# Evaluating different approaches to measuring the similarity of melodies

Daniel Müllensiefen and Klaus Frieler

Institute of Musicology, University of Hamburg,
Neue Rabenstr. 13, D-20354 Hamburg, Germany

**Abstract.** This paper describes an empirical approach to evaluating similarity measures for the comparision of two note sequences or melodies. In the first sections the experimental approach and the empirical results of previous studies on melodic similarity are reported. In the discussion section several questions are raised that concern the nature of similarity or distance measures for melodies and musical material in general. The approach taken here is based on an empirical comparision of a variety of similarity measures with experimentally gathered rating data from human music experts. An optimal measure is constructed on the basis of a linear model.

## 1 Introduction

While working on an empirical project on human memory for melodies at Hamburg University (Müllensiefen (2004), Müllensiefen and Hennig (2005)) it soon became very clear that measuring the similarity of two given melodies is an important analytical tool in setting up a prediction model for what people can remember of melodies that they just heard once. But melodic similarity is not only a key concept in memory research or music psychology, but also several of musicology's subdisciplines have a strong need for valid and reliable similarity measures for melodies. Some of these subdisciplines are ethnomusicology, music analysis, copyright issues in music, and music information retrieval. So it is not surprising that many different approaches and concepts for measuring the similarity of melodies have been proposed in the literature in the last two decades. Several techniques for computing melodic similarity have been defined 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. In combination with each of these aspects different approaches for constructing distance or similarity measures have been used with music data in the past. Some important algorithms are the edit distance, n-grams, geometric measures and hidden Markov models. In the literature there many examples of successful applications of specific similarity measures that combine an abstraction technique for a special musical aspect of melodies with a specific approach to computing similarities or distances. In the past, we dedicated several studies to the comparison of different approaches to the similarity measurement for melodies (Müllensiefen and Frieler (2004a, 2004b, 2006)) and its applications for example in folk music research (Müllensiefen and Frieler (2004c)). In this earlier work it was shown that these differently constructed similarity measures may generate very different similarity values for the same pair of melodies. And from the mathematical or algorithmical construction of the similarity measures it is by no means clear which one is the most adequate to be used in a specific research situation, like a memory model. The answer to this confusing situation was to compare the measurement values of different similarity measures to ratings from human judges that rate the similarity of melody pairs after listening to the two melodies. This paper first resumes our previous comparative work and in the discussion section we are able to adress some issues of melodic similarity from a meta-perspective.

## 2 Different approaches to measuring the similarity of melodies

To introduce a general framework for comparing different algorithms for similarity measurement it seems useful to first get a clear idea what a melody is on an abstract level. In a useful working definition that was pointed out earlier (Müllensiefen and Frieler (2004a)), a melody will be simply viewed as a time series, i.e., as a series of pairs of onsets and pitches $(t_n, p_n)$, whereby pitch is represented by a number, ususally a MIDI number, and an onset is a real number representing a point in time. A similarity measure $\sigma(m_1, m_2)$ is then a symmetric map on the space of abstract melodies $\mathcal{M}$, mapping two melodies to a value between 0 and 1, where 1 means identity. The similarity measure should meet the constraints of symmetry, self-identity, and invariance under transposition in pitch, translation in time, and tempo changes. The construction of most algorithms for measuring melodic similarity involves 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). 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, *Gaussification* (see Frieler (2004a)), classifications and contourization are the most important. Rhythmical weighting can be done for quantized 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 $T$. After a pitch projection the weighted sequence will still reflect the rhythmical structure. The concept of rhythmical weighting has been widely used in other studies (e.g. Steinbeck (1982), Juhasz (2000)). Classification is mainly used to assign a difference between pitch or time coordinates to a class of musical intervals or rhythmic durations. Other studies used this idea of classification in very similar ways (e.g. Pauws (2002)). Contourization is based on the idea that the perceptionally important notes are the extrema, the turning points of a melody. One takes this extrema (which to take, depends on the model) and substitutes the pitches in between with linear interpolated values, for example. We used linear interpolation exclusively for all of the tested contour models. The contourization idea was employed, for example, in the similarity measures by Steinbeck (1982) and Zhou and Kankanhalli (2003).

