The Study of Melodic Similarity using Manual Annotation and Melody Feature Sets

Anja Volk, Peter van Kranenburg, Jörg Garbers, Frans Wiering, Remco C. Veltkamp, Louis Grijp

Technical Report UU-CS-2008-013
July 2008

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
www.cs.uu.nl

ISSN: 0924-3275

Department of Information and Computing Sciences
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands

The Study of Melodic Similarity using Manual Annotation and Melody Feature Sets

Anja Volk, Peter van Kranenburg, Jörg Garbers,
Frans Wiering, Remco C. Veltkamp, Louis Grijp*

Department of Information and Computing Sciences, Utrecht University
*Meertens Institute, Amsterdam
The Netherlands

Abstract

This paper[1] describes both a newly developed method for manual annotation for aspects of melodic similarity and its use for evaluating melody features concerning their contribution to perceived similarity. The second issue is also addressed with a computational evaluation method. These approaches are applied to a corpus of folk song melodies. We show that classification of melodies could not be based on single features and that the feature sets from the literature are not sufficient to classify melodies into groups of related melodies. The manual annotations enable us to evaluate various models for melodic similarity.

1 Introduction

The long term goal of the WITCHCRAFT-project is to create computational methods that support folk song research.[2] This paper takes an essential step towards this goal by investigating the similarity of songs that have been classified by humans into groups of similar melodies.

For the computational modeling of melodic similarity numerous features of melody could be taken into account. However, for a specific problem such as classification only a few features might be sufficient. Hence, we need a means to evaluate which features are important. Once a similarity measure is designed that uses a single feature or a few features, we also need a means to evaluate that similarity measure.

Therefore, we have developed a manual annotation method that gathers experts judgments about the contribution of different musical dimensions to the perceived similarity. We use this method to characterize the similarity of selected folk songs from our corpus. The human perception of melodic similarity is a challenging topic in cognition research (see e.g. [1] and [4]). The establishing of the annotation data in this paper is a first step to study the similarity as perceived by humans in the special case of similarity between melodies belonging to the same melody group. We evaluate in how far available computational features contribute to the characterization of similarity between these songs.

Contribution: With these two methods we address the following questions:

[1]A shorter version of this paper will be published in the proceedings of ISMIR 2008.
[2]http://www.cs.uu.nl/research/projects/witchcraft

1. Is there a small subset of features, or even one single feature, that is discriminative for all melody groups?
2. Is the membership of a melody group based upon the same feature for all member melodies?
3. Are the feature sets provided in earlier research sufficient for classification of the melodies?

1.1 Human classification of melodies

The Meertens Institute in Amsterdam hosts and researches folk songs of the corpus *Onder de groene linde* that have been transmitted through oral tradition. Musicological experts classify these songs into groups called *melody norms* such that each group is considered to consist of melodies that have a common historic origin. Since the actual historic relation between the melodies is not known from documentary evidence, the classification is based on similarity assessments. If the similarity between two melodies is high enough to assume a plausible genetic relation between them, the two melodies are assigned to the same melody norm. In the human process of assigning melody norms some melodies receive the status of a *prototypical* melody of their norms as the most typical representative. All other melody candidates are then compared to this prototypical melody in order to decide whether they belong to this norm.

The classification of melodies into groups of related melodies is a special case of human categorization in music. In order to be able to retrieve melodies belonging to the same melody norm we have to investigate whether all melodies belonging to a melody norm share a set of common features or vary in the number and kind of characteristic features they possess. Two different views of categorization are relevant for this.

The *classical* view on categorization goes back to Aristotle and defines a category as being constituted of all entities that posses a common set of features. In contrast to this, the *modern* view claims that most natural concepts are not well-defined but rather that individual exemplars may vary in the number of characteristic features they possess. The most prominent models according to this view are Wittgenstein's *family resemblance* model (see [9]) and Rosch's *prototype* model (see [7]). Deliege in [2] and Ziv & Eitan in [11] provide arguments that the family resemblance and the prototype model are most appropriate to describe the categories built in Western classical music.

2 Similarity annotations

The study of melodic similarity in this paper contributes to the development of a search engine for the collection of Dutch folk songs *Onder de groene linde*, which contains both audio data, metadata and paper transcriptions. The test collection employed consists of 1198 encoded songs (MIDI and **kern formats) segmented into phrases. The songs have been classified into melody norms. Three experts annotated four melody norms in detail. For each melody group one expert determined a reference melody that is the most prototypical melody. All other melodies of the group were compared to the reference melody.

The annotation data consists of judgements concerning the contribution of different musical dimensions to the similarity between the melody and the prototype of its melody. In daily practice, the experts mainly perform the similarity evaluation in an intuitive way. In order to analyze this complex and intuitive similarity evaluation, we specified the musical dimensions of the annotations in close collaboration with the experts. These dimensions are rhythm, contour, motifs, form and mode. They describe important factors within the decision process of assigning melody norms according to the experts. In order to be used as a ground truth for computational algorithms we standardized the human evaluation such that numeric values are assigned to most of the dimensions. For these we distinguish three different numeric values 0, 1 and 2:[3]

0. The two melodies are not similar, hence according to this dimension a relation cannot be assumed.
1. The two melodies are somewhat similar, a relation according to this dimension is not implausible.
2. The two melodies are obviously similar, a relation according to this dimension is highly plausible.

