8 Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments

DANIEL MÜLLENSIEFEN,
CHRISTOPH-PROBST-WEG 8
20251 HAMBURG
GERMANY

KLAUS FRIELER
HOPFENSTRAßE 22
20359 HAMBURG
GERMANY

Abstract

Melodic similarity is a central concept in many sub-disciplines of musicology, as well as for many computer based applications that deal with the classifications and retrieval of melodic material. This paper describes a research paradigm for finding an 'optimal' similarity measure out of a multitude of different approaches and algorithmic variants. The repertory used in this study are short melodies from popular (pop) songs and the empirical data for validation stem from two extensive listener experiments with expert listeners (musicology students). The different approaches to melodic similarity measurement are first discussed and mathematically systematized. Detailed description of the listener experiments are given and the results are discussed. Strengths and weaknesses of the several tested similarity measures are outlined and an 'optimal' similarity measure for this specific melodic repertory is proposed.

8.1   Introduction

Melodic similarity is a key concept in several of musicology's subdisciplines. Among these are ethnomusicology (e.g. Bartók & Lord, 1951; Seeger, 1966; Kluge, 1974; Bartók, 1976; Steinbeck, 1982; Jesser, 1992, Juhász, 2000), music analysis (e.g. Meyer, 1973; Lerdahl and Jackendoff, 1983; Baroni et al. 1992; Selfridge-Field, 2003), copyright issues in music (e.g. Cronin, 1998), music information retrieval (e.g. Mongeau and Sankoff, 1990; McNab et al. 1996; Downie, 1999; Meek & Birmingham, 2002; Uitdenbogerd, 2002), and music psychology (Wiora, 1941; Schmuckler, 1999; Hofmann-Engl, 2000, 2001; McAdams & Matzkin, 2001; Deliège, 2002). An overview of motivations, research paradigms, and related concepts is given in the Volume 11 (1998) of *Computing in Musicology* (*Melodic Similarity: Concepts, Procedures, and Applications*), and the 2001 spring issue of *Music Perception* (Vol. 18, No. 3). For different research questions a variety of methodologies for measuring melodic similarity have been developed.

The motivation for the present investigation came from the area of music psychology. Following the research approaches to memory for melodies of Sloboda and Parker (1985), Kauffman and Carlsen (1989), and Dowling and colleagues (Dowling et al., 2002) a way to describe the memory representation of a melody is the goal of a current psychological research enterprise (Müllensiefen, in preparation). One necessary tool to find an adequate description of a melodic memory representation seems to be a similarity measure that relates an original melody to its (probably transformed) version in memory in a cognitively appropriate way. A proper measure for this purpose was already called for by Sloboda and Parker complaining that "there is no psychological theory of melodic or thematic identity" (Sloboda and Parker, 1985: 161).

The literature on similarity measurement for melodies of the last two decades does not suffer for the lack of measurement procedures for melodic similarity but rather from their abundance. Different techniques for defining and computing melodic similarity have been proposed to emphasize distinct aspects or elements of melodies. Among features emphasized are intervals, contour, rhythm, and tonality, often with several options to transform the musical information into numerical datasets. Current basic techniques for measuring the similarity of this type of datasets are edit-distance, *n*-grams, correlation- and difference- coefficients, and hidden Markov models (HMMs). There are many examples of successful applications of these specific similarity measures: These include McNab et al. (1996) and Uitdenbogerd (2002) for edit-distance, Downie (1999) for *n*-grams, Steinbeck (1982) and Schmuckler (1999) for correlation- and difference-coefficients, O'Maidin (1998) for a complex difference measure, and Meek and Birmingham (2002) for HMMs.

The basic question addressed in the present paper is, Which type of data and which similarity measures are cognitively most adequate? The aim of this investigation is to find the "optimal" similarity measure out of a set of basic techniques and their variants.

The "optimal" similarity measure would probably be the mean rating of a group of music experts. But as such a group of experts is not always at hand, the idea of this investigation was to model expert ratings with some of the basic measurement techniques just mentioned. So a rating experiment was conducted to compare expert ratings with the results of similarity algorithms. The "optimal" or most cognitively adequate measure would be the one that predicts the expert judgments best.

Not very many extensive studies comparing human ratings to algorithmic similarity measurement have been undertaken yet. Exceptions are Schmuckler (1999), Eerola et.al. (2001), McAdams and Matzkin (2001), Hofmann-Engl (2002), and very recently Pardo, Shifrin, and Birmingham (2004). The studies of Schmuckler (1999), McAdams and Matzkin (2001), and Pardo et al. (2004) come closest to the present approach, but the variety of similarity models and musical material employed here is far greater and closer to "ordinary" western music.

In the next section the different approaches to data transformations and similarity measures are defined and systematized. References to the original literature are made. Section 8.3 describes the rating experiment and the treatment of the collected data. Section 8.4 compares human ratings with the employed algorithmic models and proposes an optimization for a combination of different models. Section 8.5 discusses aspects of strength and weakness of the optimized model and points out musical dimensions of melodies that have not been covered by the basic models nor their combination presented here and that could be perspectives for future research.

8.2 Data Transformations and Similarity Models

For defining the general notion of a similarity measure, one has first to define what a melody is. An algorithmic or mathematically based similarity measure has to find an abstract representation of a true musical melody sounding in time and space. For our purposes a melody will be simply viewed as a time series, i.e., as a series of pairs of onsets and pitches (t_n, p_n), where pitch is represented as a number, usually a MIDI number, and an onset is given by a real number representing a point in time.

The two components of this time series will be called "rhythm" and "pitch-melody" respectively. Most of the considered similarity measures work either on pitch or rhythm alone.

Furthermore, it is useful to view rhythm or pitch-melody as a vector in a suitable n-dimensional (real) vector space or as a string in a more computer oriented sense. According to this, we discriminate different classes of similarity measures: Vector, symbolic and musical measures. The musical measures use an abstract representation of melodies as well, but they rely on more or less detailed musical knowledge rather than on more abstract properties. We will concentrate here mainly on the vector and symbolic measures.

8.2.1 Definition

A similarity measure $\sigma(m_1, m_2)$ is a symmetrical map of the space of abstract melodies M mapping two melodies on a value between 0 and 1, where 1 means identity. It should be normalized, i.e., the similarity of a melody to itself should be 1. Furthermore, it should be invariant under transposition in pitch, translation in time and under tempo changes, i.e., dilation in time.

A general (and brute-force) way to achieve the desired invariances, which we adopted to some of the measures, is to take the maximum over all possible transpositions, and/or translations/dilation. The algorithm by O'Maidin (1998) employs a similar strategy.

The space of similarity measures is convex, i.e., if one has two or more similarity measures σi, a weighted sum, $\Sigma\, w_i\, \sigma_i$ with $\Sigma\, w_i = 1$, will yield another similarity measure. This will be exploited for finding an optimal measure by means of a linear regression over our data.

Looking at this abstract definition, it is intuitively clear that the space of similarity measures is enormous. The problem is not, as stated earlier, the lack of measures but to find the cognitively most adequate ones.

All of the herein presented measures typically follow some basic construction steps. First they transform the melodies with more fundamental transformations like the interval and/or duration representation, and then they apply more elaborate ones, like Fourier transformation or fuzzifications/classifications. At a last step a standard method of correlation, like vector correlation or edit-distance is adopted.

8.2.2 Representations of Abstract Melodies

Due to the invariance properties of a similarity measure, melodies are often written in duration and interval representations. The first goes from onsets to onset, or by inter-onset-intervals [IOIs] ($\Delta\, t_n = t_{n+1} - t_n$) and/or uses integral multiples of a common minimal duration for IOIs $\Delta\, t_n = k(n)\, \Delta T$. In the latter we will speak about quantized melodies and quantized representation, which are invariant under translation/dilation by construction. The second representation uses intervals, i.e. differences of pitches $\Delta\, p_n = p_{n+1} - p_n$, instead of absolute pitch. Any similarity measure using this representation has already the required invariance under transposition.

Another fundamental representation is achieved by rhythmical weighting. Similarity measures working on pitch alone use only the sequence order, but no absolute time information, giving shorter tones the same weight as longer ones. To account for this, and if one has quantized melodies (as we always had), one can substitute every pitch in the melody by n-times the same pitch, where n is the duration in shortest time units of the tone. So, e.g., if one has the melody (in quantized representation):

$$(2, 64), (2, 66), (2, 68)$$

it becomes

$$(1,64), (1,64), (1,66), (1,66),(1,68) (1,68)$$

The concept of rhythmical weighting has been widely used in other studies (e.g. Steinbeck 1982, Juhász 2000, Hofmann-Engl 2002).

8.2.3   Transformations of Pitch

The most important transformations of pitch are contourization, fuzzification and Fourier transformation.

