Finding Occurrences of Melodic Segments in Folk Songs Employing Symbolic Similarity Measures

Janssen, B.; van Kranenburg, P.; Volk, A.

DOI
10.1080/09298215.2017.1316292

Publication date
2017

Document Version
Final published version

Published in
Journal of New Music Research

Journal of New Music Research, 2017
Vol. 46, No. 2, 118–134, https://doi.org/10.1080/09298215.2017.1316292

Routledge
Taylor & Francis Group

Finding Occurrences of Melodic Segments in Folk Songs Employing Symbolic Similarity Measures

Berit Janssen[1,2], Peter van Kranenburg[2] and Anja Volk[3]

[1]*Music Cognition Group, University of Amsterdam, Amsterdam, the Netherlands;* [2]*Meertens Institute, Amsterdam, the Netherlands;* [3]*Utrecht University, Utrecht, the Netherlands*

(Received 9 October 2015; accepted 27 March 2017)

Abstract

Much research has been devoted to the classification of folk songs, revealing that variants are recognised based on salient melodic segments, such as phrases and motifs, while other musical material in a melody might vary considerably. In order to judge similarity of melodies on the level of melodic segments, a successful similarity measure is needed which will allow finding occurrences of melodic segments in folk songs reliably. The present study compares several such similarity measures from different music research domains: correlation distance, city block distance, Euclidean distance, local alignment, wavelet transform and structure induction. We evaluate the measures against annotations of phrase occurrences in a corpus of Dutch folk songs, observing whether the measures detect annotated occurrences at the correct positions. Moreover, we investigate the influence of music representation on the success of the various measures, and analyse the robustness of the most successful measures over subsets of the data. Our results reveal that structure induction is a promising approach, but that local alignment and city block distance perform even better when applied to adjusted music representations. These three methods can be combined to find occurrences with increased precision.

Keywords: symbolic, music similarity, segments, occurrences, pattern matching

1. Introduction

A large body of computational music research has been devoted to the study of variation of folk songs in order to understand what characterises a specific folk style (e.g. Juhász, 2006; Conklin & Anagnostopoulou, 2011), or to study change in an oral tradition (e.g. Bronson, 1950; Louhivuori, 1990; Olthof, Janssen, & Honing, 2015). In particular, a very active area of research is the automatic comparison of folk song melodies, with the aim of reproducing human judgements of relationships between songs (e.g. Eerola, Jäärvinen, Louhivuori, & Toiviainen, 2001; Garbers et al., 2007; Müllensiefen & Frieler, 2007; Hillewaere, Manderick, & Conklin, 2009; Bade, Nürnberger, Stober, Garbers, & Wiering, 2009; Boot, Volk, & de Haas, 2016). Recent evidence shows that human listeners do not so much recognise folk songs by virtue of their global structure, but instead focus on the presence or absence of short melodic segments, such as motifs and phrases (Volk & van Kranenburg, 2012).

The present article compares a number of similarity measures as potential computational approaches to locate melodic segments in symbolic representations of folk song variants. We investigate six existing similarity measures suggested by studies in ethnomusicology and music information retrieval as promising approaches to find occurrences.

In computational ethnomusicology, various measures for comparing folk song melodies have been proposed: as such, correlation distance (Scherrer & Scherrer, 1971), city block distance and Euclidean distance (Steinbeck, 1982) have been considered promising. Research on melodic similarity in folk songs also showed that alignment measures can be used to find related melodies in a large corpus of folk songs (van Kranenburg et al., 2013).

As this paper focuses on similarity of melodic segments rather than whole melodies, recent research in musical pattern discovery is also of particular interest. Two well-performing measures in the associated MIREX challenge of 2014 (Velarde & Meredith, 2014; Meredith, 2014) have shown success when

Correspondence: Berit Janssen, Meertens Institute, PO Box 10855, 1001 EW Amsterdam, the Netherlands. Email: berit.janssen@gmail.com

evaluated on the Johannes Keppler University segments Test Database (JKUPDT).[1] We test whether the underlying similarity measures of the pattern discovery methods also perform well in finding occurrences of melodic segments.

The six measures investigated in this paper were used in a previous study and evaluated against binary labels of occurrence and non-occurrence. In this article, we evaluate not only whether occurrences are detected correctly, but also whether they are found in the correct position. Moreover, we evaluate on a bigger data-set, namely the Annotated Corpus of the Meertens Tune Collections, MTC-ANN 2.0 (van Kranenburg, Janssen, & Volk, 2016).

Two measures compared in our previous study–B-spline alignment (Urbano, Lloréns, Morato, & Sánchez-Cuadrado, 2011) and Implication-Realization structure alignment (Grachten, Arcos, & López de Mántaras, 2005)—were not evaluated in this study as in their current implementation, they do not allow determining the positions of occurrences in a melody.

We present an overview of the compared similarity measures in Table 1, with their abbreviation used throughout the article, and bibliographical references to the relevant papers.

We evaluate the measures by comparison to phrase annotations by three domain experts on a selection of folk songs, produced specifically for this purpose. We employ the similarity measures and the annotations to address four research questions:

- Q1 Which of the proposed similarity measures performs best at finding occurrences of melodic segments in folk songs?
- Q2 Folk songs are often notated in different octaves or keys, or in different meters, as exemplified by two variants displayed in Figure 1. How can the resulting transposition and time dilation differences best be resolved? Does a different music representation improve the performance of similarity measures?
- Q3 Can a combination of the best-performing measures improve agreement with human annotations?
- Q4 Our folk song corpus contains distinct groups of variants. How robust are the best-performing measures to such subgroups within the data-set?