For each dimension we defined a number of criteria that the human decision should be based upon when assigning the numeric values. These criteria are as concrete as necessary to enable the musicological experts to give reliable ratings that are in accordance with their intuitive assignments. However, the criteria still leave room for personal interpretation. With these criteria we developed a specific way of defining contour, rhythm, form etc. that seemed most appropriate for the given musical material.

2.1 Criteria for the similarity annotations

In this section we describe the criteria for all musical dimensions that are rated numerically.

2.1.1 Rhythm

We defined the following criteria for the comparison of two melodies with respect to their rhythmic similarity.

- If the two songs are notated in the same, or a comparable meter (e.g. 2/4 and 4/4), then count the number of transformations needed to transform the one rhythm into the other (see Figure 1 for an example of a transformation):
  - If the rhythms are exactly the same or contain a perceptually minor transformation: value 2.
  - If one or two perceptually major transformations needed: value 1.
  - If more than two perceptually major transformations needed: value 0.
- If the two songs are not notated in the same, or a comparable meter (e.g. 6/8 and 4/4), then the notion of transformation cannot be applied in a proper manner (it is unclear which durations correspond to each other). The notation in two very different meters indicates that the rhythmic structure is not very similar, hence a value of 2 is not appropriate.
  - If there is a relation between the rhythms to be perceived: value 1.
  - If there is no relation between the rhythms to be perceived: value 0.

In all cases "rhythm" refers to the rhythm of one line. Hence the songs are being compared line-wise.

[3] Differentiating more than three values proved to be an inadequate approach for the musicological experts.

Figure 1: Example of a rhythmic transformation: In the first full bar one transformation is needed to transform the rhythm of the upper melody into the rhythm of the lower melody.

2.1.2 Contour

The contour is an abstraction of the melody. Hence it remains a subjective decision which notes are considered important for the contour. From the comparison of the lines we cannot automatically deduct the value for the entire melody via the mean value. Therefore we also give a value for the entire melody that is based on fewer points of the melody and hence on a more abstract version of the melody than the line-wise comparison. We defined the following criteria:

- For the line-wise comparison:
  - Determine begin (if the upbeat is perceptually unimportant, choose the first downbeat as begin) and end of the line and 1 or 2 turning points (extreme points) in between.
  - Based on these 3 or 4 points per line determine whether the resulting contour of the lines are very similar (value 2), somewhat similar (value 1) or not similar (value 0).
- For the comparison of the global contour using the entire song:
  - Decide per line: if the pitch stays in nearly the same region choose an average pitch for this line; if not, choose one or two turning points.
  - Compare the contour of the entire song consisting of these average pitches and turning points.
  - If the melody is too long for this contour to be memorized, then choose fewer turning points that characterize the global movements of the melody.

2.1.3 Motifs

The decision to assign a certain norm to a melody is often based on the detection of single characteristic motifs. Hence it is possible that the two melodies are different on the whole, but they are recognized as being related due to one or more common motifs. We defined the following criteria:

- If at least one very characteristic motif is being recognized: value 2.
- If motifs are shared but they are not very characteristic: value 1.
- No motifs are shared: value 0.

Characteristic in this context means that the motif serves as a basic cue to recognize a relation between the melodies.

2.1.4 Mode

Concerning the tonality we distinguish the following modes: Major/Ionian, Minor/Aeolian, Dorian, Phrygian, Lydian and Mixolydian. Since a piece in D Minor might be perceived as a slight variation of the same piece in Dorian we assign in a generalization of this observation to all modes that exhibit minor characteristics the value 1 when compared with each other. The same applies to modes with major characteristics. Hence we arrange all modes into two groups. Group 1: Major/Ionian, Lydian and Mixolydian; group 2: Minor/Aeolian, Dorian, Phrygian. This leads to the following criteria:

- If the two melodies have exactly the same mode: value 2.
- If the modes of the two melodies are different but belong to the same group: value 1.
- If the modes of the two melodies belong to different groups: value 0.

2.1.5 Text

In some cases it is possible that the text is the main reason to assign two melodies to the same norm, even though the musical material does not provide clear clues about a relation between the melodies.

Therefore we examine whether the comparison of the texts of two songs suggests a relation between them. Hence we define that a value of 2 is assigned whenever the texts obviously indicate a genetic relation between the two songs, a value of 1 is assigned whenever the text might indicate a relation (but not for certain) and a value of 0 whenever no relation is obvious. The principles for the assigning of the values are as follows:

- If the text is either literally the same, or semantically the same or the *strophic form* is *characteristic* and the same (or any combination of these factors): value 2.
- If parts of the texts are literally or semantically the same, or the strophic form is the same but not very characteristic, the combination of these factors might still indicate a significant relationship: value 2.
- If only parts of the text are literally, or semantically or according to the strophic form the same (or any combination of these factors) and the partial resemblances or their combination is not very convincing: value 1.
- If none of the above cases applies: 0.

The *strophic form* is defined by the following features: number of accents per line, rhyme gender, rhyme scheme (refrain). A strophic form is *characteristic* if it contains uncommon patterns, such as uneven verse lengths and an irregular rhyme scheme. Usually characteristic forms are rare i.e. they serve just one melody type.

