Melodic similarity among folk songs: An annotation study on similarity-based categorization in music

Anja Volk and Peter van Kranenburg

Musicae Scientiae 2012 16: 317 originally published online 24 July 2012
DOI: 10.1177/1029864912448329

MUSICÆ SCIENTIÆ

Musicae Scientiae
16(3) 317–339
© The Author(s) 2012
Reprints and permission: sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1029864912448329
msx.sagepub.com
SAGE

Melodic similarity among folk songs: An annotation study on similarity-based categorization in music

Anja Volk
Utrecht University, the Netherlands

Peter van Kranenburg
Meertens Institute, Amsterdam, the Netherlands

Abstract

In this article we determine the role of different musical features for the human categorization of folk songs into tune families in a large collection of Dutch folk songs. Through an annotation study we investigate the relation between musical features, perceived similarity and human categorization in music. We introduce a newly developed annotation method which is used to create an annotation data set for 360 folk song melodies in 26 tune families. This dataset delivers valuable information on the contribution of musical features to the process of categorization which is based on assessing the similarity between melodies. The analysis of the annotation data set reveals that the importance of single musical features for assessing similarity varies both between and within tune families. In general, the recurrence of short characteristic motifs is most relevant for the perception of similarity between songs belonging to the same tune family. Global melodic features often used for the description of melodies (such as melodic contour) play a less important role. The annotation data set is a valuable resource for further research on melodic similarity and can be used as enriched "ground truth" to test various kinds of retrieval algorithms in Music Information Retrieval. Our annotation study exemplifies that assessing similarity is crucial for human categorization processes, which has been questioned within Cognitive Science in the context of rule-based approaches to categorization.

Keywords

melodic similarity, categorization, musical features, tune families

Introduction

Cognitive Science considers the assessment of similarity a fundamental human capability. Similarity plays a crucial role for many mental processes, such as learning, problem-solving, expecting and memorizing (Goldstone & Son, 2005). Similarity is notably important for

Corresponding author:
Anja Volk, Utrecht University, PO Box 80089, Utrecht, 3508TB, the Netherlands.
Email: a.volk@uu.nl

categorization processes: elements that are perceived as being similar are grouped into the same category. Most cognitive studies in the 1970s considered similarity as the basis of categorization. In the 1980s the fundamental role of similarity for categorization processes was questioned, since the construct of similarity seemed to be too vague to serve as a ground for categorization (Love, 2002). Instead, theory-based approaches to categorization have been developed (e.g., Murphy & Medin, 1985). However, "the theory-based view has itself proven vague" (Love, 2002, p. 103) and has not led to a breakthrough in understanding human categorization. In reviewing recent similarity-based approaches to categorization developed in Cognitive Science, Love (2002) concludes that understanding how similarities are processed will move our understanding of human categorization substantially forward.

Most of the studies on the relation between similarity and categorization in Cognitive Science have been carried out in other domains than music, predominantly in the domains of language and vision. However, aspects related to similarity and human categorization in the domain of music have been investigated in music cognition (e.g., Eitan & Granot, 2009), in ethnomusicology (e.g., Cowdery, 1984), and music theory (e.g., Zbikowski, 2002). These studies suggest that similarity is crucial for categorization in music (e.g., Cambouropoulos, 2001; Deliège, 2001). However, while cognitive studies have shown that listeners have an astonishing ability to build meaningful categories (Dalla Bella & Peretz, 2005), they often do not provide sufficient conclusions about how musical features are used in the human categorization process. On the other hand, studies on the role of musical features for music similarity assessment do not provide conclusions about the role of features for categorization processes in music. Hence, the relation between musical features and similarity-based categorization is hardly understood. This relation is of crucial importance in the area of Music Information Retrieval (MIR). The rapid growth of MIR over the last decades has led to an increased interest in the topic of musical similarity for computational approaches to music. The success of search engines such as Google and Yahoo has encouraged researchers to develop MIR methods that aim to assist users in searching in large collections of digitized musical data (Downie et al., 2009). Music similarity is a fundamental principle used in MIR to retrieve music. One of the most typical approaches to similarity in MIR is the use of broad categories, such as genre labels, in order to infer similarity relations between musical pieces, assuming that pieces belonging to the same category are more similar to each other than across categories. However, MIR often faces the challenge of a lack of domain knowledge on finer grained music similarity going beyond broad categories such as genres, and therefore a lack of human similarity data that computational algorithms can be tested upon.

This paper investigates the relation between perceived similarity, musical features and human categorization through a newly developed annotation method. The annotation method has been developed together with folk song experts in the context of researching oral transmission in a large collection of Dutch folk songs; the method follows the experts' daily working procedure. The method provides insights into the experts' categorization of Dutch folk songs into tune families based on the assessment of melodic similarity. The experts gave ratings on the perceived similarity of songs regarding relevant musical features. These features are rhythm, contour, motifs, and lyrics. The features have been rated numerically regarding their contribution to similarity ('0' for not similar, '1' for somewhat similar, '2' for obviously similar). By analysing the resulting annotation data set (containing annotations for 360 melodies in 26 tune families), we will show that the importance of single musical features for assessing similarity varies both between and within tune families. The annotation data set reveals that global musical features (such as contour) usually employed for describing melodic similarity are less important for categorization based on similarity than local features, namely short characteristic motifs.

The results of our annotation study provide information on important musical features of similarity gained through introspection into the human process of assessing similarity, which is unprecedented in the detail with which aspects of melodic similarity are quantized. We first describe the context of our annotation study within Musicology, Music Information Retrieval, Cognitive Science and Music Cognition in order to identify the specific contribution of this paper to these areas. We then introduce the annotation method, present an analysis of the results of the annotations and give conclusion about the insights gained within the annotation study.

Similarity and categorization

Similarity and categorization are relevant research topics within different research areas that relate to our study on melodic similarity and categorization into tune families. In the following section we will identify, for each area – Musicology, Music Information Retrieval, Cognitive Science and Music Cognition – the contribution of this paper to linking similarity and categorization in music.

Musicology: motivic-thematic relations and tune families

Many research topics in musicology are inherently related to similarity and categorization in music, such as the topic of style analysis. Prominent examples of the discussion on the link between human similarity and categorization in music are the study of motivic-thematic relations in Western classical music (e.g., Réti, 1951; Schoenberg, 1942, 1967; Webern, 1963) and the investigation of tune families within ethnomusicology (e.g., Bayard, 1950; Bronson, 1950; Cowdery, 1984; Wiora, 1941). Both examples relate to our annotation study.

Zbikowski (2002) compares musical motifs as defined by Schoenberg to the basic-level categories introduced by Rosch et al. (1976), which are defined as the highest taxonomic level at which category members have similarly perceived overall shapes. As our annotation study shows, the similarly perceived overall shape of musical motifs is of great importance for the categorization of songs into tune families. A tune family consists of folk songs that are supposed to have a common origin in history (as defined by Bayard, 1950); due to variations introduced in the course of oral transmission, these songs are more or less similar to each other. Different theories have been developed as to what kinds of variation typically are introduced in the course of oral transmission. For instance, Wiora (1941) gives a list of those changes for German folk songs, containing, for instance, changes in contour, tonality and rhythm. However, he provides no quantitative estimation on how these different factors are combined. Bronson (1950) discusses the need to determine the relative weight of different musical dimensions contributing to variation and describes the relative importance of some dimensions for a corpus of British-American folk songs. Klusen et al. (1978) explicitly test the changes in the dimensions of pitch and rhythm introduced through oral transmission in a perceptual study that mimics the process of oral transmission from one person to another. As a result, they show that the rhythmic structure is more stable than the melodic structure.

