Modelling experts' notions of melodic similarity

DANIEL MÜLLENSIEFEN* AND KLAUS FRIELER**
* Department of Computing, Goldsmiths College, University of London
** Institute of Musicology, University of Hamburg

• ABSTRACT

In this article we show that a subgroup of music experts has a reliable and consistent notion of melodic similarity, and that this notion can be measured with satisfactory precision. Our measurements enable us to model the similarity ratings of music experts by automated and algorithmic means. A large number of algorithmic similarity measure found in the literature were mathematically systematised and implemented. The best similarity algorithms compared to human experts were chosen and optimised by statistical means according to different contexts. A multidimensional scaling model of the algorithmic similarity measures is constructed to give an overiew over the different musical dimensions reflected by these measures. We show some examples where this optimised methods could be successfully applied to real world problems like folk song categorisation and analysis, and discuss further applications and implications.

1. INTRODUCTION

There are only a few studies that compare similarity algorithms for melodies with each other and with human expert judgements. The most prominent studies are: Schmuckler (1999), Eerola *et al.* (2001), McAdams and Matzkin (2001), and Hofmann-Engl (2002). So, at the outset of this study it was by no means clear whether there is an inter-subjective reliable notion of similarity between melodies, and, if there were, which algorithmic approach would model human judgements best. As there was a need for a good similarity algorithm as an analysing tool in another study on melodic memory (Müllensiefen, 2004), the first step was an exhaustive literature survey on the existing approaches to measuring the similarity of melodies (as summarised in Müllensiefen and Frieler, 2004b).

While reviewing the literature on similarity measurement of the last two decades, it became clear that it was not the lack of measurement procedures, but their abundance that was a cause for concern. Several very different techniques for defining and computing melodic similarity have been proposed that cover distinct aspects or elements of melodies. Among these aspects are intervals, contour, rhythm,

and tonality, each 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 (McNab *et al.*, 1996; Uitdenbogerd, 2002), N-grams (Downie, 1999), correlation and difference coefficients (Steinbeck, 1982; O'Maidin, 1998; Schmuckler, 1999), and hidden Markov models (Meek and Birmingham, 2002). There are plenty of examples of successful applications of these specific similarity measures.

So the questions arose, which representation and which similarity measures would be the most adequate ones. An optimal similarity measure would probably be the mean rating of a group of music experts. But such a group of experts is not always at hand, so the idea of the experiments reported in this paper was to model expert ratings with some of the measurement techniques mentioned above. Three rating experiments were conducted to compare expert judgements with the results of different similarity algorithms. The optimal measure would be the one that predicts the expert assessments best.

This paper first outlines the mathematical framework that we created to systematise the different transformation and computing approaches. Then the experiments are described and a model of the experts' ratings depending on the manipulated experimental parameters is developed. After that a more general approach to modelling is adopted that can be used in non-experimental situations where the characteristics and differences between the melodies are not created by an investigator. This modelling approach comes up with an optimised combination of different similarity algorithms from a linear regression analysis. The last section demonstrates one application of this optimised measure in the field of ethnomusicology and shows the capabilities of the software toolkit SIMILE that was developed in the course of this study to compare different algorithms and to analyse large melody collections in a comfortable and easy way.

2. SIMILARITY MEASURES AND DATA TRANSFORMATIONS

This section gives a brief overview over the mathematical framework, the transformations employed with melodic data and the different similarity algorithms that we used. A much more detailed description of the employed approaches and algorithms can be found in Müllensiefen and Frieler (2004b).

2.1. MATHEMATICAL FRAMEWORK
In order to handle the huge amount of different similarity measures found in the literature we found it necessary to develop a mathematical framework, which gives a systematisation and classification of the similarity measures in a compact and unified way, and which allows the different models to be compared with each other and with the empirical data. Furthermore, it served as a kind of construction kit and as a

source of inspiration for new similarity measures. Last but not least it was very helpful for implementing the algorithms into our software.

The first step is to define an abstract melodic space **M** as a subset of the Cartesian product of a real-valued time coordinate representing onsets and an integer- or real-valued pitch coordinate.

A similarity measure can then be defined as a map

$$s: \mathbf{M} \times \mathbf{M} \to [0,1]$$

with the following properties:
1. Symmetry: $s(m, n) = s(n, m)$
2. Self identity: $s(m, m) = 1$
3. Transposition-, Time translation- and time dilation invariance.

Transposition means translation of the pitch coordinate, *Translation* is time shift, and time *dilation* means tempo change (time stretch). These properties are inspired by psychological reality; at least approximately similarity judgements do not depend on absolute pitch and tempo of the melodies, and for many applications these are desirable or even necessary. Similarity measures form a convex set, *i.e.*, any linear combination of similarity measures, where the sum of coefficients equals 1, is again a similarity measure. This property enabled us to calculate combined, optimal measures, by means of linear regression. Furthermore, any product of two similarity measures is again a similarity measure.

2.2. DATA TRANSFORMATIONS

Most of the similarity measures involve the following processing stages:
1. Basic transformations (Representations)
2. Main Transformations
3. Computation

The most common basic transformations are projection, restriction/composition and *differentiation*. Projections can be either on the time or pitch coordinate, (with a clear preference for pitch projections found in the literature). Differentiation means using coordinate differences instead of absolute coordinates, *i.e.,* intervals and durations instead of pitch and onsets.

Among the main transformations rhythmical weighting, *fuzzification* (classification) and *contourisation* are the most important ones. Rhythmical weighting can be done for quantised melodies, *i.e.,* melodies where the durations are integer multiples of a smallest time unit *T*. Then each pitch of duration *nT* can be substituted by a sequence of *n* equal tones with duration *T* each. After a pitch projection the weighted sequence will then still reflect the rhythmical structure. The concept of rhythmical weighting has been widely used in other studies (*e.g.,* Steinbeck, 1982; Juhász, 2000; Hofmann-Engl, 2002).

Fuzzifications are based on the notion of fuzzy sets (Zadeh, 1965), *i.e.,* sets,

where an element belongs to it with a certain degree between 0 and 1. But if the basic set is decomposed into mutually disjoint subsets, the fuzzifications reduce to classifications, as they did in all our cases. Other studies exploited this idea in very similar ways (*e.g.,* Pauws 2002).

Contourisation is based on the idea that the extremes, *i.e.,* the turning points of a melody, are the most important notes. By taking the extremes (which to take depends on the model) and substituting the pitches in between with interpolation values, *e.g.,* coming from linear interpolation (which we used exclusively), one gets a contourisation of the melody. The contourisation idea was employed, for example, in the similarity measures by Steinbeck (1982) and Zhou and Kankanhalli (2003). We used two different kinds of contourisation, one as proposed by Steinbeck and one, which only excludes classical returning notes from the set of extremes.

Among the other core transformations are the ranking of pitches and Fourier transformation on contour information (following the approach of Schmuckler, 1999) or methods of assigning a harmonic vector to certain subsets (bars) of a melody, just to name a few (Krumhansl, 1990).