8.2.3.1   Contourization

The concept of contourization relies on the perceptual salience of melodic contour. The idea relies on the fact that the exact sequence of pitches is often not crucial, but the turning points of a melody are. In our model a changing tone is not taken for a local extremum if the notes immediately before and after the candidate are the same. Instead it picks out the local extremes of a pitch sequence and makes some kind of interpolation, mostly linear, between these anchor tones. The [This?] concept of contourization was employed in the similarity measures by Steinbeck (1982) and Zhou and Kankanhalli (2003).

We used two different contourization procedures—the one used by (Steinbeck 1982), and our personal one. The difference lies in the treatment of "changing tones" (a sequence of three notes in which the first and third are the same). The idea behind this is that changing tones, which always make for a local extreme, are irrelevant for contour perception. In our model a changing tone is substituted for the three events if the note before and the note after the candidate are the same. In Steinbeck's model, two tones before and after must be either strictly descending or ascending.

8.2.3.2   Fuzzification

The main idea of fuzzy logic is to allow a whole range of truth values between 0 and 1 for a logical statement, where 0 means "false" and 1 means "true". Accordingly, a fuzzy set (Zadeh, 1965) is a set, where each element belongs to this set only to some certain degree between 0 and 1. The advantage of this concept is that it offers an easy way to model

fuzziness in perception and other areas.

The idea can be carried forward to intervals. Using fuzzy concepts with intervals reflects the fact that even an experienced listener is not always able to determine a interval exactly, but has always a certain perception of the magnitude of an interval. A listener will always discriminate a step from a "skip", e.g., a second from larger intervals such as fifths and sixths.

We define certain classes of intervals and assign to each interval in the melody a vector of "belongingness" to this classes. But in fact our tested models use fuzzy sets, where each interval belongs to exactly one class, so it should be more precisely called a classification. The idea to reduce the intervals of the chromatic scale to a smaller set of interval classes is again very common in applications that use similarity measures (e.g. Pauws, 2002). We took the nine interval classes shown in Table 8.1.

Class	Intervals	Name
-4	< -7	Big leap down
-3	-7, -6, -5	Leap down
-2	-4, -3	Big step down
-1	-2, -1	Step down
0	0	Same
1	1, 2	Step up
2	3, 4	Big step up
2	5, 6, 7	Leap up
4	> 7	Big leap up

Table 8.1. Interval classes used.

The intervals are counted in semi-tones. Taking the sequence (1,64) (1,65) (1,70) (1,68) (1,65) as an example, one gets the intervallic representation

$$1, 5, -2, -3$$

and the fuzzified melody

$$1, 3, -1, -2.$$

8.2.3.3 Fourier Transform

Another method adopted from Schmuckler (1999) is that of taking the (discrete) Fourier transform of the pitch-melody, more precisely the DFT of pitch ranks, i.e., the numbering of the pitches p_n as ranks r_n starting with 0 for the lowest pitch. The idea behind this, as stated by Schmuckler (1999), is that a Fourier transform detects inherent periodicities in a signal. The complex Fourier coefficients are given by the well-known formula

$$c_n = \sum_{k=-\frac{N}{2}}^{\frac{N}{2}} r_k e^{j\omega_n k}, \omega_n = \frac{2\pi n}{N}$$

Music Query: Methods, Models, and User Studies

and the amplitudes of the real positive power spectrum from this are then $p_n = c_n c_{-n}$.

8.2.4   Transformations of Rhythm

For similarity of rhythms, a field which seems to be neglected in the literature, we had to develop methods on our own. In principle every correlational technique, whether vector or symbolical, can likewise be used for rhythm vectors or rhythm strings. As preliminary transformations we used gaussification and fuzzification.

8.2.4.1   Gaussification

The idea of gaussification is to construct a continuous, integrable function out of a set of onsets by superposition of gauss functions with a mean at the point of an onset and fixed standard deviation. So, if, t_n is a set of onsets, then

$$g(t) = \frac{1}{N} \sum_{i=0}^{N-1} e^{-\frac{(t-t_i)^2}{2\sigma^2}}$$

is called a rhythm gaussification. This transforms a n-dimensional vector t_n into an ∞-dimensional one, and as we will see later, one has to go from ordinary scalar products over to integrals.

8.2.4.2   Fuzzification

The technique of fuzzification, as explained above, can be applied to durations, too, but one has to relate the durations to a fixed duration, which we chose to be the most frequent duration (modus) d8 (of all ducations) in a melody. We used the following five classes for the fractions $\Delta T_n / d_\infty$

Class	Fraction	Name
4	$f > 3.3$	Very long
3	$1.8 < f \le 3.3$	Long
2	$0.9 < f \le 1.8$	"Normal" beat
1	$0.45 < f \le 0.9$	Short
0	$f < 0.45$	Very short

Table 8.2. Duration classes used.

This choice of classes is, of course, far from unique; it was inspired by the common categories of (binary) musical rhythm (Drake and Bertrand 2001: 24f).

8.2.5   Vector Measures

8.2.5.1   Correlation Measures

An important class of vector measures relies on the well-known correlation of n-dimensional vectors:

$$r(v,w) = \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2 w_i^2}} \in [1,-1]$$

For a similarity measure of pitch-melodies one has to ensure transposition invariance, and, furthermore, one must transform the values to the interval [0,1]. The first can be done, for example, by transposing every pitch by the mean pitch of the melody. The latter can be achieved, for example, by setting any negative value to 0, as we did in the most cases. This was done because we were not interested, unlike other investigations (e.g. Kluge 1974, Wiggins 2002, p. 308), in the degree of contrary or retrograde similarity.

Vector correlation was exploited by us by these means:

(1) Pearson-Bravais correlations of pitch-melodies (raw and rhythmically weighted, transposition by mean pitch): rawpcst, rawpcwst.

(2) Pearson-Bravais correlations of contourized melodies (unweighted, transposition by mean pitch): conspcst, conpcst.

(3) Pearson-Bravais correlations of Fourier-rank transformed melodies (weighted, unweighted): fourrst, fourrwst, fourri.

(4) Correlation of fuzzified intervals: difffuz.

(5) Correlation of fuzzified contourized pitch-melody: diffuzc.

(6) Correlation of rhythm gaussifications: rhytgaus.

(7) Harmonic correlation: harmcorr, harmcork, harmcorrc.

For the correlation of rhythm gaussifications, we have to adapt the scheme a little bit. First, one has to use integrals for the scalar products, which can be solved analytically. Second, one has to guarantee translation and dilation invariance. Translation invariance is achieved by translating each onset vector to start with $t_0 = 0$. Dilation invariance needs more sophistication in the general case. However, if one has quantized melodies, one can set the smallest time units of both rhythms to be equal and one arrives at the following formula for the scalar product of two gaussifications g and g':

$$\langle g, g' \rangle = \frac{1}{N} \sum_{n,n'} e^{-\frac{(k(n)-k'(n'))^2}{2\sigma^2}}$$

Harmonic correlation belongs rather to the field of "musical measures" and will be discussed later.

The attentive reader will have noticed that the correlation is only defined for vectors of equal dimension (length). But in practice melodies seldom have exactly the same length. To accommodate possible differences, we shift the shorter melody along the longer one and compute a similarity for each section of equal length to the shorter one.

Additionally, to account for possible missing upbeats, we shift the shorter one up to 10% of its length or 8 minimal time units, whichever is greater, to the left of the longer melody. For each of these pairs of melodies of same length we then calculate the correlations and take the maximum over all values as the true similarity.

8.2.5.2 Distance Measures

There is an natural link between distance measures on the space of melodies and similarity measures (e.g. O'Maidin 1998). If one has a distance measure $d(m,n)$ obeying translation, transposition and dilation invariance, a similarity measure can easily be obtained by

$$\sigma(m,n) = e^{-\frac{d(m,n)}{k(m,n)}}$$, or, if $k(m,n) > d(m,n)$, for all m,n.

We used just two out of this huge class of similarity measures: the mean absolute difference of intervals with different normalizations.

Set $z_i = \Delta m_i - \Delta n_i$, $\bar{z} = \frac{1}{N}\sum_i z_i$, $q_\infty = \max_i |\Delta m_i| + |\Delta n_i|$. Then $d(m,n) = N\bar{z}$ is a transposition invariant distance on pitch-melody space. The two similarity measures are given by:

$$\sigma_1(m,n) = e^{-\bar{z}}$$

$$\sigma_2(m,n) = 1 - \frac{\bar{z}}{q_\infty}$$

The first (diffexp) is merely a straightforward construct. The rationale behind the second one (diff) is to account for the size of steps or leaps of the individual melodies to be compared.

Melodies that consist of a series of large intervals and that result in a large mean absolute difference should have greater similarity values than melodies consisting of only small intervals with the same mean absolute difference.

8.2.6 Symbolic Measures

The symbolic measures view a "melody" (defined by either a series of pitches or durations) not as a vector but as a string, i.e., as a series of arbitrary symbols of finite length. Usually for strings in the computer science sense the symbols are taken to be ASCII characters. Accordingly, a string can be defined as a sequence of characters. But as we will see, the algorithm for the similarity measures used here rely only on the operation "test for equity", so arbitrary symbols like, say, real numbers are allowed. We used two common and well-known techniques: The edit-distance (or Levenshtein distance) and measures related to *n*-grams. This will be explained in the following.

