(Ludger Hofmann-Engl, chapter 2 of PhD thesis, Department of Psychology, Keele University, 2002)

Melodic similarity models: Background, Critique and Objectives

It is a rather difficult issue trying to portrait 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, interesting enough, it seems possible to group all existing approaches into four classes. Two of these classes are rooted in the cognitive science. These are the models based on 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 context, such as bio-informatics, chemistry and music and multimedia information retrieval. Most recently, transition matrices have been employed to measure melodic similarity by Hoos, Renz & Görg (2001). The author decided it would be most beneficial for the understanding of the text if we discussed each of the different approaches in general (except transition matrices as it appears that this is an approach used in music information retrieval only), and then discuss some critical remarks and the implementation of the approaches in music. The final section of this chapter will discuss some general critiques and an outlook of what a similarity model ought to deliver.

2.1. Contrast models

Central to Tversky 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:

$$S = \vartheta c + \alpha a + \beta b \qquad \text{formula 2.1.}$$

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 one 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, 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 color, than 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 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 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. However, Medin, Goldstone & Gennter (1993) demonstrated in an experiment that cognitive similarity is likely to be affected by context. These researchers found that when an object which has ambiguous features (a drawing which can be interpreted as either a 3 dimensional or a two dimensional representation) is compared with an unambiguous object (a drawing with can only be interpreted as 2 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 difficulty to Tversky model than he seemed to confess.

This deficiency is heightened by Tversky own discovery of asymmetrical similarity judgments. This is, an object A compared with an object B will not necessary 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 judgments 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 is an object with low frequency (e.g., an object we see rarely). 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 not Tel Aviv). 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 A consisting of all relevant features for a specific comparison task comparing object A and B, where we set A to be a function of the comparison order with A (A,B) ≠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 in conceptual similarity tasks and not in cognitive similarity tasks. Interpreting Tversky 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 judgment tasks: The triangle inequality, until Tversky 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*. Quite clearly, this law does not hold when we consider the following example: Given shall be 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 color, *A* and *B* 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) as possibly the most interesting one. They proposed a structure-based model. Here, feature commonalties and feature differences are replaced by alignable commonalties, 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 motor cycle, 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 motor bicycle). None-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 motor bicycle). The authors have been able to produce some evidence that alignable differences influence similarity judgments 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 isomorphism. Mathematically speaking, two objects *A* and *B* are isomorphic, if all positions of *A* can be mapped onto a corresponding unequivocal 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 motor bicycle do not fulfill this criterion, or generally speaking, isomorphism is not a requirement for us to consider two objects *A* and 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 ... A_n and A'_1, A'_2 ... A'_n while *B* might be segmented into B_1, B_2 ... B_n and B'_1, B'_2 .. 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 ... A'_n and B'_1, B'_2 ... 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 is which metal bar on the bicycle should be mapped onto which metal bar on the motor bicycle. 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 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:

$$d(x,y) = \sum w_{x_i} w_{y_i} (1 - \delta_{x_i y_i}) \qquad \text{formula 2.2.}$$

and

$$S_h(x,y) = 1 \text{ if } d(x,y) \le h \text{ and } S_h(x,y) = 0 \text{ if } d(x,y) > h \qquad \text{formula 2.3.}$$

where $d(x,y)$ is the distance between the entities (motives) x and y, w_{xi} and w_{yi} are weighting factors (which are under-defined in Cambouropoulos's work), xi and yi the ith feature of the entities x and y respectively, ν the number of features, δ_{xiyi} the Kronecker delta (with $\delta_{xiyi}=1$ for $x_i=y_i$ and $\delta_{xiyi}=0$ 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 this thesis intends to demonstrate, these features alone are insufficient to describe melodic similarity (in fact, as we will see contour is not a predictor at all). Moreover, this 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 appears to be an inferior version. Clearly, a model of similarity will have to refer to pitch (or some correlate of pitch), duration and dynamics. Still, a model could for instance 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. Finally, the findings by Egmond, Povel, & Maris (1996) are in contrast to this model. These researchers found that similarity judgments decreased with increasing transposition interval. According to the above described model, however, we find that all transpositions are treated as equivalent.