2.3. SIMILARITY MEASURES

The last stage of processing is the computation of a similarity value. The measures we used can be classified in three categories: Vector measures, symbolic measures and musical (mixed) measures, according to the computational algorithm used. The vector measures treat the transformed melodies as vectors in a suitable real vector space, hence methods like scalar products and various other means of correlation are now applicable. On the other hand the symbolic measures view melodies as strings, *i.e.,* sequences of symbols. For these, well-known measures like Edit Distance (*e.g.,* Mongeau & Sankoff, 1990) or n-gram-related ones (*e.g.,* Downie, 1999) can be used. The musical or mixed measures typically involve more or less specific musical knowledge and the computations involved are of vector or symbolic type.

Some general problems had to be solved for some models to ensure transposition and tempo invariance and to account for melodies having different lengths (*i.e.,* number of notes). If a measure is not transposition invariant a priori, one can achieve this always by taking the maximum over all similarities of all possible transpositions (within certain limits). Likewise, for models, which need the melodies to be of same length (as most of the correlation measures do), we took the maximum of all similarities of sub-melodies of the longer melody with the same length as the shorter one. This type of shifting has been proposed for example by Leppig (1987). Tempo invariance is generally no problem while using quantised melodies.

In sum, we tried to consider most of the main techniques for melodic data transformation and similarity measurement proposed in the literature of the last 15 years. Additionally, systemising these approaches led to the construction of several new similarity measures (see Frieler (in preparation) and Müllensiefen (2004) for a detailed description). At the end we implemented a total number of 48 different

similarity measures (counting all variants) into our software from which 34 were used in the analysis. A list of all 34 similarity measures with short explanations of the abbreviations used in the remainder of this text can be found in the appendix. More detailed descriptions of the similarity measures may be found in Müllensiefen and Frieler (2004b). For the experiments and for our program we used the same MIDI-files. All melodies were quantised.

3. LISTENER EXPERIMENTS

3.1. SUBJECTS, MATERIALS, AND PROCEDURE

We conducted three similarity rating experiments with human expert subjects. Experiment 1 was designed as a test-retest-experiment with one week in between the two test sessions. The aims of the experiments were a) to see if there was a consistent notion among music experts of what similarity between melodies means, and b) to generate reference data that could be used for modelling the experts' judgements with algorithmic measures. A more detailed description of the procedures for on the three experiments can be found in Müllensiefen and Frieler (2004b).

3.1.1. Subjects

A pre-test with subjects with little or no music background showed that similarity judgements from many subjects were unstable and not consistent over time. So we decided to recruit musicology students from the University of Hamburg as subjects. The subjects' musical activity background was measured by an extensive questionnaire very similar to the one employed by Meinz and Salthouse (1998). As is typical for musicology students they have a long history in music making (e.g. mean years for playing an instrument were 12; mean months of paid instrumental lessons were 71), but their most active musical phase is several years back, which is reflected in less time spent for current musical activities compared to the most active musical phase in the past.

3.1.2. Materials

To obtain results that can be applied to the domain of modern western popular music, 14 melodies from western Pop songs were chosen as stimulus material. For each melody six comparison variants with errors were constructed resulting in 84 variants of the 14 original melodies.

Five error types with their respective probabilities were defined: Rhythm errors ($p = 0.6$), pitch errors leaving the contour intact ($p = 0.4$), 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; $p = 0.2$). Each 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. As an example the reference melody *n* of the song "I want it that way" as interpreted by the *Backstreet Boys* and the variant *n6* are depicted below in fig. 1 and fig. 2. The variant *n6* contains nine rhythm errors and two modulation errors that can be easily found. Omissions of notes are counted as rhythm errors.

Figure 1.
Example of reference melody No. n: "It want it that way" by the Backstreet Boys.

Figure 2.
Example of variant n6 to reference melody No. n.

3.1.3. Procedure and Design

In the first experiment, subjects had to judge pairs of a reference melody and its six respective variants on a 7-point scale with "1" meaning very dissimilar and "7" meaning very similar/identical. In one session three, four or five different reference melodies with their respective variants were tested. So the subjects had to rate 18 to 30 pairs per session and the length of the sessions varied from 15 to 20 minutes.

The second and third experiment served as control experiments. Both experiments consisted of only one session. In the second experiment two melodies from the first experiment were chosen and presented along with the original 6 variants plus 6 resp. 5 variants, which had their origin in completely different melodies. Participants rated 24 melody pairs and the duration of the session was about 17 minutes.

As Experiment 1, the third experiment tested only the similarity of reference melodies and variants that had their origin in the reference melody. But a different error distribution for the variants was tested, and all variants were randomly transposed with respect to the reference melody. So it was possible to look for the effects of the transposition of the variants, although subjects were told to ignore the absolute pitch level of the variants. Four different reference melodies with eight variants each were tested. The duration of the test session with these 32 melody pairs was about 23 minutes.

3.2. RESULTS

To be selected as 'expert' for the upcoming data analysis a single subject had to fulfil two constraints:

1) The subject should rate the same pair of reference melody and variant similar in the test as in the retest session (Experiment 1) or to the same pair of reference melody and variant that were presented twice in a single test session (Experiment 2 and 3). A value of not less than 0.5 as measured by Kendall τ_b was the criterion to include the data of a subject.

2) Subjects should give very high similarity values to melody pairs where reference melody and variant were identical. Values of 6 or 7 on the 7-point rating scale should be assigned to these melody pairs.

Of 82 subjects participating in the first experiment, the data of only 23 could be selected on the basis of these two criteria. For experiment 2 and 3, 12 out of 16 and 5 out of 10 participants, respectively, were selected. We consider this type of reliability measurement to be an important methodological improvement compared to earlier experiments involving similarity ratings.

After selecting the experts from the tested subjects, the first and very important result is the very high correlations between the experts' judgements. The coefficient Cronbach's alpha, which reflects the degree to which several variables (here: judgements) measure a latent magnitude, reached very high values of 0.962, 0.978 and 0.948 for the three experiments respectively. These values indicate the conformity of the experts' judgements and are quite outstanding, comparing them for example with the alpha values reached in a similar situation by Lamont and Dibben (2001). The Kaiser-Meyer-Olkin Measure (KMO) reflects very similarly 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 of experiment 1 and 0.811 for the subjects of experiment 2. So, the subjects that were selected only because their individual judgement showed a high reliability and consistency, judged all very much alike. This led us to assume, that there is something like a true similarity at least for the group of western musical experts, and that experts can estimate this true similarity quite reliably. This agreement among experts on what similarity between melodies means was considered as a necessary condition for comparing and optimising the different

similarity algorithms with a set of reference data. For the modelling of the expert ratings the data of the selected experts was collapsed by taking the mean.

3.3. MODELLING EXPERTS' RATINGS

To reduce the complexity of the linear regression approach and to guide a first-step variable selection, we assumed that the experts' notion of melodic similarity would exploit not more than five dimensions: Contour information, interval structure, harmonic content, rhythm and common characteristic motives. We classified all 34 measures to one of these five dimensions according to the kind of data and algorithm used in the measure. For each dimension the Euclidean distances of the included measures to the mean of the subjects' ratings were computed, and the best measure of each dimension was taken to serve as an independent variables for a linear regression. The dependent variables to be predicted were the means of the expert subjects' ratings over all 84 and 24 variants of Experiment 1 and 2 respectively. For further variable selection, the widely used stepwise procedure was applied to the five independent variables. This regression was done separately for experiment 1 and experiment 2.

