Proceedings of the 8th International Conference on Music Perception & Cognition, Evanston, IL, 2004
ICMPC8
S.D. Lipscomb, R. Ashley, R.O. Gjerdingen, & P. Webster (Eds.)
Adelaide, Australia: Causal Productions

MELODIC SIMILARITY: APPROACHES AND APPLICATIONS

Daniel Müllensiefen, Klaus Frieler

Department of Musicology, University of Hamburg (Germany)

ABSTRACT

This paper describes the systematization, testing and optimiziation of different approaches for measuring similarities of melodies. First, a quick overview of our mathematical systematization for similarity measures, including data transformations and calculation methods is given. Behavioral data from three listener experiments is used to model experts' similarity judgments of short melodies from popular music in different contextual situations. A weighted combination of several similarity measures, representing two resp. three different sources of information, is found to explain user ratings best. As an application example one of the optimal similarity measures resulting from these three experiments is used to analyze a body of about 600 folk melodies from Luxembourg. Finally, the expert classification of the individual phrases of these melodies that has carried out in an extensive ethno-musicological study (Sagrillo, 1999) is reconstructed with the help of an optimal combination of similarity measures using logistic regression.

1. INTRODUCTION

The investigations on melodic similarity reported in this paper were carried out with the help of a comprehensive tool kit for symbolic analysis of melodies with the current name SIMILE. Among other functions it serves to determine the complexity of rhythms and to assign meter and phase information to a given rhythm, to segment long melodic passages into memorable phrases, to reduce melodic surfaces to structural representations, and last but not least to measure the similarity of melodies. The tool kit is mainly 'tuned' or parametrized for singable melodies from the popluar or folk repertoire. Only the array of algorithms and empirical investigations belonging to field of similarity measurement are explained in this article. But the general framework of testing competitive approaches and algorithms with data from human subjects and combining the most powerful ones in an 'optimal' hybrid approach is adopted for the other application areas of the toolkit as well, with great emphasis laid on the cognitive 'adequacy' in the performance of the algorithms.

The development of a tool for melodic similarity analysis has its roots in a music psychological research enterprise on melody representations in human memory (Müllensiefen, 2004). While reviewing the literature on similarity measurement for melodies of the last two decades, it became clear that it was not the lack of measurement procedures for melodic similarity, but their abundance that needed serious 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 (O'Maidin, 1998; Schmuckler, 1999), and hidden Markov models (Meek & Birmingham, 2002). In the literature there are plenty of examples of successful applications of these specific similarity measures.

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

Studies that have been comparing human ratings with algorithmic similarity measurements are for example Schmuckler (1999), Eerola et.al. (2001), McAdams & Matzkin (2001), and Hofmann-Engl (2002). The studies of Schmuckler (1999) and McAdams & Matzkin (2001) come closest to the present approach, but the variety of similarity models and musical material employed here is far greater and closer to 'ordinary' popular western music.

2. SIMILARITY MEASURES AND DATA TRANSFORMATIONS

2.1 Mathematical Framework

In order to handle the huge amount of different similarity measures found in the literature we developed a mathematical framework. This allowed us to give a systematization and classification of the similarity measures in a compact and unified way, and made it possible to compare the different models with each other and with the empirical data. Furthermore, it served as kind of a construction kit and as a source of inspiration for new similarity measures. At last 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 a map

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

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). These properties are intuitively clear from perceptional reality. 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). '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', 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. Steinbeck, 1982, Juhász, 2000, Hofmann-Engl, 2002).

Fuzzifications are based on the notion of fuzzy sets, 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 disjunct 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).

Contourization is based on the idea, that, the perceptionally important notes are the extremas, the turning points of a melody. One takes this extremas (which to take depends on the model) and substitutes the pitches in between with interpolation values, e.g. coming from linear interpolation, which we used exclusively. The contourization idea was employed, for example, in the similarity measures by Steinbeck (1982) and Zhou & Kankanhalli (2003).

Among the other core transformations were 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 next stage of processing is the computation of a similarity value. The measures we used could roughly be classified in three categories: Vector measures, symbolic measures and musical (mixed) measures, according to the computational algorithm used. The vector measure treat the transformed melodies as vectors in a suitable real vector space, where methods like scalar products and other means of correlation can be applied to. On the contrary the symbolic measures treat the melodies as strings, i.e., sequences of symbols, where well-known measures like Edit Distance (e.g. Mongeau & Sankoff, 1990) or n-gram-related measures (e.g. Downie, 1999) 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 or even completely different ways like scoring models.

