Modelling Folksong Melodies

Wiering, F.; Veltkamp, R.C.; Garbers, J.; Volk, A.; van Kranenburg, P.; Grijp, L.P.

published in
Interdisciplinary Science Reviews
2009

DOI (link to publisher)
10.1179/174327909X441081

document version
Peer reviewed version

document license
CC BY

INTERDISCIPLINARY SCIENCE REVIEWS, Vol. 34 No. 2–3, 2009, 154–171

Modelling Folksong Melodies

Frans Wiering, Remco C Veltkamp, Jörg Garbers, Anja Volk and Peter van Kranenburg

Department of Information and Computing Sciences, Utrecht University, The Netherlands

Louis P Grijp

Meertens Institute, Amsterdam, The Netherlands

In the second half of the twentieth century, ethnomusicologists assembled a collection of more than 7000 field recordings of Dutch ballads. Collectively known as *Onder de groene linde*, these recordings are preserved at the Meertens Institute in Amsterdam. Because of its size, composition and quality of metadata, *Onder de groene linde* is a unique resource for studying the musical properties of folksongs. For such study, it would be essential to search and order the songs automatically, not just using the metadata, but especially their musical content. It is the aim of the WITCHCRAFT project to design and implement methods for processing the musical content of the songs. Such a project involves two disciplines, musicology and computer science, that have different goals and methodologies. Such differences can lead to unproductive tensions, but can also be exploited in order to attain new insights that could not have been attained by the separate disciplines.

KEYWORDS Computational musicology, Music information retrieval, Folksong, Melodic similarly measure, WITCHCRAFT

Introduction

Technology has provided valuable tools for the research of traditional music[1] for more than a century. The earliest known field recordings using wax cylinders were made by Jesse Walter Fewkes in 1890 doing research among Native Americans (Pegg 2009). Recording on various media soon became a standard tool of the discipline, and so became exact measurement of musical properties using these recordings, most notably of tunings and scales. From this point of view, to employ computational methods to the analysis of folksongs may seem the next natural step. Yet despite the obvious potential of these methods, many methodological and practical issues still have to be solved. This article examines some of these within the context of a particular

DOI 10.1179/174327909X441081

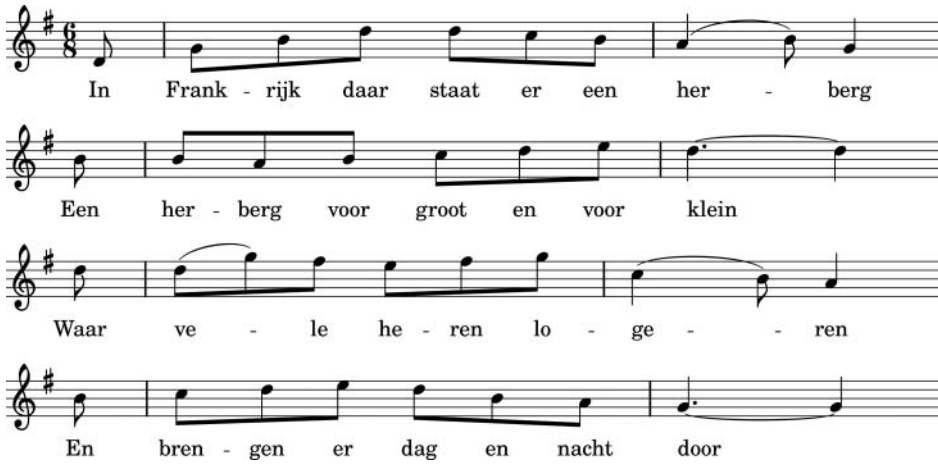

FIGURE 1 *In Frankrijk daar staat er een herberg*, NLB 73515.

repertoire, Dutch monophonic ballads, and a specific research project, WITCHCRAFT ('What Is Topical In Cultural Heritage: Content-based Retrieval Among Folksong Tunes').

Ballads are narrative songs, usually consisting of a number of strophes sung to the same melody. Figure 1 shows an example, the first strophe of the ballad *In Frankrijk daar staat er een herberg*. The story, which is rather typical for the genre, goes as follows:

Long ago, there was an inn outside the gate of a market town. Merchants meet there in the evening for drinking, food and enjoying female company. Everybody participates in the dissolute behaviour except one pious maidservant. One morning she finds a dead baby in her bed. Obviously, she has a secret lover, got pregnant and killed the baby. Therefore she is sentenced to be hanged. But then a miracle happens: angels keep her alive while she is hanging from the gallows. This proves that she is innocent and she is set free. The real culprit is soon identified: the innkeeper's daughter. After killing her illegitimate baby she tried to put the blame on the maidservant. The daughter and her mother, who invented the evil plan, were hanged.

Singing was common during manual labour until the 1930s. It helped to synchronize movement (for example in pile driving), prevented gossiping, and relieved the monotony of the work. Songs were learned by listening and participating: they were seldom written down. In this process of oral transmission, the songs underwent continuous change, both in text and in melody. As an illustration, Figure 2 contains a variant of *In Frankrijk*, which differs in both respects from the one shown in Figure 1, but yet seems closely related.

As a consequence of labour mechanization and the introduction of the radio, ballad singing has completely disappeared from the Netherlands. That we still have access to this tradition is the consequence of a long-term effort to record these songs, started by Will Scheepers in the early 1950s and continued by Ate Doornbosch. They collected around 7000 field recordings,

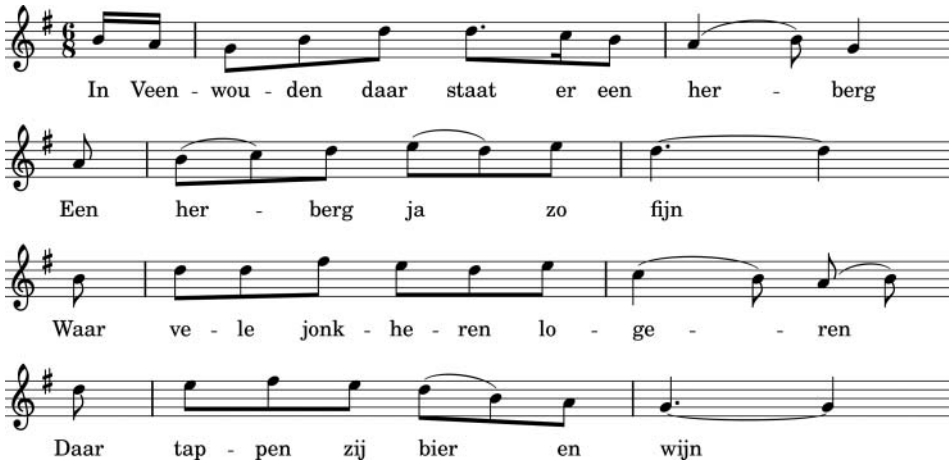

FIGURE 2 *In Veenwouden daar staat er een herberg*, NLB 73424.