The best five models for experiment 1 were (ordered according to their Euclidean distances, minimum first):
- *conEd* (Edit Distance of contourised melodies, classical contourisation algorithm)
- *rawEdw* (Edit Distance of rhythmically weighted raw pitch melodies)
- *nGrCoord* (coordinate matching based on count of distinct n-grams of melodies)
- *harmCorE* (Edit Distance of harmonic symbols per bar, obtained with the help of Krumhansl's tonality vectors)
- *rhytFuzz* (edit distance of classified length of melody tones)

And for experiment 2 (same ordering):
- *diffEd* (Edit Distance of intervals)
- *nGrSumCo* (based on count of common n-grams)
- *harmCorE* (cf. above)
- *conSEd* (Edit Distance for contourised melodies, Steinbeck's algorithm)
- *nGrCooFR* (based on count of distinct n-grams of classified note lengths)

From this input we obtained combined measures from a linear regression model, which were 28.5 % and 33.4 % better than the best single measure for each experiment. In experiment 3 where all variants were transposed and the combination and distribution of errors were altered, the combined measure from experiment 1 gave still superior results to the tested single algorithmic measures.

For predicting the similarity of the melody pairs from experiment 2 (context: variants from the original melody and variants from other melodies) a weighted combination of three different measures called *opti3* proved to be optimal:

$$opti3 = 3.027*ngrukkon + 2.502*rhytfuzz + 1.439*harmcore$$

The superior performance of the optimised hybrid measure can be seen in the following diagram in figure 3, which shows graphically the Euclidean distances
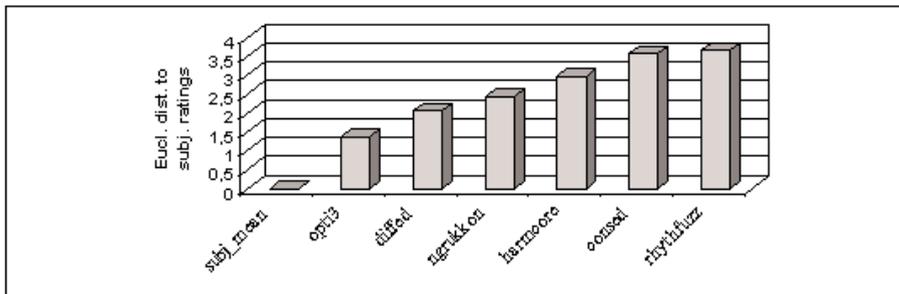
Figure 3.

Performance of different similarity measures on data from experiment 2; for explanation of acronyms see appendix.

between the mean of the subjects ratings and the predictions of the different measures, calculated over all 24 variants from experiment 2:

Interestingly, the combined model for the data of experiment 1 consisted of two measures that reflect pitch information only (*rawEdw* and *nGrCoord*), while for experiment 2 harmonic (*harmCore*) and rhythm measures (*rhythFuzz*) showed high explanatory power in addition to a pitch measure. This leads to the interpretation that in situations where the context of stimuli is heterogeneous — *i.e.,* subjects have to tell apart "right" and "wrong" variants — they make use of more sources of information like rhythmic information. For separating "right" and "wrong" variants with respect to a give reference melody *opti 3* is the optimal combination of the algorithms tested here. This hybrid measure will mainly be used in the applications presented in the sections below.

The combined or optimised models fit very well to the data. For experiment 1 there was 83 % of variance explained by the combined measure (R^2: 0.83, corrected R^2: 0.826), and for experiment 2 92 % (R^2: 0.92, corrected R^2: 0.909).

3.4. THE SPACE OF SIMILARITY MEASURES

To explore the important dimensions that differentiate between the tested similarity measures, and thus span the space of similarity measures, we performed two multidimensional scaling procedures (MDS) on the data of experiment 1 und experiment 2.

For the sake of clarity in the graphical representation, we chose the 18 measures with the least distance to the mean expert ratings to be included in the MDS along with the mean similarity ratings of the expert subjects. A 19×19 matrix of Euclidean distances between the 18 similarity measures and the experts' ratings was calculated on basis of the their similarity values for the 84 pairs of reference melody and respective variant from experiment 1. This distance matrix entered an ordinal MDS procedure. The ALSCAL algorithm was used with stress and RSQ as indicators of
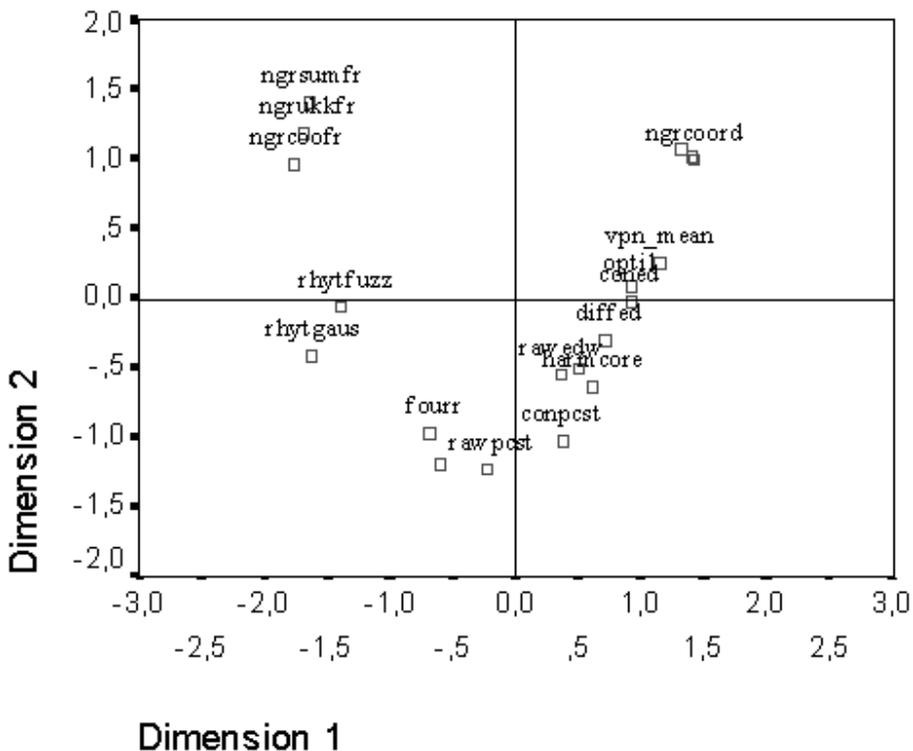
Figure 4.

Two-dimensional MDS-solution on data of experiment 1 (only 14 algorithmic measures are displayed with their names for better readability, vpn_mean denotes the mean of the subjects' ratings).

fit. As these two indicators did not show any rapid change with the reduction of dimensions and as their values for the two-dimensional solution were still quite good (Stress: 0.085; RSQ: 0.97), this two-dimensional solution was chosen and is depicted in figure 4.

A meaningful interpretation of this solution views dimension 1 as the degree to which the similarity measures incorporate rhythmical information: To the left measures from the n-gram approach are located that use the fuzzified rhythm (ending on *fr) values as data. The other two rhythmic measures *rhytGaus* (gaussification of onsets) und *rhythFuzz* (edit distance of fuzzified duration values) are located as well to the very left on this axis.

