A MANUAL ANNOTATION METHOD FOR MELODIC SIMILARITY AND THE STUDY OF MELODY FEATURE SETS

Anja Volk, Peter van Kranenburg, Jörg Garbers, Frans Wiering, Remco C. Veltkamp, Louis P. Grijp*
Department of Information and Computing Sciences, Utrecht University and *Meertens Institute, Amsterdam
volk@cs.uu.nl

ABSTRACT

This paper 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. [1] 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 expert judgments about the contribution of different musical dimensions to 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.

[1] http://www.cs.uu.nl/research/projects/witchcraft

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

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 [10]) and Rosch's *prototype*

model (see [6]). 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. [2] 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. 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.

2.1 Criteria for the similarity annotations

2.1.1 Rhythm

- 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.

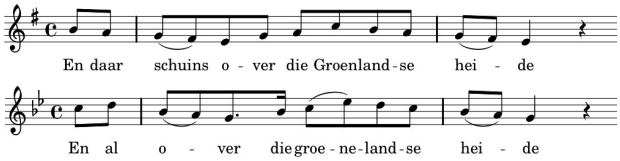

Figure 1. Example of a rhythmic transformation: In the first full bar one transformation is needed to transform the rhythm of the upper melody into the rhythm of the lower melody.

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

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.

- 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:

[2] Text often provides cues as to whether songs are recognized as being genetically related; hence this dimension is annotated too. Form is an auxiliary criterion in order to make reductions possible, such as from ABCC to ABC. Text and form are not discussed here; for their description see [8].

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

– 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.

- 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

We distinguish two groups of modes based on major vs. minor characteristics. Major/Ionian, Lydian and Mixolydian hence form one group, while Minor/Aeolian, Dorian and Phrygian from another group.

- 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.

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.

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 concerning the value 2, while *Boerinnetje* scores highest for motifs and *Bergen* for global contour. Hence the importance of the different musical dimensions regarding the similarity assignment of melodies belonging to one norm varies be-

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.

tween 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 to a considerable amount. For instance, in the melody norm *Meisje* two melodies score higher than 95% for value 2, while two melodies score lower than 20% for value 2 with corresponding high scores for value 0.

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, one melody 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 [7].
- 40 features provided by Barbara Jesser [3].
- 40 rhythm, pitch and melody features implemented in jSymbolic by Cory McKay [5].

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. For the complete list of features we refer to [8].

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 3 lists the best scoring features. For both *Boerinnetje* and *Meisje* none of the features have low values for the intersections. 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

Table 3. I_{mn} for the six best scoring features sorted according to the smallest intersection (in bold) for 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.

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 [9] 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.

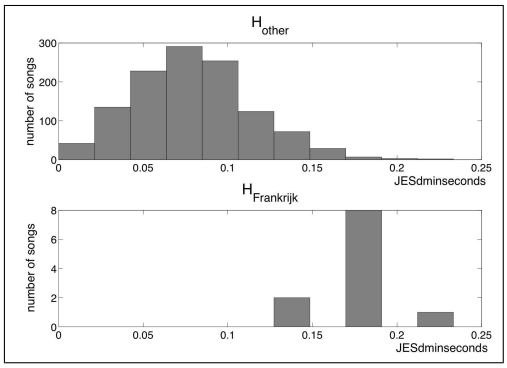

Figure 2. Unnormalized histograms for JESdminseconds for both *Frankrijk* and the other songs.

Most of the lowest values in Table 3 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 mel-

ody 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 3, $I_{mn} \geq 0.211$, 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.

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

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

[8] Volk, A., van Kranenburg, P., Garbers, J., Wiering, F., Veltkamp, R., Grijp, L. "The Study of Melodic Similarity using Manual Annotation and Melody Feature Sets", Technical Report UU-CS-2008-013, Utrecht University.

[9] 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.

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

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