which were usually broadcasted in the radio programme *Onder de groene linde* (Under the Green Linden, 1957–1994) (Grijp and Van Beersum 2008). This programme was an early example of interactive radio. Listeners were encouraged to contact Doornbosch if they knew more about the songs that were played. Doornbosch would then record their version and broadcast it. In this manner, a collection was created that documents, in addition to a now departed bit of Dutch cultural heritage, the textual and melodic variation that results from oral transmission.

Onder de groene linde is currently hosted at the Meertens Institute, a research institute that studies and documents the diversity in language and culture in the Netherlands. The collection contains not only field recordings, but also c. 5000 hand-written expert transcriptions of these songs in music notation and very rich metadata (i.e. verbal, catalogue-like descriptions of the content, provenance and context of an item). These metadata are stored in a much larger resource, the *Nederlandse Liederenbank*[2] (Dutch Song Database), containing information about c. 125,000 songs in the Dutch language dating from the Middle Ages to the present. Because of its size, composition and quality of metadata, *Onder de groene linde* is in fact a unique resource for studying the mechanism of oral transmission. And, because of the amount of melodies, it would be essential to search and order these automatically, not just using the metadata, but especially the musical content. It is the aim of the WITCHCRAFT project to design and implement methods for processing the musical content of the songs. There are three aspects to this aim:

- practical: the creation of a melody search engine, to be integrated in the Liederenbank
- scientific: the design and evaluation of methods for measuring musical similarity
- musicological: enabling research into the oral transmission of folksongs.

As will become clear in the course of this article, tensions exist between these aims, and to a certain extent, this article can be read as a case study in dealing with such tensions in an interdisciplinary project. Accordingly, this article is structured as follows. First, we present the research context of the WITCHCRAFT project. Next, our particular approach to interdisciplinarity is outlined. Then, three important aspects of the project are described: the pilot, in which many of the issues came to light and some were resolved; the design of similarity measures for melodies; and finally the evaluation of melodies. The article ends with a brief conclusion about interdisciplinary collaboration and, as this is a project in progress, plans for the near future.

Research context

The WITCHCRAFT project connects two different areas of research: folksong research and music information retrieval. In this section, these two areas will be briefly sketched. This will enable us to describe the interdisciplinary nature of the project more precisely.

In the nineteenth century, an interest in studying folksong traditions emerged in several European countries. Often the underlying motivation for this research was a desire to trace supposedly original and pure aspects of the national musical character, an aim that was thoroughly discredited by the historical developments in the twentieth century. The groundwork for folksong research consisted of the collection and publishing of large amounts of folksong melodies.

An important property of folksong traditions is the variability of songs, caused by the process of oral transmission. The capabilities of such human faculties as perception, memory, performance and creativity play an important role in the transmission of songs in this process. Performers have more or less abstract representations of songs in their memories. The only way in which others have access to a song is to listen to a performance. Research into music cognition (Peretz and Zatorre 2005) shows that the representation of a song in human memory is not 'literal'. During performance, the actual appearance of the song is reconstructed or recreated. In the process of transforming the memory representation into audible words and melody, considerable variation may occur. As long as the processes of encoding songs in, and performing songs from human memory are not sufficiently understood, one has to focus mainly on the recorded or transcribed song instances in order to infer knowledge about this kind of variation.

The variation in transcribed melodies was studied in great detail for German folksongs by Walter Wiora (1941) and for Anglo-American folksongs by Bertrand Bronson (1950). Their conclusion was that nearly every aspect of a folksong is unstable.[3] Wiora distinguishes and illustrates seven categories of change, which include changes in melodic contour and rhythm, insertion and deletion of parts, and even demolition of the entire melody. Important changes may also occur in the connection between text and melody. Sometimes the melody of a song is replaced, or a new text may be supplied to an

existing melody. This general state of flux makes it easy to understand why content-based searching can provide better access to folksongs than metadata searching.

An important concept in folksong research is the 'tune family'. It was developed by Samuel Bayard (1950) and defined as: 'a group of melodies showing basic interrelation by means of constant melodic correspondence, and presumably owing their mutual likeness to descent from a single air that has assumed multiple forms through processes of variation, imitation, and assimilation'. The corresponding term used in Dutch folksong research is 'melody norm' (*melodienorm*). A melody norm is a name assigned to a group of melodies with a presumed common historical origin. This concept was developed for dealing with songs from historical sources that lack a notated melody but instead have a tune indication. An intrinsic difficulty in applying this concept to folksongs is that there is very often no documentary evidence to reason from, so in practice, melody norm assignment is based on the assessment of musical and textual similarity or metadata.

Computational research into folksongs emerged as early as 1949, when Bronson proposed a method to represent folksongs on punch cards. A major development was the creation of the Essen folksong database in the early 1980s, which has been continuously expanded since and currently contains c. 20,000 encoded folksongs from a variety of sources and cultures.[4] Only a minor part of the research on this collection is motivated by a specific interest in folksongs as part of oral tradition. Usually, the collection is used as a convenient test bed for quantitative hypotheses about melody in general. In music information retrieval, the same situation can be observed. The melodies have been used for testing generic models of similarity and as data for online search engines such as Themefinder[5] and Meldex.[6] Several other folksong collections provide searching of the musical content, for example, the Danish Folksong Archives[7] and the Colonial Music Institute.[8] The former is the only such collection that provides some motivation from folksong research for its search methods. In general, there seems to be little attention in computational research to the specific questions of folksong research. In particular, questions regarding the understanding and modelling of tune families are barely addressed.