Dimension 2 can be interpreted as global vs. local information. The n-grams measures are located in the positive range of this dimension. The n-gram approach generally reflects only differences in short sequences of notes. In the negative half contour (*e.g., conpcst* = Pearson correlation on classical contourised melodies,

stretched to interval 0-1; *fourr* = Fourier transformation and contour ranks according to Schmuckler) and edit distance (*e.g., harmcore* = edit distance based on harmonic vectors according to Krumhansl) measures can be seen that give importance to the coherence of two melodies over their full course. As expected, the optimised measure from linear regression (*opti 1*) is located closest to the mean of subjects' ratings.

To see if this interpretation holds true with a different set of melodies (variants from the same reference melody and variants from completely different melodies) the same MDS procedure was carried out on the data of experiment 2. Again the 18 best algorithmic measures plus the subjects' mean ratings should be located in the similarity space. Again, the two-dimensional model gave high parameters of fit (stress = 0.075, RSQ = 0.98). And again, the two-dimensional solution could be very well interpreted as consisting of a rhythmical dimension and a global vs. local dimension, as can be seen from figure 5.
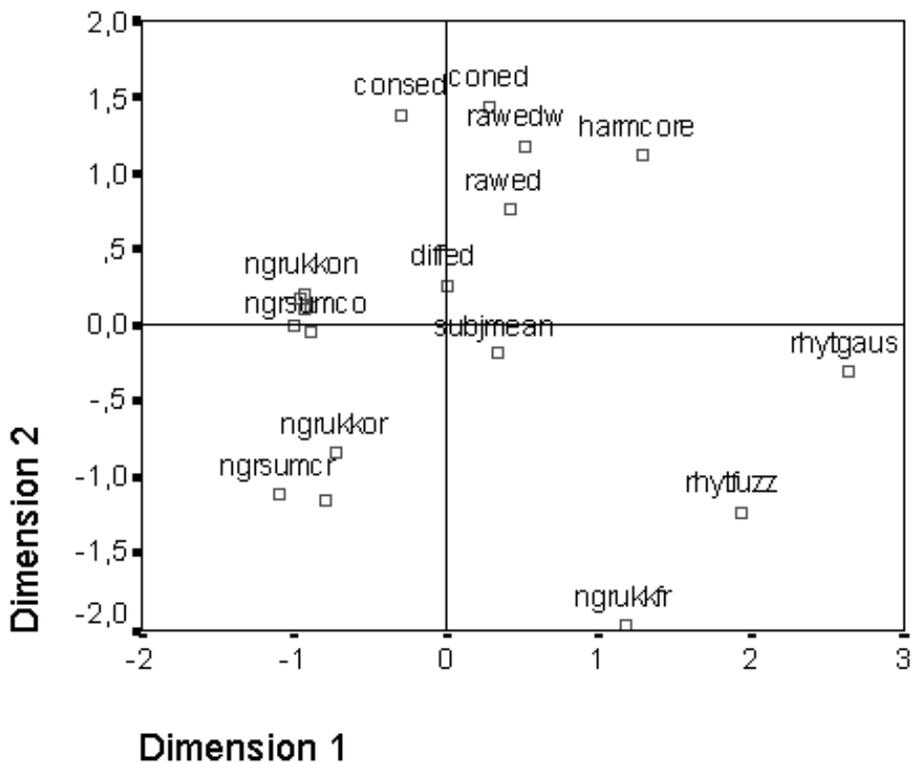

Figure 5.

Two-dimensional MDS-solution on data of experiment 2 (only 13 algorithmic measures are displayed with their names for better readability, vpn_mean denotes the mean of the subjects' ratings).

As the outcome of the two MDS procedures on different melody data give somehow similar results, it seems plausible to assume two main dimensions that differentiate the tested measures: the incorporation of rhythmic information and the reflection of local (motivic) vs. global similarity. As the solutions suggest, there is no obvious difference between contour and interval/pitch measures. This observation might be taken as a hypothesis for future research, and should be tested by a confirmatory MDS on new but comparable data.

So at least for music experts tested with rather short melodies it might be assumed for the moment, that differences in contour and intervals are about of same importance. This finds its corroboration in the optimised measure from linear regression, because there was no solution obtained that incorporated a contour measure as well as a pitch measure. As the tested measures might well be taken as a quasi-representative sample of the existing approaches to similarity measurement for melodies, this conclusion may be generalised to the broader field of existing literature on this topic.

It is furthermore interesting to observe the position of the subjects' ratings in both MDS solutions. Subjects seem to have a very balanced judgement with respect to dimension 2 (global vs. local) in both experiments. They almost give the same importance to both aspects with a slight advantage to the local pole. On dimension 1 the position of subjects differs more strongly between experiments 1 und 2. As was concluded from the outcome of the linear regression models, in experiment 2 rhythm plays a more important role for the subjects, if they have to judge in a context where melodies and variants are more heterogeneous. Subjects' ratings are located more to the pole of rhythm in experiment 2 and moved more to the pole of intervals/contour in experiment 1.

So the similarity judgement of experts seems to be a flexible concept that adapts to the specific experimental task or context of melodies. At the same time it seems to be a stable notion that can be very well agreed upon among experts in a given situation. As the good indicators of model fit and the good prediction values from the optimised models from experiment 1 and 2 show, experts' judgements can be very well predicted if the context of judging is known.

4. Similarity Analysis of a folksong collection

Now the question arose, how well do the optimised measures for melodic similarity perform in contexts beyond experimental data? Can they really substitute human experts' similarity judgements in every or at least some real situations? To test the optimal similarity measures with a different melody repertoire and to compare them to more analytical expert ratings, we referred to the comprehensive ethno-musicological study of Damien Sagrillo (1999). Sagrillo collected, analysed and classified phrases of 586 folksongs from Luxembourg from different sources manually and with the aid

of the computer, primarily for sorting purposes according to several parameters. His classification work is done with great experience in ethno-musicological practices concerning the treatment of large melody collections. He gives great emphasis to musically relevant features and details of the melodies and phrases. As we were provided with a digital copy of the melody catalogue in its classified form, we were able to test the performance of our algorithmic measures against Sagrillo's classification. As our optimised measure opti 3 came from an experiment with the context of variants and different songs, we used this measure almost exclusively for the analysis of the melodies from Luxembourg.

4.1. DOUBLETS AND VARIANTS

One crucial test for a similarity measure is the task of identifying doublets (duplicates) in a database. Unfortunately we had no complete information about doublets, but a suffix "V" in Sagrillo's catalogue index of the tunes indicated a variant for a specific melody. There were 19 of so marked tunes in the Luxembourg database, which we inspected manually. Four of them (L0027V, L0035V, S0073V, T0222V) were songs with the same lyrics, but different melodies, as stated in the remark section of the songs. For one song (K1086V) no counterpart was found, neither in the database nor by means of our similarity measure, so maybe this is a mistake of the collectors. The remaining 14 tunes were indicated as melody variants. Out of these 14 tunes 8 had an opti 3 similarity value above 0.8, 2 above 0.7 and 3 above 0.6, each with their corresponding original. Only one tune, L0039V, which was said to be a variant of L0038, had a similarity value of only 0.27. But a glance at the melodies revealed, that these two are in fact very different, *i.e.*, L0039V is much longer, with much more phrases than L0038 and a scale shift in the midst. When we tested the similarity of the beginning phrases alone, we got a value of 0.53.