The remainder of this paper is organised as follows: first, we describe our corpus of folk songs, which has annotations of phrase occurrences. Next, we give details on the compared similarity measures, and the methods used to implement the similarity measures, and to evaluate them. In Section 5, we perform an overall comparison of the six similarity measures (Q1). Section 6 addresses the influence of transposition and time dilation on the results (Q2). Section 7 introduces a combined measure based on the best-performing similarity measures and music representations (Q3), and Section 8 investigates the robustness of the best measures towards variation in the data-set (Q4). The evidence from our results leads to a number of concluding remarks and incentives for future research.

[1] http://www.music-ir.org/mirex/wiki/2014:Discovery_of_Repeated_Themes_%26_Sections_Results

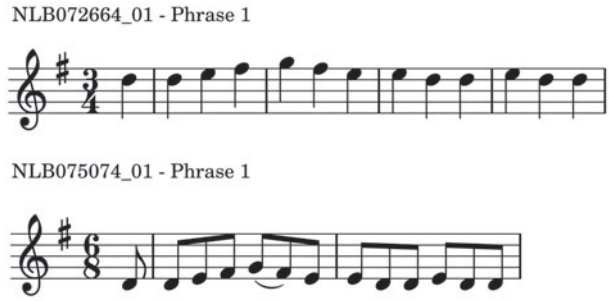

Fig. 1. The first phrase of two variants of a folk song, notated at different octaves and in different meters. Similarity comparison of the pitches and durations might lead to no agreement between the two variants, even though they are clearly very related.

2. Material

We evaluate the similarity measures on a corpus of Dutch folk songs, MTC-ANN 2.0, which is part of the Meertens Tune Collections (van Kranenburg, de Bruin, & Grijp, 2014). MTC-ANN 2.0 contains 360 melodies from Dutch oral tradition, which have mostly been transcribed from recordings, while some have been digitised from song books. Various metadata have been added to the folk songs by domain experts. The melodies have been categorised into groups of variants, or *tune families*, considered to have descended from the same ancestor melody (Bayard, 1950). We parse the **kern files as provided by MTC-ANN 2.0 and transform the melodies and segments into the required music representations using music21 (Cuthbert et al., 2010).

Even though MTC-ANN 2.0 comprises very well documented data, there are some difficulties to overcome when comparing the digitised melodies computationally. Most importantly, the transcription choices between variants may be different: where one melody may have been notated in 3/4, and with a melodic range from D4 to G4, another transcriber may have chosen a 6/8 meter, and a melodic range from D3 to G3, as shown in Figure 1. This means that notes which are perceptually very similar might be hard to match based on the digitised transcriptions. Musical similarity measures might be sensitive to these differences, unless they are transposition or time dilation invariant, i.e. work equally well under different pitch transpositions or meters.

For the corpus of 360 melodies categorised into 26 tune families, we asked three Dutch folk song experts to annotate similarity relationships between phrases within tune families. The annotators all have a musicological background, and had worked with the folk song collection for at least some months previous to the annotation procedure. They annotated 1891 phrases in total. The phrases contain, on average, nine notes,

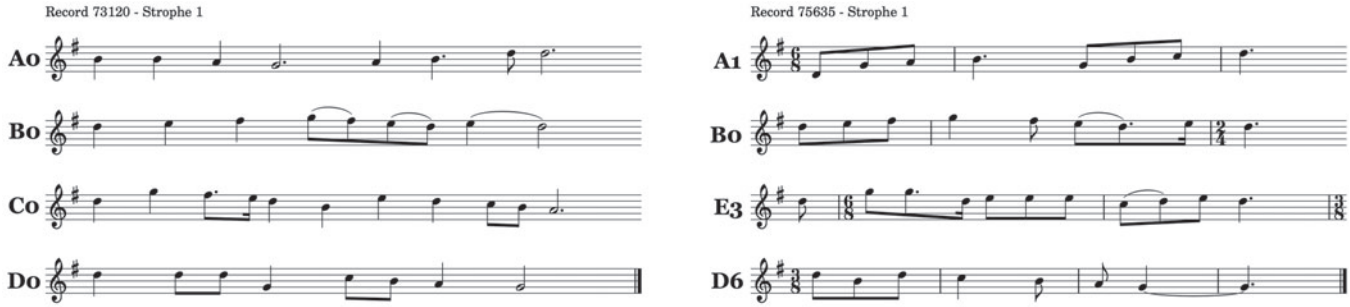

Fig. 2. An example for two melodies from the same tune family with annotations. The first phrase of each melody is labelled with the same letter (A), but different numbers, indicating that the phrases are 'related but varied', the second phrase is labelled B0 in both melodies, indicating that the phrases are 'almost identical'.

with a standard deviation of two notes. The data-set with its numerous annotations is publicly available.[2]

For each tune family, the annotators compared all the phrases within the tune family with each other, and gave each phrase a label consisting of a letter and a number. If two phrases were considered 'almost identical', they received exactly the same label; if they were considered 'related but varied', they received the same letter, but different numbers; and if two phrases were considered 'different', they received different letters (cf. an annotation example in Figure 2).

The three domain experts worked independently on the same data, annotating each tune family separately, in an order that they could choose themselves. To investigate the subjectivity of similarity judgements, we measure the agreement between the three annotators on pairwise phrase similarity using Fleiss' Kappa, which yields $\kappa = 0.76$, constituting substantial agreement.