Approach

Potentially, much benefit can be drawn from an interdisciplinary approach that brings together folksong research and music information retrieval. In practice, such collaboration is difficult as the values, objectives and methods of the areas involved are different. Folksong research investigates research questions about music by means of primarily qualitative methods from the humanities and social sciences, in order to explain or enhance the music as a valuable cultural asset. Music information retrieval research primarily treats music as a special type of data with interesting properties. These properties are investigated by means of quantitative, empirical approaches. The cultural

values of music, though considered an important given, are not the object of investigation. Instead, these often function as 'ground truths' against which computational methods can be evaluated.

In order to bridge this apparent gap, a connecting role is needed (Kranenburg *et al*. 2007). This role is fulfilled by computational musicology. Briefly speaking, computational musicology tries to answer research questions about music by means of computational methods. It provides a common ground where the values and objectives of folksong research coexist with the methods of music information retrieval. This role is directly related to an interesting methodological issue in this project. On the one hand, in order to be able to retrieve variants of melody in a meaningful way, one needs to know the mechanism of oral transmission; on the other, the purpose of retrieval methods is precisely to help to understand this mechanism. This has led to an incremental approach where, starting from simple assumptions about melodic similarity and experts' intuitions about oral transmission, methods and models are gradually refined in an ongoing process. A crucial step in this process has been the creation and annotation of an expert ground truth, which is described below.

Pilot project

At the beginning of the project, an experimental search engine for folksongs was created. Given the fact that the songs are available as audio recordings, the most logical approach to searching the songs would seem to be to digitize these recordings and to perform the actual retrieval on the sound files. This proved not to be an option for two reasons. One is that audio-based methods are not sufficiently developed yet to support reliable matching of the kinds of high-level features one needs for retrieving related melodies (for the state-of-the-art, cf. Casey *et al*. 2008). The other reason is the quality of the recordings. Although the technical quality is often already problematic, the real problem resides in the singers: their worn-out, uneducated voices are often unstable as regards pitch and rhythm. Even for specialists, it is sometimes hard to determine their musical intention. This is where the existing transcriptions come in handy, for the experts who prepared these have already solved this problem to the best of their abilities. These transcriptions can be encoded in a searchable music format relatively easily.

Figure 3 shows the high-level architecture of the search engine. A user creates a musical query in the user interface. This query is compared to items in the music database, and the items that best resemble the query are shown in the interface as a ranked list. The comparison is performed by the similarity measure, which calculates a score by comparing the features extracted from the query to those of each of the database items. Such features are for example pitch, interval, note duration or, to mention one at a slightly higher conceptual level, melodic contour. From a research perspective, the choice of features and the design of similarity measures present the greatest challenges. For users, the usability of the interface and the overall effectiveness of the system are important issues.

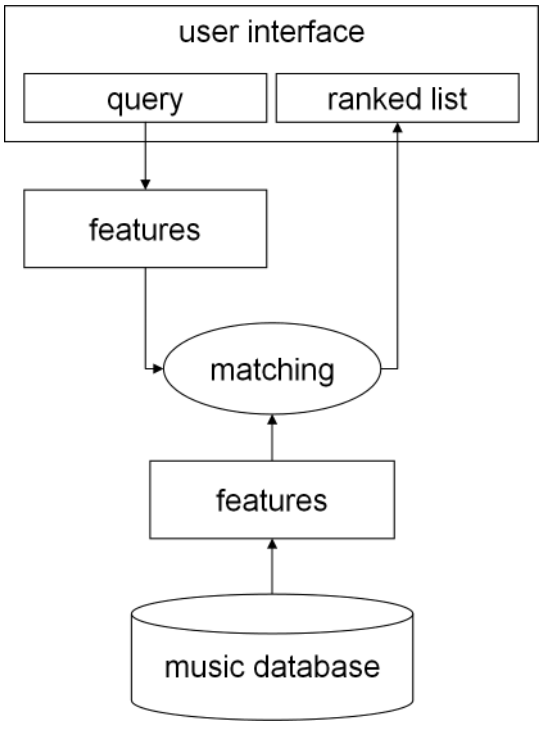

FIGURE 3 High-level architecture of a music information retrieval system.

Our experimental system was created from components of the YahMuugle platform[9] (Bosma *et al.* 2006). The user interface consisted of a simple software keyboard for the input and a ranked list of notated melodies for the output (see Figure 4). As a similarity measure, we used an implementation of the Earth Mover's Distance (EMD), which is described in the next section. A test corpus of 140 melodies was selected from the handwritten transcriptions. It was encoded and classified into 15 different melody norms. Hereafter the capitalized expression 'Test Corpus' will be used to refer to this particular set.

For the encoding of the melodies, two problems needed to be solved: the choice of encoding format and the software environment. Unfortunately, there is a wealth of encoding formats for music, each with specific advantages and limitations (for a dated overview, cf. Selfridge-Field 1997). We decided to use three different formats and to implement a workflow that enables the conversion between these. Data is entered in a subset of LilyPond.[10] LilyPond provides a well-documented text-based format, excellent rendering of music notation, and automatic conversion to the MIDI format. MIDI is easily exchangeable between applications, but as it was primarily designed as a communication protocol between electronic musical instruments, it supports only a restricted set of musical properties. Therefore conversion to another format is also provided, namely to Humdrum, a very rich analytical encoding of music, which however is not easy to enter and visualize (Selfridge-Field 1997).