8.2.6.1 Edit-Distance

The main idea behind the concept of edit-distance is to take the minimum number of operations ("edits") needed to transform one string into the other as a similarity measure for strings. The allowed operations are insertion, deletion and substitution. The calculation of edit-distance is done with a well-known dynamic programming algorithm. See Mongeau and Sankoff (1990) or Uitdenbogerd (2002) for details of the algorithm.

It is clear that the maximal possible edit-distance of two strings is equal to the length of the longer string, which enables us to define a similarity measure $\sigma(s_1, s_2) = 1 - \frac{d_e(s_1, s_2)}{\max(|s_1|, |s_2|)}$, where $|s|$ denotes the length of string s. We used this edit-distance in several ways:

(1) Edit-distance for raw melodies (rhythmically weighted and unweighted): rawed, rawedw. (Here we had to take the maximum over all transpositions.)

(2) Edit-distance for contourized melodies (Steinbeck-contourization and our own): consed, coned. (Again, we had to take the maximum over all transpositions.)

(3) Edit-distance for intervals: diffed.

(4) Edit-distance for fuzzified rhythms: rhytfuzz.

(5) Edit-distance of harmonic strings: harmcore.

8.2.6.2 *n*-grams

An *n*-gram is simply a string of length *n*. Strings of different lengths are denoted 3-grams, 4-grams and so forth. To make for a similarity measure of strings, one questions about the distribution of substrings of fixed length, the *n*-grams, in two to be compared strings. We used three different ways to account for a similarity measure: The Sum Common, the Count Distinct (or Coordinate) and the Ukkonen measure. An i*n*-depth discussion of *n*-grams as representations of melodies can be found in Downie (1999) and Uitdenbogerd (2002).

Sum Common Measure. Let *s* and *t* be two strings. We write s_n for the set of distinct *n*-grams in a string *s*. The Sum Common Measure sums the frequency of *n*-grams τ occurring in both strings.

$$c(s,t) = \sum_{\tau \in s_n \cap t_n} f_s(\tau) + f_t(\tau)$$

where $f_s(\tau)$ and $f_t(\tau)$ denote the frequencies of the *n*-gram τ in string *s* and *t* resp. The maximum frequency of an *n*-gram in a string *s* is |*s*| - *n+1*, so the maximum value of the Sum Common measure is |*s*| + |*t*| - *2(n-1)*. A similarity measure is then given by

$$\sigma(s,t) = \frac{c(s,t)}{|s| + |t| - 2(n-1)}$$

Count Distinct (Coordinate Matching) Measure. The Count Distinct Measure resembles much the Sum Common Measure, the only difference is, that we do not sum the frequencies of the common n-grams, but just count them.

The Count Distinct Measure of the above example would then simply be 2, because there are two common n-grams.

For normalization we divide this by the maximum count of distinct n-grams of either string and arrive by the following similarity measure

$$\sigma(s,t) = \frac{\displaystyle\sum_{\tau \in s_n \cap t_n} 1}{\max(\# s_n, \# t_n)}$$

The Ukkonen Measure. The Ukkonen measure is kind of opposite to the Sum Common Measure for it sums differences of the frequencies of the n-grams not occurring in both strings. The formula is:

$$u(s,t) = \sum_{\tau \in s_n \cup t_n} \left| f_s(\tau) - f_t(\tau) \right|$$

For making a similarity measure we normalize by the maximum possible number of n-grams and subtract this from 1:

$$\sigma(s,t) = 1 - \frac{u(s,t)}{|s| + |t| - 2(n-1)}$$

Application: We combined this three measures with four different melody representations:

(1) n-grams with pitch numbers as symbols (taking the maximum over all transpositions): ngrsumco, ngrcoord, ngrukkon.

(2) n-grams with fuzzified intervals: ngrsumcf, ngrcoorf, ngrukkof.

(3) n-grams with the alphabet S, D, U for intervals, assign "S" if the interval is the prime, "D" for an descending and "U" for an ascending interval1: ngrsumcr, ngrcoorr, ngrukkor.

(4) n-grams for fuzzified rhythm: ngrsumfr, ngrcoofr, ngrukkfr.

For each variant we also took the maximum over n-gram lengths 3 to 8.

---

[1] This alphabet is sometimes called the Parsons Code and is, for example, used in *The Dictionary of Tunes and Musical Themes* (Parsons 1975

8.2.7 Musical Measures: Harmonic Correlations

From the class of musical measures we defined only measures for harmonic correlation here. There are actually some very interesting musical measures that look for similarities in several musical dimensions simultaneously, but they will be the subject of future investigations.

We used four different measures for harmonic correlation, all of them based on the tonality vector of Krumhansl. The main idea behind all the four measures is to assign to each bar a tonality vector, which could be either major or minor. Hence, one gets a (vector of) harmonic vector(s) or a harmonic string, to which the usual techniques could be applied.

Krumhansl's Tonality Vector. Krumhansl and Schmuckler discovered (Krumhansl 1990, Krumhansl and Kessler 1982) that to each of the 12 semi-tones of the modern equally tempered scale can be assigned a numerical value measuring its significance or relative strength for a given tonality. They proposed two 12-dimensional vectors, one for major and one for minor scales. The values are:

T_M = (6.33, 2.23, 3.48, 2.33, 4.38, 4.09, 2.52, 5.19, 2.39, 3.66, 2.29, 2.88) (Major)

T_m = (6.33, 2.68, 3.52, 5.38, 2.60, 3.53, 2.54, 4.75, 3.98, 2.69, 3.34, 3.17) (Minor)

The nth position in the vectors stands for the value of the nth semi-tone (modulo 12) above a given base tone. For example, a pitch of class "E" has a relative significance for C-Major of 4.38 (4th semi-tone in major), whereas for D-minor it has only 3.52 (2nd semi-tone in minor).

For a bar, the relative strength of each tone in the bar (weighted by its duration) is computed for each of the 2 possible modes (major or minor) and 12 possible base tones giving two 12-dimensional vectors H_i and h_i.

For example, given a bar (3,C) (1, D) (2, E) (2, C) (in IOI-pitch representation) the value for C-Major (0th component of H_i) would be:

$$3 \cdot 6.33 + 1 \cdot 3.48 + 2 \cdot 4.38 + 2 \cdot 6.33 = 43.89$$

and for d-Minor (2th component of h_i)

$$3 \cdot 3.34 + 1 \cdot 3.52 + 2 \cdot 2.60 + 2 \cdot 3.34 = 25.42$$

Harmonic Vector Correlation I. For each corresponding bar of the melodies two 12-dimensional harmonic vectors (major and minor) and their correlations are computed. (If one melody is shorter than the other, we simply ignored the supernumerary bars.) Next we computed the average correlations for all bars, again for major and minor separately. The maximum of the two values is the harmonic vector correlation I.

Harmonic Vector Correlation II. Instead of computing the vector correlation of corresponding bars for each mode separately and averaging the single correlations, one can use the 24-dimensional vectors directly. One

gets a vector of these vectors for each bar of each melody, and for these vectors-of-vectors one can calculate the usual correlation.

Harmonic Edit-Distance. We also computed a single tonality value for each bar as the key, which had the maximum value of the 24 possible keys, taking values 0-11 as major keys and values 12-23 as minor keys. This gave a "harmonic string" for each melody for which we computed the edit-distance and got a harmonic similarity with the usual normalization.(See above.)

Harmonic Circle Correlation. A more elaborate version of correlating the 24- dimensional tonality vectors is based on the idea of the Circle of Fifths to reflect the fact that similarity of keys is in correspondence to their relative position in the Circle. Therefore we retrieved first a harmonic vector for a melody by finding the maximum of the tonality vector like we did for the harmonic string. This gave a value ranging from 0 to 23 for each bar. Next this value was transformed in a relative position on the circle of fifths by using a angular variable in steps of $\omega = \frac{2\pi}{12}$. We arbitrarily set $\varphi_0 = Db = 0 * \omega$, $\varphi_1 = Ab = \omega$ and so on up to $\varphi_{11} = F\# = 11\omega$ for the major keys. The minor keys followed the same structure, with respect to their major parallels, but the angles were shifted by $\omega/2$, giving $\varphi_{12} = Bb^m = 0.5\omega$, $\varphi_{13} = F^m = 1.5\omega$ up to $\varphi_{23} = Eb^m = 11.5\omega$.

With the help of this transformation we now defined the correlation of two tonalities as the cosine of the difference of their angles:

$$r_i = \cos(\phi_i^1 - \phi_i^2)$$

This choice comes from the scalar product of two vectors on the unit circle in 2-dimensional space. The total correlation is then defined to be

$$r = \frac{1}{N}\sum_i^N r_i$$

where N is the number of bars and we set negative values to 0.