2.2. The distance model

The second approach towards a similarity measure 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 categories (for instance: weight, color, 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), 3 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 kg - 4 kg = 1 kg, in color 660 nm - 460 nm = 200 nm and in frequency 2000 Hz - 200 Hz = 1800 Hz. Thus the similarity will be a function of the mean difference. Although the author used a physical example for the purpose of clearness, Shepard constructs for the similarity measure an "abstract psychological space". 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 similarity *S* will be proportional to *I*:

$$S \propto I \qquad \text{formula 2.4.}$$

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

Shepard's model is usually written in the form:

$$d(x,y) = \left(\sum_{k=1}^{D} |x_k - y_k|^p \right)^{1/p} \qquad \text{formula 2.5.}$$

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 similarity to 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 model is not only supported by this study, but by several studies conducted by Shepard (for instance that the length of time it takes participants to make same/different judgments about pairs of shapes, one in standard position and the other rotated, is proportional to the degree of rotation). However, it seems there are two major problems with Shepard's model. 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 in 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 is dependent on the comparison order, possibly in the form $w_{k_{xy}}$ for comparison of the object *x* with object *y* and w_{kyx} for comparison of the object y with the object x. Note, this is admitted an unpleasantly as it increases the complexity of the model substantially, at least Shepard's model is able to incorporate asymmetries directly in contrast to Tversky's model. Still, if we understand both models as cognitive and not as conceptual similarity models such weighting might not be necessary.

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 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:

$$d(x,y) = \frac{\sum_{k=1}^{n} |x_k - y_k|}{n} \qquad \text{formula 2.6.}$$

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 kth 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 analysis 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:

$$\text{difference} = \sum_{k=1}^{n} \left| p_{1k} - p_{2k} \right| w_k ws_k \qquad \text{formula 2.7.}$$

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, ws_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 ws_k gives weight according to metrical stress. Thus, an upbeat note might fetch the value $ws_k = 4$, while a down beat might fetch the value $ws_k = 2$. 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 34 time 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*, *e*, *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 $ws_k = 2$ for the first beat of a bar and $ws_k = 1$ 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 $ws_k = 4$ for the first beat of a bar and $ws_k = 1$ 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 with 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:

$$\text{difference} = \sum_{k=1}^{n} \left| p_{1k} - p_{2k} - m \right| w_k ws_k \qquad \text{formula 2.8.}$$

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, ws_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), who demonstrated that transposition is a factor in certain conditions of melodic similarity. Thus, an algorithm will have to be able to accommodate such conditions.

2.3. Dynamic Programming

A more recent approach to similarity seems to have emerged from the biological sciences, where scientist endeavor to analyze 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 hemoglobin to the first 11 amino acids of the human beta hemoglobin:

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 hemoglobin.

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:

$$S = am - bi - cs \qquad \text{formula 2.9.}$$

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.

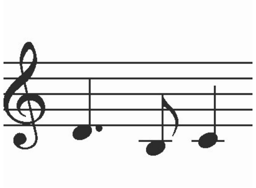
The resemblance of this model with Tversky's model is striking. Equating the amount of matches with the count of common features, we find that indels (amount of operations) corresponds to features present in one object but not the other, whereas substitutions are a cross between common features (both 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 (e.g., that his model does not consider structural similarities), applies to this model.

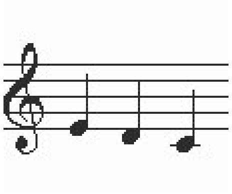
2.3.1. Dynamic Programming in music

The dynamic programming approach has been utilized by Mongeau and Sankoff (1990) for melodic comparison, although they present their model in slightly different form (separating the indels into the components delete and addition). In order to apply dynamic programming the authors have to regard a melody as a sequence of tones t_1, t_2, ..., 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 $S_1 = t_{11}, t_{12}, ..., t_{1n} = (p_{11}, d_{11}), (p_{12}, d_{12}) ... (p_{1n}, d_{1n})$ with a sequence $S_2 = t_{21}, t_{22}, ..., t_{2n} = (p_{21}, d_{21}), (p_{22}, d_{22}) ... (p_{2n}, d_{2n})$. The authors then produce 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: $I_{\mathrm{ij}} = (| p_{1i} - p_{2j} | + | d_{1i} - d_{2j} |) /2$, where 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:

$$\begin{array}{cccc} I_{11} & I_{12} & \ldots & I_{n} \\ I_{21} & I_{22} & \ldots & I_{2\mathrm{n}} \\ \ldots & \ldots & \ldots & \ldots \\ I_{\mathrm{m}1} & I_{\mathrm{m}2} & \ldots & I_{mn} \end{array}$$