The annotation was organised in this way to guarantee that the task was feasible: checking for instances of each phrase in a tune family in all its variants (27,182 comparisons) would have been much more time-consuming than assigning labels to the 1891 phrases, based on their similarity. Moreover, the three levels of annotation facilitate evaluation for two goals: finding only almost identical occurrences, and finding also varied occurrences. These two goals might require quite different approaches. In the present paper, we focus on finding 'almost identical' occurrences.

3. Compared similarity measures

In this section, we present the six compared similarity measures, describing the music representations used for each measure. We describe the measures in three subgroups: first, measures comparing equal-length note sequences; second, measures comparing variable-length note sequences; third, measures comparing more abstract representations of the melody.

Some measures use note duration next to pitch information, whereas others discard the note duration, which is the easiest way of dealing with time dilation differences. Therefore, we distinguish between music representation as *pitch sequences*, which discard the durations of notes, and *duration-weighted pitch sequences*, which repeat a given pitch depending on the length of the notes. We represent a crotchet or quarter note by 16 pitch values, a quaver or eighth note by 8 pitch values and so on. Onsets of small duration units, especially triplets, may fall between these sampling points, which shift their onset slightly in the representation. Structure induction requires a music representation in *onset, pitch* pairs.

In order to deal with transposition differences in folk songs, van Kranenburg et al. (2013) transpose melodies to the same key using pitch histogram intersection. We take a similar approach. For each melody, a pitch histogram is computed with MIDI note numbers as bins, with the count of each note number weighted by its total duration in a melody. The pitch histogram intersection of two histograms h_s and h_t, with shift σ, is defined as

$$PHI(h_s, h_t, \sigma) = \sum_{k=1}^{r} \min(h_{s,k+\sigma}, h_{t,k}), \qquad (1)$$

where k denotes the index of the bin, and r the total number of bins. We define a non-existing bin to have value zero. For each tune family, we randomly pick one reference melody and for each other melody in the tune family we compute the σ that yields a maximum value for the histogram intersection, and transpose that melody by σ semitones. This process results in *pitch-adjusted sequences*.

To test how the choice of reference melody affects the results of pitch histogram intersection, we performed the procedure 100 times, with a randomly picked reference melody per tune family in every iteration. We compare the resulting pitch differences between tune family variants with pitch differences as a result of manually adjusted pitches, available through the MTC-ANN-2.0 data-set. We compare all 2822 pairs of tune family variants. On average, pitch histogram intersection adjusts 93.3% of the melody pairs correctly, so the procedure succeeds in the vast majority of cases. The standard deviation of the success rate is 2.4%, which is low enough to conclude that it does not matter greatly which melody is picked as a reference melody for the pitch histogram intersection procedure.

[2] http://www.liederenbank.nl/mtc/

Table 1. An overview of the measures for music similarity compared in this research, with information on the authors and year of the related publication.

Abbreviation	Similarity measure	Authors
CD	Correlation distance	(Scherrer & Scherrer, 1971)
CBD	City-block distance	(Steinbeck, 1982)
ED	Euclidean distance	(Steinbeck, 1982)
LA	Local alignment	(van Kranenburg et al., 2013)
SIAM	Structure induction	(Meredith, 2014)
WT	Wavelet transform	(Velarde & Meredith, 2014)

3.1 Similarity measures comparing equal-length note sequences

To describe the following three measures, we refer to two melodic segments q and p of length n, which have elements q_i and p_i. The measures described in this section are distance measures, such that lower values of $dist(q, p)$ indicate higher similarity. Finding an occurrence of a melodic segment within a melody with a fixed-length similarity measure is achieved through the comparison of the query segment against all possible segments of the same length in the melody. The candidate segments with maximal similarity to the query segment are retained as matches, and the positions of these matches within the match melody are saved along with the achieved similarity. The implementation of the fixed-length similarity measures in Python is available online.[3] It uses the *spatial.distance* library of scipy (Oliphant, 2007).

Scherrer and Scherrer (1971) suggest correlation distance to compare folk song melodies, represented as duration-weighted pitch sequences. Correlation distance is independent of the transposition and melodic range of a melody, but in the current music representation, it is affected by time dilation differences.

$$dist(q, p) = 1 - \frac{\sum_{i=1}^{n}(q_i - \bar{q})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n}(q_i - \bar{q})^2}\sqrt{\sum_{i=1}^{n}(p_i - \bar{p})^2}} \quad (2)$$

Steinbeck (1982) proposes two similarity metrics for the classification of folk song melodies: city block distance (Equation 3) and Euclidean distance (Equation 4). He suggests to compare pitch sequences with these similarity measures, next to various other features of melodies such as their range and the number of notes in a melody (p. 251f.). As we are interested in finding occurrences of segments rather than comparing whole melodies, we compare pitch sequences, based on the pitch distances between each note in the sequence.

$$dist(q, p) = \sum_{i=1}^{n} |q_i - p_i| \quad (3)$$

$$dist(q, p) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (4)$$

[3] https://github.com/BeritJanssen/MelodicOccurrences

City block distance and Euclidean distance are not transposition invariant, but as they are applied to pitch sequences, time dilation differences have minor influence. All the equal-length measures in this section will be influenced by variations introducing more notes into a melodic segment, such as melodic ornamentation. Variable-length similarity measures, discussed in the following section, can deal with such variations more effectively.