2.1.6 Form

We consider the form of a melody not necessarily as an important factor for classification, since we can find melodies of very different line numbers and forms within the same melody norm. However, for testing purposes and in order to make reductions possible, such as from ABCDCD to ABCD, knowing the form is very valuable for the computational similarity measures.

- Annotate the form of both songs in letters (e.g. ABBCC).
- Annotate AA (no apostrophe) if the second line is a literal repetition.
- Annotate AA if the second line is a repetition with variation due to a difference in the number of text syllables: a note is subdivided due to an additional syllable in the text.
- Annotate AA' (with apostrophe) if the second line is a repetition with some variation, especially it there is an other ending ("ouvert-clos").

- Annotate AA' also in the special case that A' is a pitch transposition of A.

3 Experiment on creating annotations

From the set of 1198 encoded melodies 4 melody norms containing 11–16 melodies each have been selected to be annotated by three musicological experts for an initial experiment on the similarity annotation. These are the melody norms *Frankrijk buiten de poorten 1* (short: *Frankrijk*), *Daar was laatst een boerinnetje* (short: *Boerinnetje*), *Daar was laatst een meisje loos 1* (short: *Meisje*) and *Toen ik op Neerlands bergen stond* (short: *Bergen*). For each melody norm one musicological expert determined the reference melody. Similarity ratings were assigned to all other melodies of the same norm with respect to the reference melody. In a first stage of the experiment *Frankrijk* and *Boerinnetje* were annotated, in a second stage *Meisje* and *Bergen*. After the first stage the results were discussed with all experts.

3.1 Agreement among the experts

Table 1 gives an overview of the agreement among the three experts for all musical dimensions using three categories. Category A counts the number of total agreement, i.e. all three experts assigned the same value. Categories PA1 and PA2 count the number of partial agreements such that two experts agreed on one value while the third expert chose a different value. In PA1 the difference between the values equals 1 (e.g. two experts assigned a 1 while one expert assigned a 2). In PA2 the difference between the values equals 2 (e.g. two experts assigned 0 while one expert assigned a 2). Category D counts the cases in which all experts disagree.

Melody Norm	A	PA1	PA2	D
Frankrijk	58.7	38.1	1.6	1.6
Boerinnetje	50.8	42.6	0.5	6.1
Meisje	70.4	27.6	1	1
Bergen	77.5	18.5	1.1	2.9
Average	64.3	31.7	1.1	2.9

Table 1: Comparison of agreement among three experts: A for total agreement, PA1 and PA2 for partial agreement D for disagreement (see section 3.1 for further details). Numbers are percentages.

Both the percentage of disagreement in category D and the percentage of partial agreement PA2 containing both values for *not similar* and *very similar* are quite low. The category of total agreement A comprises the majority of the cases with 64.3%. Moreover, comparing the values obtained for *Frankrijk* and *Boerinnetje* to those for *Meisje* and *Bergen* reveals that the degree of agreement is much higher within the second stage of the experiment after the discussion of the results of the first stage. Hence, this experiment indicates that the musical dimensions have been established in such a way that there is considerable agreement among the musical experts as to how to assign the similarity values.

Melody Norm	*Frankrijk*			*Boerinnetje*			*Meisje*			*Bergen*		
Value	0	1	2	0	1	2	0	1	2	0	1	2
Rhythm	0	1.3	98.7	11.2	51.6	37.2	3.3	8.2	88.5	3.5	15.8	80.7
Global contour	0	31.7	68.3	12.8	48.7	38.5	33.3	13.3	53.4	2.5	10.3	87.2
Contour per line	5.6	52.5	40.9	41.9	26.4	31.7	20.7	31.8	47.5	4.8	22.5	72.7
Motifs	0	36.6	63.4	0	20.5	79.5	13.3	16.7	70	0	17.9	82.1
Mode	13.3	13.3	83.4	0	0	100	0	0	100	0	0	100

Table 2: Distribution of the assigned values within each dimension per melody norm as percentages.

3.2 Comparing dimensions across melody norms

Table 2 lists the distribution of the assigned values within each musical dimension for all melody norms. In three melody norms the dimension *mode* receives in 100% of the cases the value 2, since all melodies of the norm belong to the same mode. However, mode as an isolated dimension can hardly function as a discriminative variable for the classification of the melodies. In the following we study the values for the other musical dimensions.

Both *Frankrijk* and *Meisje* score highest for rhythm (98.7% and 88.5% for value 2), while *Boerinnetje* scores highest for motifs (79.5%) and *Bergen* for global contour (87.2%). For *Bergen* the dimensions motifs and rhythm receive noticeably high scores for value 2 as well (both above 80%), while for *Frankrijk* all other dimensions than rhythm score below 70% for value 2.

Hence, the importance of the different musical dimensions regarding the similarity assignment of melodies belonging to one norm varies between the norms. Moreover, in most of the cases single dimensions are not characteristic enough to describe the similarity of the melodies belonging to one melody norm.