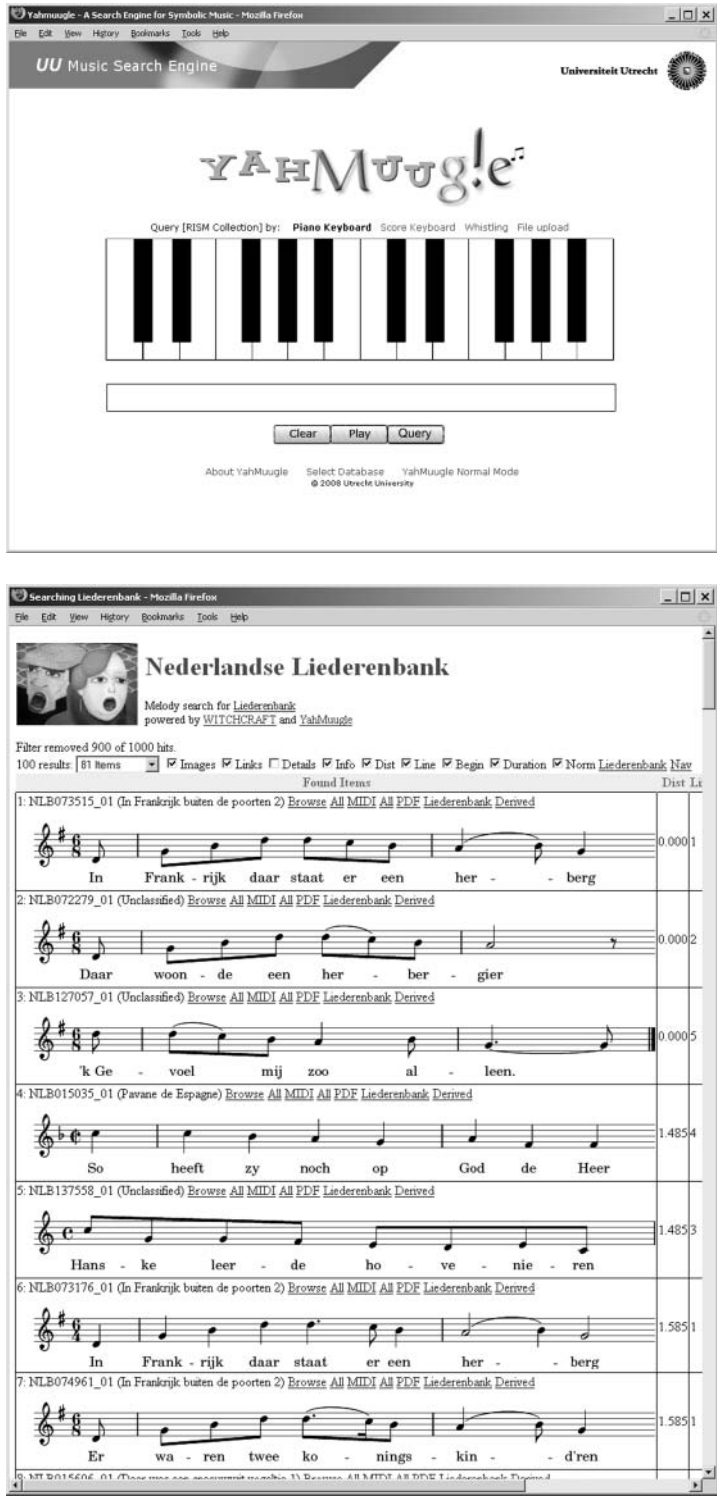

FIGURE 4 Standard YahMuugle interface (a) software keyboard, (b) ranked list.

To support data entry, we created the WitchCraftEditor, which integrates a number of existing components, including LilyPond and a text editor. Data entry is still text-based, which may seem less intuitive than using a graphical user interface. In practice, data entry is at least as fast as for example the music printing program Finale, including correction of errors that appear after conversion to music notation. The editor also allows the annotation of a very important musical property, the subdivision of a melody in phrases. The WitchCraftEditor exists in a standalone and an online version.

This pilot project had three important outcomes. First, from a qualitative evaluation, it became clear that the EMD gave reasonable but not very good results for the Test Corpus, and that there was ample opportunity for improving the matching. Second, results differed considerably between melody norms, but the data set was too small to draw any firm conclusions. Therefore it was decided to greatly enlarge the collection at this stage, but such large-scale data production is of course also an essential step towards the creation of a functional system. To date (March 2009), the collection contains some 3900 folksong melodies and 1900 instrumental melodies. Most of the data entry was done in a parallel project, Dutch Songs as Musical Content. Musicology students also contributed a significant number of encodings. Finally, an evaluation method was lacking that could provide quantitative figures on the one hand and on the other, could provide insight in the musical properties that caused a similarity measure to perform in a particular way.

Similarity measures

In this section, we describe four similarity measures that have been designed for and/or used on the encoded folksongs. Each of them presents a different solution to the problem of modelling and comparing melodies.

The EMD had proved to be an interesting similarity measure in an earlier project (Typke *et al.* 2007). It is a geometrical method that calculates the amount of effort that is needed to transform one pattern of weighted points into another. It can be applied to melodies by representing them in a two-dimensional space with dimensions time and pitch. Each note is a point in this space and is given a weight that corresponds to its importance in the melody. In the current approach, note duration is used for the weight. What happens when two melodies are compared is best visualized as follows. One set of notes is considered as a pattern of heaps of earth, the other as a pattern of holes in the ground. For these two patterns, the minimum effort needed to fill the holes with the earth is calculated. This effort is a measure of how similar the two patterns are: the closer holes and heaps are to each other and the more similar the amounts of earth and the capacities of the holes, the smaller the effort needed to fill the holes with earth (and the more similar the represented melodies are). Figure 5 gives an illustration of the flow from the heaps (upper row) to the holes (lower row). The EMD between the two melodies is small, and perceptually they are indeed very similar. The musical interpretation of the EMD is that it models the similarity of melodic contours.

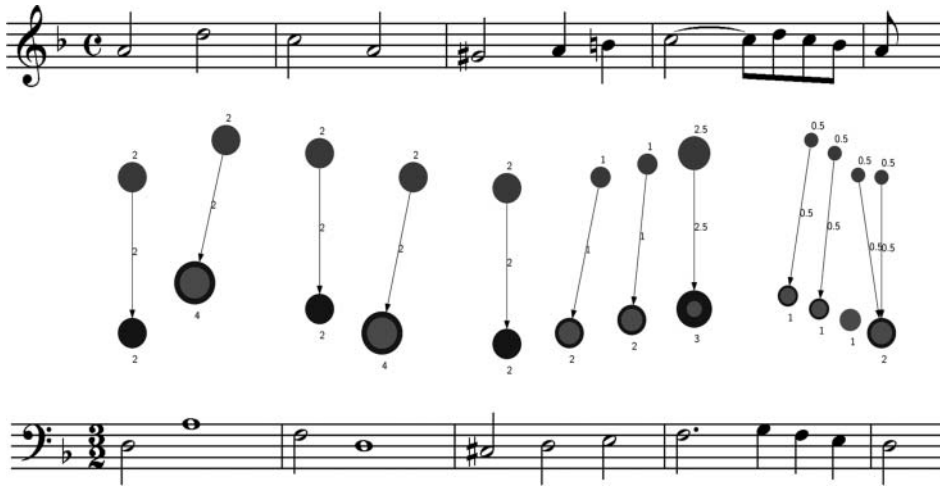

FIGURE 5 Weight flow in the Earth Mover's Distance. The melodies are from J.S. Bach's *Kunst der Fuge* (BWV 1080).