3.2 Similarity measures comparing variable-length note sequences

To formalise the following two measures, we refer to a melodic segment q of length n and a melody s of length m, with elements q_i and s_j. The measures described in this section are similarity measures, such that higher values of $sim(q, s)$ indicate higher similarity. The implementation of these methods in Python is available online.[3]

Mongeau and Sankoff (1990) suggest the use of alignment methods for measuring music similarity, and they have been proven to work well for folk songs (van Kranenburg et al., 2013). We apply local alignment (Smith & Waterman, 1981), which returns the similarity of the segments within a given melody which matches the query best.

To compute the optimal local alignment, a matrix A is recursively filled according to equation 5. The matrix is initialised as $A(i, 0) = 0, i \in \{0, \ldots, n\}$, and $A(0, j) = 0, j \in \{0, \ldots, m\}$. $W_{insertion}$ and $W_{deletion}$ define the weights for inserting an element from melody s into segment q, and for deleting an element from segment q, respectively. $subs(q_i, s_j)$ is the substitution function, which gives a weight depending on the similarity of the notes q_i and s_j.

$$A(i, j) = \max \begin{cases} A(i-1, j-1) + subs(q_i, s_j) \\ A(i, j-1) + W_{insertion} \\ A(i-1, j) + W_{deletion} \\ 0 \end{cases} \quad (5)$$

We apply local alignment to pitch-adjusted sequences. In this representation, local alignment is not affected by transposition differences, and it should be robust with respect to time dilation. For the insertion and deletion weights, we use $W_{insertion} = W_{deletion} = -0.5$, and we define the substitution score as

$$subs(q_i, s_j) = \begin{cases} 1 & \text{if} \quad q_i = s_j \\ -1 & \text{otherwise} \end{cases} . \quad (6)$$

The insertion and deletion weights are chosen to be equal, and to be smaller than the weight of a substitution with a different pitch; substitution with the same pitch is rewarded. Effectively, this means that the alignment matrix will have non-zero values only if substitutions with the same pitch occur.

The local alignment score is the maximum value in the alignment matrix A. This maximum value can appear in more than one cell of the alignment matrix due to phrase repetition. This means that several matches can be associated with a

given local alignment score. To determine the positions of the matches associated with the maximum alignment score, we register for each cell of the alignment matrix whether its value was caused by insertion, deletion or substitution. We backtrace the alignment from every cell containing the maximal alignment score, which we take as the end position of a match, continuing until encountering a cell containing zero, which is taken as the begin position of a match.

We normalise the maximal alignment score by the number of notes n in the query segment, which gives us the similarity of the detected match with the query segment.

$$sim(q, s) = \frac{1}{n} \max_{i,j}(A(i, j)) \tag{7}$$

Structure induction algorithms (Meredith, 2006) formalise a melody as a set of points in a space defined by note onset and pitch, and perform well for musical pattern discovery (Meredith, 2014). They measure the difference between melodic segments through so-called translation vectors. The translation vector $\mathbf{T}$ between points in two melodic segments can be seen as the difference between the points q_i and s_j in onset, pitch space. As such, it is transposition invariant, but will be influenced by time dilation differences.

$$\mathbf{T} = \begin{pmatrix} q_{i,onset} \\ q_{i,pitch} \end{pmatrix} - \begin{pmatrix} s_{j,onset} \\ s_{j,pitch} \end{pmatrix} \tag{8}$$

The maximally translatable pattern (MTP) of a translation vector $\mathbf{T}$ for two melodies q and s is then defined as the set of melody points q_i which can be transformed to melody points s_j with the translation vector $\mathbf{T}$.

$$MTP(q, s, \mathbf{T}) = \{q_i | q_i \in q \wedge q_i + \mathbf{T} \in s\} \tag{9}$$

We use the pattern matching method SIAM, defining the similarity of two melodies as the largest set match achievable through translation with any vector, normalised by the length n of the query melody:

$$sim(q, s) = \frac{1}{n} \max_{\mathbf{T}} |MTP(q, s, \mathbf{T})| \tag{10}$$

The maximally translatable patterns leading to highest similarity are selected as matches, and their positions are determined through checking the onsets of the first and last notes of the MTPs.

3.3 Similarity measures comparing abstract representations

Wavelet transform converts a pitch sequence into a more abstract representation prior to comparison. We apply wavelet transform to each query segment q and melody s in the data-set prior to searching for matches.

Velarde et al. (2013) use wavelet coefficients to compare melodies: melodic segments are transformed with the Haar wavelet, at the scale of quarter notes. The wavelet coefficients indicate whether there is a contour change at a given moment in the melody, and similarity between two melodies is computed through city block distance of their wavelet coefficients. The method achieved considerable success for pattern discovery (Velarde & Meredith, 2014).

We use the authors' Matlab implementation to compute wavelet coefficients of duration-weighted pitch sequences. An example for an excerpt from a melody and the associated wavelet coefficients can be found in Figure 3. In accordance with Velarde and Meredith's procedure, we use city block distance to compare wavelet coefficients of query segment and match candidates, retaining similarity and position information of matches as described in Section 3.1.

Through the choice of music representation and comparison of the wavelet coefficients, this is an equal-length similarity measure sensitive to time dilation; however, it is transposition invariant.

4. Evaluation

For the evaluation, we distinguish three concepts: *match*, *instance* and *occurrence*. A *match* is a note sequence in a melody s at which maximum similarity with the query segment q is achieved, as detected by one of the similarity measures. An *occurrence* is a match whose similarity score exceeds a given threshold. An *instance* of a query phrase in a melody is given if the annotators indicate that a query phrase q is found within a given melody s. There can be multiple matches, occurrences and instances of a query phrase in a given melody due to phrase repetitions.