The best musical feature (excluding mode) of *Boerinnetje* scores 79% for value 2, the other musical dimensions score below 40%. From this perspective, the melodies of *Boerinnetje* seem to form the least coherent group of all four melody norms. While *Frankrijk* receives the highest rating in a single dimension for value 2, all other dimensions score relatively low. *Bergen* scores in all dimensions above 72% for the value 2. Hence these melodies seem to be considerably similar to the reference melody across all dimensions. For *Meisje* two dimensions receive scores above 70% for value 2, on the other hand three dimensions have considerably high scores (between 13% and 33%) for the value 0. Hence this norm contains melodies with both very similar and very dissimilar aspects.

Comparing the contribution of the musical dimensions reveals that the contour scores for only one melody norm (*Bergen*) above 70% for value 2. Both rhythm and motifs score above 70% for value 2 in three out of four cases. Hence rhythm and motifs seem to be more important than contour for the human perception of similarity in these experiments.

3.3 Similarity within melody norm

As a measurement for the degree of similarity of each melody within the norm to the reference melody we calculated the average over the dimensions rhythm, global contour, contour per line and motifs. The results show, that the degree of similarity within the norm can vary with considerable amount. For instance, in the melody norm *Meisje* two melodies (NLB073517-01 and NLB111465-01, see Table 3) score higher than 95% for value 2, while two melodies score lower than 20% for value 2 with corresponding high scores for value 0 (NLB071449-01 and NLB139121-01). The degree of similarity of the melodies within the groups *Frankrijk*, *Boerinnetje* and *Bergen* is listed in Tables 5 to 7 in the appendix.

value	0	1	2
NLB070321-01	12.5	22.9	64.6
NLB070560-01	4.2	8.3	87.5
NLB071374-01	0	12.5	87.5
NLB071449-01	56.3	25	18.7
NLB071734-01	4.2	14.6	81.2
NLB072923-01	16.4	8.3	75
NLB073517-01	0	0	100
NLB111465-01	0	4.2	95.8
NLB139116-01	39.6	37.5	22.9
NLB139121-01	43.3	41.7	15

Table 3: Degree of similarity of all melodies of the group *Meisje* to the reference melody NLB070412-01 averaged over all dimensions as percentages.

The evaluation of single dimensions shows that also within these single features the degree of similarity to the reference melody varies. For instance, *Meisje* scores for the dimension rhythm on average 88.5% for value 2. However, melody NLB071449-01 scores for rhythm only 42% for value 2 and 33% for value 0. Hence we conclude, that there is not one characteristic (or one set of characteristics) that all melodies of a melody norm share with the reference melody.

3.4 Discussion

From sections 3.2 and 3.3 we conclude that both across and within the melody norms the importance of the musical dimensions for perceived similarity varies.

There is not one characteristic (or one set of characteristics) that all melodies of a melody norm share with the reference melody. Therefore, the category type of the melody norms cannot be described according to the classical view on categorization, but rather to the modern view. This agrees with the studies in [2] and [11] on categorization in Western classical music.

4 Evaluating Computational features

This section complements the preceding one by an evaluation of computational features related to melodic similarity.

4.1 Global Features

We evaluate the following three sets of features:

- 12 features provided by Wolfram Steinbeck [8], listed in Table 8.

- 40 features provided by Barbara Jesser [3], listed in Table 9.
- 40 rhythm, pitch and melody features implemented in jSymbolic by Cory McKay *et al.* [5], listed in Table 10.

The sets of Steinbeck and Jesser were specifically assembled to study groups of folk songs within the Essen Folk Song Collection that are related through the process of oral transmission. Because our corpus consists of folk song melodies, the evaluation of especially these two feature sets is important to get an indication of the value of computational features in general. McKay's set has a general purpose.

All features for which absolute pitch is needed (e.g. Steinbeck's Mean Pitch) were removed because not all melodies in our corpus have the same key. Also the multidimensional features from the set of jSymbolic were removed because they are primarily needed to compute other features. Thus we have 92 features, which are characterized as 'global' because for each feature an entire song is represented by only one value.

These features can be considered aspects of the musical dimensions that were chosen for the manual annotations. For example, features like the fraction of descending minor seconds, the size of melodic arcs and the amount of arpeggiation contribute to contour, but they do not represent the holistic phenomenon of contour exhaustively.

4.2 Feature evaluation method

For the four melody norms that were examined in the previous sections, the discriminative power of each individual feature is evaluated. The songs are divided into two groups: one group contains the songs from the melody norm under consideration and the other group all other songs from the test collection. The intersection of the normalized histograms of both groups is taken as a measure for the discriminative power of a feature:

$$I_{mn} = 1 - \frac{\sum_{i=1}^{n} \left| H_{mn}[i] - H_{other}[i] \right|}{\sum_{i=1}^{n} H_{mn}[i]}$$

where $H_{mn}[i]$ is the value for bin i of the histogram of the songs belonging to the melody norm mn and H_{other} is the histogram for all other songs. Both histograms have n bins, with the same edges. For the nominal features n is the number of possible values, and for real valued features, $n = 11$, which is the size of the smallest class.

The smaller the intersection, the larger the discriminative power of the feature. The intersection therefore indicates whether a search algorithm that makes use of a certain feature could be successful or not retrieving the songs of the melody norm from the entire corpus.