The EMD has some desirable properties, both from a computational and a musical viewpoint. It is robust to all sorts of transformations such as local pitch and duration changes and ornamentation. Patterns do not have to be melodies: they can be chords, for example. Finally, the EMD supports partial matching, for example of a phrase against a complete melody. Its main disadvantages are that strange matches may occur when the number of notes in the two patterns is very different and that it is a computationally expensive method. The latter can be remedied by normalizing the weights so that the total weight of each pattern equals 1. This allows efficient indexing, but the partial matching property is lost (Vleugels and Veltkamp 2002).

An alternative approach represents melodies as graphs. Vertices in the graph represent pitch, the edges the intervals between pitches (Figure 6). Efficient computational methods exist that allow the retrieval of graphs that resemble a given graph. One such method was evaluated on the Test Corpus, performing considerably better than the EMD (Pinto *et al.* 2007). There are two drawbacks to the graph-based approach. One is that it is hard to give a precise musical interpretation of it especially because, in the indexing, the pitch information is lost and only the connectivity of the graph is represented. The other drawback is that other important melodic information is also not represented, most notably duration.

A method that only uses rhythmic features is Inner Metric Analysis (IMA; Volk *et al.* 2007; Volk 2008). This method measures the contribution of each rhythmic event to the formation of rhythmic-metric patterns. The general idea is to search for chains of equally spaced events with a certain minimum length. Such chains can partially overlap when the distance between events is different or when there is a phase difference between them. The metric weight of the event is a function of the number of chains it belongs to and of

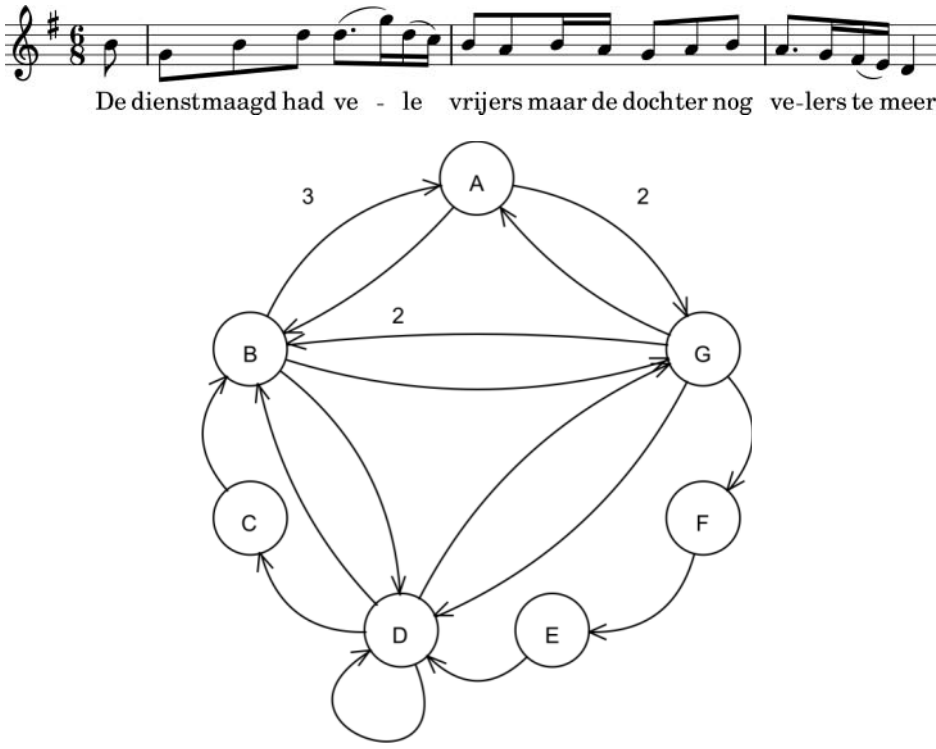

FIGURE 6 A melodic phrase and its graphical representation (*In Frankrijk buiten de poorten*, NLB 72272-2).

their lengths. The higher the weight, the more prominent the event is in the metrical structure of the melody. IMA patterns correlate well with human perception and music-theoretical insights (Volk 2009). Finally, the similarity between two melodies is measured by calculating the correlation between their IMA patterns (see Figure 7). Initial experiments on the Test Corpus showed this measure to be slightly more effective than *rhytGauss*, a well-known measure from the literature (Müllensiefen and Frieler 2006).

Another approach, currently under investigation by Peter van Kranenburg, uses sequence alignment methods to establish the similarity between melodies. Melodies are represented as strings, for example of characters representing pitches. In more complex representations, durations are also accounted for. The general idea of sequence alignment is to minimize the total penalty incurred by substitutions, deletions and gap insertions. When applied to folksongs, the aim is to choose the penalties in such a way that the more likely a change is to occur in oral transmission of melodies, the smaller the penalty. Figure 8 gives an example of the alignment of two melodies. It illustrates a problem that often occurs in the comparison of two variant melodies, namely that melodic fragments are inserted or deleted, or that their length is reduced or extended. In other words, the temporal process in the two melodies is

