

OPTIMIZING MEASURES OF MELODIC SIMILARITY FOR THE EXPLORATION OF A LARGE FOLK SONG DATABASE

Daniel Müllensiefen

University of Hamburg
Department of Systematic Musicology

Klaus Frieler

University of Hamburg
Department of Systematic Musicology

ABSTRACT

This investigation aims at finding an optimal way of measuring the similarity of melodies. The applicability for an automated analysis and classification was tested on a folk song collection from Luxembourg that had been thoroughly analysed by an expert ethnomusicologist. Firstly a systematization of the currently available approaches to similarity measurements of melodies was done. About 50 similarity measures were implemented which differ in the way of transforming musical data and in the computational algorithms. Three listener experiments were conducted to compare the performance of the different measures to human experts' ratings. Then an optimized model was obtained by using linear regression, which combines the output of several measures representing different musical dimensions. The performance of this optimized measure was compared with the classification work of a human ethnomusicologist on a collection of 577 Luxembourg folksongs.

1. INTRODUCTION

Melodic similarity is a very valuable concept for analyzing large melody databases. Especially with comprehensive folksong collections, one often wants to answer questions like: Which melodies are variants of one another? Which melodies are duplicates with minor differences only? How can melodies be grouped together according to their similarity and do they reflect inherent relationships?

The study of folk song collections along this approach has a long tradition in ethnomusicology and goes far back before the computer age [1, 2, 8, 9, 10, 22, 23]. Now, with the computer as a convenient tool for the treatment of large data collections and the recent advances in the field of music information retrieval, our interest was, whether a grouping of folk song melodies according to their similarity could be done algorithmically and to which extent the quality approaches that of a human expert.

In case the results indeed show little difference compared to the work of a human expert such an algorithm could become a useful tool for ethnomusicologists for gaining an overview on the similarity relations in a large melody collection.

As reference sample we used a catalogue of 3312 phrases taken from 577 Luxembourg folk songs. They were assembled from 5 different sources, which were sorted, analyzed, and partially annotated by ethnomusicologist Damien Sagrillo [20]. His classification work was carried out with great experience in ethnomusicological treatments 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.

Reviewing the literature on similarity measurement for melodies of the last two decades the biggest concern was actually not the lack of measurement procedures for melodic similarity but rather their abundance. Several very different techniques have been proposed for defining and computing melodic similarity. They all cover distinct aspects or elements of melodies, e.g. intervals, contour, rhythm, and tonality, and each with several ways of transforming the musical information into numerical datasets. The basic techniques for measuring the similarity of this type of datasets are edit distance, n-grams, correlation and difference coefficients, and hidden Markov models. In the literature there are plenty of examples of successful applications of these specific similarity measures: For example McNab et al. [14] and Uitdenbogerd [24] for edit distance and n-grams, also Downie [4] for n-grams, Steinbeck [23] and Schmuckler [21] for correlation and difference coefficients, O'Maidin [18] for a complex difference measure and Meek & Birmingham [15] for HMMs.

This study evaluates for which type of data which similarity measures are cognitively most adequate. We first conducted three listener experiments to find an optimized similarity measure out of a set of basic techniques and their variants. The expert ratings generated in these experiments were compared with the output of the similarity algorithms implemented in the software toolkit SIMILE. An optimization was done using a linear regression model to combine measures that reflect melodic information from distinct dimensions.

Finally the model was tested on the similarity classification in Sagrillo's catalogue.

2. APPROACHES TO MEASURING MELODIC SIMILARITY

2.1. Mathematical Framework

We developed a mathematical framework in order to handle the huge amount of different similarity measures that can be found in the literature. This allowed us to systematically classify the similarity measures in a compact and unified way. It also simplified the comparison of different models with one another other and with empirical data. Furthermore, it served as kind of a construction kit and as a source of inspiration for new similarity measures. Finally, it was very helpful for implementing the algorithms into our software.

We define the "melodic space" M as a subset of the Cartesian product of a (real-valued) time coordinate (representing onsets) and a (integer- or real-valued) pitch coordinate.