8.2.8   Implementation of the Models

We implemented a total of 48 models, counting all variants, of which 39 were used in this study. The implementation was done in C/C++ with GCC under Linux, and was also ported to Win32-platforms. It comprises over 5,000 lines of code. As input files we used .CSV-files, which were generated by extraction from ordinary MIDI-Files.

8.3   Experiments

8.3.1   Experiment 1

The idea of this study is to pick the similarity measures of the ones presented in the last section that best predicts or approximates the similarity judgments of human music experts. For that reason two constraints were applied to the tested sample of subjects:

> (1) Their judgments should be consistent over time.

> (2) They should recognize identical melodies as highly similar.

Fulfilling these criteria a subject is expected to give reliable and stable similarity judgments that can be modeled algorithmically.

Subjects. A pretest with subjects with little or no music background showed that similarity judgments from many subjects were unstable and not consistent over time. Judgments of these subjects tended to be influenced by many no*n*-musical factors such as position of the comparison item in the sequence of items and session length. As a consequence for the main study, only musicology students from introductory courses at the University of Hamburg were recruited as subjects. In all 82 subjects participated. Of these 82 subjects, the data of only 23 could be selected on the basis of the aforementioned criteria. The subjects' musical background was measured by an extensive questionnaire very similar to the one employed by Meinz and Salthouse (1998). Typically musicology students have a long history in music making (e.g. the mean number of years for playing an instrument was 12; the mean number of months of paid instrumental lessons was 71), but their most active musical phase is several years anterior, which is reflected in less time spent for current musical activities when compared to a more active musical phase in the past.

Materials. To obtain ecologically valid results 14 existing melodies from western popular songs were chosen as stimulus material. Among these melodies were songs like "As long as you love me" by the Back Street Boys, "Summer is Calling" by Aquagen, and "From Me to You" by the Beatles. All melodies were between seven and ten bars long (15-20 sec.). The melodies were selected according to several criteria: They should contain at least three different phrases and two thematically distinct motives. They should have a radio-like, popular character and they should be unknown to the subjects to precluded effects from previous knowledge. In fact, some of the melodies were known to a few participants, as was evidenced by the questionnaire. But the ratings of the subjects who knew the songs did not differ from the other subjects' ratings in any respect. So data from these melodies and from these subjects were kept in the study.

For each melody six comparison variants with "errors" were constructed, resulting in 84 variants of the 14 original melodies. The error types and their distribution were done according to the literature on memory errors for melodies (Sloboda and Parker, 1985; Oura and Hatano, 1988; Zielinska and Miklaszewski, 1992; McNab et al., 1996; Meek & Birmingham, 2002; Pauws, 2002).

Five error types with their respective probabilities were defined: Rhythm errors (p=0.6), pitch errors not changing pitch contour (p=0.4), pitch errors changing the

contour (p=0.2), errors in phrase order (p=0.2), modulation errors (pitch errors that result in a transition into a new tonality; p=0.2). Every error type had three possible degrees: 3, 6, and 9 errors per melody for rhythm, contour and pitch errors, and 1, 2, and 3 errors per melody for errors of phrase order and modulation. For the construction of the individual variants, error types and degrees were randomly combined, except for the two types of pitch errors that were never combined in a single variant, to evaluate their influence separately. As a result 50% of the variants had between 4 and 12 errors in sum, with summed errors ranging from 0 to 16. As an example the test melody D, the chorus melody of the dance title "Wonderland" (as interpreted by Passion Fruit), is depicted in its original form (Figure 8.1) and its variant D1 (Figure 8.2), containing 3 rhythm errors (note repetition and deletions are counted as rhythm errors) and 9 contour errors (accumulating mostly in Bars 7 and 8).

Figure 8.1. "Wonderland" by Passion Fruit, original version.

Figure 8.2. "Wonderland" by Passion Fruit, version D1.

Basically, the types and frequencies of errors in the test material are of fundamental importance to the comparison of different similarity models. Because of the uni-dimensional nature of most of the simple similarity measures discussed above, these measures perform quite differently according to the type and frequency of error (the error dimensions) that a particular set of melodies for comparison contains. So the errors were chosen according to the domain in which the optimal similarity measure should operate. In this case this domain is the reproduction of popular melodies from memory.

Procedure. Subjects were instructed to rate the similarity of pairs of melodic variants on a 7-point-scales (with 7 representing maximal similarity). To make the task more realistic they were asked to imagine that the first member of a comparison pair was a reference that could be played by a music teacher on a piano. The second member of each pair should represent a sung rendition of the same melody by a student. Sometimes the rendition could contain many errors, sometimes only few errors, and in some cases it could be without any error. With their ratings subjects should give "grades" to the imaginary student according to the "severeness" of the errors in sum. They were encouraged to make use of the whole range of the rating scale. None of the subjects mentioned that they were unable to perform the task or that they did not understand it.

Each trial run consisted of a first exposure to the original reference melody to familiarize the subjects with it. After 4 seconds of silence, six pairs each consisting of the reference melody and a different variant were played to the subjects. The members of the pairs were separated with 2 seconds of silence, the pairs were separated from each other with the announcement of the next pair and 4 seconds of silence. There was a break of 20 seconds after each trial, where the subjects had to indicate on the rating sheet, if they knew the reference melody and if so, to write down the title of the song. One test session consisted of 3 or 5 trial runs each with a different reference melody and took 17 to 23 minutes.

Subjects were tested in groups in their normal classroom environment. The melodies were played from CD over suitable loudspeakers with a piano sound at a comfortable listening level (around 65 db). After the test session the subjects had to fill out the extensive questionnaire concerning their previous and current musical activities.

The design of the whole experiment was a test-retest-design: Subject groups were tested in one week and retested one week later. The design of the retest was identical to the test, but involved changing all but one reference melody. So, for example, one subject group was tested in Week 1 with test melodies A, B, and C and in week 2 with D, E, and A. In this way it was possible to compare the judgments of melody A from Weeks 1 and 2 for each subject. Subjects were informed of the retest going to take place one week later, but they were told that they would be re-tested exclusively with different melodies.

Results. The rating data of the subjects had to meet three criteria: subjects should have attended both test sessions, their ratings of variants containing 0 errors (identical to original) should be at least 6 in 85% of the cases, and the correlation of their ratings for the same variants from week 1 to week 2 should not be less than 0,5 as measured by Kendall's τ_b. Data of 23 subjects remained in the analysis. Of course different parameters or numerical values for the latter two selection criteria could have been chosen, but on this point there is no orientation in the literature. For example, the judgments of the subjects tested by Schmuckler (1999) and by McAdams and Matzkin (2001) do not seem to have been tested for reliability and/or consistency at all.

The 23 selected subjects may well be called "music experts" not only for their reliable and consistent similarity judgments, but also because of their musical activities. To give just a few statistics, none of them had been playing an instrument for less than 4 years (mean: 12.52 years), none was making music for less than 4 hours per week in his/her most active musical phase (mean: 21.35 hours/week), and only two had less than 6 months of paid instrumental lessons in their life (mean: 71.91 months).

Obviously, modeling the subjects' similarity judgments with algorithms only makes sense if the ratings of different subjects are quite similar, i.e. the inter-subject reliability is high. This would mean that there is something like a "true" similarity value for a given comparison pair, and that subjects' ratings over- or underestimate this true value only slightly. To test this hypothesis among other measures Cronbach's α was calculated. This measure reflects how well all subjects' ratings measure a latent unidimensional factor ("true" similarity). For the two subject groups α-values of 0.962 and 0.978 were obtained. The Kaiser-Meyer-Olkin measure (KMO) reflects the global coherence in a correlation matrix and is frequently used to evaluate solutions in factor analysis. For the present correlation matrix of the subjects' ratings it yields values of 0.89 and 0.94 for the two

tested groups. These values indicate a very high intersubject reliability. They are clearly higher than the α-values (around 0.84) obtained by Lamont and Dibben (2001: 253) in a comparable situation. Form this result it can be inferred that there is something like a true or cognitive adequate similarity value for the comparison of melody pairs, at least for the population of "music experts".

Given the type of data collected in the experiment, many further results could be obtained, for example the dependency of the similarity ratings on the errors types and degrees and the error position, the dependency of judgment reliability and stability on musical expertise, and the influence of the original melodic structure on the ratings. These results will be the subject of a detailed, more psychological oriented analysis in the future.

8.3.2 Experiment 2

In tests prior to this experiment it was observed that some of the above described similarity measures tended to overestimate the similarity of melodies that do not come from a common original. Similarity values of up to 0.5 for completely different melodies were found. The idea of Experiment 2 was to collect expert similarity ratings for pairs of reference melody and respective variants and reference melodies and variants that have their origin in different reference melodies. In this sense Experiment 2 served as a control experiment for dissimilar material.