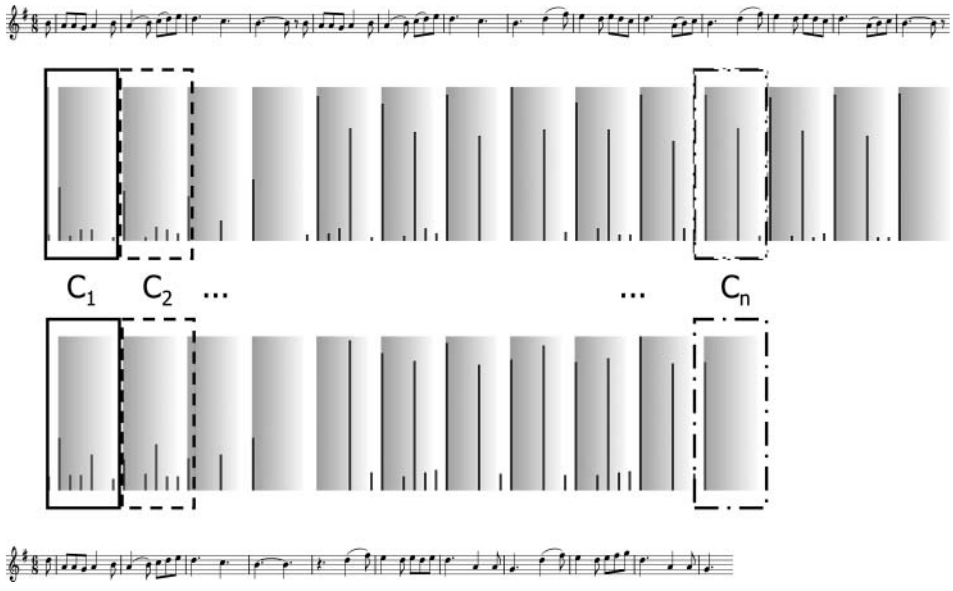

FIGURE 7 Comparing IMA profiles of NLB 72368 (top) and NLB 72452 (bottom).

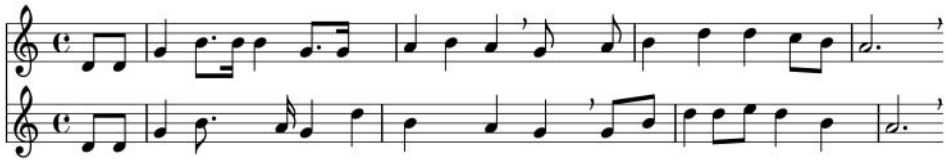

FIGURE 8 Aligning two melodies. For the alignment, longer note values were subdivided into shorter ones. Slurs indicate the subdivision.

considerably different. This makes the application of the EMD problematic, as an assumption underlying this measure is that the overall temporal process is the same for related melodies. Sequence alignment methods do not rely on this assumption, and moreover by providing different (combinations of) penalty raters that determine the cost of substitutions, deletions and gap insertions, searching can be adapted to different types of musical similarity. The ultimate goal is to define a combination that adequately deals with oral variation.

A variation of the sequence alignment approach is to use a group of melodies as the query (Garbers and Wiering 2008). In this way, one can use songs with the same melody norm to retrieve unknown melodies that may also be assigned this norm. An application scenario for this method is to start with one melody, form a group from the query and the relevant items found, and repeat the process until the group consists of all relevant items. One would then have also created a formal model of this group.

Melody retrieval tasks and evaluation

The above discussion of several similarity measures inevitably leads to questions about evaluation. Is there a single best measure, or have measures such different properties that it is better to see them as each other's complements? And how can one compare the performance of similarity measures? Finally, probably the hardest problem is what we mean by musical similarity.

General information retrieval offers a standard approach to the comparison of retrieval methods. First, a database of items (documents, melodies, etc.) is selected or created. Then one defines a (large) number of queries on this database. For each of these, the perfect answer set from the database is then created. Ideally, this is done by creating a *ground truth*, that is, by having human specialists compare all items in the database to the queries; in practice, this is not feasible and only a subset of the items is examined. Usually, the judgement is binary: an item is considered to be either relevant or irrelevant to the query. Independently of the human assignment, the same queries are answered by the retrieval methods. For each method, the output is compared to the human evaluation of the queries by applying a number of standard performance measures (Manning *et al.* 2008).

To a considerable extent, this scenario can be adapted to suit the folksong domain. The current collection of encoded melodies is an excellent database for testing, as it is sufficiently large and carefully prepared. There is a ground truth available, namely the melody norms that were assigned by the experts to most of the melodies. The assignment generally involves a binary judgement whether the melody norm is applicable or not (in practice, it is sometimes necessary to make a conditional assignment). Moreover, a melody cannot have more than one melody norm.

As an example of performance evaluation, we discuss an important measure, the precision-recall graph. Given a set of retrieval results, precision is the proportion of relevant documents in that set. Recall is the proportion of relevant items found. There is a trade-off between the two. Obviously, an optimal recall can be realised by including all documents in the result set, but then precision will be very low. Small result sets have a much better precision, but at the expense of recall. Precision-recall graphs visualize this trade-off. The idea is to use the ranked lists to create result sets of different sizes and to calculate precision and recall for these sets. Their sizes are usually expressed as proportions of the maximum recall, and for each size the corresponding precision is calculated.

Figure 9 shows such a graph, comparing two sequence alignment methods on the set of 360 melodies described below. The task was to retrieve melodies with the same melody norm as the query. For each method, results of all queries are averaged to provide a global assessment of the method. These graphs can be interpreted as follows. Ideally, the precision equals 1 at all levels of recall, meaning that at any set size, there are only relevant items in the set. This rarely happens for real-world cases. The best performing measure is the one whose curve comes most close to the top right corner of the area: for this measure, the result sets contain the largest proportion of