Cowdery (1984) provides a new perspective on the concept of tune family, arguing that not all members of a tune family need to have one single ancestor in history, since tune families can "blend into each other" (p. 496) over the course of oral transmission. Among the three principles that Cowdery discusses for determining tune families, the "recombining" principle stresses the importance of motifs: melodies are composed from material from the same "pool" of melodic motifs. According to this principle, not only global similarity aspects should be considered for comparing melodies (such as contour, rhythm or tonality), but the role of motifs within the process of oral transmission needs to be studied. Sagrillo (1999) discusses different existing classification systems in folk song research that use different dimensions of the melodies depending on the specific corpus investigated (e.g., Koller, 1902; Krohn, 1903; Stief, 1990; Bartók's and Kodály's system, see a description by Suchoff, 1981); he classifies a corpus of 500 Luxembourgian folk songs by determining phrase similarity.

Deliège (2007) distinguishes the two prominent examples of similarity studied in musicology (motivic-thematic relations in Western classical music and the study of tune families in ethnomusicology) as *internal* similarity relations (within a musical work) versus *external* similarity relations (similarity between different songs belonging to one tune family). Our study contributes to modelling external similarity relations, while we find close connections to typical phenomena of internal similarity relations, namely the importance of motivic relations. The identification of musical features considered important for similarity in our annotation study relates to Wiora's list of typical variations. The quantification of the contribution of different musical features to perceived similarity within the annotation data proves the importance of motifs for assessing the similarity between songs of the tune families, which has been proposed by Cowdery (1984).

Music Information Retrieval:The lack of ground-truth-data on similarity and categorization

Although the term *Music Information Retrieval* was introduced as early as 1966 (see Kassler, 1966), intensive research in this discipline started to prosper in the late 20th century, following a massive digitization of music. The availability of large amounts of digitized musical data created the need to develop methods that enable users to search within large collections of music. Thus, the notion of *music similarity* became a central concept in realizing the access to musical collections: for a given query of the user, similar pieces within the collection are determined. However, MIR often faces the problem of a lack of *ground truth*: what musical pieces should be considered as being similar, so that similarity measures can be developed to automatically detect these similarities?

In this paper we describe the generation of an annotation data set for 360 melodies belonging to a collection of more than 6000 digitized Dutch folk songs. This data set contains ratings of musicological experts on the similarity between melodies that have been classified into tune families. The resulting annotated corpus provides valuable ground-truth-data on fine-grained music similarity for research in MIR.

Cognitive Science:The role of similarity for categorization

The relation between similarity and categorization is a major topic in Cognitive Science (Hahn & Ramscar, 2001). While in the 1970s most cognitive studies considered similarity as the basis of categorization, the fundamental role of similarity for categorization processes was questioned in the 1980s (Love, 2002). For instance, Murphy & Medin (1985) argue, that "many things appear to be similar just because they are members of the same category" and therefore it is an open issue whether similarity explains why the category was formed. Instead, Murphy and Medin (1985) emphasize the important role of people's theories about the world for categorization processes. Rips (1989) and Smith and Sloman (1994) demonstrated through experiments a dissociation between categorization and similarity, showing that "categorization can be done in two ways, by similarity and by rule" (Smith & Sloman, 1994, p. 377). However, Goldstone & Son (2005) argue that we use similarity as a default method to reason about a domain if we do not have specific knowledge about it. Our annotation study on categorization into tune families contributes to investigating the role of similarity in situations in which domain knowledge is not available for categorization, but the experience of similarity is.

There are diverging concepts on categorization in Cognitive Science, most prominent being the "classical" (or "digital") and "modern" ("analog") view on categorization. The classical view goes back to Aristotle and defines a category as being constituted of all entities that possess a set of properties or satisfy a set of conditions (Sutcliffe, 1993). The "modern" ("analog") view on the other hand suggests that most natural concepts are not well defined but rather based on relationships that are only generally true. Individual exemplars may vary in the number of characteristic features they possess. For instance, Wittgenstein's model of "family resemblance" (Wittgenstein, 1953) assumes that though it is possible that not one feature is shared by each individual within the family, it is still possible to recognize a visible family resemblance.

Rosch has demonstrated that most natural categories humans use cannot be defined by a single set of critical features. Rosch (1973) instead favours the "prototype model": categories are located around prototypes in such a way that some members of a category are more central than others; the prototype is the most central member of a category. Rosch and Mervis (1975) argue that there is a close relationship between family resemblance and prototype theory. Items viewed as most prototypical of one category are those with least family resemblance to other categories and which bear family resemblance to other members of the categories to a large extent. Our annotation study shows that tune families are most adequately described with the family resemblance and prototype model, rather than the classical view on categoriztion.

Music cognition

While similarity is a prominent research topic in Cognitive Science in general, investigations in the domain of music are rather underrepresented in comparison to, for instance, the domains of language and vision. However, recent experimental studies in music have investigated categorization tasks regarding motivic-thematic relations, and a number of studies have investigated musical similarity in the special case of melodic similarity. Our annotation study contributes to linking studies on music similarity and categorization.

Categorization in music based on motivic-thematic relations. Recent empirical studies in music cognition have demonstrated that even non-experienced listeners have an astonishing ability to build meaningful categories in music in the context of motivic-thematic relations in Western classical music. Hence, the musical categories are defined by motifs or thematic sections that belong to different principle themes. For instance, Melen and Wachsmann (2001) and Koniari et al. (2001) showed that infants from 6 to 10 months and children aged 10 to 11 respectively are able to form categories of musical motifs in pieces by Schubert and Diabelli. McAdams et al. (2004) have shown that both experts and novices are able to build categories within a contemporary piece by Roger Reynolds that correspond to thematic sections in the piece. Ziv and Eitan (2007) demonstrated that listeners build categories that agree fairly well with music theoretic concepts of categorization into principal themes in a piece by Beethoven. Deliège (1996, 1997, 2001) investigated in a number of studies the interrelation of cue abstraction, categorization and similarity in classical music. The computational model introduced by Cambouropoulos (2001) that automatically builds clusters based on melodic cues has successfully been applied to motifs from Bach's violin sonata as provided by Deliège (1996, 1997). The model supports Deliège's hypothesis on human categorization in music.

However, there is little consensus in these studies as to what musical features enabled the categorization. The distinction between "surface" and "deep" features of the music and the investigation of their different roles for categorization processes in music is a major topic in perceptual studies on categorization. This alludes to the distinction between "surface" and "deep" similarity discussed in Cognitive Science (e.g., Medin & Ortony, 1989; Vosniadou & Ortony, 1989). However, the application of this concept to musical features yields contradicting concepts of "surface" and "deep" features. Therefore, these terms remain rather problematic (see also comments by Cambouropoulos (2010) on this topic). For instance, McAdams et al. (2004) distinguish between "surface" and "structural" features according to the hierarchical structures formalized by Lerdahl & Jackendoff (1983) and indicate that "structural" features are considered to occur on "higher hierarchical levels of the musical structure". They list as examples for surface features tempo, event density, harmonic density, register, melodic and rhythmic contour and articulation, while "underlying harmony" and "hierarchically important pitch distributions" should be considered as "structural" features. On the other hand, Lamont and Dibben (2001) define "surface" features according to music theoretic notions of features of motifs (see Meyer, 1973; Réti, 1951), such as changes of texture, orchestration, register and pace. They are opposed to "deep" features of motifs, such as the derivation and fragmentation of the original pitch and rhythm information. This approach is also used by Ziv and Eitan (2007), and Eitan and Granot (2009). Hence, in order to serve as a useful distinction for describing categorization processes in music, surface and deep features need a clear formalization first. Furthermore, other distinctions of features discussed in Cognitive Science for similarity and categorization processes (such as the distinction between local and global features, see, e.g., Navon, 1977) need to be examined to see whether they provide meaningful alternatives for the characterization of musical features.

Our annotation study differs to a great extent from the experimental setup in previous studies. The musicological experts who carried out the annotations were involved in designing the annotation method, especially in defining the musical features considered relevant for the perceived similarity. The experts were very familiar with the music involved, and both audio files of the original recordings and musical notations were available for deciding on the perceived similarity of songs belonging to the same tune family. The decision on the similarity between two songs was thus grounded in a detailed knowledge of the musical context, namely a large collection of folk songs. No time constraint restricted their decision on the role of different musical features to perceived similarity. Our experimental setup hence allowed for a deep reflection on perceived similarity, so that parameters often classified as "secondary" (such as articulation, timbre or dynamics) do not play a role in our study.