For computing the similarity of melodies several basic techniques have been described in the literature. Most of these techniques have their origin in application areas other than music, e.g. text retrieval and comparing gene sequences. But for most of them it has been shown that an adaption for musical data is possible. It is impossible to explain these techniques here in detail, so the reader should refer to the following publications or may find a summary in Müllensiefen and Frieler (2004). Among the most prominent techniques for computing melodic similarity are the edit distance algorithm (McNab et al (1996) Uitdenbogerd (2002)), n-grams (Downie (1999)), correlation and difference coefficients (O'Maidin (1998), Schmuckler (1999)), hidden Markov models (Meek and Birmingham (2002)), and the so-called earth mover distance (Typke et al (2003)).

As is described in Müllensiefen and Frieler (2004a) we implemented 48 similarity measures into a common software framework. These 48 similarity measures were constructed as meaningful combinations of basic and main transformations plus a specific computing technique.

## 3 Experimental evaluation of melodic similarity measures

### 3.1 Experimental design

We conducted three rating experiments in a test-retest design. The subjects were musicology students with longtime practical musical experience. In the first experiment the subjects had to judge the similarity of 84 melody pairs taken from western popular music on a 7-point scale. For each original melody six comparison variants with errors were constructed, resulting in 84 variants of the 14 original melodies. The error types and their distribution were constructed according to the literature on memory errors for melodies (Sloboda and Parker (1985), Oura and Hatano (1988), Zielinska and Miklaszewski (1992)). Five error types with differing probabilities were defined: rhythm errors, pitch errors leaving the contour intact, pitch errors changing the contour, errors in phrase order, modulation errors (pitch errors that result in a transition into a new tonality). For the construction of the individual variants, error types and degrees were randomly combined.

The second and third experiment served as control experiments. In the second experiment two melodies from the first experiment were chosen and presented along with the original six variants plus six resp. five variants, which had their origin in completely different melodies. The third experiment used the same design as the first one, but tested a different error distribution for the variants and looked for the effects of transposition of the variants.

### 3.2 Stability and correlation of human ratings

Only subjects who showed stable and reliable judgments were taken into account for further analysis. From 82 participants of the first experiment 23 were chosen, which met two stability criteria: They rated the same pairs of reference melody and variant highly similar in two consecutive weeks, and they gave very high similarity ratings to identical variants. For the second experiment 12 out of 16 subjects stayed in the analysis. 5 out of 10 subjects stayed in the data analysis of the third experiment. We assessed the between-subject similarity of the judgements in the three experiments using two different, i.e. Cronbach's alpha and the Kaiser-Meyer-Olkin measure (Kaiser (1974)). The inter-personal jugdments of the selected subjects showed very high correlations:

- As an indicator of the coherence of the estimations of the latent magnitude 'true melodic similarity' Cronbach's alpha reached values of 0.962, 0.978, and 0.948 for subjects' ratings of the three experiments respectively.
- The Kaiser-Meyer-Olkin measure reached values as high as 0.89, 0.811, and 0.851 for the three experiments respectively.

This high correlation between the subjects' ratings led us to assume, that there is something like an objective similarity at least for the group of 'western musical experts', from which we took a sample.

### 3.3 Optimisation of similarity measures

It is an old assumption in music research that for the perception and mental computation of melodies all musical aspects play a role to a certain degree. We therefore considered melodic similarity to work on five musical dimensions: contour information, interval structure, harmonic content, rhythm and characteristic motives. For each dimension the euclidean distances of the included measures to the mean subjects' ratings were computed, and the best measure for each dimension was pre-selected to serve as an input for a linear regression. This regression was done for the data of all three experiments separately and used the step-wise variable selection procedure. The best five similarity measures for experiment 1 were (ordered according to their euclidean distances, minimum first):

- *coned* (edit distance of contourized melodies)
- *rawEdw* (edit distance of rhythmically weighted raw pitch sequences)
- *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 Carol Krumhansl's tonality vectors (Krumhansl (1990))
- *rhytFuzz* (edit distance of classified length of melody tones)

From this input we obtained a linear combination of the two measures *rawEdw* and *nGrCoord* for the data from experiment 1, which was 28.5% better than the best single measure for that experiment in terms of the euclidean distance from the subjects ratings over all 84 melody pairs. The model reached an corrected $R^2$ value of 0.826 and a standard error of 0.662. Given these results the optimisation within the linear model can be seen as successful. As the experimental task and the constructed melodic variants to be rated differed systematically in experiment 2 and 3 different similarity measures were preselected for the five musical dimensions and linear regression lead to weighted combinations of similarity measures that were different for each experiment.