Starting with a distance $I_{\mathrm{i}1}$ in the first column a sequence of distance is constructed and summated to an overall distance $D = I_{\mathrm{i}1} + I_{\mathrm{j}2} + ... + I_{kp} + I_{\mathrm{l}(p+1)} ... + ... I_{mn}$, with $i \leq j \leq k \leq p \leq \mathrm{l} \leq p+1 \leq m \leq n$. The starting point $I_{\mathrm{i}1}$ 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:

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

As we can see, the minimal distance is given by $D_{\mathrm{min}} = I_{11} + I_{12} + I_{33} = 2$. 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 tone 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_n of S_1, without any further deletion or addition, this will mean that tone $t_{1n\text{-}1}$ 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. For our example above, we find that tone $t_1 = (d, 3/8)$ of S_1 was mapped onto tone $t_1 = (e, 1/4)$ as well as onto the first quaver time of the tone $t_2 = (d, 1/4)$ of S_2, while the tone $t_2 = (b, 1/8)$ of S_1 was mapped onto the second half of the tone $t_2 = (d, 1/4)$. Tone t_3 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.

There is no doubt in the mind of the author that this is a powerful 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 judgment decreases its application 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. Clearly, the model will classify retrogrades and inversions generally not as being similar, although this might be, as will be shown later, in favor of this model from a cognitive point of view.

2.4 Transition matrices

The most recent approach to modeling melodic similarity is based on a concept which was firstly introduced by Fucks (1965). Fucks intended to find a measurement 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	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

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	0.33	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

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 values of possible parameters. The author concludes that this model has no future.

2.5. Critiques on the conception of similarity

Summarizing the features of the three above discussed models, we find particularly one 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 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) 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 judgments. In fact, the difficulties with all these models are so substantial that we might ask the question whether a model on melodic similarity is attainable or at least desirable.

This question, whether similarity measures are attainable and desirable, was challenged 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 judgment anyone often (or ever) makes”. Their argument, so they claim, is supported by the fact that so far nobody, 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 a useful model. Although the exposition of several existing models above renders such a claim an over-generalization, the none-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 the question to our attention of what we are actually seeking. True, if these authors are correct with their opinion that similarity judgments are hardly ever made, then there is no need for the development of a model, indeed. However, this seems not to be the case. The author will give eight examples which will involve some form of similarity judgment: (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) the classification and retrieval of melodic material in a data base, (e) a composer producing a variation of a theme, and (f) a judge deciding whether a copyright infringement suit over a motive should be granted or not. 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 aware of any similarity judgment, while the ethnomusicologist is or should be interested in cognitive processes. Admitted, the strategy and result of similarity judgments 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 is it attainable and if so in what 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 *A*s, because of its similarity to this category, requires the existence of the category of *A*s 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 above formulated argument will also imply that an appropriate similarity model would ideally consider categorization. Without going into 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 thesis. 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 functional model which has been shown to be 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 at this point that the absence of a theory of melodic categorization increases the need for a melodic similarity model initially independent of categorization. However, in order to establish other features this model should incorporate, we will consider a second critique by Goodman.