We evaluate each of the 1890 phrases in the data-set as query segments. Using the various similarity measures, we detect for each query segment q, per tune family, its matches in every melody s, excluding the melody from which the query segment was taken. As we are interested in the positions of the matches, we then determine which notes belong to the match. We assign to each note in a melody belonging to a match the similarity score of that match; the other notes receive an arbitrary score which for each measure exceeds the largest (CD, CBD, ED, WT) or smallest (LA, SIAM) similarity values of all matches.

Different thresholds on the similarity measures determine which notes are selected as constituting occurrences. Notes from matches with similarity values below (for the distance measures CD, CBD, ED, and WT) or above (for LA and SIAM) are considered as belonging to occurrences. We vary the similarity threshold for each measure stepwise from the matches' minimum similarity to maximum similarity, and for each step compare the retained occurrences to the human annotations.

We evaluate the occurrences against the annotations of 'almost identical' instances of the query segments in all melodies from the same tune family. As we would like to know which instances of query phrases most annotators agree on, we combine the three annotators' judgements into a *majority vote*: if for a given query segment q in one melody t, two or more annotators agree that a phrase p with exactly the same label (letter and number) appears in another melody s of the same

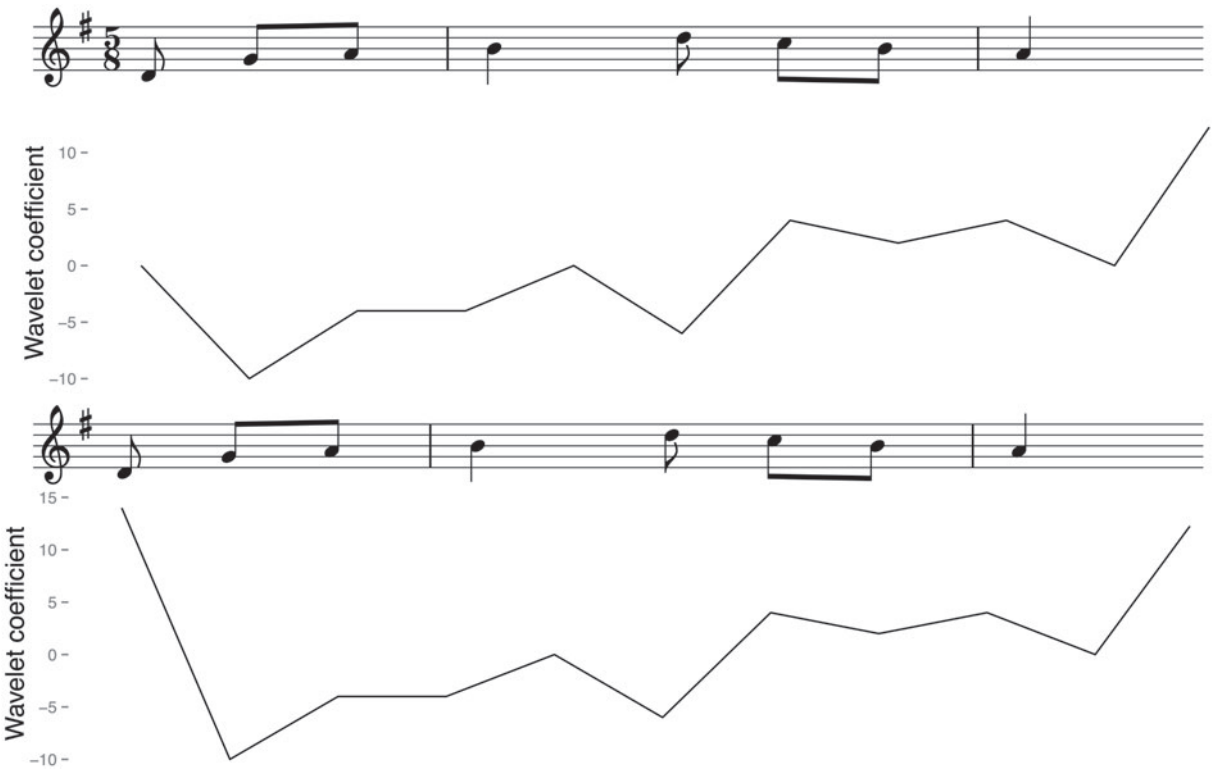

Fig. 3. The first two phrases of a melody from the tune family 'Daar ging een heer 1', with the values of the Haar wavelet coefficient underneath.

tune family, we consider phrase p's notes to constitute an instance of query segment q in s.

Conversely, if there is no such phrase in melody s to which two or more annotators have assigned exactly the same label as q, the notes of melody s do not represent any instances of that phrase. This means that the phrases considered 'related but varied' are not treated as instances of the query segment for the purpose of this study. The query phrases are compared against a total of 1,264,752 notes, of which 169,615 constitute instances of the query phrases.

All the notes which annotators consider to constitute instances of a query phrase are positive cases (P), all other notes are negative cases (N). The notes that a similarity measure with a given threshold detects as part of an occurrence are the positive predictions (PP), all other notes are negative predictions (NP). We define the intersection of P and PP, i.e. the notes which constitute an occurrence according to both a similarity measure with a given threshold and the majority of the annotators, as true positives (TP). True negatives (TN) are the notes which both annotators and similarity measures do not find to constitute an occurrence, i.e. the intersection of N and NP. False positives (FP) are defined as the intersection of N and PP, and false negatives (FN) as the intersection of P and NP.