A similarity measure is then a map

$$s : M \times M \rightarrow [0,1]$$

with the following properties:

1. Symmetry: $s(m,n) = s(n,m)$
2. Self identity: $s(m,m) = 1$
3. Transposition-, Translation- and Dilation invariance.

'Transposition' means translation in the pitch coordinate, translation is time-shift and 'dilation' means tempo change (time warp). Though these properties hold only approximately for human similarity judgments, they facilitate implementation and comparison. 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, optimized measures, by means of linear regression. Furthermore, any product of two similarity measure is again a similarity measure.

Most of the similarity measures involved the following processing stages:

1. Basic transformations (Representations)
2. Main Transformations
3. Computation

2.2. Transformations

The most common basic transformations are projections and 'differentiations'. Projections can act on either the time or pitch coordinate, (with a clear preference for pitch projections). 'Differentiation' means using differences between coordinates instead of absolute coordinates, i.e. intervals and durations instead of pitch and onsets.

Among the main transformations rhythmical weighting, fuzzifications (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. [6, 9, 23].

Fuzzifications are based on the notion of fuzzy sets, i.e. sets in which an element belongs to it with a certain degree between 0 and 1. However, the fuzzifications reduce to classification if the basic set is decomposed into mutually disjunct subsets. This was true for all our cases. Other studies exploited this idea in similar ways e.g. [19].

Gaussification is a method to construct an integrable function from a set of discrete time-points [5]. This enables a comparison between two rhythms by means of the scalar product of two functions. This can be viewed as a measure for rhythmic similarity.

Contourization is based on the idea, that the perceptually important notes are the extrema, i.e. the turning points of a melody. This extremum is taken (the exact choice depends on the model) and the pitches in between are substituted with interpolated values, e.g., from a linear interpolation (in fact we used this exclusively). The idea of contourization was e.g. employed in the similarity measures by Steinbeck [23] and Zhou & Kankanhalli [26].

Among the other core transformations available are the ranking of pitches and Fourier transformation on contour information (following the approach of Schmuckler, [21]) or methods of assigning a 'harmonic vector' like Krumhansl's tonality vector [11] to certain subsets (bars) of a melody, just to name a few.

2.3. Similarity Computations

The next stage of processing is the computation of a similarity value. The measures we used can roughly be classified in three categories: Vector measures, symbolic measures and musical (mixed) measures, depending on the computational algorithm. The vector measure treats the transformed melodies as vectors in a suitable real vector space such that methods like scalar products and other means of correlation can be applied

to. The symbolic measures on the contrary treat the melodies as strings, i.e. sequences of symbols, where well-known measures like edit distance (see e.g. [16]) or n-gram-related measures (see e.g. [4]) can be used. The musical or mixed measures typically involve more or less specific musical knowledge and the computation can be from either the vector or the symbolical realm.

The symbolical measures based on edit distance and n-grams has proven to be the best throughout. We used edit distance for many different alphabets (raw pitch, interpolated pitch, intervals, fuzzified intervals, fuzzified durations, implicit tonality) and we employed solely global alignment and the simplest cost function.

We applied three different n-gram approaches (Sum Common, Coordinate Matching, Ukkonen [24]) for different alphabets (intervals, interval categories, interval directions, fuzzified rhythms) with a constant n-gram length of 3.

Some general problems had to be solved for some models to ensure transposition and tempo invariance and to account for melodies having different lengths (number of notes). If a measure is not transposition invariant a priori, in principle the maximum over all similarities of all possible transpositions by an integer number of semitones within an octave can be taken, like O'Maidin [18] proposed. Likewise, for models, which require 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 [13]. Tempo invariance is generally no problem when using quantized melodies.

In summary, the techniques for melodic data transformation and pattern matching/similarity measurement employed in this study incorporate the major approaches in this field of the last 15 years. Additionally, systemizing these approaches led to the construction of several new similarity measures (see [5, 16] for a detailed description). We implemented in our software a total number of 48 different similarity measures, counting all variants out of which 39 were used in the analysis. A complete list with short descriptions of the various measures is found in the appendix. We used the same MIDI-files as program input that were used for the experiments. All melodies were quantized.

3. LISTENER EXPERIMENTS

3.1. Experimental Design

We conducted three rating experiments in a test-retest-design. The subjects of the tests were musicology students with longtime practical musical experience. In the first experiment the subjects had to judge 14 melodies taken from western popular music to six systematically derived variants of each on a 7-point

scale. 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 or 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.