We also examined all 49 pairs with similarity values above 0.6. These pairs can be roughly classified in

1. Doublets (same or near same melody and same or near same title): 37 pairs
2. Parodies (same or near same melody, but different title and probably different lyrics): 10 pairs
3. Recitatives: 2 pairs

The recitatives are a special case of songs, which are typically written without meter, consist almost completely of tone repetitions and have usually small tone range. The high similarity values here are to be expected, because all the measures involved in opti3 give high ratings for long sequences of tones of same length and pitch.

Some songs could be found with 3 or more variants. One example is a song called "De Malbrough", which can also be found in the collection from Lorraine. Furthermore, it is highly similar to the well-known (English) song "He's a jolly good fellow". This gives rise to the idea, that a broad cross-cultural survey of folk song databases could reveal songs that are known in different cultures or even some

melodic archetypes. Speed and convenience of automated similarity analysis could make new studies and research possible, like, *e.g.,* a melodic migration map, which would be too cumbersome to do manually.

4.2. THREE EXAMPLES

A look at the distribution of the 171,405 similarity values from a complete comparison of all melodies in the database shows that - very much like in a normal distribution-, very low and very high similarity values are extremely rare (fig. 6).
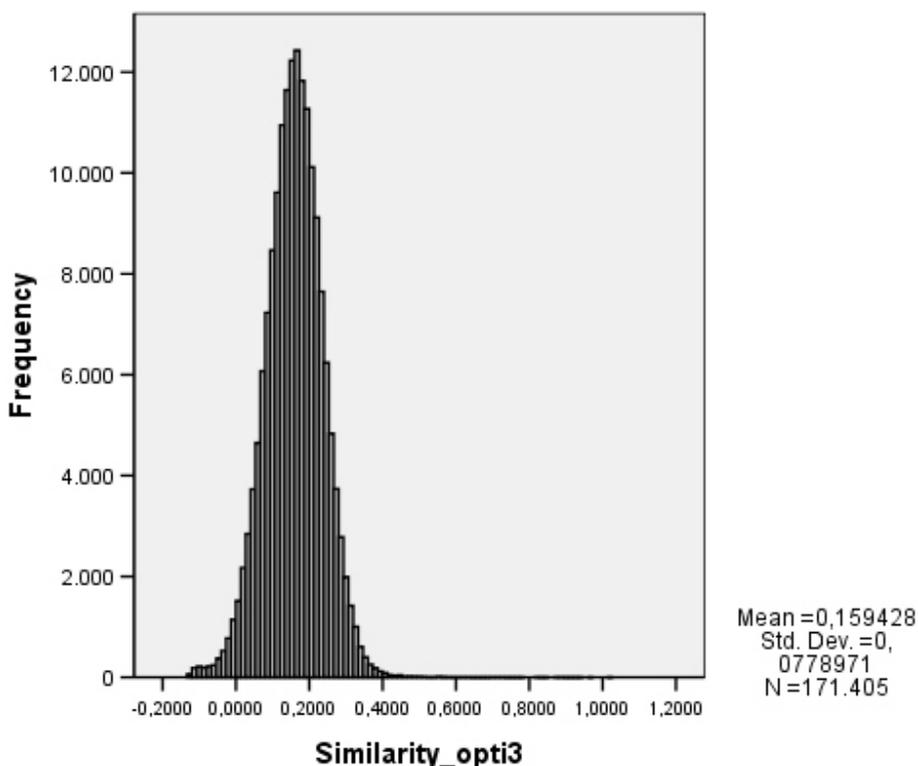

Figure 6.
Frequency distribution of similarity values (optimised measure opti3).

High similarity values above 0.4 can be expected in only 1 % of the cases. So our optimal measure can be used to scan a large database of songs for doublets, variants and other kind of interesting relationship in a reliable and valid way, and additionally in relatively short time, which would otherwise hardly be possible. To illustrate this, we will now have a short analytical look at three selected pairs of folk tunes with high

similarity, whereby we additionally gain some insight in the functioning of the *opti 3* measure.

"Die beiden Hasen" (K1185) vs "Plauderei an der Linde" (T0216)

The notation of these two songs is given in figure 7 and figure 8 respectively.

Figure 7.
Melody of "Die beiden Hasen" (K1185).

Figure 8.
Melody of "Plauderei an der Linde" (T0216).

The similarity value for these tunes with our optimal measure *opti 3* is 0.634. One easily sees that these two melodies are nearly identical, despite the fact that K1185 is notated in triple meter, and T0216 in duple meter. The melodies differ only in 5 notes, which are mostly passing tones. One might wonder why — facing this high structural accordance — the similarity value is that low. To understand this one has to look at the single values of the combined measure: *nGrUkkon* scores 0.71, *rhytFuzz* 0.95, but *harmCorE* is just 0.1. In contrast to that, the best single measure *diffEd* gives a similarity value of 0.87. Here the fact that the two tunes are in different meters comes into play, because *harmCorE* is calculated on a bar-wise base, which explains the low value. But the similarity value is nevertheless exceptionally high.

"Jetzt reisen wir zum Tor hinaus" (K0083) vs "Eng ongeheiesch Freiesch" (T0228)

Figure 9 and figure 10 give their melodies in conventional music notation.

Figure 9.
Melody of „Jetzt reisen wir zum Tor hinaus" (K0083).

Figure 10.
Melody of „Eng ongeheiesch Freiesch" (T0228).

Their opti*3* similarity value is 0.473. One first observes that K0083 is 4/4 meter, while T0228 is 6/8 meter. The melodies have different lengths, K0083 is 12-bars long and T0228 has 10 bars. A closer look on the first 6 bars of the two tunes reveals a nice structural relationship. They both start with an upbeat from the dominant to the tonic, which is repeated 5, resp. 4 times, then a 3-2-1 figure follows and the phrase ends on a long note on the second step of the scale. This takes 3 bars in K0083 but just 2 in T0228, because of the different meters. This phrase is repeated in T0228, but not in K0083. The next phrase of both songs is built like the first: upbeat followed by the same melodic motif, but now on the second step. The remaining phrases of both songs are not so clearly in accordance. But one sees that both rise up (to the third in K0083 and to the fourth in T0228) starting a falling sequential motion towards the tonic, though K0083 avoids ending on the tonic;

instead it has two extra bars, which form some kind of extended ending. Here we have a mixture of identifiable common phrases in the beginning and clear deviants of phrases at the end. The single measures of opti*3* give the following values, which reflect this observations: *nGrUkkon* 0.386, *rhytFuzz* 0.73 and *harmCorE* 0.5. One sees that the value of opti3 stems essentially from harmonic and rhythmical congruencies. The *diffEd* value is very near to the opti3 value: 0.49.

"Ist denn Liebe ein Verbrechen" (T0262) vs. "Ehestandslehren" (T0385)
These melodies are depicted in fig. 11 and fig. 12.

Figure 11.
Melody of „Ist denn Liebe ein Verbrechen" (T0262).

Figure 12.
Melody of „Ehestandslehren" (T0385).

Their opti3-similarity value is 0.462. Both melodies are written in the same meter and both consist of two 4-bar-phrases starting with an upbeat. The rhythmical structure is rather simple in both songs, using only the patterns of two eight-notes followed by either two-quarter notes or a pair of dotted quarter note and eight note. However, the melodic contour is quite different in the first three bars of the first phrase. One could say: when T0262 moves up, T0385 moves down. The second phrases are more similar: Rising up to the tonic in the octave, they fall down along the seventh and fourth step of the scale to the low tonic (T0385) or to the third (T0262).