We summarise the relationship between true positives and false positives for each measure in a receiver-operating characteristic (ROC) curve with the threshold as parameter and the axes defined by true positive rate (*tpr*) and false positive rate (*fpr*). The greater the area under the ROC curve (AUC), the better positive cases are separable from negative cases.

We would like to know the optimal similarity threshold for each measure to retrieve as many as possible notes annotated as instances correctly (high recall), and retrieving as few as possible irrelevant notes (high precision). A common approach to strike this balance is the F1-score, the harmonic mean of precision and recall. However, as our data have a strong bias (86.6%) towards negative cases, the F1-score is not an adequate criterion, as it focuses on true positives only. Therefore, we evaluate both positive and negative cases with sensitivity, specificity, positive and negative predictive values, and optimise the similarity threshold with respect to all these values through Matthews' correlation coefficient (Matthews, 1975).

Sensitivity, or recall, is equal to the true positive rate. It is defined as the number of true positives, divided by all positive cases, i.e. the number of notes correctly detected as part of occurrences, divided by all notes considered by annotators to constitute instances of query phrases.

$$SEN = \frac{TP}{P} \tag{11}$$

Specificity, or true negative rate, is defined as the number of true negatives, divided by all negative cases, i.e. is the number of notes which are correctly labelled as not belonging to an occurrence, divided by all notes considered by annotators to not belong to any occurrences.

$$SPC = \frac{TN}{N} = 1 - fpr \tag{12}$$

The positive predictive value, or precision, is defined as the number of true positives, divided by all positive predicted

cases, i.e. the number of all relevant notes labelled as part of an occurrence, divided by all notes detected to constitute occurrences by the similarity measure.

$$PPV = \frac{TP}{PP} \tag{13}$$

The negative predictive value is defined as the number of true negatives, divided by all negative predicted cases, i.e. the number of notes correctly labelled as not belonging to an occurrence, divided by all notes not constituting an occurrence according to the similarity measure.

$$NPV = \frac{TN}{NP} \tag{14}$$

To maximise both true positive and true negative rates, i.e. sensitivity and specificity, their sum should be as large as possible. The same goes for the positive and negative predictive values, the sum of which should be as large as possible. Powers (2007) suggests the measures informedness and markedness, which are zero for random performance, and one for perfect performance:

$$INF = SEN + SPC - 1 \tag{15}$$

$$MRK = PPV + NPV - 1 \tag{16}$$

Moreover, informedness and markedness are the component regression coefficients of Matthews' correlation coefficient ϕ, which is a good way of describing the overall agreement between a predictor and the ground truth (Powers, 2007). $\phi = 1.0$ for perfect agreement between ground truth and predictors, $\phi = 0.0$ for random performance, and $\phi = -1.0$ if there is a complete disagreement between ground truth and predictors, such that every positive case is a negative prediction, and vice versa.

$$\phi = \sqrt{INF \cdot MRK} \tag{17}$$

4.1 Glass ceiling

As our ground truth is defined as the majority vote of three annotators, we analyse the agreement of the three annotators with the majority vote. This gives us an indication of the 'glass ceiling' of the task, or how much agreement with the ground truth is maximally achievable. If the annotators do not perfectly agree on occurrences in our data-set, it is not realistic to expect that a similarity measure can achieve perfect agreement with the current ground truth (Flexer & Grill, 2016).

Table 2 shows that all annotators show similar agreement (measured by Matthews' correlation coefficient) with the annotators' majority vote. There are individual differences, however: for example, annotator 3 shows lower sensitivity, which is counter-balanced by a higher positive predictive value. This means that this annotator misses some of the occurrences on which the two other annotators agree, but finds almost no spurious occurrences.

The closer the compared similarity measures get to the annotators' agreement with the majority vote of $\phi \simeq 0.86$, the better we take them to be at finding occurrences of melodic segments in folk song melodies.

4.2 Baselines

Next to the best possible performance, we would like to know what a very naive approach would do, and introduce two baselines: one which considers every note as part of an occurrence (*always*), leading to perfect sensitivity, and a baseline which considers no note as part of an occurrence (*never*), leading to perfect specificity. The positive predictive value of *always* and the negative predictive value of *never* reflect the aforementioned bias towards negative cases; the respective other predictive values are zero as there are no negative predictions for *always*, and no positive predictions for *never*. As informedness is 0.0 in both cases, Matthews' correlation coefficient also leads to $\phi = 0.0$, meaning both have random agreement with the ground truth.

5. Comparison of similarity measures

Presently, we compare the previously described six similarity measures applied to the music representations for which they were proposed. The results suggest some answers to our first research question (Q1), i.e. which of the measures best serves the purpose of finding correct occurrences of melodic segments in folk songs.

5.1 Results

Figure 4 shows the ROC curves of the six compared measures, which reflect the true positive rate versus the false positive rate of the measures over a range of similarity thresholds. The higher and sharper the 'elbow' in the upper left corner, the better a measure can separate between positive and negative cases. Chance-level performance would be on the diagonal connecting zero true and false positive rates to full true and false positive rates.

The straightness of the curves on the right is caused by the fact that a considerable amount of the notes annotated as instances are not found by the measures. The ROC curve interpolates between considering all matches found by a given measure as occurrences, and considering all notes in the data-set as constituting occurrences, leading to $tpr = fpr = 1.0$.