Some general problems had to be solved for some models to ensure transposition and tempo invariance or to account for melodies having different lengths (number of notes). If a measure is not transposition invariant a priori, one can principally take the maximum over all similarities of all possible transpositions. 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 quantized melodies.

In sum, the techniques for melodic data transformation and pattern matching / similarity measurement employed in this study resume 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 Frieler (2004) and Müllensiefen (2004) for a detailed description). We implemented in our software a total number of 48 different similarity measures, counting all variants, from which 39 were used in the analysis. As input to our program served the MIDI-files, which we used in the experiments. All melodies were quantized.

3. LISTENER EXPERIMENTS

3.1 Design, Materials, and Procedure

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 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 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.

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 stayed in the data analysis of the third experiment.

The inter- and intrapersonal jugdments 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 led us to assume, that there is something like a 'true' similarity' at least for the group of 'western musical experts', which is a necessary condition for comparing algorithmic vs. human judgments.

3. 2 Results

Besides the comparative and explorative aims, this study set out to get an 'optimal' measure from the considered algorithms.

Therefore melodic similarity was assumed to work on five dimensions: Contour information, interval structure, harmonical 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 for each dimension was taken to serve as an input for a linear regression. This regression was done for the data of both experiments separately.

The best five models for experiment 1 were (ordered according to their euclidean distances, minimum first):

- **coned** (Edit Distance of contourized melodies, own contourization 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 contourized melodies, Steinbeck's algorithm)
- **nGrCooFR** (based on count of distinct n-grams of classified note lengths)

From this input we obtained combined measures, which were 28.5% and 33.4% better than the best single measure for each experiment. The superior performance of the 'optimized' hybrid measure **opti3** can bee seen from the following diagram:

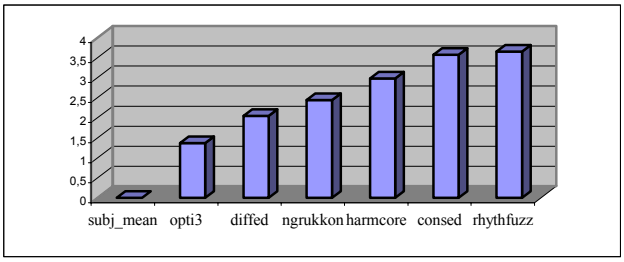


Fig. 1: performance of different similarity measures on data from experiment 2

Interestingly, the combined model for the data of experiment 1 consisted of two measures that reflect pitch information only, while for experiment 2 harmonic and rhythm measures 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 'real' and 'wrong' melodies - they make use of more sources of information like rhythmic information. These combined or 'optimized' models fit very well to the data. For experiment 1 there was 83 % of variance explained by the combined measure, and for experiment 2 even 92%.

4. SIMILARITY ANALYSIS OF A FOLKSONG COLLECTION

To test the optimal similarity measure from experiment 2 with a different melody repertoire and to compare to more analytical

expert ratings than those that can be done in a listening experiment, we referred to the comprehensive ethno-musicological study of Damien Sagrillo (1999). Sagrillo analyzed and classified phrases of 577 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 coined by great experience with 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 optimized measure **opti3** 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 Luxemburg.

4.1 Distribution of Similarity Values

As our conception of similarity is one of a multidimensional feature that reflects the sum of various independent sources of information about melodies, its frequencies should be distributed approximately normally. We received a close to normal distribution when we plotted the 254,910 similarity values against their frequencies that resulted from a complete comparison of 585 and 435 folksongs from Luxembourg and Lorraine respectively (fig. 2).

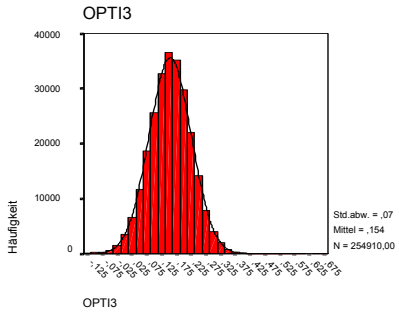


Figure 2: Frequency distribution of similarity values for 1020 Melodies from Luxembourg and Lorraine

Probably because of the restricted range of the similarity measure, a Kolmogorov-Smirnov-test for normal distribution failed to reach significance.