Perceptual studies on melodic similarity. Understanding music similarity is crucial for basic music processes such as grouping, segmentation or expectancy (Toiviainen, 2007). The perception of melodic similarity as a specific case of musical similarity has been the main focus within perceptual studies that investigate the role of musical features for assessing similarity. A typical approach to measure the perceived similarity between two melodies is the pairwise rating of similarity on a given scale, such as between 1–7 or 1–9 (e.g., Eerola et al., 2001; McAdams & Matzkin, 2001; Müllensiefen & Frieler, 2007). Novello et al. (2011) used triadic ratings, so that listeners had to decide which two songs out of three were most similar and least similar, respectively. Typke et al. (2005) asked participants to order a given list of musical items into a ranked list according to their similarity to one musical query. Musical materials used for similarity experiments include folk songs of different ethnic origin (Eerola et al., 2001), melodies and their music-theoretic transformation according to surface and reduced structure (McAdams & Matzkin, 2001), short excerpts from pop songs from different genres (Novello et al., 2011) and popular songs along with variants obtained by systematic transformations (Müllensiefen & Frieler, 2007).

The search for relevant features underlying the human similarity assessment in these different studies does not allow coherent conclusions. McAdams & Matzkin (2001) conclude that transformations that respected the reduced structure (as defined by Lerdahl & Jackendoff, 1983) were judged significantly more similar to their original than those that did not. Using multi-dimensional scaling, Novello et al. (2011) concluded that the most important axes in the perceptual space of the participants as they assessed for similarity were "slow–fast", "vocal–non-vocal", "synthetic–acoustic". Similarity measures based on statistical and descriptive variables extracted form the musical material in Eerola et al. (2001) were regressed upon the similarity ratings of the listeners for all pairs of melodies. The overall prediction rate for the statistical variables was rather low, the descriptive variables were somewhat better predictors of melodic similarity. The authors do not discuss whether the similarity rankings of the participants reflected the distinct ethnic folk song styles from the material used in the experiment, so that melodies belonging to the same style were judged as more similar to each other than to melodies of other styles. Müllensiefen & Frieler (2007) tested 34 algorithmic similarity measures via a regression model on listeners' ratings. An optimal fit between algorithmic similarity measures and listeners' ratings was achieved by using a combination of a rhythmic similarity measure, a harmonic measure and an n-gram-related measure. The study does not discuss whether the similarity ratings reflect the categories of systematic transformations introduced into the pop songs (such as transformations in rhythm or tonality) by the authors.

In summary, the investigation of the perception of melodic similarity concentrates on features listeners attend to when rating melodic similarity and on how these features contribute to the overall similarity rating. From these studies, no general conclusions can be drawn about what features are important cues for listeners to decide on the similarity between melodies.

Categorization and similarity in music perceptual studies. Studies on categorization tasks discussed under 'Categorization in music based on motivic-thematic relations' investigate the ability of listeners to recognize a musical category by assessing similarity. However, they often fail to conclude what musical features listeners used in the categorization process. On the other hand, studies of melodic similarity described under 'Perceptual studies on melodic similarity' concentrate on features listeners used in assessing similarity, but do not connect the results of the similarity ratings to specific categories within the music stimuli (such as folk music styles of different ethnic origins). Our annotation study links the study of categorization in music to the study of features that are relevant for assessing similarity.

A manual annotation study for tune families

In this section we present an annotation method developed for the deconstruction of experts' evaluation of melodic similarity of melodies belonging to the same tune family. Our approach to melodic similarity in folk songs investigates categorization based on perceived similarity.

Musicological experts' knowledge of categorization is investigated regarding the features underlying their similarity assessment. In what follows we describe the context of the annotation study, the annotation method, and the evaluation of the results of the annotation study.

The context of the annotation study

The Meertens Institute hosts and researches Dutch folk songs contained in the corpus *Onder de groene linde* (Grijp, 2008) that have been transmitted through oral tradition. The collection contains over 7000 audio recordings of Dutch folk songs from the 1950s till the 1980s. Since most of the recorded songs were sung from memory, considerable variation occurs between variants of the same tune. Many of these recordings were transcribed into musical notation at the Meertens Institute. Within the WITCHCRAFT[1] project, around 2500 of these song transcriptions have been digitized (along with more than 3500 songs from written sources) and are thus available for computational processing.

The melodies in the collection have been classified by musicological experts[2] at the Meertens Institute into tune families so that each tune family 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 classified into the same tune family. According to the distinction in categorization by similarity or by rule in Smith and Sloman (1994), this form of categorization clearly falls into categorization by similarity.

The domain experts base their similarity assessment in most cases on an intuitive, holistic decision. Where there are doubts, single features of the melodies are examined to achieve a decision on the classification of the song. However, no systematic approach on how musical features contribute to the similarity assessment by the domain experts has been developed in their daily work. In the human process of classifying into tune families some melodies receive the status of a *prototypical* melody of their tune family as the most typical representative. This is the melody that – after hearing all melodies – most distinctively stays in the mind as being the most characteristic version. All other melody candidates are then compared to this prototypical melody in order to decide whether they belong to this tune family.

In the context of the WITCHCRAFT project, we aimed at developing computational methods for these melodies that would allow us to search automatically for all members of a given tune family in a corpus of 6000 digitized melodies. Hence, we were in need of the best possible understanding of the similarity relations as perceived by the domain experts. At the beginning of the WITCHCRAFT project, there was no formalized concept on how to describe the similarity between the melodies. In order to be able to develop computational methods for the similarity of the melodies, we therefore had to develop a method which would allow us to make the implicit, intuitive criteria of the experts as explicit as possible.

Starting from the work method of the domain experts, we first established with the experts what musical features play an important role in their intuitive, holistic similarity assessment. As a result, melodic contour, rhythm, lyrics, and motifs were chosen as relevant features contributing to the similarity assessment. We then defined together with the experts criteria for numerical ratings for each of the features that best reflect the experts' working method. Hence, in our annotation method, for each pair of melodies, numerical similarity ratings are given for each feature, so that the relative importance of different musical features of similarity is explicitly represented in the annotations. We directly involved domain experts in the process of

Table 1. Overview of iterations in designing the annotation method.

Experiment 1	Iteration 1:	Annotation of tune families *Frankrijk* and *Boerinnetje*.
	Iteration 2:	Annotation of tune families *Meisje* and *Bergen*.
	First analysis:	Analysis of the agreement among the annotators and relative importance of the features.
Experiment 2	Iteration 3:	Annotation of 26 tune families; creation of the Annotated Corpus.
	Second analysis:	Analysis of relative importance of the features.

designing the experimental setup and in the design of the annotation system. It is reasonable to assume that the given ratings reflect a deep understanding of the underlying similarity relations by the musicological experts.

The annotation method

The annotation method has been designed in an iterative way. In this section, we describe the results of the three final iterations of this process. Three experts annotated both in the first and second iteration two tune families. In the third iteration 26 other tune families were annotated in detail. After the first and second iteration the annotations were compared and the annotation method was adapted where necessary. After the second and the third iteration we used the annotation values to draw several conclusions about the relative importance of the various features of melodic similarity, referred to as the first and second experiment in the remainder of this section. Table 1 shows a schematic overview of the process.

For an exhaustive similarity description of the entire corpus, annotating the similarity of each pair of melodies in the corpus would have been ideal. However, annotating 360 melodies in a pairwise manner would have been far too time-consuming. Instead, we stayed close to the daily working practice of the experts by making use of their concept of a prototypical melody for the annotation procedure. For each tune family the expert who originally assigned the melodies to the tune families determines which melody they perceive as the most prototypical within the classification process. This melody receives the status of reference melody in our annotation study. All other melodies of the tune family are compared to this reference melody and the level of similarity for several musical features is rated, which comes close to the daily working practice of the experts. By comparing all melodies to a reference melody, the comparisons have meaning for the tune family as a whole, since the reference melody is supposed to represent the tune family as a whole.[3]

Relations between tune families are not annotated using this method. However, it could be that all melodies in a tune family are very similar in terms of rhythm, but that they are also rhythmically similar to many melodies from other tune families since their rhythm is very generic. To deal with this problem, we included several additions to the annotations system after the second iteration, which are described below in 'Second experiment on similarity annotation'.