Normalization of the histograms is needed for the intersection to get comparable values between 0 and 1. Because the four melody norms all have very few melodies compared to the entire corpus, this involves heavy scaling. As a consequence, the intersection value only serves as an indicator for the achievable recall of a retrieval system using the feature. If both $H_{mn}[i] > 0$ and $H_{other}[i] > 0$ the absolute number of songs in $H_{other}[i]$ is almost certainly larger. Therefore, to get an indication of the precision as well, the absolute values of H_{other} should be considered.

4.3 Results

Table 4 lists the best scoring features. For both *Boerinnetje* and *Meisje* none of the features have low values. According to the annotation data the similarity of the melodies in these norms to their respective reference melody is less obvious; *Boerinnetje* is the least characteristic of all melody norms, while *Meisje* contains melodies with both very similar and dissimilar aspects.

Feature	I_F	I_B	I_M	I_N
JESdminsecond	**0.068**	0.764	0.445	0.686
STBAmbitus	0.739	0.720	0.622	**0.183**
Range	0.739	0.720	0.622	**0.183**
JESprime	**0.197**	0.575	0.574	0.719
Repeated_Notes	**0.197**	0.575	0.574	0.719
JESmeter	**0.211**	0.540	0.632	0.269
Stepwise_Motion	**0.227**	0.667	0.474	0.566
Chromatic_Motion	**0.250**	0.788	0.500	0.644
JESdstep	**0.251**	0.637	0.525	0.567
Pitch_Variety	0.685	0.566	0.451	**0.253**
JESnumlines	**0.258**	0.428	0.400	0.582
JESafifth	**0.263**	0.749	0.717	0.749
STBDurationLineCorrespondence	0.579	0.714	0.709	**0.288**
STBFractionStressed	0.520	0.388	0.532	**0.304**
Amount_of_Arpeggiation	**0.318**	0.531	0.555	0.699
Triple_Meter	0.810	**0.323**	0.684	0.810
Combined_Strength_of_Two_Strongest_Rhythmic_Pulses	0.600	**0.329**	0.475	0.433
Distance_Between_Most_Common_Melodic_Intervals	**0.358**	0.776	0.892	0.920
Most_Common_Melodic_Interval_Prevalence	**0.377**	0.861	0.485	0.630
Polyrhythms	0.799	**0.378**	0.895	0.482
Strength_Ratio_of_Two_Strongest_Rhythmic_Pulses	0.456	0.392	0.540	**0.379**
STBFractionEqualDurations	0.570	0.794	0.683	**0.385**
JESaminthird	**0.387**	0.733	0.601	0.450

Table 4: I_{mn} for the best scoring features sorted according to the smallest intersection (in bold) of any of the melody norms *Frankrijk* (F), *Boerinnetje* (B), *Meisje* (M) and *Bergen* (N). The prefixes JES- and STB- mean that the feature is in the set of Jesser or Steinbeck. The other features are from jSymbolic.

We observe that the best feature for *Frankrijk*, JESdminsecond, has quite high values for the other melody norms, which means that it is only discriminative for *Frankrijk*. This feature measures the fraction of melodic intervals that is a descending second. Apparently a large number of descending minor seconds is a distinctive characteristic of *Frankrijk*, but not of the other melody norms. Melodic samples are shown in Figure 1 and the histograms for this feature are shown in Figure 2. While for the normalized histograms the largest bin of $H_{Frankrijk}$ is much larger than the corresponding bin of H_{other}, the absolute values are 7 for H_{other} and 8 for $H_{Frankrijk}$. This means that a retrieval engine using only this feature would achieve a quite low precision.

The annotations suggest that rhythm contributes most to the similarity of the songs in the melody norm *Frankrijk*. Furthermore, the investigation of a set of melody norms using a rhythmic similarity approach in [10] indicates that the melodies of *Frankrijk* are rhythmically more similar to each other than to melodies of other norms. However, none of the rhythmic features of the three sets is discriminative.

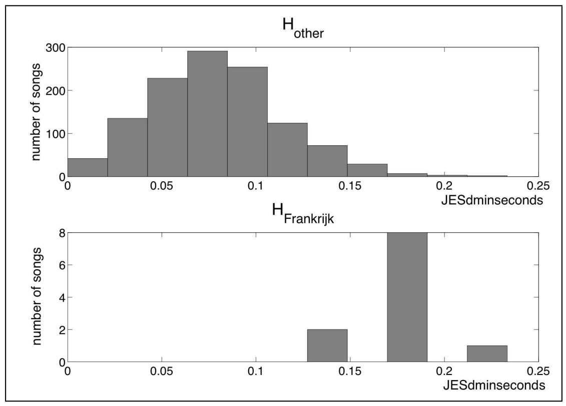

Figure 2: Unnormalized histograms for JESdminseconds for both *Frankrijk* and the other songs.

Most of the lowest values in Table 4 are for *Frankrijk*. STBAmbitus and Range (which are actually the same feature, but from different sets) receive low values for *Bergen*. According to the annotation data, *Bergen* is the only melody norm with high ratings for both the global contour and the line-wise contour. Range is an aspect of contour. The melodies of *Bergen* typically have a narrow ambitus.

For all other features not shown in Table 4, $I_{mn} \geq 0.387$, which indicates that these are not discriminative.

4.4 Discussion

The evaluation of the individual features from the three feature sets shows that there is no single feature in the current set that is discriminative for all four melody norms. Most of the few features that proved discriminative are only so for *Frankrijk*. Therefore, it is not even the case that we find per melody norm a good feature. None of the three sets of features is sufficiently complete for this.