4.2 Doublets and variants

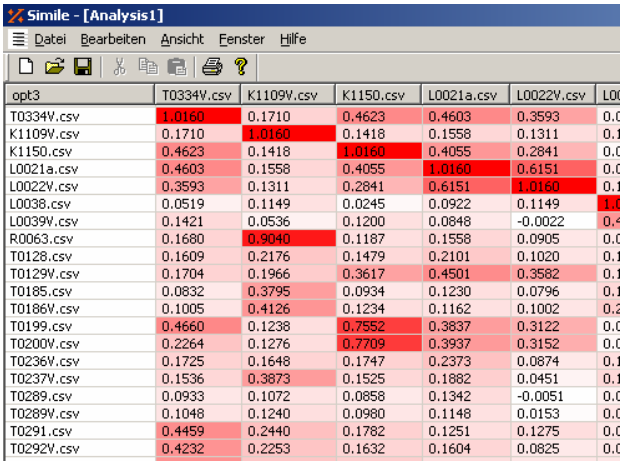
One crucial test for any similarity measure is the task of identifying doublets in a database. Unfortunately, we had no complete information about doublets, but a suffix 'V' in Sagrillo's catalogue index of the tunes indicated a variant to a specific melody. There were 19 of so marked tunes in the Luxembourg database, which we inspected manually. Four of them (L0027V, L0035V, S0073V, T0222V) were songs with the same lyrics, but different melodies, which was indicated in the remark section of the songs. For one song (K1086V) there was no pendant found, neither in the database nor by means of our similarity measure, so maybe it is a mistake of the collectors. The last 14 tunes were indicated as melody variants. Out of these 14 tunes 8 had an **opti3** 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, e.g., 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. 'Psalms': 2 pairs

The so-called 'psalms' are a special case of songs, which are typically written without meter, consist almost completely of tone repetitions and have usually small tone range. Here the high similarity values might be purely accidental, because all the measures involved in **opti3** give high ratings for long sequences of tones of same length and pitch. This in fact is a true similarity, but one might argue that this type of songs obey some different kind of musical logic than the other folk tunes. 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. Inspecting it, it turned out that it is highly similar (by ear) to the well-known (English) song 'He's a jolly good fellow'.

The spotting out of highly similar tunes was performed with the help of the graphical interface of the similarity software (SIMILE) that gives shades of deeper red to more similar pairs of melodies, as in shown in fig. 3:



opt3	T0334V.csv	K1109V.csv	K1150.csv	L0021a.csv	L0022V.csv	L0
T0334V.csv	1.0160	0.1710	0.4623	0.4603	0.3593	0.0
K1109V.csv	0.1710	1.0160	0.1418	0.1558	0.1311	0.1
K1150.csv	0.4623	0.1418	1.0160	0.4055	0.2841	0.0
L0021a.csv	0.4603	0.1558	0.4055	1.0160	0.6151	0.0
L0022V.csv	0.3593	0.1311	0.2841	0.6151	1.0160	0.1
L0038.csv	0.0519	0.1149	0.0245	0.0922	0.1149	1.1
L0039V.csv	0.1421	0.0536	0.1200	0.0848	-0.0022	0.4
R0063.csv	0.1680	0.9040	0.1187	0.1558	0.0905	0.0
T0128.csv	0.1609	0.2176	0.1479	0.2101	0.1020	0.1
T0129V.csv	0.1704	0.1966	0.3617	0.4501	0.3582	0.1
T0185.csv	0.0832	0.3795	0.0934	0.1230	0.0796	0.1
T0186V.csv	0.1005	0.4126	0.1234	0.1162	0.1002	0.2
T0199.csv	0.4660	0.1238	0.7552	0.3837	0.3122	0.0
T0200V.csv	0.2264	0.1276	0.7709	0.3937	0.3152	0.0
T0236V.csv	0.1725	0.1648	0.1747	0.2373	0.0874	0.1
T0237V.csv	0.1536	0.3873	0.1525	0.1882	0.0451	0.1
T0289.csv	0.0933	0.1072	0.0858	0.1342	-0.0051	0.0
T0289V.csv	0.1048	0.1240	0.0980	0.1148	0.0153	0.0
T0291.csv	0.4459	0.2440	0.1782	0.1251	0.1275	0.0
T0292V.csv	0.4232	0.2253	0.1632	0.1604	0.0825	0.0

Figure 3: screenshot of the similarity software SIMILE.