The annotation data consist of judgments concerning the contribution of different musical features to the similarity between the melody under consideration and the reference melody. These features are rhythm, contour, motifs, and lyrics. In order to enable its use as reference data to design computational algorithms, we standardized the human evaluation by assigning numeric values to most of the features. We distinguished three different numeric values 0, 1 and 2:

0. The two melodies are not similar according to this feature.
1. The two melodies are somewhat similar according to this feature.
2. The two melodies are obviously similar according to this feature.

Differentiating more than three values proved to be an inadequate approach for the musicological experts since the exact thresholds for choosing the right value are very hard to determine. For example, when adding just one level, resulting in 0, 1, 2, and 3 as possible values, in many concrete situations, it is not clear how to discern between levels 1 and 2 exactly. The three-level approach is more appropriate since it only has the two extreme cases (not similar and obviously similar) and a mid-level in between.

In the next section explicit criteria are provided for rating the different features.

Criteria for the annotations

For each feature we iteratively defined a number of criteria that the human decision should be based upon when assigning the numeric values that best reflect the intuitive assignments of the experts. These criteria are as concrete as necessary to enable the musicological experts to give consistent ratings that are in accordance with their intuitive assignments. In the following subsections these criteria are given in detail for the musical features (rhythm, contour and motifs). These are the definitions that have been established after the first iteration. After the second iteration several additions were made, as described under 'Second experiment on similarity annotation', but the definitions have not been changed. For a detailed description of the criteria of the extra-musical feature *lyrics* we refer to (Van Kranenburg, 2010). Moreover, criteria were established on rating features based on the comparison of single phrases and of entire melodies. In the comparison of pairs of phrases, not all pairs of phrases between two songs were annotated. Instead, the annotator was free to choose those pairs of phrases to annotate that were most reasonable to compare.

Rhythm. We established the following criteria for the comparison of a pair of phrases from two melodies with respect to their rhythmic similarity.

- If the two songs are notated in the same, or in 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 in 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 the rhythmic similarity of pairs of individual phrases is annotated (local rhythm).

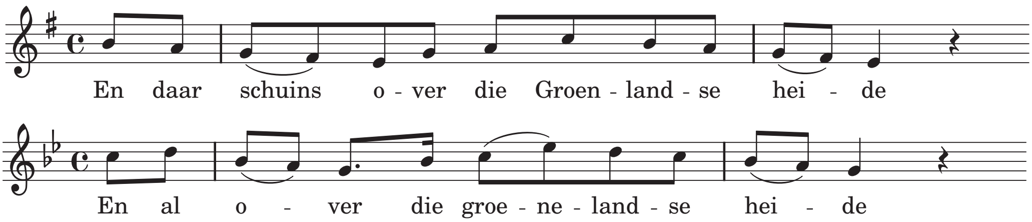

En daar schuins o - ver die Groen - land - se hei - de
En al o - ver die groe - ne - land - se hei - de

Figure 1. Example of a rhythmic transformation: in the first full bar one minor transformation is needed to transform the rhythm of the upper melody into the rhythm of the lower melody. The corresponding value for rhythmic similarity is hence 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 phrases 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 the start (if the upbeat is perceptually unimportant, choose the first downbeat as the start) and end of the line and one or two turning points (extreme points) in between.
  - Based on these three or four points per line determine whether the resulting contours 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.

Motifs. The decision to categorize a melody into a certain tune family 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.

First experiment on similarity annotations

For an initial experiment on the similarity annotations, four tune families containing 11–16 melodies have each been selected to be annotated by three musicological experts. These are the tune families *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 tune family one musicological expert determined the reference melody. Similarity ratings were assigned to all other melodies of the same tune family with respect to the reference melody. In a first iteration, *Frankrijk* and *Boerinnetje* were annotated, in a second iteration *Meisje* and *Bergen* (see Table 1 for an overview of iterations in designing the annotation method). After the two iterations the results were discussed with all experts.

Agreement among the experts. Table 2 gives an overview of the agreement among the three experts for all musical features using Fleiss's kappa (Fleiss & Cohen, 1973). The overall agreement is 0.74, which is considered a substantial agreement. Concerning the different features, agreement is highest for rhythm. The agreement increased between the first and second iterations from 0.67 overall agreement in the first round (considered a substantial agreement) to 0.81 in the second round (considered an almost perfect agreement). For all musical features individually the agreement between the first and second iteration has increased as well. We considered the agreement among the experts to be high enough that in the third iteration each tune family could be annotated by only one expert.

Comparing features across tune families. Table 3 lists the distribution of the assigned values within each musical feature and the lyrics feature for all tune families. The feature *lyrics* receives on average high scores for value 2 (81.2%).

For both *Frankrijk* and *Meisje*, the highest score for rhythm is value 2, while *Boerinnetje* scores highest for motifs and *Bergen* for global contour. Hence, the similarity of melodies belonging to one tune family differs with regard to the contribution of the different musical features to the overall similarity. Moreover, in most of the cases single features are not characteristic enough to describe the similarity of the melodies belonging to one tune family.

Table 2. Fleiss's kappa for the agreement among three experts.

	Fleiss's kappa
All features of all tune families	0.74
Frankrijk, all features	0.71
Boerinnetje, all features	0.66
Meisje, all features	0.78
Bergen, all features	0.82
Feature rhythm for all tune families	0.8
Feature contour for all tune families	0.72
Feature motifs for all tune families	0.62
All features in first iteration	0.67
All features in second iteration	0.81
Rhythm in first iteration	0.73
Rhythm in second iteration	0.86
Contour in first iteration	0.65
Contour in second iteration	0.78
Motifs in first iteration	0.51
Motifs in second iteration	0.72

Table 3. Distribution of the assigned values within each feature per tune family as percentages.

	Frankrijk			*Boerinnetje*		
Value	0	1	2	0	1	2
Rhythm/l	0	1.3	98.7	11.2	51.6	37.2
Contour/g	0	31.7	68.3	12.8	48.7	38.5
Contour/l	5.6	52.5	40.9	41.9	26.4	31.7
Motifs	0	36.6	63.4	0	20.5	79.5
Lyrics	26.7	6.6	66.7	5.1	33.3	61.5

	Meisje			*Bergen*			Average		
Value	0	1	2	0	1	2	0	1	2
Rhythm/l	3.3	8.2	88.5	3.5	15.8	80.7	4.5	19.2	76.3
Contour/g	33.3	13.3	53.4	2.5	10.3	87.2	12.1	26	61.9
Contour/l	20.7	31.8	47.5	4.8	22.5	72.7	18.3	33.3	48.2
Motifs	13.3	16.7	70	0	17.9	82.1	3.3	22.9	73.8
Lyrics	0	3.3	96.7	0	0	100	7.9	10.8	81.2

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

Comparing the contribution of the musical features reveals that the contour scores above 70% for value 2 for only one tune family (*Bergen*), which results in a considerably low average. Both rhythm and motifs score above 70% for value 2 in three out of four cases, resulting in an average of 76.3% for value 2 for the feature rhythm and 73.8% for value 2 for the feature motifs. Hence, rhythm and motifs seem to be more important than contour for the experts' perception of similarity in these experiments.

Similarity within tune families. By comparing the ratings for the individual songs within a tune family, we get an indication of the variation in the importance of the musical features for perceived similarity.

As a measurement for the degree of similarity of a melody to the reference melody we count the number of occurrences of the three rating values and express it as percentage. For this we only use the musical features: rhythm, global contour, contour per line and motifs. The results show that the degree of similarity within the family can vary by a considerable amount. For instance, in the tune family *Meisje* two melodies (with Song IDs 73517_01 and 111465_01, see Table 4) score higher than 95% for value 2, while two melodies score lower than 20% for value 2 with corresponding high scores for value 0 (Song IDs 71449_01 and 139121_01).