In the manual annotations we observed that motifs are important for recognizing melodies. There are many kinds of motifs: a rhythmic figure, an uncommon interval, a leap, a syncopation, and so on. Therefore it is not possible to grasp the discriminative power of motifs in only a few features. Besides that, global features are not suitable to reflect motifs, which are local phenomena. This is an important shortcoming of the approach based on global features.

It proves difficult to find clear links between the musical dimensions used in the manual annotations and the computational features. The two approaches reside on different levels of abstraction. Computational features have to be computed deterministically. Hence, low level and countable characteristics of melodies are more suited than the more intuitive and implicit concepts that are used by the human mind. Nevertheless, computational features provided complementary insights to the manual annotations, such as the characteristic descending minor second for *Frankrijk*.

5 Concluding Remarks

With the results of both approaches, we are able to provide answers to the questions stated in the introduction. First, there is no single feature or musical dimension that is discriminative for all melody norms. Second, it is not guaranteed that one single feature or musical dimension is sufficient to explain the similarity of each individual melody to the melody norm. Third, although two of the sets of computational features were specifically assembled for folk song melodies, none of the involved sets provides features that are generally useful for the classification task at hand. A next step would be to evaluate subsets of features instead of individual ones. Although these might prove more discriminative than single features, the importance of the dimension 'motifs' indicates strongly that local model-based features are needed rather than adding more global statistic ones.

The manual annotation of melodic similarity proved a valuable tool to analyze the complex and intuitive similarity assessment of the experts by specifying the constituent parts that contribute to the specific perception of melodic similarity that underlies folksong classification. Therefore a larger set of such annotations is now being created. The annotation data can also be used to evaluate similarity measures that are based on one or more of the musical dimensions.

Acknowledgments. This work was supported by the Netherlands Organization for Scientific Research within the WITCHCRAFT project NWO 640-003-501, which is part of the CATCH-program. We thank musicologists Ellen van der Grijn, Mariet Kaptein and Marieke Klein for their contribution to the annotation method and for creating the annotations.

References

[1] Ahlbäck, S. *Melody beyond notes.* PhD thesis Göteborgs Universitet, 2004.

[2] Deliege, I. "Prototype effects in music listening: An empirical approach to the notion of imprint", *Music Perception*, 18 (2001), 371–407.

[3] Jesser, B. *Interaktive Melodieanalyse.* Bern, 1991.

[4] Müllensiefen, D. & Frieler, K. "Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgements", *Computing in Musicology*, 13 (2004), 147–177.

[5] McKay, C. & Fujinaga, I. "Style-independent computer-assisted exploratory analysis of large music collections", *Journal of Interdisciplinary Music Studies*, 1 (2007), 63–85.

[6] McKay, C. *Automatic Genre Classification of MIDI Recordings.* [MA Thesis]. McGill University, Montreal, 2004.

[7] Rosch, E. "Natural Categories", *Cognitive Psychology*, 4 (1973), 328–350.

[8] Steinbeck, W. *Struktur und Ähnlichkeit: Methoden automatisierter Melodieanalyse.* Kassel, 1982.

[9] Wittgenstein, L. *Philosophical investigations.* London, 1953.

[10] Volk, A., Garbers, J., Van Kranenburg, P., Wiering, F., Veltkamp, R., Grijp, L. "Applying Rhythmic Similarity based on Inner Metric Analysis to Folksong Research", *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, 2007, 293–296.

[11] Ziv, N. & Eitan, Z. "Themes as prototypes: Similarity judgments and categorization tasks in musical contexts", *Musicae Scientiae*, Discussion Forum 4A, 99–133, 2007.

6 Appendix

value	0	1	2
NLB072267-02	60.4	33.3	6.2
NLB072268-01	0	50	50
NLB072268-02	0	52.1	46.9
NLB072274-01	0	8.3	91.7
NLB072277-01	0	16.7	83.3
NLB073784-02	0	15.8	84.2
NLB074211-01	4.2	50	45.8
NLB074211-03	0	16.7	83.3
NLB074341-01	3.1	43.8	53.1
NLB074937-01	0	17.7	82.3

Table 5: Degree of similarity of all melodies of the group *Frankrijk* to the reference melody NLB072275-01 averaged over the dimensions rhythm, contour per line, global contour, and motifs as percentages.

value	0	1	2
NLB072379-01	0	2.1	97.9
NLB072570-01	3.1	24	72.9
NLB072672-01	29.2	33.3	37.5
NLB072722-01	4.2	11.4	84.4
NLB073066-01	27.1	44.8	28.1
NLB074172-01	28.8	52.3	18.9
NLB074200-01	13.5	60.4	26.1
NLB074212-01	25	44.8	30.2
NLB074740-01	20.6	50	29.4
NLB075085-01	1.2	22.5	76.3
NLB075085-03	5	21.3	73.7
NLB148976-01	26	53.5	20.5
NLB149150-01	30.6	58.3	11.1

Table 6: Degree of similarity of all melodies of the group *Boerinnetje* to the reference melody NLB074966-01 averaged over the dimensions rhythm, contour per line, global contour, and motifs as percentages.

value	0	1	2
NLB071985-01	9.7	50	40.3
NLB072020-01	0	0	100
NLB072836-01	0	0	100
NLB073625-01	0	1.4	98.6
NLB073947-01	0	0	100
NLB074329-01	0	0	100
NLB074709-01	5.5	26.4	68.1
NLB075151-01	0	6.9	93.1
NLB075248-01	2.8	23.6	73.6
NLB076111-01	9.6	56.4	34
NLB076111-02	1.9	40.4	57.7
NLB076848-01	0	2.8	97.2
NLB076853-01	5.6	8.3	86.1

Table 7: Degree of similarity of all melodies of the group *Bergen* to the reference melody NLB071076-01 averaged over the dimensions rhythm, contour per line, global contour, and motifs as percentages.