For each measure, we report the area under the ROC curve to numerically represent the difference between the curves in Figure 4. Moreover, we select the similarity threshold which maximises Matthews' correlation coefficient, and report the associated ϕ, sensitivity, specificity, positive and negative predictive values. These measures are summarised in Table 3.

Table 3 shows that all of the compared measures agree much better with the ground truth than the baselines (*always* and *never*), but do not reach the level of the annotator agreement with the majority vote (cf. Table 2). Of the six measures, wavelet transform (WT) achieves least agreement with the

Table 2. The glass ceiling (top), or the annotators' agreement with the majority vote, and the majority vote agreement of the baselines (bottom), assuming every note (*always*) or no note (*never*) to be an occurrence. We report Matthews' correlation coefficient (ϕ) for the overall agreement, and the associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

	ϕ	SEN	SPC	PPV	NPV
Annotator					
Annotator1	0.877	0.900	0.982	0.887	0.985
Annotator2	0.865	0.913	0.976	0.855	0.986
Annotator3	0.861	0.815	0.993	0.947	0.972
Baseline					
always	0.0	1.0	0.0	0.134	0.0
never	0.0	0.0	1.0	0.0	0.866

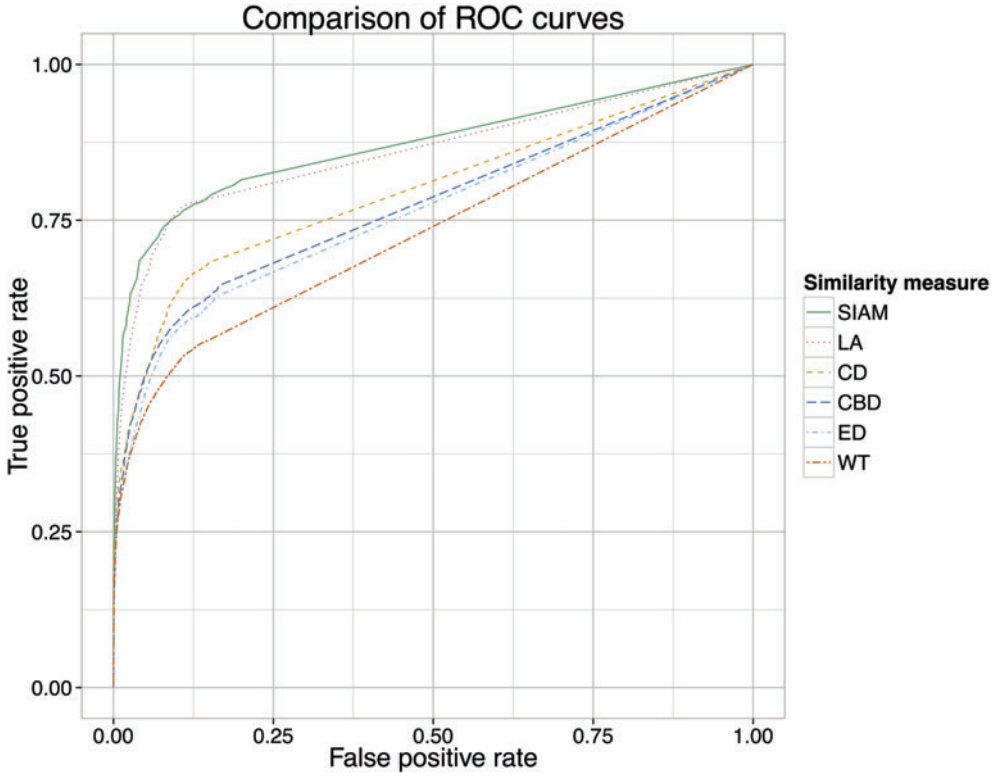

Fig. 4. The ROC curves for the various similarity measures, showing the increase of false positive rate against the increase of the true positive rate, with the threshold as parameter.

Table 3. Results of the compared similarity measures: area under the ROC curve (AUC), maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
WT	0.731	0.459	0.367	0.976	0.703	0.909
ED	0.764	0.468	0.482	0.948	0.589	0.922
CBD	0.774	0.499	0.425	0.973	0.708	0.916
CD	0.797	0.503	0.414	0.977	0.732	0.915
LA	0.859	0.621	0.646	0.956	0.695	0.946
SIAM	0.870	0.665	0.632	0.973	0.787	0.945

annotators, followed by the distance measures suggested in the field of ethnomusicology (ED, CBD and CD). Local alignment (LA) and structure induction (SIAM) agree best with the majority vote and achieve Matthews' correlation coefficients

of around $\phi = 0.621$ and $\phi = 0.665$, respectively. This is still much lower than the annotator agreement, but shows that the measures find most relevant occurrences, while producing less spurious than relevant results.

5.2 Discussion

With the present results, the distance measures Euclidean distance and city block distance (ED, CBD) do not seem to be promising candidates for finding occurrences of melodic segments in melodies. Still, while they do not achieve high agreement as measured in ϕ, they perform widely above the baselines. The relatively higher success of correlation distance (CD) is most likely to be attributed to the more fine-grained music representation in the form of duration-weighted pitch sequences, which reflect the duration of the notes.