4.3 Three Examples

As was seen from the distribution of similarity values in the collection from Luxembourg, 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 relatively short time, which would otherwise hardly be possible. To illustrate this, we will now have a short analytical look on three selected pairs of folk tunes with high similarity, whereby we additionally gain some insights in the functioning of the **opti3** measure.

'Zwei Hasen' (K1185) vs 'Plauderei an der Linde' (T0216) This two songs can be seen in fig. M1 and fig.M2 in the appendix. The similarity value for of these tunes with our optimal measure **opti3** is 0.634. One easily sees that these two melodies are nearly identical, despite the fact, that K1185 is set in triple meter, and T0216 is notated 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 **diffEd** gives a similarity value of 0.87, as one might have expected. 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)

The two songs can be seen in fig.M3 and fig.M4 in the appendix. Their **opti3** 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 10. A closer look on the first 6 bars of the two tunes reveals a nice structural relationship. They both start with a 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 a 'natural' 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, though in different meters and clear deviants of phrases in the end. The single measures of **opti3** give the following values, which reflect this observations: **nGrUkkon**: 0.386, **rhytFuzz** 0.73 and **harmCorE** 0.5. One might argue that the value of **opti3** stems essentially from harmonical 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.M5 and fig.M6 in the appendix. 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 quite simple in both songs, using only the patterns of two eight-notes followed by either two quarter-notes or a 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) respectively.

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.

Conclusion 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).

4.4. Algorithmic and Expert Classification

A 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 hierachical levels of similarity grouping. His method for classification was numerical sorting of the phrases according to several gross criteria followed by a very careful analysis 'by hand' of the phrases. We simply used the grouping on one classification level as criterium 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 (Reiceiver Operating Curves) from Signal Detection theory to evaluate the solutions. Due to computing limiations, we worked on a sample of 52,724 melody comparisions coming from 438 phrases classified by Sagrillo in 21 groups.

We first tested the performance of our **opti3** measure. But as the opti3 measures was optimized for longer melodic lines, it perfomed quite poorly on the short phrases of Sagrillo's catalogue (usually only 1-3 bars). We received an AUC value of only 0.676. So an optimization for the new empirical melodic enity 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 n-grams on raw pitch values)
- Harmony: **harmCorr** (Correlation measure for tonality values based on Krumhansl's tonality vector)
- Rhythm: r**hytFuzz** (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 diagramm and its AUC value of 0.845 which can be interpreted as 'excellent' according to Hosmer & Lemeshow (2000).

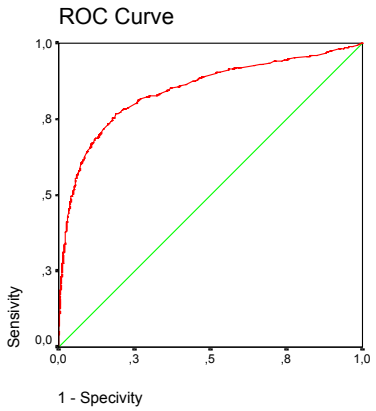


Figure 4: ROC curve of optimized measure for phrase classification

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

Still the detection of the class members is not perfect. But an inspection of Sagrillo's groups showed that his similarity classification is rather of a continous nature than one of actual groups. So especially in large groups the first and the last members posses generally low similarity values in our optimized model. A more sophisticated approach would be to use all levels of his hierachical classification or the proximity of the phrases in his ordered catalouge as dependent variable in the regression model. This will be a task of the near future.