### 3.4 Applications

We used our similarity measures in several analytical tasks on a folk song collection that was investigated thoroughly by an expert ethnomusicologist (Sagrillo (1999)). For example we filtered out successfully variants and exact copies of melodies in a catalogue of about 600 melodies from Luxembourg using the optimised similarity measure from our experiment 3 (Müllensiefen and Frieler (2004c)). This specific linear combination of similarity measures was chosen because the experimental taks the subjects had to fullfill in experiment 3 came closest to the duty of finding highly similar melodies. A second application within this folk song research (Müllensiefen and Frieler (2004c)) was to predict if two given melodic phrases from the total of 3312 phrases in the catalogue belong to the same group as classified by Sagrillo. For this task, we again pre-selected the best five out of 48 similarity measures (this time according to their *area under curve* values after drawing a ROC curve for each similarity measure) and we subsequently used logistic regression to predict for each melody pair if the two melodies belonged to the same group or not. Further applications that we tested so far, are the measurement of melodic similarity in cases of plagiarism in pop songs where one melody is assumed to be an illegal copy of a previously existing melody, and the ordering of short melodic phrases from classical music (incipits) according to similarity criteria.

## 4 Discussion

Having worked very intensely for three years on the measurement of similarity between melodies, we came across several conceptual issues that are hardly discussed in the literature. We would like to pose the respective questions here and answer with some tentative hypotheses, but the field is still very open to discussion.

- **Homogeneity of human similarity judgements** For all experiments we conducted so far, we found very high correlations of similarity judgements between subjects (Cronbach's $\alpha$ with values $> 0.9$). This is not to forget that we always selected subjects on the basis of their within-subject reliability, i.e. subjects had to rate the same melody pair in two consequent weeks alike, and they should rate identical melodies as highly similar. The interesting fact is that subjects selected according to their within-subject reliability show a very high between-subjects correlation. The only bias that entered our selection procedure for subjects was the natural requirement that subjects should rate identical melodies as highly similar. But in their judgments of non-identical melodies, subjects were completely free to give their subjective evaluation of the rhythmical, pitch, and contour differences between the melodies. It could have turned out that some of the reliable subjects rated differences in the rhythmical structure as much more severe than others or that contour errors would have been of different importance to different subjects. This would have resulted in lower correlations as reflected by the between-subjects correlation measures, which we actually did not find. These high between-subject correlations could be interpreted as if there is a latent but clear inter-personal notion of melodic similarity that each subject tries to estimate in an experimental situation. This assumption of an inter-personal agreement on what melodic similarity actually is, lays the conceptual foundation for the statistical modelling of human similarity perception of melodies.
- **Human notion of melodic similarity may change** Although there seems to be a consensus on what is similar in melodies, this consensed notion may make different use of the information in the various musical dimensions. For example for melodies that are all very similar because they were constructed as variants from each other like in expriment 1 it was possible to model subjects ratings exclusively with similarity measures that exploit pitch information only. Whereas in experiment 2 where some of the to be compared melodies were drawn at random from a larger collection and were therefore very dissimilar, subjects' ratings could be modeled best including similarity measures that reflect rhythmical information and implicit harmonic content. So obviously, humans show an adaptive behaviour to different tasks, different stylistic repertoires, and different contexts of experimental materials. For modelling subjects' ratings there are two solutions to this adaptive behaviour:

1. Find a general similarity measure that works well in most situations, but be aware that it might not be the optimal measure to model a specific task with a specific repertoire of melodies.
2. Try to gather test data of that specific situation and run an optimisation on that test data before predicting similarity in that domain.

