vol 24.2, 235-249

Markman, A. B. & Gentner D. (1993). Structural Alignment during Similarity Comparisons. *Cognitive Psychology* 25, 431 - 467

Marvin, W. & Laprade, P. A. (1987). Relating Musical Contours: Extensions of a Theory for Contour. *Journal of Music Theory*, vol. 31.2, 225-267

Medin, D.L., Goldstone, R.L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, vol 100, 254-278

Medin, D. L., and A. Ortony. (1989). Psychological essentialism. In: S. Vosniadou and A. Ortony (ed.), *Similarity and Analogical Reasoning*. Cambridge: Cambridge University Press, 179-195

Nattiez, J.-J. (1975). *Fondements d'une sémiologie de la musique*. Paris

McNab, R. J., Smith, L. A., Witten, I. H., Henderson, C. L., & Cunningham, S. J. (1996). Towards the digital music library: Tune retrieval from acoustic input. In *Proceedings of Digital Libraries '96*. ACM.

Müllensiefen, D. & Frieler, K. (2004a). Measuring melodic similarity: Human vs. algorithmic judgments. In: *Proceedings of CIM04*, Graz

Müllensiefen, D. & Frieler, K. (2004b). *Music Query: Methods, Models, and User Studies* In: *Computing in Musicology* 13, 147-176, MIT Press

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology* vol 23:94-140

Posner, M. I. & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, vol 77, 353-363

Rosch, E. & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. *Cognitive Psychology*, vol. 7, 573-605

- Rosner, B. (1984). Review of F. Lerdahl and R. Jackendoff, *A Generative Theory of Tonal Music*. *Music Perception*, vol 2.2., 275-290
- Shepard, R. N. (1987). Toward a universal law of generalization of psychological science. *Science*, vol 237, 1317-1323
- Shepp, B. E. & Swartz, K. B. (1976). Selective attention and the processing of integral and non-integral dimensions: A developmental study. *Journal of experimental child psychology*, vol 22, 73-85
- Smith, L. A, McNab, R. J, Witten, I. H. (1998). Sequence-Based Melodic Comparison: A Dynamic-Programming Approach. In *Melodic Similarity - Concepts, Procedures, and Applications* (ed. Hewlett & Selfridge-Field), MIT Press, 101-108
- Smith, L. (1989a). A model of perceptual classification in children and adults. *Psychological Review*, vol 96.1, 125-144
- Smith, L. (1989b). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosnuadou & A. Orthonoy (Eds.), *Similarity and analogical reasoning* (pp. 146-178). New York: Cambridge University Press.
- Smith, L.B. & Kemler, D.G. (1977). Developmental trends in free classification: Evidence for a new conceptualization of perceptual development. *Journal of Experimental Child Psychology*, vol 24, 279-298
- Suzuki, H. & Ohnishi, H. & Shigemasu K. (1992). Goal-Directed Processes in Similarity Judgment. Online: <http://www.nime.ac.jp/~ohnishi/EduPsy/CogSci9x/cogsci92.html>, Tokyo Institute of Technology
- Pienimäki, A. (2002). Indexing Music Databases Using Automatic Extraction of Frequent Phrases. In *Proceedings of ISMIR 2002*, Paris, 25-30
- Toivainen, P. & Eerola, T. (2002). A computational model of melodic similarity based on multiple representations and self-organizing. In *Proceedings of ICMPC 7*, Sydney, 236-239
- toussaint, G. (2004). A comparison of rhythmic similarity measures. In: *Proceedings of ISMIR 2004*, Barcelona, Spain
- Tversky, A. (1977). Features of Similarity. *Psychological Review*, vol 84, 327-352
- Typke R., Giannopoulos P., Veltkamp R. C., Wiering F., Oostrum van R. (2003). Using Transportation Distances for Measuring Melodic Similarity. In *Proceedings of ISMIR 2003*,

Washington, D.C.

Uitdenbogerd, A. & Zobel, J., (1998). Melodic Matching Techniques for Large Musical Databases.

In: Proceedings. ACM International Multimedia Conference, Orlando, Florida

Whorf, B. L. (1941). Languages and logic, In: Language, Thought and Reality: Selected papers by

Benjamin Lee Whorf (ed. Carroll). MIT Press (1956), Cambridge, MA, 233-245