Subjects. The subjects were 16 musicology students from an undergraduate course; 11 of them were tested in one group, 5 were tested individually. There were no observable effects of testing in groups vs. individual testing.

Material. Two of the melodies of experiment 1 were chosen as reference melodies. The variants for comparison consisted of the same six variants as in Experiment 1 plus six or five [five or six ??] variants from other reference melodies that seemed to be overestimated in the their similarity by some of the algorithmic models. Unlike Experiment 1, every variant was transposed to a key different from the reference melody, so that the subjects could not make use of absolute pitch information for their ratings.

Procedure. Instructions and procedure were very similar to those of Experiment 1 with two exceptions: One trial with one reference melody consisted of 12 comparison pairs, and there was no retest session one week later. There were only two trials in one test session. To test reliability and stability, subjects should again rate identical variants highly similar and a comparison pair in one trial was repeated. The two identical comparison pairs should be rated with not more than 1 point difference.

Results. According to the two criteria, 12 of the 16 subjects were selected as music experts and their data stayed in the analysis. Again the measures of intersubject reliability, KMO and Cronbach's α, yielded very high values of 0.811 and 0.9788 respectively. The music experts of the control experiment seemed also to estimate the "true" similarity values quite well. Like the music experts of experiment 1, they had a highly active musical background. The results of the comparison between these human expert judgments and the tested algorithmic models are displayed in the following section.

8.4   Algorithmic vs. Human Judgments

According to an ANOVA with error type (interval vs. contour) as factor and rhythm, modulation, and phrase order errors as covariates, there was no significant difference (p=0.709) between the similarity ratings for variants with interval and contour errors. Thus, further analysis treated variants containing these two types of errors equally.

8.4.1   Modeling Experts' Ratings with Linear Regression

To model the similarity ratings of the subjects and thus find the optimal similarity measure, the information of the several dimensions or parameters contained in the melodies must be combined to yield an effective measure (see Selfridge-Field, 1998). The information contained in single-line melodies and relevant for human memory and similarity judgments can be classified in five dimensions: Intervals, contour, rhythm, implied harmonic content, and characteristic motives. Each of the similarity measures explained above can be viewed as to measures the similarity of a melody and its variant along one of these five dimensions. A classification of the similarity measures is shown in Table 8.3.

Dimension	Definition	Measures
Interval	Difference, correlation, or symbolic measures operating on the sequence of pitches or intervals, or their fuzzified values	diff, diffexp, diffed, diffuz, rawed, rawedw, rawpcst, rawpcwst
Contour	Correlation and symbolic measures operating on the sequence of substituting contour values	consed, constpcst, coned, conpcst, fourrst, fourrwst
Rhythm	Correlation or symbolic measures operating on the sequence of fuzzified rhythm values or gaussified onset points	rhythfuzz, rhythgaus, ngrcoorfr, ngrsumfr, ngrukkfr
Harmonic content	Correlation or symbolic measures operating on the sequence of harmonically weighted pitch values	harmcorr, harmcork, harmcore, harmcorc
Characteristic motives	Symbolic measures operating on sub-sequences of interval values or their directions or fuzzified substitutes	ngrsumco, ngrukkon, ngrcoord, ngrsumcr, ngrukkor, ngrcoorr, ngrsumcf, ngrukkof, ngrcoorf

Table 8.3. Melodic dimensions and tested measures.

As it is probable that human music experts make use of the information on several dimensions simultaneously, an optimal algorithmic model of the human ratings would encompass measures from several dimensions in a linear combination. So the optimization process takes two steps:

(1) For a given set of melodies and variants choose for every dimension the measure that has minimal Euclidean distance to the subjects ratings. These are the "best" measures.

(2) With these five "best" measures perform a linear regression analysis to find the optimal combination and the optimal weights for the indi-

vidual measures so that subjects' ratings are best explained by the linear combination. The criteria for this step were: a positive sign for the weight of the factor (measure), a level of significance of $p<0.05$ for each factor, the corrected R^2 should be maximal and the standard error should be minimal for the regression model.

This analysis was done for the three contexts of the 84 comparison pairs of Experiment 1, the 13 pairs with "real" variants that were manipulations of the reference melody in control experiment 2, and all 24 comparison pairs of control experiment 2.

Main experiment. For the main Experiment 1, the "best" measures with their respective Euclidean distance to experts ratings are: coned (5.29), rawedw (5.63), ngrcoord (5.94), harmoncore (6.18), rhythfuzz (10.43). Distances ranged from 5.29 to 12.8. Distances to all measures are found in the appendix.

Linear regression analysis with these measures yielded the best model according to the above described criteria involving only two measures, rawedw and ngrcoord. Interestingly, in combination with other measures the overall best measure, coned, was not able to support explicative power to the model anymore, so that the p-value to its β-weight became insignificant in combination with rawedw or ngrcoord. Any model including coned yielded a lower overall fit than the one involving rawedw and ngrcoord.

The overall fit of the model is quite high: $R = 0.911$, $R^2 = 0.830$, corrected $R^2 = 0.826$, standard error of estimated values 0.66. This means that 83% of the variance in the rating data of the subjects is explained by this model, and the mean deviation for the estimated values is 0.66 points on the 7-point-scale. The standardized β-beta-weights for the two factors are: rawedw (β = 0.543), ngrcoord (β = 0.497). The linear combination to best predict the subjects' ratings on the 7-point-scale is:

$$\sigma_{best} = 3{,}355 \cdot rawedw + 2{,}852 \cdot ngrcoord$$

With this optimized similarity model we found a Euclidean distance to the subjects' ratings of 3.789. This means that the optimized model is by 28.5% better than the best single similarity measure tested (coned). This superiority of the optimized measure opti1 is shown in Figure 8.3.

Real variants in control experiment. For the 13 variants that had their origin in the reference melody in the control experiment the results were slightly different at first glance. The "best" measures from the five dimensions were: diffed (1.3), ngrsumco (1.88), harmcore (1.98), consed (2.11), ngrcoofr (3.09). Euclidean distances ranged from 1.3 to 5.96. A table with all the distances is found in the appendix.

The best model from regression analysis contained the two measures ngrsumco and harmcore. Very high values of fit were found for that model: $R = 0.960$, $R^2 = 0.922$, corrected $R^2 = 0.906$, standard error of estimated values 0.37. Thus, 92% of the variance in the rating data of the subjects was explained by this model, and the mean deviation for the estimated values is 0.37 points on the 7-point-scale.
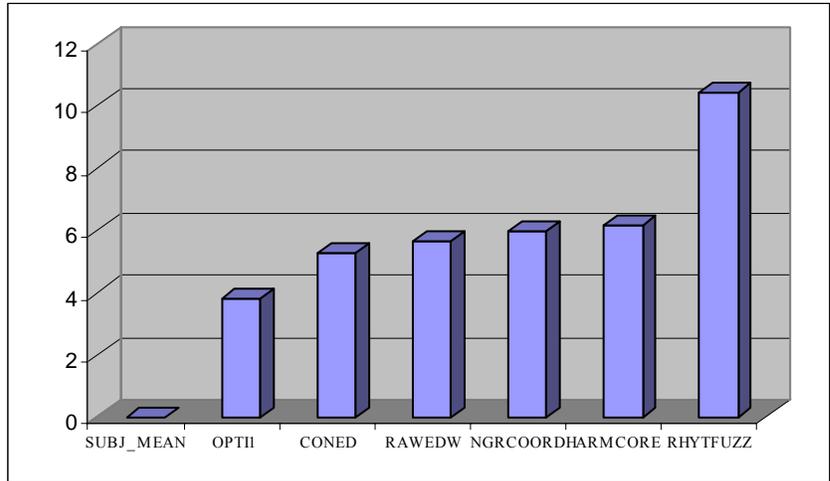
Figure 8.3. Performance of different similarity measures on data from experiment 1.

To check the validity of the result of experiment 1, a second row of regression analysis was performed using the best measures from the first experiment but with the data from the 13 'real' variants from the second. Again rawedw and ngrcoord in combination gave the best result. The model fit was also quite high: R = 0.946, R^2 = 0.895, corrected R^2 = 0.874, standard error of estimated values 0.43. At the same time, the regression model with measures from experiment 2 on the data of experiment 1 found high results as well: R = 0.884, R^2 = 0.781, corrected R^2 = 0.776 and standard error = 0.75.