The evaluation of single features also shows that the degree of similarity to the reference melody varies. For instance, *Meisje* scores for the feature rhythm on average 88.5% for value 2 (see Table 3). However, melody 71449_01 scores for rhythm only 42% for value 2 and 33% for

Table 4. Degree of similarity of all melodies of the group *Meisje* to the reference melody 70412_01 averaged over all features as percentages. Song ID numbers according to the cataloguing of the folk songs at the Meertens Institute.

Song ID	0	1	2
70321_01	12.5	22.9	64.6
70560_01	4.2	8.3	87.5
71374_01	0	12.5	87.5
71449_01	56.3	25	18.7
71734_01	4.2	14.6	81.2
72923_01	16.4	8.3	75
73517_01	0	0	100
111465_01	0	4.2	95.8
139116_01	39.6	37.5	22.9
139121_01	43.3	41.7	15

value 0. It appears that there is not one characteristic (or one set of characteristics) that all melodies of a tune family share with the reference melody to the same extent.

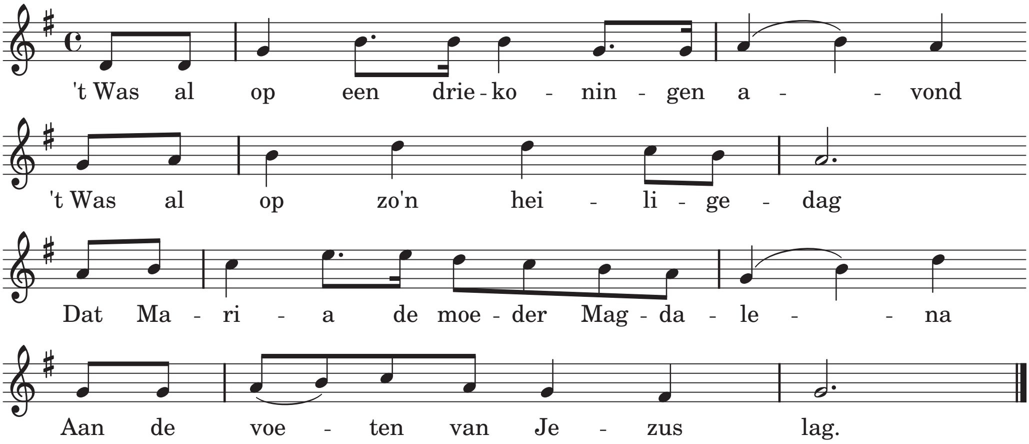
Discussion. From the two sections ‘Comparing features across tune families’ and ‘Similarity within tune families’, we conclude that both across and within the tune families the importance of the musical features for perceived similarity varies.

There is not one characteristic (or one set of characteristics) that all melodies of a tune family share with the reference melody. Therefore, the category type of the tune families cannot be described according to the classical view on categorization requiring all members of the category to possess a common set of defining features, but rather to the modern view (such as the family resemblance or prototype models).

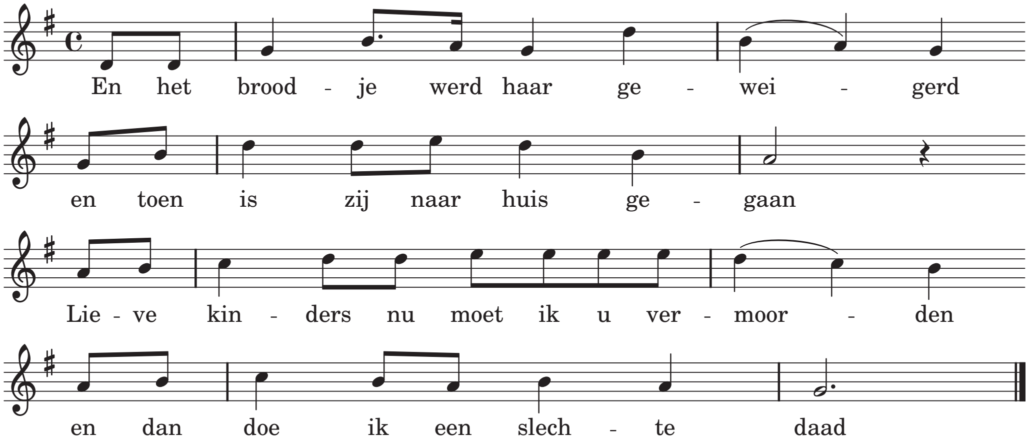
Second experiment on similarity annotation

For the third iteration (see overview iterations in Table 1), 360 melodies, grouped into 26 tune families, were selected by a musicological expert as a representative set out of over 6000 Dutch folk song melodies, which have been encoded at the Meertens Institute both from ethnomusicological transcriptions of field recordings and from written sources of folk songs. The size of each tune family was between 10 and 20 melodies. These 360 songs were selected to form a relatively small subset that is representative for the collection as a whole with regard to the variations that occur between the melodies of a tune family. Starting from her detailed knowledge on the entire folk song collection, the musicological expert selected 26 tune families so that the diversity in similarity relations within these tune families reflected the diversity in similarity relations in the entire corpus. The results of studying this subset are therefore expected to be indicative for the results that would be obtained when studying the entire corpus. An additional constraint for the selected 26 tune families was that considerable variation had to occur among the melodies that belong to the same tune family. ‘Easy’ tune families for which value 2 could be expected to be assigned to all features were thus excluded. A computational model performing well on this subset can therefore be expected to perform well on the entire corpus. We refer to this subset as the *Annotated Corpus*. The contents of this corpus can be found in Table 7, in the Appendix.

Melody A:


't Was al op een drie - ko - nin - gen a - - vond
't Was al op zo'n hei - li - ge - dag
Dat Ma - ri - a de moe - der Mag - da - le - - na
Aan de voe - ten van Je - zus lag.

Melody B:


En het brood - je werd haar ge - wei - gerd
en toen is zij naar huis ge - gaan
Lie - ve kin - ders nu moet ik u ver - moor - den
en dan doe ik een slech - te daad

Global annotations		**Local annotations**			
		Phrase from A	Phrase from B	Contour	Rhythm
Global contour	**1**	1	1	**1**	**2**
Global rhythm	**2!**	2	2	**2**	**2**
Motifs	**0**	3	3	**1**	**2**
Lyrics	**0**	4	4	**0**	**2**

Figure 2. Example annotations for two melodies from the tune family *Driekoningenavond*.

Table 5. Distribution of the assigned values per feature for 360 melodies.

	absolute values			values in %			doubts in %	classification key in %
	0	1	2	0	1	2		
Rhythm/global	9	62	143	4.2	29.0	66.8	3.3	22.4
Rhythm/local	87	491	1069	5.3	29.8	64.9	1.0	0
Contour/global	6	116	211	1.8	34.8	63.4	3.6	0.6
Contour/local	120	623	904	7.3	37.8	54.9	1.0	0
Motifs	8	31	293	2.4	9.3	88.3	0	9.3
Lyrics	34	10	245	11.8	3.5	84.8	0	2.4

The analysis of the results of the iterations as described in the previous section led to a number of modifications concerning the annotation in the third iteration. We consider the agreement of the experts in the second iteration according to Fleiss's kappa with 0.81 (generally interpreted as "almost perfect agreement") sufficient. Therefore, in the third iteration, each tune family was annotated by only one expert. Since motifs play a very important role for the classification of melodies (according to the verbal descriptions of the musicologists and the results of the numeric evaluation of features within the first experiment), the location and size of characteristic motifs were annotated in the third iteration, delivering a list of motifs considered important for the classification. Moreover, the musicological experts annotated doubtful cases, in which the criterion as defined under 'Criteria for the annotations' for assigning a certain value did not correspond to the experts' intuitive rating of this feature. Thus, the validity of the defined criteria concerning the similarity assignments of the experts was tested. Furthermore, if one of the features had been of particularly importance for the decision to include the melody in the tune family, the annotator had the option to identify that feature as a *classification key*. Finally, an optional feature was added for global rhythm. In addition to the rhythmical rating of pairs of phrases (local rhythm), the annotators rated the rhythmic similarity of the two entire melodies according to the same criteria, namely comparing all phrases and then assigning the average over all lines as the value for global rhythm. This feature was added as optional, since the annotation for the local rhythm of all single lines allows the derivation of the value for global rhythm in an automatic manner for those cases for which global rhythm was not annotated manually. An example for the annotations for two melodies from the tune family *Driekoningenavond* is shown in Figure 2.