This analysis is again reflected in the single measure values: *nGrUkkon* is 0.21,

rhytFuzz is 0.84 and *harmCorE* is 0.625, whereas the *diffEd* value is only 0.3, which is rather low.

One can learn from the considerations above that opti3 is in fact an optimal measure in the sense that it forms the optimal compromise for a large number of cases. For comparison: The best single measure *diffEd* gave in one case a clearly higher result (K1185/T0216), in another case a similar result (K0083/T022) and in the third case a clearly lower value (T0262/T0385).

4.3. ALGORITHMIC AND EXPERT CLASSIFICATION

Sagrillo split his 577 melodies in 3312 phrases and ordered these phrases by first filtering according to a few gross criteria like melodic range and eventually by manually fine-tuning the ordering according to his analytical expert notion of similarity. That way he obtained a one-dimensional ordering of the phrases, being notated in a succession of lines, and he marked groups of phrases belonging together. These group marks separated the ordering of the catalogue. So for any pair of phrases it is known whether the two phrases belong to the same group or not. In a previous study (Müllensiefen and Frieler, 2004a) it was shown that it is possible to reconstruct this group membership with a logistic regression model that involved the Edit Distance of the raw pitches, the edit distance of the contourised melody line (Steinbeck's contourisation algorithm) and the Ukkonen measure from the n-gram approach. In that sample 88,6 % of the 52,724 phrase pairs were classified like in Sagrillo's catalogue. An empirical ROC Curve for this model is displayed in figure 13. The ROC curve represents the ability of the combined (logistic) similarity measure to discriminate between phrases from the same group and phrases from different groups. The y-axis (termed "Sensitivity") stands for the hit rate, or the correct classification where Sagrillo's judgement was "same group". The x-axis (1 — "Specificity") is the false alarms rate, or the incorrect classification by the model where Sagrillo judged two phrases to come from the same group. Generally, a good agreement between model and external classification is represented by a curve bent to the upper left corner. The straight diagonal represents the discrimination of pure chance. This means that correct and incorrect classifications by the model have the same frequency. To capture the predictive power of the model in one number, typically the area under curve (AUC) is calculated. AUC values range from 1 (perfect prediction) to 0,5 (chance prediction). For our model the corresponding AUC value was 0.845 and indicated a high discriminating potential of the hybrid similarity measure.

As mentioned in the previous publication (Müllensiefen & Frieler, 2004a), the similarity of the pairs within each group varies considerably with the distance of lines between two phrases. So, for example, the first and last phrase of a large group can have similarity values of < 0.1. So it would be interesting to reconstruct the differences in lines between the phrases of each pair.

The problem of modelling the line differences in Sagrillo's catalogue as a function
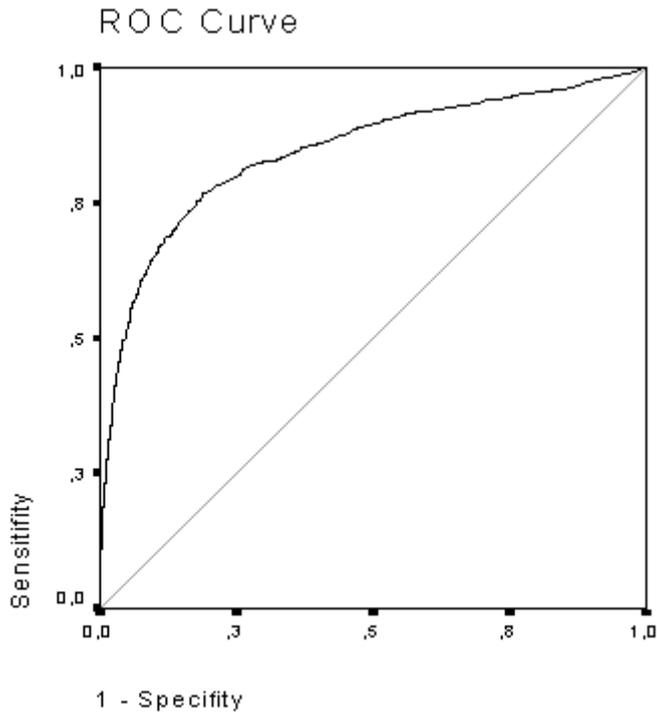
ROC Curve

Figure 13.
ROC curve of combined measures for phrase group discrimination.

of different similarity measures lies in the specific frequency distribution of line differences. From the structure of the set of all differences in a contiguous set of numbers it follows that smaller differences occur more often than larger ones. The relation between the line difference value and its frequency is actually: $f(d) = 2(n - d)/n(n - 1)$, where n is the number of lines.

Consequently, the line differences of a sample of 44 850 phrase pairs show a left skewed distribution as can be seen in figure 14.

With this distribution linear models will fail because the frequency of the residuals will deviate systematically from a normal distribution, which is one of the constraints of models in the general linear approach. The distribution of the residuals from a linear model will be extremely left skewed again. Another problem with the linear ranking of phrases in the catalogue is that in some cases doublets or almost identical copies of the same phrases follow one another. In other cases the difference between one phrase and its follower is rather large. In clear cases a group boundary is marked in the catalogue. So to make sense out of Sagrillo's ordered catalogue, a model is needed that a) can cope with a left skewed distribution and that b) takes the musical characteristics of the phrases into account.
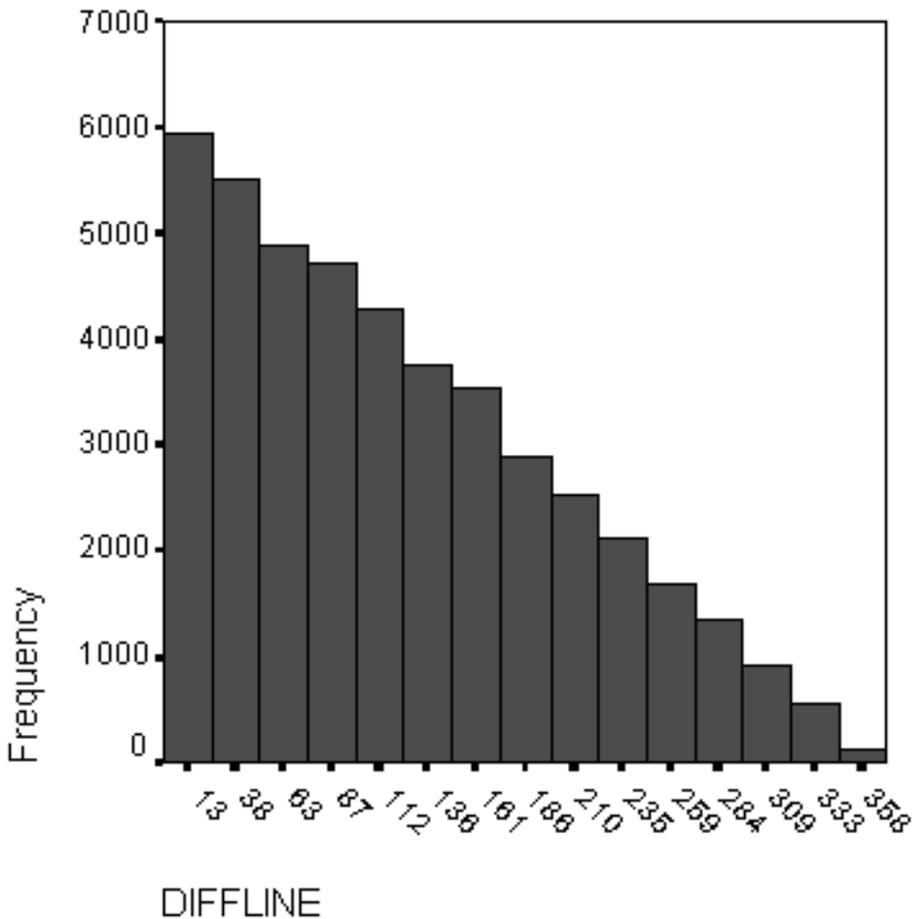
Figure 14.
Frequency distribution of line differences of a sample from Sagrillo's phrase catalogue.