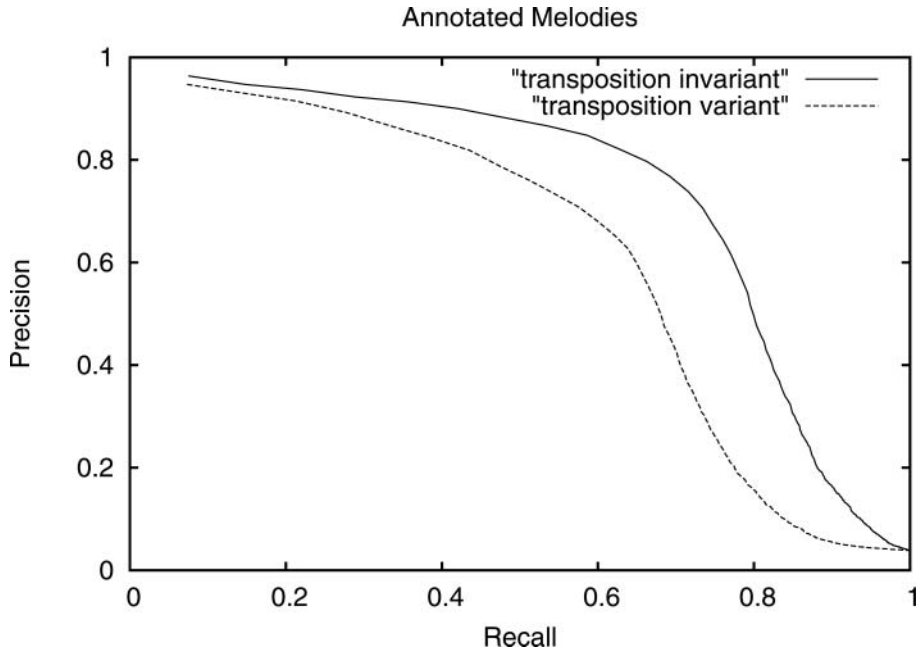

FIGURE 9 Precision-recall graph for two sequence alignment methods.

relevant items. It is also interesting to look at the left end of the curve. The higher this is, the more often the first item is relevant. In our example, the transposition-invariant method performs best in both respects.

Useful though quantitative evaluation may be, there are several problems in the procedure outlined above. To consider relevance as binary is already problematic for textual documents; in the case of folksong retrieval, relevance is equated with melodic similarity, which in itself is a very complex and not very well understood concept (Müllensiefen and Frieler 2004). It is certainly not binary by nature.[11] Finally, though evaluation against a given ground truth measures the performance of a retrieval method, it does not lead to musical insight. Rather such a ground truth acts as a hermetic boundary between music information retrieval and folksong research. The important question here is whether retrieval methods are developed merely to *reproduce* the ground truth, or to *evaluate* the theories that are behind that ground truth (Kranenburg *et al*. 2007). From our point of view, choosing the second option, that is, assuming the computational musicology role, will obviously lead to a better understanding of the music, which in turn will lead to better approaches to music information retrieval.

We have explicitly taken up this role in an extended investigation of the process of melody norm assignment (Volk *et al.* 2008). For this, we developed a manual annotation method that gathers expert judgements about the contribution of different musical dimensions to perceived similarity. In daily

practice, the experts mainly perform the similarity evaluation by comparing candidate melodies to a *reference melody*. To create the annotations, a number of melody norms with sufficient members were selected; one expert determined the prototypical melodies. The first set of four annotated melody norms involved judgements by all three experts; after having determined that there is sufficient consistency between expert judgements, the remaining norms were divided between them and thus annotated only once. In total, 360 melodies from 26 melody norms were annotated.

In order to analyze the 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 text. Text and form are auxiliary criteria; the focus was on the other three criteria.

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. The value 0 indicates that the two melodies are not similar according to this dimension, so that a relation cannot be inferred from it. The value 1 indicates a weak similarity and a possible relationship, the value 2 an obvious similarity and a highly plausible relationship. 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.

An analysis of these annotations has taught us several things, notably:

- even within melody norms, the similarity between members and the reference melody may be based on different dimensions
- motifs, that is, local, distinctive melody patterns, often play an important role in similarity assignments
- groups cohere in different ways: some are more compact, others more loose.

Generally, there is not one characteristic (or set of characteristics) that melodies with the same melody norm share with the reference melody. The concept of melody norm thus does not represent the classical view on categorization, which defines a category as being constituted of all entities that possess a common set of features. It appears that most natural categories are not as strongly defined, but rather use a family resemblance model (Wittgenstein 1953), in which individual exemplars may vary in the choice and number of characteristic features they share with each other. This view on categorization closely resembles classification concepts from machine learning (Manning *et al.* 2008).

Melody norms are thus based on family resemblance. An important consequence for melody classification and the study of similarity measures is that there is no single feature that can adequately mimic human performance. Instead, for optimal retrieval, it must be possible to select one or more

appropriate methods from a larger set. For algorithm development, the practical use of the annotations is that the performance of similarity measures can be related to musical characteristics. Also, when a measure targets a specific characteristic, an appropriate subset can be selected. This has already proved useful in the design of alignment methods.

Conclusion and future work

An important drive of the WITCHCRAFT project has been the urgent need that researchers from the Meertens Institute felt to provide access to their folksong collection. This practical aim inevitably created a tension with the more theoretical, computer science goals of the project, which first of all involve the design of robust algorithms with provable properties. The tension was made productive by consciously developing the computational musicology position within the project as a middle ground where computational methods and musical goals meet. Important outcomes of these are the analysis and modelling of intuitive musical concepts, first of all the melody norm. Similarity methods can be designed in this project on the basis of more and better-described data than is usually the case. And for the Meertens Institute, it is important that the availability of search tools legitimises investment in further data entry and provides new forms of access and thus new forms of folksong research.

As a first step towards the creation of a music search engine for the Nederlandse Liederenbank, the YahMuugle search engine has been integrated into the Liederenbank. At this stage, the similarity measure is still the Earth Mover's Distance, but this will change in the course of 2009, when other measures will be added as well. At the end of the WITCHCRAFT project, we plan to offer two interfaces: one for general users, offering the generally best-performing similarity measure and a simple search interface, and one offering a full suite of query entry methods and similarity measures designed to support the domain experts in their research. While the research into melodic similarity measures was motivated first of all by their wishes, the similarity measures have been designed with more general applications in mind and will form the basis of future MIR projects.

Acknowledgements

This research was supported by the Netherlands Organisation for Scientific Research (NWO) within the WITCHCRAFT project NWO 640-003-501, which is part of the CATCH-programme. YahMuugle research and software development was enabled by the Dutch ICES/KIS III bsik project MultimediaN. Editor development and data entry was partly supported by SenterNovem project DMBO06026, Dutch Songs as Musical Content. We thank musicologists Ellen van der Grijn, Mariet Kaptein and Marieke Klein for their contribution to the annotation method and for creating the annotations.