Discussion. Table 5 shows the distribution of the assigned values per feature for all 360 melodies annotated, as well as the number of doubts expressed and the number of annotated key features. For all features doubts have been expressed very rarely, demonstrating that the criteria defined under 'Criteria for the annotations' coincide in most cases with the intuitive assignments of the experts. Single features have hardly been indicated as being the main reason for the classification of a melody, as the low values for classification key show. Hence, the similarity assessment is in most cases a multi-dimensional process. In those cases where a single key for the classification was indicated, global rhythm most often served as the key factor. Comparing the values across the different features shows that the rhythm of melodies within a tune family is still considered more similar than the contour. However, in comparison to motifs, rhythm plays a less prominent role than in the first experiment. Among the musical features, motifs receive the highest scoring for value 2 with 88.3%. This relates to the tune family concept

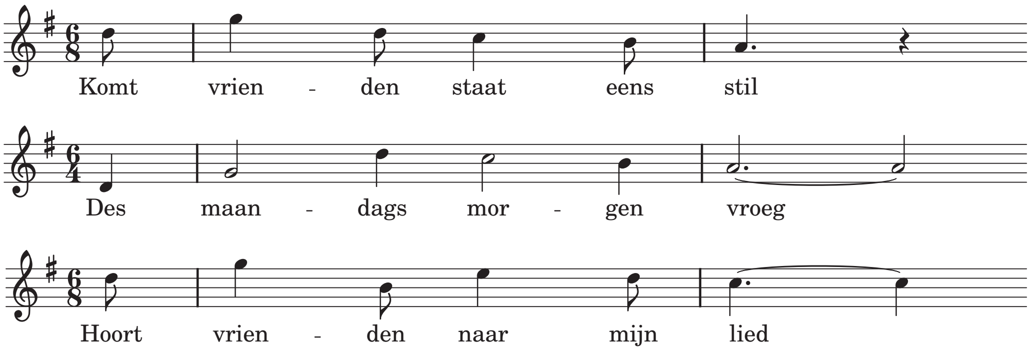

Komt vrien - den staat eens stil
Des maan - dags mor - gen vroeg
Hoort vrien - den naar mijn lied

Figure 3. First lines of three melodies belonging to tune family no. 22 (*Nood*); the reference melody is displayed on top.

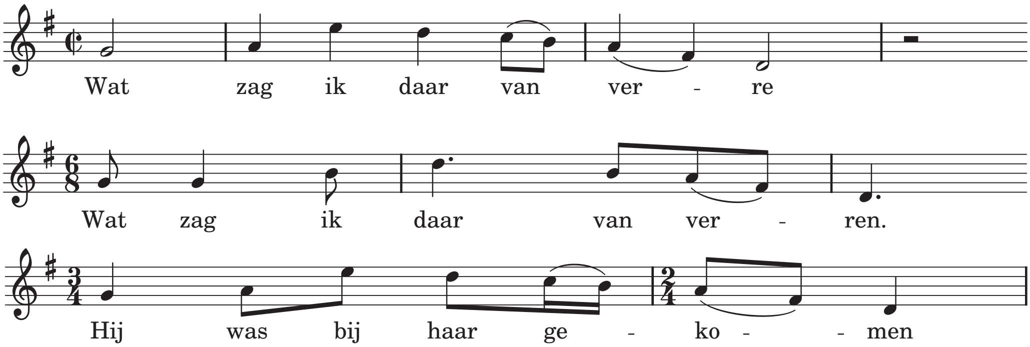

Wat zag ik daar van ver - re
Wat zag ik daar van ver - ren.
Hij was bij haar ge - ko - - men

Figure 4. First lines of three melodies belonging to tune family no. 25 (*Verre*); the reference melody is displayed on top.

described by Cowdery (1984) who puts high emphasis on local motifs for determining tune families. However, though motifs receive the highest similarity scorings, they are only in 9.3% of the cases indicated as the classification key. Other than global rhythm, which has the potential to serve as a classification key in a small number of tune families, motifs hardly serve as an independent reason for the classification into tune families, demonstrating that other features of the melodies also contribute to the classification process.

Table 6 shows the distributions of scores among the various tune families. Again, there are considerable differences between the tune families. This indicates that the experts' evaluation of melodic similarity is not a one-dimensional process. The various features of similarity are not equally important in every case. For instance, the members of the tune family *Nood* (no. 22 in Table 6) are considered to be more similar to the reference melody concerning the feature *local rhythm* than concerning the feature *local contour* (value 2 receives 95.6% for local rhythm and 52.9% for local contour according to Table 6). On the other hand, tune family *Verre* (no. 25 in Table 6) scores higher on value 2 with 76.6% on local contour than on local rhythm (with 53.1% assigned to value 2). Figures 3 and 4 display for each of three examples the first lines of melodies belonging to tune family 22 and 25 respectively. These examples demonstrate that for tune family 22, the rhythmic structure is indeed rather stable, while the contour changes; for

Table 6. Distribution of the assigned values within each feature per tune family in the second experiment.[4]

N	Rhythm/global			Rhythm/local			Contour/global			Contour/local			Motifs			Lyrics		
	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2	0	1	2
1	0	60	40	11.7	36.7	51.7	0	86.7	13.3	11.7	61.7	26.6	0	0	100	20	0	80
2	–	–	–	2.8	63.9	33.3	0	18.2	81.8	0	61.1	38.9	0	36.4	63.6	0	0	100
3	5	45	50	4.8	41.6	53.6	3.7	18.5	77.8	9.6	24	66.4	0	14.8	85.2	0	0	100
4	–	–	–	0	54.5	45.5	0	6.3	93.8	1.5	22.7	75.8	0	0	100	0	0	100
5	–	–	–	0	68.5	31.5	0	11.1	88.9	0	24.1	75.9	0	22.2	77.8	0	0	100
6	25	50	25	35.1	33.3	31.6	0	41.7	58.3	3.5	47.4	49.1	0	8.3	91.7	0	0	100
7	0	75	25	0	57.1	42.9	0	37.5	62.5	7.1	39.3	53.6	0	37.5	62.5	62.5	0	37.5
8	0	40	60	0	40	60	0	13.3	86.7	8.3	23.3	68.3	6.7	6.7	86.7	20	6.7	73.3
9	–	–	–	0	10	90	0	10	90	0	30	70	0	10	90	0	0	100
10	12.5	18.8	68.8	23.4	22.1	54.5	6.3	68.8	25	19.5	50.6	29.9	0	12.5	87.5	0	0	100
11	–	–	–	0	37.5	62.5	0	0	100	0	39.3	60.7	0	0	100	0	0	100
12	0	18.2	81.8	0	6.4	93.6	0	54.5	45.5	3.8	46.2	50	0	0	100	0	0	100
13	–	–	–	0	34.4	65.6	0	41.7	58.3	0	50	50	0	0	100	–	–	–
14	0	30	70	10	13.3	76.7	0	30	70	33.3	30.0	36.7	20	0	80	20	60	20
15	10	30	60	17.2	20.7	62.1	20	70	10	31	55.2	13.8	10	10	80	0	0	100
16	0	28.6	71.4	6.3	31.3	62.5	0	100	0	21.9	62.5	15.6	0	0	100	0	0	100
17	–	–	–	0	26.6	73.4	0	6.3	93.8	0	23.4	76.6	0	13.3	86.7	81.3	0	18.8
18	0	45.5	54.5	6	18	76	0	63.6	36.4	28	32	40	36.4	0	63.6	36.4	9.1	54.5
19	5.9	5.9	88.2	11.1	2.5	86.4	11.8	11.8	76.5	13.6	18.5	67.9	0	0	100	0	0	100
20	10	20	70	7.5	23.8	68.8	0	70	30	2.5	46.3	51.3	0	20	80	0	0	100
21	–	–	–	2.9	64.3	32.9	0	64.3	35.7	8.6	44.3	47.1	0	14.3	85.7	0	0	100
22	0	0	100	0	4.4	95.6	0	57.1	42.9	7.4	39.7	52.9	0	0	100	0	0	100
23	0	12.5	87.5	0	26.3	73.7	0	56.3	43.8	5.3	27.6	67.1	0	18.8	81.3	0	12.5	87.5
24	–	–	–	0	18.4	81.6	0	0	100	0	52	48	0	0	100	0	0	100
25	0	15.4	84.6	7.8	39.1	53.1	0	0	100	1.6	21.9	76.6	0	0	100	0	0	100
26	0	6.3	93.8	0	15.5	84.5	0	31.3	68.8	4.8	41.7	53.6	0	18.8	81.3	23	0	75