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. This type of reliability measurement is considered an important methodological improvement compared with earlier experiments involving similarity ratings. For the second experiment 12 out of 16 subjects stayed in the analysis. 5 out of 10 subjects remained in the data analysis of the third experiment.

The inter- and intrapersonal judgments of the selected subjects showed very high correlations on various measures (e.g. the coefficient Cronbach's alpha reached values of 0.962, 0.978 and 0.948 for the three experiments respectively). This supports the assumption that something like a 'true' similarity exists, at least for the group of western musical experts. This is of course a necessary prerequisite for the comparison between automated algorithmic and human judgments.

3.2. Results

To get an overview over the performance and the differentiation of the 39 similarity measures in relation to the subjects' mean ratings (vpn_mean), multi-dimensional scaling (MDS; for details on algorithms and model options see [3], [12]) was used to display the results graphically. The euclidean distances between all 39 similarity measures over all melody pairs from experiment 2 (variants from original vs. from different melodies) were computed. 18 measures that showed the least distance to the subjects' means and that could be representative for all the 39 measures were selected for the MDS (13 are displayed with their names on the following graph). With these 18 measures (and the mean of subjects' ratings) an MDS model was computed that used only the ordinal information in the distance data. The usual MDS criteria, RSQ (=portion of the variance explained) and stress, were used as indicators of fit (stress = 0.075, RSQ = 0.98). A two-dimensional solution was chosen that is represented by figure 1.

A meaningful interpretation of this solution views dimension 1 as the degree to which the similarity measures incorporate rhythmical information: To the right a measure from the n-gram approach are located that uses the fuzzified rhythm (*ngukkfr*) values as data.

- a) The description of the almost normal distribution of the 171,405 similarity values between all melodies in the collection,
- b) The analysis of ‘interesting cases’ in which the similarity values of different dimensions (melody, harmony, rhythm) differ significantly,
- c) The spotting out of doublets and variants and
- d) The classification of melodies into groups or families according to Sagrillo’s catalogue.

Only steps c) and d) will be covered in the two following subsections.

4.1. Duplicates and variants

A crucial test for any similarity measure is the task of identifying identical or almost identical melodies in a database. Unfortunately, we had no complete information about identical melodies, but a suffix ‘V’ in database of the tunes indicates a variant to a specific tune. There were 19 of such marked songs in the Luxembourg database, which we inspected manually. Apart from the 5 tunes marked with a ‘V’ that had the same lyrics but a different melody, the remaining 14 melodies had similarity values of 0.6 or higher according to the *opti3* measure.

We also examined all melody pairs in the database with similarity values above 0.6 (49 melody pairs). These pairs can be roughly classified in

1. ‘Duplicates’ (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. ‘Psalms’: 2 pairs

The so-called ‘psalms’ are special types of songs which are typically written without meter, consist almost completely of tone repetitions and have usually small tone range. Some songs could be found with 3 or more variants. One example is a song called ‘De Malbrough’, which can also be found in a collection from Lorraine. Inspecting it, it turned out that it is highly similar to the well-known (english) song ‘He’s a jolly good fellow’.

4.2. Algorithmic and Expert Classification

The final task was the reconstruction of Sagrillo’s classification of the 3312 phrases from the Luxembourg melodies. Apart from the indication of variants (see above) Sagrillo used two hierarchical levels of similarity grouping. He firstly sorted the phrases numerically according to several gross criteria and then performed a very careful analysis ‘by hand’. We simply used the grouping on one classification level as criterion of a greater similarity (0=not member of the same group, 1=member of the same group). We used logistic regression to model Sagrillo’s classification with our similarity measures and the Area under Curve (Receiver Operating Curves) from Signal Detection theory to evaluate the solutions. Due to computing limitations, we

worked on a sample of 52,724 melody comparisons coming from 438 phrases classified by Sagrillo in 21 groups.