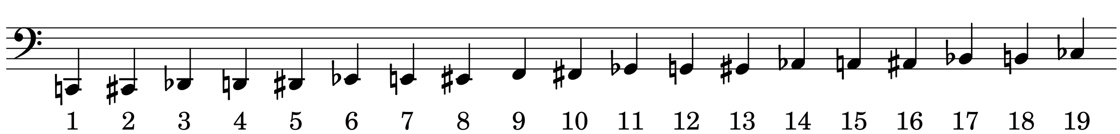

Figure 3: Base-19 pitch representation used by Wolfram Steinbeck.

Table 8: The set of features that is defined by Wolfgang Steinbeck [8]. *MeanPitch* was not used in our experiment.

Feature	**Description** (page numbers refer to [8])
MeanPitch	Mean of the pitches in de melody. For all pitch-based features the base-19 pitch representation depicted in Figure 3 has been used. The pitches are weighted according to their length (p.156ff).
StdPitch	Standard deviation of the pitch (p.156ff).
Ambitus	Difference between the highest and lowest pitch in the melody (p.155).
MeanInterval	Mean of the size of the intervals. The intervals between the phrases are not taken into account (p.165ff).
StdInterval	Standard Deviation of the size of the intervals (p.165ff).
ChangingDirection	The fraction of the intervals that cause a change of direction (p.149f).
MeanSteepness	The steepness is the deviation in pitch between two turning points divided by the duration. This feature is the mean of these steepnesses (p.173ff).
FractionStressed	The sum of durations that start on a stressed beat as fraction of the total duration (p.178ff).
FractionDottedDuration	The fraction of transitions between pitches that has duration quotient 3:1 (p.152ff).
FractionHalfDuration	The fraction of transitions between pitches that has duration quotient 2:1 or 1:2 (p.152ff).
FractionEqualDurations	The fraction of transitions between pitches that has duration quotient 1:1 (p.152ff).
PitchLineCorrelation	The correlation of the pitch contours of the individual lines. For each line the maximum of the correlations with the other lines is taken. Of these values the mean is computed (p.299ff, p.93).
DurationLineCorrespondence	Similarity of the sequence of durations. This is computed in the same way as the previous feature, but instead of correlation the fraction of durations that corresponds is taken (p.299ff).

Table 9: The features defined by Jesser [3] used in the experiment.

Feature	**Description**
prime	fraction of the melodic intervals that is a prime.
aminsecond	fraction of the melodic intervals that is an ascending minor second.
amajsecond	fraction of the melodic intervals that is an ascending major second.
aminthird	fraction of the melodic intervals that is an ascending minor third.
amajthird	fraction of the melodic intervals that is an ascending major third.
afourth	fraction of the melodic intervals that is an ascending perfect fourth.
aaugfourth	fraction of the melodic intervals that is an ascending augmented fourth.
afifth	fraction of the melodic intervals that is an ascending perfect fifth.
aminsixth	fraction of the melodic intervals that is an ascending minor sixth.
amajsixth	fraction of the melodic intervals that is an ascending major sixth.
aminseventh	fraction of the melodic intervals that is an ascending minor seventh.
amajseventh	fraction of the melodic intervals that is an ascending major seventh.
aoctave	fraction of the melodic intervals that is an ascending perfect octave.
ahuge	fraction of the melodic intervals that is larger than an ascending octave.
dminsecond	fraction of the melodic intervals that is a descending minor second.
dmajsecond	fraction of the melodic intervals that is a descending major second.
dminthird	fraction of the melodic intervals that is a descending minor third.
dmajthird	fraction of the melodic intervals that is a descending major third.
dfourth	fraction of the melodic intervals that is a descending fourth.
daugfourth	fraction of the melodic intervals that is a descending augmented fourth.
dfifth	fraction of the melodic intervals that is a descending perfect fifth.
dminsixth	fraction of the melodic intervals that is a descending minor sixth.
dmajsixth	fraction of the melodic intervals that is a descending major sixth.
dminseventh	fraction of the melodic intervals that is a descending minor seventh.
dmajseventh	fraction of the melodic intervals that is a descending major seventh.
doctave	fraction of the melodic intervals that is a descending perfect octave.
astep	fraction of the melodic intervals that is an ascending step.
aleap	fraction of the melodic intervals that is a ascending leap.
dstep	fraction of the melodic intervals that is a descending step.
dleap	fraction of the melodic intervals that is a descending leap.
shortestlength	shortest duration such that all durations are a multiple of this shortest duration, except for triplets.
doublelength	fraction of the notes with duration of twice the shortest duration.
triplelength	fraction of the notes with duration of three times the shortest duration.
quadruplelenght	fraction of the notes with duration of four times the shortest duration.
dotted	fraction of the notes that is dotted.
triplets	fraction of the notes that belongs to a triplet.
meter	the meter.
hasmeterchanges	‘yes’ if there are meter changes, ‘no’ otherwise.
numlines	number of lines.
numpitchclasses	number of distinct pitch classes.