tune family 25, on the contrary, the rhythmic structure varies to a great extent between the melodies, while the general contour per line is preserved. Hence, the process of oral transmission leads to variations introduced to the melodies that are of very different type. Our annotation study does not support the findings by Klusen et al. (1978) that rhythmic structure in general is preserved more in a stable way in the process of oral transmission than pitch structure. The annotated data set provides a rich source for future research on the oral transmission of folk songs in understanding the underlying mechanisms of variation. For instance, one might think of investigating whether for a given melody it is possible to predict which of the features will be better preserved based on the characteristics of that melody, or the extent to which this is linked to the context in which the melody is used (such as within dances).

Conclusion

In this paper we have presented an annotation method that facilitates the study of categorization based on the similarity of melodies that are related through oral transmission. According to our annotation study, the following aspects of melody are important for establishing similarity: contour (both per phrase and for the entire song), rhythm (both per phrase and for the entire song), and motifs. In individual cases, the relative importance of these features varies to a large extent. However, in general the recurrence of characteristic motifs seems most important. Motifs are local phenomena, while the other features describe melodies, or individual phrases, globally. Hence, the identification of features that are important for the similarity assessment which underlies the categorization of the songs into tune families and the quantification of their contribution to the similarity ratings give important insights into the link between feature-based similarity and categorization. By making implicit criteria of domain experts for similarity evaluations explicit, we obtain insights into the complexity and multidimensionality of experts' evaluation of melodic similarity. Both characteristics of tune families, namely local motifs discussed by Cowdery (1984), and global characteristics discussed by Wiora (1941) (such as contour), are relevant for the corpus *Onder de groene linde*. In terms of the different types of categorization discussed in Cognitive Science, the classification into tune families is best described using the prototype model and family resemblance.

The annotations confirm that the categorization of the melodies into tune families is based on musical similarity, not on other information. The actual historic relationship of the songs is not known, hence membership to a tune family cannot be concluded based on historic origin. The only extra-musical information is provided by the lyrics of the songs. However, the annotation of the classification key shows that the lyrics hardly ever provide the main reason for the classification. Hence, in most cases the perceived music similarity is crucial for the classification process. Thus, our annotation study contributes to underpinning the crucial role of similarity as a basis for classification. This has been questioned in Cognitive Science in favour of rule-based approaches to classification (see, e.g., Love, 2002; Rips, 1989; Smith & Sloman, 1994). In only very few instances did one feature serve as a classification key, in most cases more than one feature was important for the similarity assessment. Most often, global rhythm served as a classification key. The preservation of rhythmic-metric characteristics in tune families that are rather heterogeneous with respect to other musical features has been described in Selfridge-Field (2007) in the context of dance music. Selfridge-Field (2007) speculates whether this might hint at the existence of a link between gesture and memory; however this has to be investigated in future research.

The musicological experts assess similarity in their daily work routine in an intuitive and holistic manner. The annotation method was successfully established as a means to explicate this intuitive decision, as the low level of doubts annotated reveals. It has often been argued that similarity is context-dependent (Cambouropoulos, 2009; Müllensiefen & Frieler, 2007; Novello et al., 2011). In our study, the musical context was pretty well controlled since the similarity ratings were given in the context of classification of folk songs into tune families.

The Annotated Corpus that results from this study is a valuable resource for further research on melodic similarity. Similarity relations between melodies are described in detail, both quantitatively (ratings) and qualitatively (features). These annotations can be used as enriched 'ground truth' to test retrieval algorithms. For example, in (Van Kranenburg, 2010) the Annotated Corpus has successfully been used to evaluate a computational alignment approach to the similarity of folksongs. The large number of annotated motifs (1426 motif occurrences of 104 motif classes) can be used to test algorithms that detect recurring patterns.

Acknowledgments

This work was supported by the Netherlands Organization for Scientific Research within the WITCHCRAFT-project NWO 640-003-501. Anja Volk is supported by the Netherlands Organization for Scientific Research, NWO-VIDI grant 276-35-001. Peter van Kranenburg is supported by the Royal Netherlands Academy of Arts and Sciences, through the Computational Humanities funding program. We want to thank members of the WITCHCRAFT-team (Jörg Garbers, Frans Wiering, Louis C. Grijp, and Remco C. Veltkamp) and the documentalists from the Meertens Institute (Ellen van der Grijn, Mariet Kaptein and Marieke Klein) for contributing to the establishing of this ground truth. We thank Bas de Haas for comments on an earlier draft of this article.

Notes

1. http://www. cs.uu.nl/research/projects/witchcraft
2. They are Ellen van der Grijn, Mariet Kaptein and Marieke Klein. They are appointed as documentalists for the song collection of the Meertens Institute, all three have an academic masters degree in Musicology.
3. The application of a computational approach to the similarity between the melodies in (Van Kranenburg, 2010) has shown that in most cases the reference melodies have been assigned in a reasonable way by the expert. Tune family members can be successfully retrieved by determining the similarity of the melody to the reference melody.
4. For tune family 13, annotations concerning the feature Lyrics are missing. The feature Global rhythm was an optional feature and was therefore not in all cases annotated.

References

Bayard, S. (1950). Prolegomena to a study of the principal melodic families of British-American folk song. *Journal of American Folklore, 63*(247), 1–44.

Bronson, B. H. (1950). Some observations about melodic variation in British-American folk tunes. *Journal of the American Musicological Society, 3*, 120–134.

Cambouropoulos, E. (2001). Melodic cue abstraction, similarity, and category formation: A formal model. *Music Perception, 18*(3), 347–370.

Cambouropoulos, E. (2009). How similar is similar? *Musicae Scientiae*, Discussion Forum 4B, 7–24.

Cambouropoulos, E. (2010). The musical surface: Challenging basic assumptions. *Musicae Scientiae*, Special Issue, 131–147.

Cowdery, J. R. (1984). A fresh look at the concept of tune family. *Ethnomusicology, 28*(3), 495–504.

Dalla Bella, S., & Peretz, I. (2005). Differentiation of classical music requires little learning but rhythm. *Cognition, 96*, B65–B78.

Deliège, I. (1996). Cue abstraction as a component of categorisation processes in music listening. *Psychology of Music, 24*, 131–156.

Deliège, I. (1997). Similarity in processes of categorisation: Imprint formation as a prototype effect in music listening. In *Proceedings of the Interdisciplinary Workshop on Similarity and Categorisation* (pp. 59–65). University of Edinburgh.

Deliège, I. (2001). Prototype effects in music listening: An empirical approach to the notion of imprint. *Music Perception, 18*(3), 371–407.

Deliège, I. (2007). Similarity relations in listening to music: How do they come into play? *Musicae Scientiae*, Discussion Forum 4A, 9–37.