5. REFERENCES

1. Downie, J. Stephen, Evaluating a Simple Approach to Musical Information retrieval: Conceiving Melodic N-grams as Text. PhD thesis, University of Western Ontario, 1999
2. Eerola, T., Järvinen, T., Louhivuori, J., and Toiviainen, P. (2001). "Statistical Features and Perceived Similarity of Folk Melodies." Music Perception, Vol. 18, No. 3, 2001, p275-296.
3. Frieler, Klaus, Mathematische Musikanalyse - Theorie und Praxis, PhD thesis, University of Hamburg (in preparation), 2004
4. Hofmann-Engl, Ludger, "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
5. Hosmer, David W., and Lemeshow, Stanley, Applied Logistic Regression, Wiley, New York, 2000.
6. Krumhansl, Carol L., Cognitive foundations of musical pitch. New York: Oxford University Press, 1990
7. McAdams, Stephen, and Matzkin, Daniel, "Similarity, Invariance, and Musical Variation". The Biological Foundations of Music. Ed. Robert J. Zatorre & Isabelle Peretz. New York Academy of Sciences, New York, 2001, p62-74.
8. Mongeau, Marcel, and Sankoff, David, "Comparision of Musical Sequences". Computers and the Humanities 24, 1990, p161-175.
9. Müllensiefen, Daniel. *Varianz und Konstanz von Melodien in der Erinnerung*. PhD thesis, University of Hamburg (in preparation), 2004
10. O`Maidin, Donncha, "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
11. Pauws, Steffen, "Cuby hum: A Fully Operational Query by Humming System". ISMIR 2002 Conference Proceedings, IRCAM, 2002, p187-196.
12. Sagrillo, Damien, Melodiegestalten im luxemburgischen Volkslied: Zur Anwendung computergestützter Verfahren bei der Klassifikation von Volksliedabschnitten, Holos, Bonn, 1999.

13. Schmuckler, Mark A, “Testing Models of Melodic Contour Similarity." Music Perception Vol. 16, No. 3, 1999, p109-150.

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

15. Uitdenbogerd, Alexandra L, Music Information Retrieval Technology. PhD thesis, RMIT University Melbourne Victoria, Australia, 2002

16. Zadeh, Lofti., "Fuzzy sets". Inf. Control, 1965, p338-353.

17. Zhou, Yongwei & Kankanhalli, Mohan S. "Melody alignment and Similarity Metric for Content-Based Music Retrieval". Proceedings of SPIE-IS&T Electronic Imaging. SPIE Vol. 5021, 2003, p112-121.

Appendix: Melody Examples

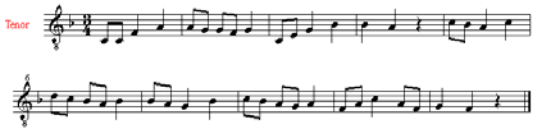


Figure M1 – K1185, ‘Die beiden Hasen’

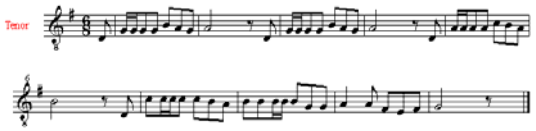


Figure M2 – T0216, ‘Plauderei an der Linde’

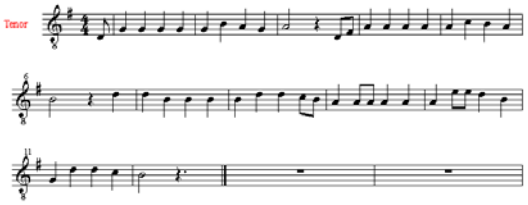


Figure M3 – K0083, ‘Jetzt reisen wir zum Tor hinaus’

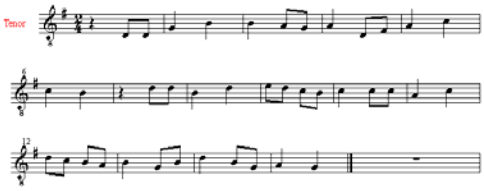


Figure M4 – T0228, ‘Eng ongeheiesch Freiesch’

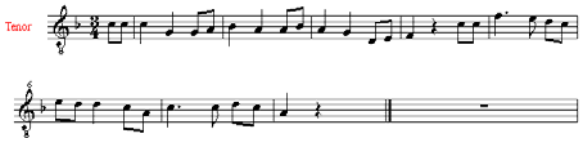


Figure M5 – T0262, ‘Ist denn Liebe ein Verbrechen’

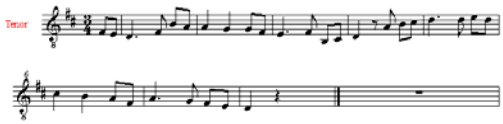


Figure M6 – T0385, ‘Ehestandslehren’