The standardized β-weights for both models were approximately the same for each data set. Weighting rawedw about 1.15 times more than ngrcoord, and weighting ngrsumco about the same as harmcore. So both models seem to give valid estimations of the subjects' ratings for the similarity of "real"variants and their respective reference melodies. But there are two reasons to assume the model including rawedw and ngrcoord resulting from experiment 1 the superior one: Firstly, the model including rawedw and ngrcoord was found to fit better for a larger data set (Experiment 1). Secondly, the difference of the corrected R^2 values to the second model for the data of Experiment 2 was smaller (0.906 - 0.874 = 0.032) than the other way around (0.826 - 0.776 = 0.05). So to model the similarity ratings of melodies and their variants by music experts, the above stated linear combination including rawedw and ngrcoord is believed to be the optimal model, but with slightly different weights and a constant due to the overall shifting of the ratings towards the pole of maximum similarity:

$$\sigma_{best} = 2{,}254 + 2{,}61 \cdot rawedw + 1{,}72 \cdot ngrcoord$$

"Real" and "wrong" variants of the control experiment. For all 24 comparison pairs of the control experiment, including "real" and "wrong" variants, the five best measures were: diffed (2.04), ngrukkon (2.44), harmcore (2.98), consed (3.57) and rhythfuzz (3.65). Distances ranged from 2.04 to 7.73 as can been seen in the appendix. The best regression model was obtained with three measures: ngrukkon, rhythfuzz, harmcore. Again, the model estimated the subjects' ratings very well: R = 0.96, R2 = 0.921, corrected R2 = 0.909, standard error of estimated values 0.49. A second try with the measures from the main experiment data set, rawedw and ngrcoord, yielded a clearly worse result with a corrected R2 of 0.826. So the best linear combination for estimating the subjects' ratings on the 7-point scale is:

$$\sigma_{best} = 3{,}027 \cdot ngrukkon + 2{,}502 \cdot rhythfuzz + 1{,}439 \cdot harmcore$$

Again, with this optimized model we achieved a much more better result than for any of the single measures. The Euclidean distance was 1.403, which is about 33.4% better than diffed. This is depicted in Figure 8.4.
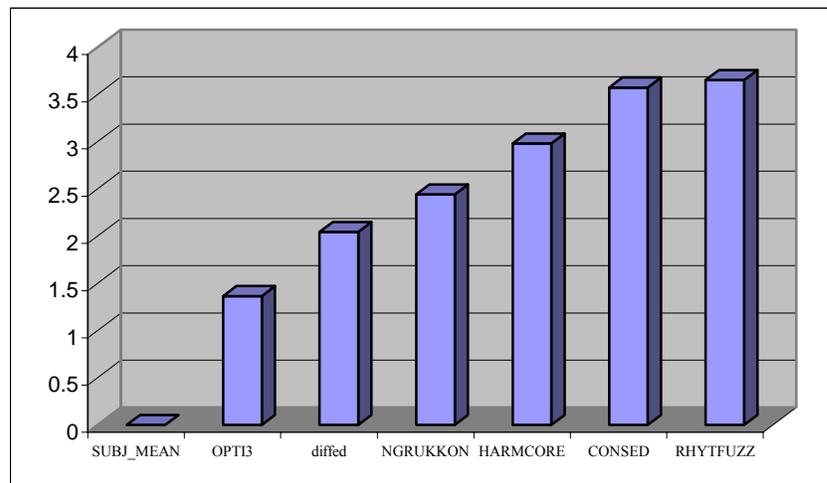

Figure 8.4. Performance of different similarity measures on data from Experiment 2.

Obviously, for the full data set of the control experiment, information from very different sources is needed to model the subjects' ratings. It seems very plausible that subjects make use of easy-to-detect dimensions like rhythm and harmonic content when the task is to tell apart different songs from variants pertaining to the same song. It is also interesting to note that from the *n*-gram measures the Ukkonen distance performed best here, because it is the only *n*-gram measure that counts the differences between two symbol sequences rather than the elements in common.

In any case, it seems that human music experts change their judgment strategy for melodic similarity ratings according to the task. With the more subtle differences in judging the similarity of "real" variants the make more use of individual pitch information, whereas with "wrong" variants to be judged they rely more on rhythm and differences in interval successions.

General observations about the performance of the employed measures Apart from finding only the best combination of measures, the comparison with human judgments can teach us something about the general performance characteristic of the tested data transformations and similarity measures.

Looking at the data of the main experiment, we can ask, Which measures perform best for the different types of transformed melodic data? A series of rank tests and inspections of the rank positions relative to the to the subjects' ratings was undertaken. Generally, contour measures came closer to the subjects' ratings than their raw pitch counterparts. This holds true as much for the edit-distance as for the weighted correlational measures.

For pitch and interval measures the no*n*-parameterical Kruskal-Wallis-Test was performed on the ranks of distance from the rating data. The test was significant with p=0.005. Measures based on edit-distance performed best with a mean rank of 3, while *n*-gram measures had a mean rank of 6. Correlational measures and difference measures both stayed far behind with a mean rank of 13.5 in both cases.

A slightly different picture appeared when comparing measures for contour data. Edit-distance measures again produced the result that was by far the best, with a mean rank of 1.5, while correlation and Fourier based measures shared the middle position, with a mean rank of 8 both. Additionally, the *n*-grams measures that contained only direction information (Parsons alphabet) showed the poorest performance with a mean rank of 12.33. The test statistic approached the usual significance level with a p-value of 0.071.

Among the three different measures using the *n*-gram concept, coordinate matching (ngrcoord) was found to give the best result (distance to ratings in main experiment: 5.94), although the difference from the sum-common (6.16) and the Ukkonen measure (6.08) is not large. This is consistent with the observations by Uitdenbogerd and Zobel (1999, 2000) who found coordinate matching and sum-common to be the best overall *n*-gram measures.

In our data the *n*-grams based on intervals gave better results than *n*-grams using fuzzified interval classes and much better results than interval directions only, which proved to be the poorest measure of this family. The superiority of the *n*-grams consisting of intervals was true as well for the data of the control experiment. In any case *n*-grams of length

three gave the highest similarity values and which were interpreted as the true similarity values. This superiority of short *n*-grams comes from the fact, that shorter *n*-grams have a higher probability to match, because of the lesser number of possible *n*-grams.

Concerning the rhythm measures, the one using edit-distance on the sequence of fuzzified rhythm classes was best with both data sets (distance to ratings in control experiment: 3.65), outperforming the Gaussian measures (3.81) and the *n*-grams fuzzy rhythm measures (3.92 to 4.64). As for the interval-based *n*-gram measures in the control experiment, the best rhythm *n*-gram measure was the Ukkonen measure, which counts the differences in rhythmic *n*-grams in both melodies.

Compared to over all similarity measures that were employed with durationally weighted and unweighted options, these two types of options reached the same mean rank in a U-Test. So it is not possible to decide whether weighted or unweighted measures perform better in general. But looking at the distance table for the main experiment in the appendix produces a clearer picture for the individual measures: The edit-distance on the raw pitch information performed better with the weighting for durations, while correlation measures for raw pitch data tend to perform better without weighting. Lastly, the Fourier measures seem slightly better with weighting.

8.5   Discussion

The aims of these study were to systematize measures for melodic similarity and to find the "optimal" similarity measures. The optimal measure should mirror the similarity ratings of human music experts most closely. To this end two rating experiments were conducted. In the first experiment, music experts were asked to judge the similarity between a constant reference melody and several variants differing in type and number of errors. The second experiment collected human rating data for a constant reference melody and "real" variants proceeding from that reference melody as well as "wrong" variants having their origin in a different melody.

As results several observations that are central to the investigation and the modeling of human similarity judgments were made. The music experts showed a very high correlation among their similarity ratings. An inter-subject correlation as high as 0.978 (as measured by Cronbach's α) was found. This high reliability was corroborated by ratings of the same variants that were done by a different subject group in a different experimental context in Experiment 2. The significant correlation between these two groups' ratings of the same variants was 0.862. This leads to the conclusion that there is something like "true" similarity and that at least subjects with a strong musical background (music experts) can reliably estimate this true similarity. A reliable and stable similarity representation is a necessary condition for modeling human judgments.

The modeling started from the idea that comparing different aspects of the melodies would give better results than just employing one algorithmic measure. With the means of regression analysis an optimal compound measure was found for the main experiment. This measure includes the durationally weighted edit-distance and the coordinate matching of *n*-grams of length 3 based on intervals. This compound similarity measure came 28.5% closer to the human ratings than the best single measure. For the different task in experiment 2 this combined measure still gave good results, but a combination of the Ukkonen measure for counting *n*-grams of intervals, the edit-distance of sequences of duration classes, and the edit-distance of the sequence of main tonalities of the individual bars was able to explain even more of the variance in the subjects ratings (92%) and was 33.4% better than the best single measure.

From this outcome it seems obvious that for different experimental tasks, different types and lengths of melodies, and probably different subject populations, the individual similarity measures and their specific combinations vary in their predictive power. So, for applications employing melodic similarity measures this means that the optimal combination and optimal weights for that specific task (e.g. query-by humming routines, symbolic searches in a melody database, the establishment of relations between melodic variants in ethno-musicological corpora) can be obtained. At the same time it would be interesting to find a compound similarity measure that is most robust and most sensible under a broad range of different conditions, at least in one music culture. But to find this "best" similarity model for example for western popular music there is still much more experimental work to be carried out.