Downie, J., Byrd, D., & Crawford, T. (2009). Ten years of ISMIR: Reflections on challenges and opportunities. In *Proceedings of the 10th International Society on Music Information Retrieval Conference (ISMIR)* (pp. 13–18). Kobe.

Eerola, T., Järvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception, 18*(3), 275–296.

Eitan, Z., & Granot, R. (2009). Primary versus secondary musical parameters and the classification of melodica motives. *Musicae Scientiae*, Discussion Forum 4B, 139–179.

Fleiss, J. & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33*, 613–619.

Goldstone, R. L. & Son, J. (2005). Similarity. In K. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp. 13–36). Cambridge: Cambridge University Press.

Grijp, L. P. (2008). Introduction. In L. P. Grijp & I. van Beersum (Eds.), *Under the green linden – 163 Dutch ballads from the oral tradition* (pp. 18–27). Amsterdam: Meertens Institute + Music & Words.

Hahn, U. & Ramscar, R. (2001). *Simillarity and categorization*. New York: Oxford University Press.

Kassler, M. (1966). Toward musical information retrieval. *Perspectives of New Music, 4*(6), 59–76.

Klusen, E., Moog, H., & Piel, W. (1978). Experimente zur mündlichen Tradition von Melodien. *Jahrbuch für Volksliedforschung, 2*, 11–32.

Koller, O. (1902). Die beste Methode, um Volks- und volksmäßige Lieder nach ihrer melodischen (nicht textlichen) Beschaffenheit lexikalisch zu ordnen. *Sammelbände der Internationalen Musikgesellschaft, 4*(1), 1–15.

Koniari, D., Predazzer, S., & Melen, M. (2001). Categorization and schematization processes used in music perception by 10- to 11- year old children. *Music Perception, 18*(3), 297–324.

Krohn, I. (1903). Welche ist die beste Methode, um Volks- und volksmäßige Lieder nach ihrer melodischen (nicht textlichen) Beschaffenheit lexikalisch zu ordnen? *Sammelbände der Internationalen Musikgesellschaft, 4*(4), 643–660.

Lamont, A. & Dibben, N. (2001). Motivic structure and the perception of similarity. *Music Perception, 18*(3), 245–274.

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge: MIT Press.

Love, B. C. (2002). Similarity and categorization: A review. *AI Magazine, 18*(3), 103–105.

McAdams, S. & Matzkin, D. (2001). Similarity, invariance, and musical variation. *Annals of the New York Academy of Sciences, 930*, 62–67.

McAdams, S., Vieillard, S., Houix, O., & Reynolds, R. (2004). Perception of musical similarity among contemporary thematic materials in two instrumentations. *Music Perception, 22*(2), 207–237.

Medin, D. & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 179–195). Cambridge: Cambridge University Press.

Melen, M. & Wachsmann, J. (2001). Categorization of musical motifs in infancy. *Music Perception, 18*(3), 325–346.

Meyer, L. (1973). *Explaining music: Essays and exploration*. Berkeley: University of California Press.

Müllensiefen, D. & Frieler, K. (2007). Modelling experts' notions of melodic similarity. *Musicae Scientiae*, Discussion Forum 4A, 183–210.

Murphy, G. L. & Medin, D. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289–316.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology, 9*, 353–383.

Novello, A., McKinney, M., & Kohlrausch, A. (2011). Perceptual evaluation of intersong similarity in Western popular music. *Journal of New Music Research, 40*, 1–26.

Réti, R. (1951). *The thematic process in music*. New York: Macmillan.

Rips, L. (1989). Similarity, typicality, and categorization. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 21–59).

Rosch, E. (1973). Natural categories. *Cognitive Psychology, 4*, 328–350.

Rosch, E., Mervis, C., Gray, W., Johnson, D., & P.Boyes-Braem (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 328–493.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573–605.

Sagrillo, D. (1999). *Melodiegestalten im luxemburgischen Volkslied: Zur Anwendung computergestützter Verfahren bei der Klassifikation von Volksliedabschnitten*. Bonn: Holos-Verlag.

Schoenberg, A. (1942). *Models for beginners in composition*. New York: Schirmer.

Schoenberg, A. (1967). *Fundamentals of musical composition*. London: Faber & Faber.

Selfridge-Field, E. (2007). Social dimensions of melodic identity, cognition, and association. *Musicae Scientiae*, Discussion Forum 4A, 77–97.

Smith, E. E. & Sloman, S. (1994). Similarity- versus rule-based categorization. *Memory and Cognition, 22*, 377–386.

Stief, W. (1990). Probleme der Volksmusikforschung. In H. Braun (Ed.), *Neue Ordnung im Melodienkatalog des Deutschen Volksliedarchivs* (pp. 324–329). Bern.

Suchoff, B. (1981). *The Hungarian Folksong*. Albany: State University of New York Press.

Sutcliffe, J. P. (1993). Concept, class, and category in the sense of Aristotle. In I. van Mechelen, J. Hampton, R. S. Michalski, & P. Theuns (Eds.), *Categories and Concepts* (pp. 33–65). London: Academic Press.

Toiviainen, P. (2007). Editorial. *Musicae Scientiae*, Discussion Forum 4A, 3–6.

Typke, R., de Hoed, M., de Nooijer, J., Wiering, F., & Veltkamp, R. C. (2005). A ground truth for half a million musical incipits. *Journal of Digital Information Managment, 3*(1), 34–39.

Van Kranenburg, P. (2010). *A computational approach to content-based retrieval of folk song melodies* (doctoral dissertation). Utrecht University.

Vosniadou, S. & Ortony, A. (1989). Similarity and analogical reasoning: A synthesis. In S. Vosniadou, & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 1–17). London: Cambridge University Press.

Webern, A. (1963). *The path to the new music*. London: Theodor Presser.

Wiora, W. (1941). Systematik der musikalischen Erscheinungen des Umsingens. *Jahrbuch für Volkslied-forschung, 7*, 128–195.

Wittgenstein, L. (1953). *Philosopical investigations*. London.

Zbikowski, L. (2002). *Conceptualizing music*. Oxford: Oxford University Press.

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

Appendix

Table 7. Abbreviations for names of tune families.

Number	Short	Original	Size
1	*Heer*	*Daar ging een Heer 1*	16
2	*Jonkheer*	*Daar reed een jonkheer 1*	12
3	*Ruiter 1*	*Er reed er eens een ruiter 1*	27
4	*Ruiter 2*	*Daar was laatstmaal een ruiter 2*	17
5	*Maagdje*	*Daar zou er een maagdje vroeg opstaan 2*	10
6	*Dochtertje*	*Een Soudaan had een dochtertje 1*	13
7	*Lindeboom*	*Een lindeboom stond in het dal 1*	9
8	*Zoeteliefjes*	*En er waren eens twee zoeteliefjes*	16
9	*Herderinnetje*	*Er was een herderinnetje 1*	11
10	*Koopman*	*Er was een koopman rijk en machtig 1*	17
11	*Meisje*	*Er was een meisje van zestien jaren 1*	15
12	*Vrouwtje*	*Er woonde een vrouwtje al over het bos*	12
13	*Femme*	*Femmes voulez vous eprouver*	13
14	*Halewijn 2*	*Heer Halewijn 2*	11
15	*Halewijn 4*	*Heer Halewijn 4*	11
16	*Stavoren*	*Het vrouwtje van Stavoren 1*	8
17	*Zomerdag*	*Het was laatst op een zomerdag*	17
18	*Driekoningenavond*	*Het was op een driekoningenavond 1*	12
19	*Stad*	*Ik kwam laatst eens in de stad*	18
20	*Stil*	*Kom laat ons nu zo stil niet zijn 1*	11
21	*Schipper*	*Lieve schipper vaar me over 1*	15
22	*Nood*	*O God ik leef in nood*	8
23	*Soldaat*	*Soldaat kwam uit de oorlog*	17
24	*Bruidje*	*Vaarwel bruidje schoon*	11
25	*Verre*	*Wat zag ik daar van verre 1*	15
26	*Boom*	*Zolang de boom zal bloeien 1*	18