A statistical approach that can deal with a left skewed distribution and that involves dichotomous data as well as metrical covariates is *survival analysis* or *event-history analysis.* In survival analysis, typical questions are about time spans that can be measured between a defined starting point and the onset of an event as measured in a dichotomous variable (*e.g.,* how many days a hard disk will work until it breaks down, how many weeks will a person released from prison survive without becoming a recidivistic criminal *etc.*). A technique in survival analysis that can evaluate the influence of different covariates on the onsets of events is Cox regression. The basic form of the Cox regression model is:

$$H\left(t\right) = h_0\left(t\right) \cdot e^{\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}$$

202

In this equation h (t) signifies the hazard rate (the momentary tendency for an event to occur), h_0 (t) is the baseline hazard rate for an event to occur at time t when there is no influence from other covariates. x_1, x_2, … x_n are covariates that are influential for the onset of events and β_1, β_2, … β_n are their weights. In the classical Cox model the baseline hazard rate cannot be estimated, but the influence of the covariates may be tested for statistical significance.

Transferred to our problem the question modelled with Cox regression is: How strong is the influence of the different algorithmic similarity measures on the change of group membership (from 1 to 0) with increasing line distance? So, small line distances will make it more probable that phrases come from the same group, and distances of 70 or more lines within one group simply do not occur in the catalogue.

For variable selection we calculated the AUC-score for all of the 39 similarity measures and took five measures with the highest scores representing roughly the four quadrants found in the MDS models explained above. These selected measures were

• *rawEdw* (edit Distance of raw pitches, weighted with durations)
• *harmCorr* (maximum of mean of barwise correlation of 12-dimensional minor and major Krumhansl-tonality- vectors)
• *conSEd* (Edit Distance for contourised melodies, Steinbeck's algorithm)
• *rhythFuzz* (edit distance of classified tone durations)

To find the optimal model with these variables we used a backward selection algorithm based on the likelihood ratio as implemented in SPSS. The optimal model is summarised in the following tables:

Omnibus tests of model coefficients [a]

-2 Log-Likeli hood	Overall (value)			change from previous step			change from previous block		
	Chi-square	df	significance	Chi-square	df	significance	Chi-square	df	significance
775785,4	507,873	3	,000	495,445	3	,000	495,445	3	,000

a. starting block no. 0, starting Log-Likelihood-function: -2 Log-Likelihood: 776280,843

Figure 15.

Tests for significance of Cox regression model coefficients.

Variables in the equation

	B	SE	Wald	df	significance	Exp(B)
RAW EDW	,574	,044	166,782	1	,000	1,775
CONSED	,241	,030	62,480	1	,000	1,272
HARM CORR	2,149	,367	34,222	1	,000	8,576

Figure 16.

Coefficients of the Cox regression model.

The goodness of fit of the model is highly significant according to the χ2 statistic (p < 0.001). This is in accordance with the very high significance of the calculated coefficients from the Wald statistic (all values of p < 0.001).

As can be seen from the second table (fig. 16), the measure *rhythfuzz* (edit distance of fuzzified tone durations) is not in the equations. This might be due to the fact that Sagrillo cared less for rhythmical similarity when deciding where to mark a group boundary in his phrase catalogue. In fact it is known from folk song analysis that one melodic idea often appears in different rhythmic patterns. From the remaining three measures *harmcorr* proves to have the strongest influence on group membership. The coefficient of 2.149 can be interpreted in terms of risk: If the value of *harmcorre* drops from 1 to 0, the probability of two given melody phrases belonging to the same group decreases by a factor of about 8.6 when line distance and the other two similarity measures are equal ($e^{2.149}$ = 8.576). An inspection of Sagrillo's catalogue actually underlines the impression that he gave importance to structural and harmonically important notes and not so much to melodic details. Consequently, the weighted edit distance of raw pitches and the contour similarity account to a lesser degree for group membership of the phrases.

5. CONCLUSION

The results of our study have some implications for a theory of melodic similarity (and maybe also for a theory of melodic cognition and production) and for practical applications. First of all, a certain group of subjects termed "western music experts", showed very high reliability and stability in their judgements of melodic similarity in our rating experiments. We believe that this can be generalised to ecological situations and different cultural backgrounds, so that the notion of melodic similarity is a well-defined concept with great inter-subjective transferability, i.e., it makes sense to speak of objective melodic similarity, and that it is measurable, at least in principle. Starting from this finding, it turned out that algorithmic measurements of melodic similarity in accordance to human assessments is in fact possible by using optimal measures for different rating contexts. Furthermore, the contents of the algorithms, which perform best, could (or should) have impact on a theory of human melody processing, e.g., the fact, that symbolic measures outperform vector measures throughout. However, this waits for further elaboration.

As we have demonstrated in this paper, the application of automated similarity measurements to the field of ethno-musicology performed very well, supporting our findings and maybe opening up new perspectives in folk music research.

But the possession of automated methods for similarity measurements has also several practical consequences, not only for the ethno-musicological field. In any case, everywhere where large digital databases of melodies are present automated

similarity judgements are of great help, if not inevitable. Last but not least, our findings could give an objective base to certain aspects of musical copyright issues, like plagiarism.

<div align="center">

LIST OF SYMBOLS

</div>

From the greek alphabet:
Tau (τ)
Beta (β)
Chi (χ)

Address for correspondence:
Daniel Müllensiefen
Department of Computing
Goldsmiths College, University of London
New Cross Road, New Cross
London SE14 6NW, UK
e-mail: d.mullensiefen@gold.ac.uk

Klaus Frieler
Institute of Musicology
University of Hamburg
Neue Rabenstrasse 13,
20354 Hamburg, Germany
e-mail: kgf@omniversum.de

• REFERENCES

Downie, J. S. (1999). *Evaluating a simple approach to musical information retrieval: Conceiving melodic n-grams as text.* PhD thesis. University of Western Ontario

Eerola, T., Järvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception, Vol. 18 (3),* 275-96.

Frieler, K. (in preparation). *Mathematische Musikanalyse — Theorie und Praxis,* PhD thesis, University of Hamburg.

Hofmann-Engl, L. (2002). Rhythmic similarity: A theoretical and empirical approach. In C. Stevens, D. Burnham, G. McPherson, E. Schubert, & J. Renwick (eds). *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney 2002* (CD-rom). Adelaide: Causal Productions.

Juhasz, Z. (2000). A model of variation in the music of a Hungarian ethnic group. *Journal of New Music Research, 29 (2),* 159-72.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch.* New York: Oxford University Press.