Notes

[1] One of the possible definitions of the discipline of ethnomusicology. Loosely defined, traditional music includes both non-western music and music that is transmitted from performer to performer rather than composed as an act of individual creation. Since ethnomusicology is nowadays first of all seen as the study of music from an anthropological perspective — and a disputed term generally — it will not be used in this article; instead, the more restrictive term 'folksong research' is used to describe the study of traditional vocal music.

[2] http://www.liederenbank.nl (4/3/09).

[3] Wiora emphatically states: 'Alles an der Beschaffenheit einer Melodie ist veränderlich'.

[4] http://www.esac-data.org/ (4/3/09).

[5] http://www.themefinder.org/ (4/3/09).

[6] http://www.nzdl.org/musiclib (4/3/09).

[7] http://www.dafos.dk/melodies-online/melody-codes-and-code-searching.aspx (4/3/09).

[8] http://www.colonialdancing.org/Easmes/ (4/3/09).

[9] A standalone demo version is available at http://yahmuugle.cs.uu.nl (6/3/09).

[10] http://www.lilypond.org (4/3/09).

[11] See Typke *et al.* (2006) for an approach that uses a ranked ground truth in the evaluation of melodic similarity. Such approaches are often prohibitively labour-intensive.

Bibliography

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

Bosma, Martijn, Remco C. Veltkamp and Frans Wiering. 2006. Muugle: A modular music information retrieval framework. *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2006)*, 330–1.

Bronson, Bertrand H. 1949. Mechanical help in the study of folk song. *The Journal of American Folklore* 62(244): 81–6.

Bronson, Bertrand H. 1950. Some observations about melodic variation in British-American folk tunes. *Journal of the American Musicological Society* 3: 120–34.

Casey, Michael, Remco C. Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes and Malcolm Slaney. 2008. Content-based music information retrieval: current directions and future challenges. *Proceedings of the IEEE* 96(4): 668–96.

Garbers, Jörg, and Frans Wiering. 2008. Towards structural alignment of folk songs. *Proceedings of the Eighth International Conference on Music Information Retrieval (ISMIR 2008)*, 381–6.

Grijp, Louis P. and Ineke van Beersum, editors. 2008. *Under the green linden. 163 Dutch ballads from the oral tradition recorded by Ate Doornbosch a.o.* Meertens Instituut. (with 9 CDs and DVD).

Kranenburg, Peter van, Jörg Garbers, Anja Volk, Frans Wiering, Louis P. Grijp and Remco C. Veltkamp. 2007. Towards integration of MIR and folk song research. *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR 2007)*: 505–8. Extended version: Utrecht University, Technical Report UU-CS-2007-016, http://www.cs.uu.nl/research/techreps/repo/CS-2007/2007-016.pdf (6/3/2009).

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press.

Müllensiefen, Daniel and Klaus Frieler. 2004. Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgements. *Computing in Musicology* 13: 147–77.

Müllensiefen, Daniel and Klaus Frieler. 2006. The Simile algorithms documentation 0.3. http://doc.gold.ac.uk/isms/mmm/SIMILE_algo_docs_0.3.pdf. (4/3/09)

Pinto, Alberto, Reinier H. van Leuken, M. Fatih Demirci, Frans Wiering and Remco C. Veltkamp, 2007. Indexing music collections through graph spectra. *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR 2007)*, 153–6.

Pegg, Carole, Helen Myers, Philip V. Bohlman and Martin Stokes. Ethnomusicology. *Grove Music Online*, http://www.oxfordmusiconline.com/subscriber/article/grove/music/52178pg2 (4/2/09).

Peretz, Isabelle, and Robert J. Zatorre. 2005. Brain organization for music processing. *Annual Review of Psychology* 56: 89–114.

Selfridge-Field, Eleanor, ed. 1997. *Beyond MIDI: The handbook of musical codes*. Cambridge, MA: MIT Press.

Typke, Rainer, Remco C. Veltkamp and Frans Wiering. 2006. A measure for evaluating retrieval techniques based on partially ordered ground truth lists. *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, 1793–6.

Typke, Rainer, Frans Wiering and Remco C. Veltkamp. 2007. Transportation distances and human perception of melodic similarity. *Musicae Scientiae, Discussion Forum 4A*, 153–82.

Vleugels, Jules and Remco C. Veltkamp. 2002. Efficient image retrieval through vantage objects. *Pattern Recognition* 35(1): 69–80.

Volk, Anja, Jörg Garbers, Peter van Kranenburg, Frans Wiering, Remco C. Veltkamp and Louis P. Grijp. 2007. Applying rhythmic similarity based on inner metric analysis to folksong research. *Proceedings of the Seventh International Conference on Music Information Retrieval (ISMIR 2007)*, 293–6. Extended version: Utrecht University, Technical Report UU-CS-2008-013, http://www.cs.uu.nl/research/techreps/repo/CS-2008/2008-013.pdf (6/3/2009).

Volk, Anja. 2008. Persistence and change: local and global components of metre induction using inner metric analysis. *Journal of Mathematics and Computation in Music*, 2(2): 99–115.

Volk, Anja, Peter van Kranenburg, Jörg Garbers, Frans Wiering, Remco C. Veltkamp and Louis P. Grijp. 2008. A manual annotation method for melodic similarity and the study of melody feature sets. *Proceedings of the Eighth International Conference on Music Information Retrieval (ISMIR 2008)*, 101–6.

Volk, Anja. 2009. The study of syncopation using Inner Metric Analysis: Linking theoretical and experimental analysis of metre in music. *Journal of New Music Research* 38 (forthcoming).

Wiora, Walter. 1941. Systematik der musikalischen Erscheinungen des Umsingens. *Jahrbuch für Volksliedforschung* 7: 128–95.

Wittgenstein, Ludwig. 1953. *Philosophical investigations.* Basil Blackwell.

Notes on Contributors

Correspondence to: Frans Wiering, Department of Information and Computing Science (ICS), Utrecht University, P.O. Box 80089, NL-3508 TB, Utrecht, The Netherlands.

Email: frans.wiering@cs.uu.nl