- **Distance vs. similarity measures for melodies** To our knowledge all studies in the literature that deal with the comparision of melody pairs make exclusive use of similarity measures to conceptualise the relationship between two given melodies. Distance measures are never used for example for clustering or ordering of melodies. This seems to reflect an intuitive cognitive approach towards the processsing of comparable melodies. Humans seem to make sense out of melodies that differ in only a few notes. Obviously music listeners are used to relate them to each other effortlessly. But unrelated melodies that differ strongly in most musical dimensions are hard to relate. From our data it was clear that the subjects were much better at differentiating small changes on the rating scale when the two melodies were quite similar as when they had little in common. This might be interpreted as a reflection of the distribution of similarity values in large melody collections. As was outlined in Müllensiefen and Frieler (2004b) the distribution of about 250.000 similarity values between about 700 folk song phrases show a gauss-like distribution, but the shape of the curve was much steeper. Almost all previous studies with the exception of Kluge (1974) use similarity measures that are bounded between 0 and 1. Kluge's special research interest lead him to consider negatively correlated melodies as well and his similarity measure was therefore bounded between -1 and 1. Among our own 48 similarity measures we used several measures based on vector correlations and we tried both variants: Measures between -1 and 1 and measures where we set all negative correlation values to 0. In comparison with our experimental rating data almost always the variants with limits of 0 and 1 showed a superior performance than their -1/1 analogues.

## 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.

FRIELER, K. (2004). Beat and Meter Extraction Using Gaussified Onsets. In: *Proceedings of the 5th International Conference on Music Information Retrieval. Barcelona: Universitat Pompeu Fabra, 178-183.*

JUHASZ, Z. (2000): A Model of Variation in the Music of a Hungarian Ethnic Group. *Journal of New Music Research, 29/2, 159–172.*

KAISER, H.F. (1974): An Index of Factorial Simplicity. *Psychometrika, 39, 31-36.*

KLUGE, R. (1974): *Faktorenanalytische Typenbestimmung an Volksliedmelodien.* Leipzig: WEB Deutscher Verlag für Musik.

KRUMHANSL, C. (1990): *Cognitive Foundations of Musical Pitch.* New York: Oxford University Press.

MCNAB, R.J., SMITH, L. A., WITTEN, I.H., HENDERSON, C.L. and CUNNINGHAM, S.J. (1996). Towards the Digital Music Library: Tune Retrieval from Acoustic Input. In: *Proceedings ACM Digital Libraries, 1996.*

MEEK, C. and BIRMINGHAM, W. (2002): Johnny Can't Sing: A Comprehensive Error Model for Sung Music Queries. In: *ISMIR 2002 Conference Proceedings, IRCAM, 124–132.*

O'MAIDIN, D. (1998): A Geometrical Algorithm for Melodic Difference in Melodic Similarity. In: W.B. Hewlett & Eleanor Selfridge-Field. *Melodic Similarity: Concepts, Procedures, and Applications. Computing in Musicology 11.* MIT Press, Cambridge, 1998.

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

MÜLLENSIEFEN, D. and FRIELER, K. (2004a): Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgments. *Computing in Musicology, 13, 147–176.*

MÜLLENSIEFEN, D. and FRIELER, K. (2004b): Melodic Similarity: Approaches and Applications. In: S. Lipscomb, R. Ashley, R. Gjerdingen and P. Webster (Eds.). *Proceedings of the 8th International Conference on Music Perception and Cognition (CD-R).*

MÜLLENSIEFEN, D. and FRIELER, K. (2004c): 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. Barcelona: Universitat Pompeu Fabra, 274–280.*

MÜLLENSIEFEN, D. and HENNIG, CH. (2005): Modeling Memory for Melodies. In: *Proceedings of the 29th Annual Conference of the German Society for Classification (GfKl).* Springer, Berlin.

OURA, Y. and HATANO, G. (1988): Memory for Melodies among Subjects Differing in Age and Experience in Music. *Psychology of Music 1988, 16, 91–109.*

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, 109–150.*

SLOBODA, J.A. and PARKER, D.H.H. (1985): Immediate Recall of Melodies. In: I. Cross, P. Howell and R. West. (Eds.). *Musical Structure and Cognition.* Academic Press, London, 143–167.

STEIBECK, W. (1982): *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse.* Bärenreiter, Kassel.

TYPKE, R., GIANNOPOULOS, P., VELTKAMP, R.C., WIERING, F., and VAN OOSTRUM, R. (2003): Using Transportation Distances for Measuring Melodic Similarity. In: *ISMIR 2003:Proceedings of the Fourth International Conference on Music Information Retrieval, 107–114.*

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

ZIELINSKA, H. and MIKLASZEWSKI, K. (1992): Memorising Two melodies of Different Style. *Psychology of Music 20, 95-111.*