We considered earlier in the context of Tversky's similarity model Goodman's critique 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 judgments are context dependent (Goldstone, 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 context of pets. Barsalou (1983) also showed that seemingly highly dissimilar objects receive a high similarity rating when put into specific context (e.g., jewelry 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 context of

$1,000,000 there does not seem to be 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 possibility of how to interpret Barsalou's findings. While our example involves a one-dimensional quantity comparison, the comparison of children and jewelry involves far less clear defined features. In fact, it seems an almost infinite amount of quantitative and qualitative features can be assigned to both children and jewelry, so that context is desperately needed to select the relevant features. If no context is given, we might speculate that participants in an experiment will create their own context. Still, it remains an unsolved issue how relevant this observation will be in the context of melodic similarity. Comparing two melodies is supposed to be a cognitive and perceptual and almost automatic process, while the comparison of children and jewelry is a judgment of conceptual similarity. Finally, the comparison of two melodies seems to involve a limited number of features (such as pitch, dynamics, tone-color 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 is the transformation of the main theme of a composition. Similarly, we might expect that the change of a specific rhythm such as √≈might be more significant if the same rhythm surrounds the specific rhythm such as √≈√≈√≈. However, it seems to the author that at this point a contextual melodic similarity model is unattainable. True, that this will put some constraints onto a model, but a context free model will at least produce some testable hypotheses. Still, even if we developed a context free model disregarding harmonic implications, Goodman's argument holds true in as much that a frame of reference will be required and also that this frame of reference could depend on the features of - let us say - two isolated non-harmonic melodies Thus, a model will have to be developed so 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 organization of language (for instance the temporal structure). Consequently, we would also expect that such interdependency of language and culture will affect similarity judgments. 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 from: "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, it does not demonstrate as Goldstone (1994) seems to imply, that syntactic similarities will induce semantic similarities which then evokes cognitive similarities. However, we might argue in a Shepardian fashion, that cultural difference will lead to difference in categorization, which will then be reflected somehow in the user’s language and hence similarity judgments and language are expected if not to be causally related so still to be correlated. Consequently, we would expect 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 not be valuable 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 judgments are dependent on expertise. For instance Suzuki, Ohnishi & Shigemasu (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 expert 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 judgments depending on expertise. However, it is such a common feature of music psychological experiments (compare Deutsch, 1999) to generate expertise dependent data, that we might suspect that the same might hold true for melodic similarity. Further, we might expect that musical experts will deliver more accurate similarity judgments, meaning data with smaller variance, than musical novice 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 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, they seem to be systematic rather them random. Thus, they can be integrated in a wider model.

2.6. Features of a desirable melodic similarity model

Summarizing, we conclude that a melodic similarity model is desirable, will be context independent, will have to identify relevant features, will have to include empirical constants and will have to incorporate some conceptual flexibility. The author further hopes to have to demonstrate that all the reviewed models 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: as long as the issue of melodic similarity is not approached more systematically, it seems unlikely that we will find more than some models which incorporate heavy theoretical assumption (such as transpositional invariance) without reasoning as the authors seem to be unaware of these theoretical assumptions. The author feels, that if we asked the question, what are the relevant features and how are these features compared when rating the similarity between two melodies, seems far more promising. Such a questioning will necessitate a systematic identification of the relevant features and their precise definitions. It also will require a conceptual analysis of the cognitive process involved in similarity ratings. The author also argues in favor of a proposals as put forward by Palmer (1983), to consider similarity as related to the complexity of the transformation process required in mapping a given object onto another object. Such an assumption seems reasonable in the context of experiments involving classification tasks, and where similarity is correlated to reaction time (Goldstone & Medin, 1994): Decreasing similarity between objects increases the reaction time it takes to classify these objects, as the complexity of the task increases. Palmer's approach seems also well motivated from an evolutionary point of view: in the situation of identifying an animal (like the rattle snake) fast and one-dimensional decisions are required (yes or no). Thus, we might assume that cognitive similarity should be integrated into the comparison process. In case the complexity of the comparison process exceeds a certain threshold the animal will be classified as harmless (just a slowworm) and if the complexity undercuts the threshold the animal will be classified as dangerous (it is a rattle snake). Lower complexity also increases reaction time, which is needed in such a dangerous situation.

Investigating complexity of a comparison requires on a formal level a transformation mechanism which maps one object onto the other object. Thus, relating this approach to melodic similarity, we will have to establish a transformation mechanism, mapping any melody onto any other melody. It is our goal to establish such a similarity transformation as the conceptual framework for a possible similarity model. However, we propose not to establish one specific model but a family of models which are all related to this framework. This will guarantee the adaptability which has been made clear is necessary. Finally, a specific model will be put to the test in an experimental situation where it can be compared with existing models and approaches.