Apart from the experimental data that is needed to complete this investigation, it is also clear that the number of tested similarity measures, although already quite large, is still incomplete. So the next steps should be to incorporate some more measures of the same classes that have been systematized here. Among these measures are the TF-IDF from the *n*-gram family that was used by Downie (1998) and Uitdenbogerd (2002) and an *n*-gram measure that combines intervals and rhythm classes like the one proposed by Doraisamy and Rüger (2001). Also, more sophisticated measures for some aspects like harmonic content are possible, like the ones described by Temperley (1997) or Krumhansl and Toiviainen (2001). A melodic dimension that has not been used so far is meter or metric information that might serve to weight pitch information. Accent models like the ones by Boltz and Jones (1989) or Monahan et al.(1987) may give a more complete and a more musical picture of the entire information that melodies contain. A very promising class of similarity measures are based upon the Hidden Markov Model approach (e.g. Prado and Birmingham, 2003). This class of measures is desperately waiting to be compared in its predictive power to the measures tested in this report.

Finally, the greatest simplification of this study is that the two melodies to be compared are of approximately the same length. But in most applications that use melodic similarity measures, a shorter query is compared to a number of much longer reference melodies. So most of the tested measures have to be adjusted to this type of task. With some measures like edit-distances the best adjustment is already known (local alignment in the case of edit-distance). With others like difference or correlational measures there are several possibilities to realize this adjustment. One promising solution is the segmentation to individual phrases and their indexing of both the query and the reference melodies. But again for segmentation there are several possibilities such as accent or Gestalt laws, or models that look for repetition of melodic or rhythmic content. The choice of an adequate strategy would again depend on the type of task and data to be analyzed and compared for similarity.

Appendices

1. Table of Model Names and Abbreviations

Abbreviation	Model
RAWED	Raw pitch edit-distance
RAWEDW	Raw pitch edit-distance weighted
RAWPCST	Raw pitch P-B. corr, weighted, 0-1
RAWPCWST	Raw pitch P-B. Corr. weighted, 0-1
CONSED	Contour (Steinbeck) edit-distance
CONSPCST	Contour (Steinbeck), P-B. corr., 0-1
CONED	Fraunhofer qbh-measure (June 2003)
CONPCST	Raw pitch edit-distance weighted
FOURRST	Fourier (ranks), weighted, 0-1
FOURRWST	Fourier (ranks), weighted, 0-1
FOURRI	Fourier (ranks, intervals)
DIFFED	Intervals (edit-distance)
DIFF	Intervals (mean difference)
DIFFEXP	Intervals (mean difference, exp.)
DIFFFUZ	Intervals (fuzzy), edit-distance
DIFFFUZC	Intervals (fuzzy contour)
NGRSUMCO	*n*-grams Sum Common
NGRUKKON	*n*-grams Ukkonen
NGRCOORD	Coordinate Matching (count distinct)
NGRSUMCR	Sum Common (interval direction)
NGRUKKOR	*n*-grams Ukkonen (interval direction)
NGRCOORR	*n*-grams Coord. Match. (interval direction)
NGRSUMCF	*n*-grams Sum Common (fuzzy)
NGRUKKOF	*n*-grams Ukkonen (fuzzy)
NGRCOORF	*n*-grams Count distinct (fuzzy)
NGRSUMFR	*n*-grams sum common (fuzzy rhythm)
NGRUKKFR	*n*-grams Ukkonen (fuzzy rhythm)
NGRCOOFR	*n*-grams Coord. Match. (fuzzy rhythm)
RHYTGAUS	Rhythm (gaussified onset points)
RHYTFUZZ	Rhythm (fuzzy), edit-distance
HARMCORR	Harmonic correlation (type I)

Abbreviation	Model
HARMCORK	Harmonic correlation (type II)
HARMCORE	Harmonic correlation (edit-distance)
HARMCORC	Harmonic correlation (circle)

2. Euclidean Distances between tested Similarity Measures and Subjects' Ratings

Experiment 1.

SUBJMEAN	0	DIFFFUZ	8.894
OPTI1	3.789	RAWPCST	9.182
CONED	5.292	RAWPCWST	9.481
RAWEDW	5.633	FOURRWST	9.720
RAWED	5.804	CONSPCST	9.844
NGRCOORD	5.940	NGRCOORR	9.923
NGRUKKON	6.084	FOURRST	10.105
NGRSUMCO	6.165	NGRUKKOR	10.132
HARMCORE	6.176	DIFF	10.173
DIFFED	6.301	NGRSUMCR	10.380
NGRUKKOF	6.684	RHYTFUZZ	10.431
NGRCOORF	6.718	RHYTGAUS	10.695
NGRSUMCF	7.211	HARMCORC	11.047
CONSED	7.295	NGRUKKFR	11.053
HARMCORK	7.586	NGRCOOFR	11.115
DIFFFUZC	7.744	NGRSUMFR	11.183
CONPCST	7.819	HARMCORR	12.796
DIFFEXP	7.943		

Experiment 2 ("real" variants only)

SUBJMEAN	0	NGRUKKOF	2.024	RAWPCST	3.015
OPTI2	1.403	CONSED	2.105	FOURRWST	3.039
DIFFED	1.296	DIFFFUZ	2.217	NGRCOOFR	3.089
RAWED	1.593	CONSPCST	2.331	RAWPCWST	3.218
RAWEDW	1.771	HARMCORK	2.342	NGRUKKFR	3.418
NGRSUMCO	1.880	CONED	2.346	NGRSUMFR	3.506
NGRSUMCF	1.900	NGRUKKOR	2.500	CONPCST	3.819
NGRCOORF	1.964	NGRCOORR	2.561	RHYTFUZZ	3.881
HARMCORE	1.984	NGRSUMCR	2.574	HARMCORC	4.269
NGRUKKON	1.986	DIFFEXP	2.702	RHYTGAUS	4.319
DIFF	1.991	DIFFUZC	2.834	HARMCORR	5.960
NGRCOORD	2.019	FOURRST	2.895		

Experiment 2 (all variants)

SUBJMEAN	0	HARMCORE	2.977	HARMCORC	4.94265794
OPTI3	1.403	NGRSUMCR	3.301	RAWPCWST	4.95775163
DIFFED	2.042	NGRCOORR	3.330	DIFFFUZC	5.25830586
RAWED	2.196	CONSED	3.572	FOURRWST	5.3029391
NGRUKKON	2.439	CONED	3.629	FOURRST	5.45782753
NGRCOORD	2.505	RHYTFUZZ	3.649	RAWPCST	5.49469773
NGRSUMCO	2.513	RHYTGAUS	3.805	CONPCST	5.55753837
NGRCOORF	2.5886	NGRUKKFR	3.91766035	DIFF	5.98850934
NGRUKKOF	2.599	HARMCORK	4.07437482	CONSPCST	6.06599364
RAWEDW	2.6612	NGRSUMFR	4.12671297	HARMCORR	7.72778626
NGRUKKOR	2.788	DIFFEXP	4.15603486	HARMCORC	4.94265794
NGRSUMCF	2.801	DIFFFUZ	4.2713198	RAWPCWST	4.95775163

References

Baroni, Mario, Rossana Dalmonte, and Carlo Jacoboni. (1992). "Theory and Analysis of European Melody". *Computer Representations and Models in Music*. Ed. Alan Marsden and Anthony Pople (San Diego: Academic Press), 187-206.

Bartók, Béla, and Albert B. Lord. (1951). Serbo-Croatian Folk Songs: Texts and Transcriptions of Seventy-Five Folk Songs from the Milman Parry Collection and a Morphology of Serbo-Croatian Folk Melodies. New York: Columbia University Press.

Bartók, Béla. (1976). "Why and How Do We Collect Folk Music?". *Béla Bartók Essays,* ed. Benjamin Suchoff (London: Faber & Faber), 9-24.

Cronin, Charles. (1998). "Concepts of Melodic Similarity in Music-Copyright Infringement Suits." *Melodic Similarity: Concepts, Procedures, and Applications* (*Computing in Musicology* 11), ed Walter B. Hewlett and Eleanor Selfridge-Field (Cambridge: MIT Press), 187-210.

Deliège, Irène. (2001). "Introduction : Similarity Perception, Categorization, Cue Abstraction". *Music Perception*, 18/3, 233-243.

Dewitt, Lucinda A. & Crowder, Robert G. (1986). "Recognition of Novel Melodies after Brief Delays,". *Music Perception*, 3/3, 259-274.

Dowling, W. Jay & Fujitani, Diane S. (1971). "Contour, Interval, and Pitch Recognition in Memory for Melodies," *The Journal of the Acoustical Society of America*, 2/2, 524-531.

Dowling, W. Jay (1978). "Scale and Contour: Two Components of a Theory of Memory for Melodies," *Psychological Review*, 85/4, 341-354.

Dowling, W. Jay, Seyeul Kwak, and Melinda W. Andrews. (1995). "The time course of recognition of novel melodies". *Perception & Psychophysics*, 57 (2), 136-149.