Table 10: The features defined by Cory McKay [6, Ch. 4] that are used in the experiment.

Feature	**Description as given by Cory McKay [6, Ch. 4]**
Amount of Arpeggiation	Fraction of horizontal intervals that are repeated notes, minor thirds, major thirds, perfect fifths, minor sevenths, major sevenths, octaves, minor tenths or major tenths.
Average Melodic Interval	Average melodic interval (in semi-tones).
Changes of Meter	Set to 1 if the time signature is changed one or more times during the recording.
Chromatic Motion	Fraction of melodic intervals corresponding to a semi-tone.
Combined Strength of Two Strongest Rhythmic Pulses	The sum of the frequencies of the two beat bins of the peaks with the highest frequencies.
Direction of Motion	Fraction of melodic intervals that are rising rather than falling.
Distance Between Most Common Melodic Intervals	Absolute value of the difference between the most common melodic interval and the second most common melodic interval.
Dominant Spread	Largest number of consecutive pitch classes separated by perfect 5ths that accounted for at least 9% each of the notes.
Duration of Melodic Arcs	Average number of notes that separate melodic peaks and troughs in any channel.
Harmonicity of Two Strongest Rhythmic Pulses	The bin label of the higher (in terms of bin label) of the two beat bins of the peaks with the highest frequency divided by the bin label of the lower.
Interval Between Strongest Pitch Classes	Absolute value of the difference between the pitch classes of the two most common MIDI pitch classes.
Interval Between Strongest Pitches	Absolute value of the difference between the pitches of the two most common MIDI pitches.
Melodic Fifths	Fraction of melodic intervals that are perfect fifths.
Melodic Octaves	Fraction of melodic intervals that are octaves.
Melodic Thirds	Fraction of melodic intervals that are major or minor thirds.
Melodic Tritones	Fraction of melodic intervals that are tritones.
Most Common Melodic Interval	Melodic interval with the highest frequency.
Most Common Melodic Interval Prevalence	Fraction of melodic intervals that belong to the most common interval.
Most Common Pitch Class Prevalence	Fraction of Note Ons corresponding to the most common pitch class.
Most Common Pitch Prevalence	Fraction of Note Ons corresponding to the most common pitch.
Number of Common Melodic Intervals	Number of melodic intervals that represent at least 9% of all melodic intervals.
Number of Common Pitches	Number of pitches that account individually for at least 9% of all notes.

Table 10: The features defined by Cory McKay [6, Ch. 4] that are used in the experiment.

Feature	**Description as given by Cory McKay [6, Ch. 4]**
Number of Moderate Pulses	Number of beat peaks with normalized frequencies over 0.01.
Number of Relatively Strong Pulses	Number of beat peaks with frequencies at least 30% as high as the frequency of the bin with the highest frequency.
Number of Strong Pulses	Number of beat peaks with normalized frequencies over 0.1.
Pitch Class Variety	Number of pitch classes used at least once.
Pitch Variety	Number of pitches used at least once.
Polyrhythms	Number of beat peaks with frequencies at least 30% of the highest frequency whose bin labels are not integer multiples or factors (using only multipliers of 1, 2, 3, 4, 6 and 8) (with an accepted error of +/- 3 bins) of the bin label of the peak with the highest frequency. This number is then divided by the total number of beat bins with frequencies over 30% of the highest frequency.
Quintuple Meter	Set to 1 if numerator of initial time signature is 5, set to 0 otherwise.
Range	Difference between highest and lowest pitches.
Relative Strength of Most Common Intervals	Fraction of melodic intervals that belong to the second most common interval divided by the fraction of melodic intervals belonging to the most common interval.
Relative Strength of Top Pitch Classes	The frequency of the 2nd most common pitch class divided by the frequency of the most common pitch class.
Relative Strength of Top Pitches	The frequency of the 2nd most common pitch divided by the frequency of the most common pitch.
Repeated Notes	Fraction of notes that are repeated melodically.
Size of Melodic Arcs	Average melodic interval separating the top note of melodic peaks and the bottom note of melodic troughs.
Stepwise Motion	Fraction of melodic intervals that corresponded to a minor or major second.
Strength of Second Strongest Rhythmic Pulse	Frequency of the beat bin of the peak with the second highest frequency.
Strength of Strongest Rhythmic Pulse	Frequency of the beat bin with the highest frequency.
Strength Ratio of Two Strongest Rhythmic Pulses	The frequency of the higher (in terms of frequency) of the two beat bins corresponding to the peaks with the highest frequency divided by the frequency of the lower.
Strong Tonal Centres	Number of peaks in the fifths pitch histogram that each account for at least 9% of all Note Ons.
Triple Meter	Set to 1 if numerator of initial time signature is 3, set to 0 otherwise.