Leppig, M. (1987). Tonfolgenverarbeitungen in Rechenautomaten: Muster und Formen. *Zeitschrift für Musikpädagogik 42,* 59-65.

McAdams, S., & Matzkin, D. (2001). Similarity, invariance, and musical variation. In R. J. Zatorre, & I. Peretz (eds). *The Biological Foundations of Music.* New York: New York Academy of Sciences, 62-74.

McNab, R.J., Smith, L.A., Witten, I.H., Henderson, C.L., & Sally Jo Cunningham. (1996). Towards the digital music library: Tune retrieval from Acoustic Input. *Proceedings ACM Digital Libraries.*

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities 24,* 161-75.

Müllensiefen, D. (2004). *Varianz und Konstanz von Melodien in der Erinnerung: Ein Beitrag zur musikpsychologischen Gedächtnisforschung.* PhD thesis, University of Hamburg.

Müllensiefen, D., & Frieler, K. (2004a). Optimizing measures of melodic similarity for the exploration of a large folk song database. In*: Proceedings of the 5th International Conference on Music Information Retrieval.* Universitat Pompeu Fabra, Barcelona.

Müllensiefen, D., & Frieler, K. (2004b). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgements. *Computing in Musicology 13,* 147-76.

O`Maidin, D. (1998). A geometrical algorithm for melodic difference in melodic similarity. *Melodic Similarity: Concepts, Procedures, and Applications. Computing in Musicology 11.* 65-72. Walter B. Hewlett & Eleanor Selfridge-Field (eds). Cambridge, MA: MIT Press.

Pauws, S. (2002). Cuby hum: A fully operational Query by Humming system. *ISMIR 2002 Conference Proceedings.* IRCAM, 187-96.

Sagrillo, D. (1999). *Melodiegestalten im luxemburgischen Volkslied: Zur Anwendung computergestützter Verfahren bei der Klassifikation von Volksliedabschnitten.* Bonn: Holos.

Schmuckler, M. A (1999). Testing models of melodic contour similarity. *Music Perception 16 (3),* 109-50.

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

Uitdenbogerd, A. L (2002). *Music Information Retrieval Technology.* PhD thesis. RMIT University Melbourne Victoria, Australia.

Zadeh, L. (1965). Fuzzy sets. *Inf. Control*, 338-53.

Zhou, Y. & Kankanhalli, M. S. (2003). Melody alignment and similarity metric for content-based music retrieval. *Proceedings of SPIE-IS&T Electronic Imaging.* SPIE *Vol. 5021*, 112-21.

Appendix
List of similarity measures

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	Contour (classical), edit distance
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 dictinct)
NGRSUMCR	Sum Common (interval direction)
NGRUKKOR	n-grams Ukkonnen (interval dir.)
NGRCOORR	n-grams Coord. Match. (interval dir.)
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)
HARMCORK	Harmonic correlation (type II)
HARMCORE	Harmonic correlation (edit distance)
HARMCORC	Harmonic correlation (circle)

• **Modelando las nociones sobre similitud melódica de expertos**

En este trabajo mostramos cómo un subgrupo de expertos en música posee una noción fiable y consistente sobre similitud melódica, y que esta noción se puede medir con una precisión satisfactoria. Nuestras mediciones nos capacitan para modelar los índices de similitud de dichos expertos musicales mediante medios automatizados y algorítmicos. Ha sido matemáticamente sistematizado e implementado un gran número de mediciones algorítmicas de similitud. Se eligieron los mejores algoritmos de similitud de los expertos y se optimizaron mediante medios estadísticos de acuerdo con diferentes contextos. Se construyó un modelo multidimensional de las medidas de similitud algorítmica para ofrecer una visión general de las diferentes dimensiones musicales reflejadas por estas mediciones. Mostramos algunos ejemplos donde estos métodos optimizados se pueden aplicar con éxito a problemas concretos como la categorización y el análisis de la canción folklórica, y discutimos futuras aplicaciones e implicaciones.

• **Modellare le nozioni di similarità melodica degli esperti**

Nel presente articolo mostriamo come un sottogruppo di esperti di musica abbia una nozione attendibile e coerente di similarità melodica, e come questa nozione si possa misurare con soddisfacente precisione. Le nostre misurazioni ci permettono di modellare le valutazioni di similarità da parte di esperti di musica attraverso strumenti automatici ed algoritmici. Un gran numero di misure algoritmiche di similarità presenti nella letteratura sono state matematicamente sistematizzate e perfezionate. I migliori algoritmi di similarità messi a confronto con esperti umani sono stati scelti ed ottimizzati con mezzi statistici in relazione ai differenti contesti. Si è costruito un modello scalare multidimensionale delle misure algoritmiche di similarità per offrire una panoramica delle differenti dimensioni musicali riflesse da tali misure. Mostriamo alcuni esempi dove questi metodi ottimizzati si possono applicare con successo a problemi concreti come la categorizzazione e l'analisi dei canti popolari, e ne discutiamo ulteriori applicazioni ed implicazioni.

• **Modélisation de la conception de la similarité mélodique chez des experts**

Nous montrons dans cet article qu'un sous-groupe d'experts en musique a une conception fiable et cohérente de la similarité mélodique, et que cette conception peut être mesurée avec une précision satisfaisante. Nos mesures nous ont permis de modéliser les classifications de similarité données par des experts en musique à l'aide de moyens automatisés et algorithmiques. Nous avons systématisé mathématiquement un grand nombre de mesures algorithmiques de similarité trouvées dans la littérature et nous avons mis en œuvre le résultat. Les meilleurs algorithmes de similarité comparés aux résultats d'experts humains ont été choisis et optimisés par des moyens statistiques et dans différents contextes. On a construit

un modèle d'échelle multidimensionnel des mesures algorithmiques de similarité afin d'avoir une vision générale des différentes dimensions musicales que reflètent ces mesures. Nous montrons des exemples où ces méthodes optimisées ont pu être utilisées pour des problèmes concrets, comme la catégorisation et l'analyse de chants folkloriques ; nous étudions les implications de ce travail ainsi que d'autres applications possibles.

• **Die Modellierung der Vorstellung von melodischer Ähnlichkeit bei Experten**

Dieser Beitrag zeigt anhand experimentell gewonnener Daten, dass Musikexperten eine valide und reliable Vorstellung von melodischer Ähnlichkeit besitzen und dass dieses kognitive Konzept mit zufriedenstellender Genauigkeit gemessen werden kann. Mit Hilfe der erhobenen Daten werden zunächst die Ähnlichkeitseinschätzungen der Musikexperten algorithmisch modelliert. Dafür wird eine große Zahl von Ähnlichkeitsalgorithmen aus der Literatur mathematisch systematisiert, implementiert und für verschiedene Anwendungskontexte optimiert. Anhand eines multidimensionalen Skalierungsmodells der verwendeten algorithmischen Ähnlichkeitsmaße wird aufgezeigt, in welchen Dimensionen sich die benutzten Ähnlichkeitsmaße unterscheiden. Wir zeigen schließlich an einer Reihe von Beispielen, wie die optimierten Ähnlichkeitsmaße für echte musikwissenschaftliche Fragestellungen eingesetzt werden können, z.B. in der Volksliedanalyse und -kategorisierung. Weitere Anwendungsfelder und Implikationen des hier präsentierten Ansatzes werden diskutiert.