We first tested the performance of our *opti3* measure. But it performed quite poorly on short phrases (usually only 1-3 bars) since it had been optimized for longer melodic lines. We received an AUC value of only 0.676. So an optimization for the new empirical melodic entity of phrases seemed necessary. This was done in an analogous manner to the optimization process described in 3.2: We calculated the AUC scores for any of the 39 similarity measures and picked the measure for every information dimension that discriminated best. The five best measures for discriminating the phrases were:

- Pitch/interval: *rawEd* (Edit distance of raw pitch values)
- Contour: *consEd* (Edit Distance of contourized pitch values, contourization according to Steinbeck, 1982)
- Short motives: *nGrUkkon* (Ukkonen measures for 3-grams of intervals)
- Harmony: *harmCorr* (Correlation measure for tonality values based on Krumhansl’s tonality vector)
- Rhythm: *rhytFuzz* (Edit distance of classified duration values)

We found an optimal model including *rawEd*, *consEd*, and *ngrUkkon*, with *rawEd* having the greatest weight in the logistic regression term. This model classified 88.6% of the 52,724 phrase pairs correctly (92.4% of the non-class members and 61.1% of the class members). This model showed a good overall discrimination power as can be seen by its ROC diagram and its AUC value of 0.845 which can be interpreted as ‘excellent’ according to [7].

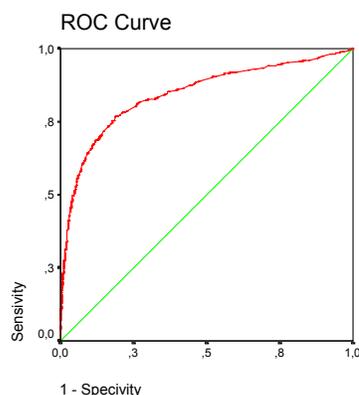


Figure 4: ROC curve of optimized measure for phrase classification

It is possible to give more weight to the detection of class-members by choosing a different cut-off value for the logistic regression function (at the cost of assuming a higher percentage of misclassified non-class members).

With a cut-off value of 0.133 we classified 72.3% of the class members correctly (85.1% of the non-class members correct, 83.5% correct overall).

However, the detection of the class members is still not perfect but an inspection of Sagrillo's groups showed that his similarity classification is rather of a continuous nature than one of actual groups. So especially in large groups the first and the last members possess generally low similarity values in our optimized model. A more sophisticated approach would be to use all levels of his hierarchical classification or the proximity of the phrases in his ordered catalogue as dependent variable in the regression model. This is planned for the near future.

5. SUMMARY

As this study has focused on the classification and exploration of a folksong collection, the chosen methods gave satisfying and promising results with strong implications. The strikingly simple idea of evaluating and gauging a large number of melodic similarity measures from the literature with the help of empirical research and mathematical systematization enabled us to develop a successful tool. Furthermore, our empirical work seems to prove that the concept of melodic similarity is a stable and well-defined notion for human music experts.

Due to the high importance of melodic similarity in many areas of music research and engineering, our tool is ready and waiting for many other applications, e.g., in Query-by-Humming systems, as a tool for melodic memory research, or as starting point for cognitive models of human melodic similarity judgments. Some of these, as well as refinements, optimization and evaluations of the tool in other domains, will be done in the future.