Dowling, W. Jay, Tillmann, Barbara, and Dan F. Ayers (*2002).* "Memory and the Experience of Hearing Music,". *Music Perception*, 19/2, 249-276.

Downie, J. Stephen (1999). "Evaluating a Simple Approach to Musical Informa-
tion Retrieval: Conceiving Melodic N-grams as Text." Ph. D. thesis, Univer-
sity of Western Ontario.

Drake, Carolyn, and Daisy Bertrand.(2001). "The Quest for Universals in Tempo-
ral Music Processing," *The Biological Foundations of Music,* ed. Robert J. Za-
torre and Isabelle Peretz (New York: New York Academy of Sciences), 17-27.

Edworthy, Judy. (1983). "Towards a Contour-Pitch Continuum Theory of Mem-
ory for Melodies," *The Acquisition of Symbolic Skills,* ed. D. Rodgers and J.
Sloboda (New York: Plenum), 263-271.

Eerola, T., T. Järvinen, J. Louhivuori, and P. Toiviainen (2001). "Statistical Fea-
tures and Perceived Similarity of Folk Melodies," *Music Perception*, 18/3, 275-
296.

Frieler, Klaus. (2004) "Mathematische Musikanalyse: Theorie und Praxis," Ph.D.
thesis, University of Hamburg (in preparation).

Hofmann-Engl, Ludger. (2001). "Towards a Cognitive Model of Melodic Similar-
ity,". *ISMIR 2001 Conference Proceedings* (Bloomington, IN).

Hofmann-Engl, Ludger. (2002). "Rhythmic Similarity: A Theoretical and Empiri-
cal Approach," *Proceedings of the 7th International Conference on Music Percep-
tion and Cognition*, Sydney 2002 [CD-ROM], ed. C. Stevens, D. Burnham, G.
McPherson, E. Schubert, J. Renwick. Adelaide, Causal Productions.

Idson, Wendy L., and Dominic W. Massaro (1978). "A Bidimensional Model of
Pitch in the Recognition of Melodies," *Perception and Psychophysics*, 24/6, 551-
565.

Jesser, Barbara (1990). *Interaktive Melodieanalyse: Methodik und Anwendung compu-
tergestützter Analyseverfahren in Musikethnologie und Volksliedforschung: typolo-
gische Untersuchung der Balladensamlung des DVA.* Bern: Peter Lang.

Jones, Mari Riess, and Marilyn Boltz (1989). "Dynamic Attending and Responses
to Time," *Psychological Review*, 96/3, 459-491.

Juhasz, Zoltán (2000). "A Model of Variation in the Music of a Hungarian Ethnic
Group,". *Journal of New Music Research*, 29/2, 159-172.

Kauffman, William H., and James C. Carlsen (1989). "Memory for Intact Music
Works: The Importance of Music Expertise and Retention Interval". *Psycho-
musicology,* 8/1, 3-20.

Kluge, Reiner (1974). *Faktorenanalytischen Typenbestimmung an Volksliedmelodien*.
Leipzig: VEB Deutscher Verlag für Musik.

Krumhansl, Carol L., and E. J. Kessler (1982). "Tracing the Dynamic Changes in
Perceived Tonal Organization in a Spatial Representation of Musical Keys,"
Psychological Review, 89, 334-368.

Krumhansl, Carol L. (1990). *Cognitive Foundations of Musical Pitch*. New York: Ox-
ford University Press.

Krumhansl, Carol L., and Petri Toiviainen (2001). "Tonal Cognition". in *The Bio-
logical Foundations of Music,* ed. Robert J. Zatorre & Isabelle Peretz (New
York: New York Academy of Sciences), 77-91.

Lamont, Alexandra, and Nicola Dibben, Nicola. (2001). "Motivic Structure and

the Perception of Similarity,". *Music Perception,* 18/3, 245-274.

Lerdahl, Fred, and Ray Jackendoff. (1983). *A Generative Theory of Tonal Music*. Cambrdige: MIT Press.

Massaro, Dominic W., Howard J. Kallman, and Janet L. Kelly (1980). "The Role of Tone Height, Melodic Contour, and Tone Chroma in Melody Recognition". *Journal of Experimental Psychology: Human Learning and Memory,* 6/1, 91-105.

McAdams, Stephen, and Matzkin, Daniel. (2001). "Similarity, Invariance, and Musical Variation" in *The Biological Foundations of Music,* ed.. Robert J. Zatorre and Isabelle Peretz (New York: New York Academy of Sciences), 62-74.

McNab, Rodger J., Lloyd A. Smith, Ian H. Witten, Clare L. Henderson, and Sally Jo Cunningham (1996). "Towards the Digital Music Library: Tune Retrieval from Acoustic Input,". *Proceedings ACM Digital Libraries*.

Meek, Colin, and Birmingham, William (2002). "Johnny Can't Sing: A Comprehensive Error Model for Sung Music Queries," *ISMIR 2002 Conference Proceedings*, IRCAM, 124-132.

Meyer, Leonard B. (1973). *Explaining Music: Essays and Explorations*. Chicago: University of Chicago Press.

Monahan, Caroline B., Roger A. Kendall, and Edward C. Carterette (1987). "The Effect of Melodic and Temporal Contour on Recognition Memory for Pitch Change," *Perception and Psychophysics,* 41/6, 576-600.

Mongeau, Marcel, and David Sankoff. (1990). "Comparision of Musical Sequences". *Computers and the Humanities,* 24, 161-175.

Müllensiefen, Daniel. (2004). *Varianz und Konstanz von Melodien in der Erinnerung.* PhD thesis, University of Hamburg (in preparation).

Narmour, Eugene (1990). *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model.* Chicago: The University of Chicago Press.

O`Maidin, Donncha (1998). "A Geometrical Algorithm for Melodic Difference in Melodic Similarity,".*Melodic Similarity: Concepts, Procedures, and Applications. (Computing in Musicology 11),* ed. Walter B. Hewlett and Eleanor Selfridge-Field. Cambridge: The MIT Press.

Oura, Yoko, and Giyoo Hatano (1988). "Memory for Melodies among Subjects Differing in Age and Experience in Music,". *Psychology of Music,* 16, 91-109.

Pardo, Brian, Jonah Shifrin, and William Birmingham (2004). "Name that Tune: A Pilot Study in Finding a Melody from a Sung Query," *Journal of the American Society for Information Science and Technology* 55/4.

Pauws, Steffen (2002). "Cuby Hum: A Fully Operational Query-by-Humming System". *ISMIR 2002 Conference Proceedings*, IRCAM, 187-196.

Scherrer, Deborah K., and Philip H. Scherrer (1971). "An Experiment in the Computer Measurement of Melodic Variations in Folksong," *Journal of American Folklore*, 84.

Schmuckler, Mark A. (1999). Testing Models of Melodic Contour Similarity." *Music Perception* Vol. 16, No. 3, 109-150.

Seeger, Charles (1966). "Versions and Variants of the Tunes of 'Barbara Allen'," (*Selected Reports in Ethnomusicology*, 1/1*)*.

Sloboda, John A., and David H. H. Parker (1985). "Immediate Recall of Melodies,". *Musical Structure and Cognition,* ed*.* Ian Cross, Peter Howell, and Robert West (London: Academic Press), 143-167.

Steinbeck, Wolfram (1982). *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse* (*Kieler Schriften zur Musikwissenschaft*, XXV). Kassel, Basel, London: Bärenreiter.

Temperley, David (1997). "An Algorithm for Harmonic Analysis,". *Music Perception,* 15/1, 31-68.

Uitdenbogerd, Alexandra, and Zobel, Justin (1999). "Matching Techniques for Large Music Databases". *Proceedings of the ACM Multimedia Conference*, Orlando, Florida, 57-66.

Uitdenbogerd, Alexandra, and Justin Zobel (2000). "Music Ranking Techniques Evaluated,". *ISMIR 2000 Conference Proceeding*, Plymouth, MA.

Uitdenbogerd, Alexandra L. (2002). "Music Information Retrieval Technology." Ph. D. thesis, RMIT University Melbourne Victoria, Australia.

Wiggins, Geraint A. (2002). "Hourses for Courses, or How to Change an Algorithm to do what you Need," *ISMIR 2002 Conference Proceedings*, IRCAM, 308.

Wiora, Walter (1941). "Systematik der musikalischen Erscheinungen des Umsingens," *Jahrbuch für Volksliedforschung,* 7, 128-195.

Zadeh, Lofti (1965). "Fuzzy Sets,". *Inf. Control*, 338-353.

Zhou, Yongwei, and Mohan S. Kankanhalli (2003). "Melody Alignment and Similarity Metric for Content-Based Music Retrieval,". *Proceedings of SPIE-IS&T Electronic Imaging,* SPIE, 5021, 112-121.

Zielinska, Halina, and Kacper Miklaszewski (1992). "Memorising Two Melodies of Different Style,". *Psychology of Music*, 20, 95-111.