6. REFERENCES

- [1] Bartók, B. & Lord, A.B. *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, 1951.
- [2] Bartók, B. "Why and How Do We Collect Folk Music?". *Béla Bartók Essays*. Ed. Benjamin Suchoff. London: Faber & Faber, 1976, 9-24.
- [3] Borg, I & Lingo, J.C. *Multidimensional similarity structure analysis*. New York: Springer, 1987.
- [4] Downie, J. S. *Evaluating a Simple Approach to Musical Information retrieval: Conceiving Melodic N-grams as Text*. PhD thesis, University of Western Ontario, 1999
- [5] Frieler, K. *Mathematische Musikanalyse - Theorie und Praxis*. PhD thesis, University of Hamburg (in preparation), 2004
- [6] Hofmann-Engl, L. "Rhythmic Similarity: A theoretical and empirical approach". *Proceedings of the 7th International Conference on Music Perception and Cognition, Sydney 2002*. Ed. C. Stevens, D. Burnham, G. McPherson, E. Schubert, J. Renwick. Adelaide, Causal Productions, 2002
- [7] Hosmer, D. W. & Lemeshow, S. *Applied Logistic Regression*. Wiley, New York, 2000.
- [8] Jesser, B. *Interaktive Melodieanalyse: Methodik und Anwendung computergestützter Analyseverfahren in Musikethnologie und Volksliedforschung: typologische Untersuchung der Balladensammlung des DVA*. Bern: Peter Lang, 1990.
- [9] Juhasz, Z. "A Model of Variation in the Music of a Hungarian Ethnic Group". *Journal of New Music Research*, 29, No. 2, 2000, 159-172.
- [10] Kluge, R. *Faktorenanalytischen Typenbestimmung an Volksliedmelodien*. Leipzig: VEB Deutscher Verlag für Musik, 1974.
- [11] Krumhansl, C. L. *Cognitive foundations of musical pitch*. New York: Oxford University Press, 1990.
- [12] Kruskal, J.B. & Wish, M. *Multidimensional scaling*. Beverly Hills: Sage, 1978.
- [13] Leppig, M. "Musikuntersuchungen in Rechenautomaten". *Musica* 41/2, 1987, p140-150.
- [14] McNab, R. J., Smith, L.A., Witten, I.H., Henderson, C.L. & Cunningham, S.J. "Towards the Digital Music Library: Tune retrieval from Acoustic Input". *Proceedings ACM Digital Libraries*, 1996.
- [15] Meek, C. & Birmingham, W. "Johnny Can't Sing: A Comprehensive Error Model for Sung Music Queries." *ISMIR 2002 Conference Proceedings*, IRCAM, 2002, p124-132.
- [16] Mongeau, M. & Sankoff, D. "Comparision of Musical Sequences". *Computers and the Humanities* 24, 1990, p161-175.
- [17] Müllensiefen, D. *Variabilität und Konstanz von Melodien in der Erinnerung*. PhD thesis, University of Hamburg (in preparation), 2004
- [18] O'Maidin, D. "A Geometrical Algorithm for Melodic Difference in Melodic Similarity". *Melodic Similarity: Concepts, Procedures, and Applications. Computing in Musicology 11*. Ed.

Walter B. Hewlett & Eleanor Selfridge-Field.
Cambridge: MIT Press, 1998

- [19] Pauws, S. "Cuby hum: A Fully Operational Query by Humming System". *ISMIR 2002 Conference Proceedings*, IRCAM, 2002, p187-196.
- [20] Sagrillo, D. *Melodiegestalten im luxemburgischen Volkslied: Zur Anwendung computergestützter Verfahren bei der Klassifikation von Volksliedabschnitten*. Holos, Bonn, 1999.
- [21] Schmuckler, M. A. "Testing Models of Melodic Contour Similarity." *Music Perception* Vol. 16, No. 3, 1999, p109-150.
- [22] Seeger, Ch. "Versions and Variants of the Tunes of 'Barbara Allen'". *Selected reports in ethnomusicology* Vol.I, No. 1, 1966.
- [23] Steinbeck, W. *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse. Kieler Schriften zur Musikwissenschaft XXV*. Kassel, Basel, London: Bärenreiter, 1982
- [24] Uitdenbogerd, A. L. *Music Information Retrieval Technology*. PhD thesis, RMIT University Melbourne Victoria, Australia, 2002
- [25] Zadeh, L. "Fuzzy sets". *Inf. Control*, 1965, p338-353.
- [26] Zhou, Y. & Kankanhalli, M. S. "Melody alignment and Similarity Metric for Content-Based Music Retrieval". *Proceedings of SPIE-IS&T Electronic Imaging*. SPIE Vol. 5021, 2003, p112-121.

NGRSUMCO	n-grams Sum Common
NGRUKKON	n-grams Ukkonnen
NGRCOORD	Coordinate Matching (count distinct)
NGRSUMCR	Sum Common (interval direction)
NGRUKKOR	n-grams Ukkonnen (interval dir.)
NGRCOORDR	n-grams Coord. Match. (interval dir.)
NGRSUMCF	n-grams Sum Common (fuzzy)
NGRUKKOF	n-grams Ukkonnen (fuzzy)
NGRCOORDF	n-grams Count distinct (fuzzy)
NGRSUMFR	n-grams sum common (fuzzy rhythm)
NGRUKKFR	n-grams Ukkonnen (fuzzy rhythm)
NGRCOORDFR	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)

7. APPENDIX: TABLE OF EMPLOYED SIMILARITY MEASURES

Abbreviation	Model
RAWED	Raw pitch edit distance
RAWEDW	Raw pitch edit distance, weighted
RAWPCST	Raw pitch P-B. corr, 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 edit distance weighted
CONPCST	Contour, P-B. corr., 0-1
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)