It is surprising that the performance of wavelet transform (WT) lies below the other compared similarity measures, as in our previous study (Janssen, van Kranenburg, & Volk, 2015) which evaluated occurrences without taking their positions into account, it performed better than the distance measures. The low sensitivity, mainly responsible for the low maximal ϕ, is caused to a large extent by undetected phrase repetitions. As wavelet coefficients represent contour change in the pitch sequence, identical phrases with the same pitch sequence representation may have different wavelet transforms, depending on notes preceding the first note of a phrase, as illustrated in Figure 3. Therefore, in only 10% of the melodies with more than one instance of a given query phrase, wavelet finds more than one occurrence.

Local alignment (LA) and structure induction (SIAM) perform better than the before-mentioned measures. One reason for this might be that they are both variable-length similarity measures, and therefore deal with slight rhythmic variation and ornamentation more effectively. Moreover, both are transposition invariant: local alignment due to the pitch adjustment performed on the pitch sequence, structure induction due to the fact that it finds transpositions between pitches by definition.

From the present results, it is not possible to differentiate whether the best-performing measures do well because their comparison method is effective, or because of the music representations they use. It also seems that duration information might improve performance, as SIAM and CD, with duration information, perform comparatively well. Moreover, in respect to duration, time dilation differences might still affect the results negatively, and a music representation which attempts to correct these differences might improve results of the best measures even further.

The next section therefore compares different music representations for the compared measures, which gives clearer insights as to which of the observed differences in the present comparison are due to the measures themselves, and which differences can be overcome with different music representations. This also allows us to perform another comparison of the similarity measures with optimised music representations.

6. Dealing with transposition and time dilation differences

The automatic comparison of folk song melodies is impeded by transposition and time dilation differences of the transcriptions, as illustrated in Figure 1. It remains an open question which music representation can best resolve these differences (research question Q2 in the introduction). Therefore, we compare seven different music representations here, applied to each of the similarity measures as appropriate.

6.1 Music representations

In the previous section, four similarity measures used a pitch sequence (P) as music representation, which does not resolve transposition differences, and does not take the duration of notes into account. To solve the problem of transposition differences, two approaches are conceivable: a music representation consisting of sequences of pitch intervals (PI), i.e. sequences of differences between successive pitches, and pitch-adjusted sequences (PA), as described and used for local alignment in the previous section.

With respect to the representation of duration, we have already seen the use of pitch and onset tuples (PO) for structure induction, and duration-weighted pitch sequences (DW) for correlation distance and wavelet transform in the previous section. The latter representation can of course also be combined with pitch adjustment, and the resulting representation (PADW) will be compared, too.

To solve the problem of time dilation differences, we test whether time dilation differences can be corrected through automatic comparison of duration value frequencies, analogous to pitch adjustment. To this end, we calculate duration histograms, in which seven duration bins are filled with the count of each duration. Only durations which are in 2:1 integer ratios are considered, as other durations, such as punctuated rhythms or triplets, would not allow easy scaling. The smallest considered duration is a hemidemisemiquaver, or 64th note, and all doublings of this duration are considered up to a semibreve, or whole note. Analogous to Equation 1, we define the duration histogram intersection of two duration histograms h_t and h_s, with a total number of r duration bins k:

$$DHI(h_t, h_s, \sigma) = \sum_{k=1}^{r} \min(h_{t,k+\sigma}, h_{s,k}), \qquad (18)$$

For each tune family, we randomly pick one reference melody and for each other melody in the tune family we compute the shift σ that yields a maximum value for the histogram intersection, and use that σ to calculate the multiplier of the onsets of melody t with relation to melody s:

$$Mult(t, s) = 2^{\sigma} \qquad (19)$$

We also tested the influence of the randomly picked reference melodies on the results of duration histogram intersection by running the procedure 100 times, and comparing against annotated duration adjustments. Of the 2822 pairs of tune

family variants, 66.5% were adjusted in the same way as annotated. This means that a third of the pairs are adjusted incorrectly, so it is an open question whether duration adjustment improves results, in spite of its rather high error rate. At any rate, the low standard deviation of 1.3% of the success rate means that it does not matter greatly which melodies are picked as reference melodies.

The result of this procedure leads us to a music representation which is pitch and duration adjusted (DA). We also make use of the metadata of the Annotated Corpus to find out the hand-adjusted (HA) optimal transposition and time dilation of each melody. Hand-adjustment is not feasible for a large collection of folk songs, but is a useful reference for comparison with the automatically adjusted music representations.

Wavelet transform and structure induction (WT, SIAM) are defined for specific representations, namely a duration-weighted pitch sequence (DW) and pitch/onset tuples (PO), respectively. As such, not all music representations are applicable for these measures. For WT, only duration-weighted pitch sequences and adjustments thereof are tested (DW, PADW, DA, HA). For SIAM, the duration adjustment and hand adjustment (DA, HA) are applied to the pitch/onset tuples, which differ slightly from the DA and HA representations in the other measures, in which duration weighed pitch sequences are adjusted.

6.2 Results

From Figure 5, it can be seen that music representation has considerable influence on the success of the similarity measures. Overall, most music representations show better performance than the pitch sequence representation (P). The only exception is the pitch interval representation (PI): attempting to resolve transposition differences between songs through pitch intervals deteriorates performance.

Duration information (DW) improves the performance of some distance measures and local alignment (LA, CD, CBD, ED), as does pitch adjustment (PA). A combination of the two (PADW) improves these measures even further. Duration adjustment (DA) of the duration-weighted sequences gives a slight advantage for some measures (CBD, LA), but does not seem to affect the other measures much (ED, CD, WT, SIAM).

The difference with the hand-adjusted (HA) representation, resulting in the best performance for all measures, shows that automatic adjustment is not completely able to resolve transposition and time dilation differences. A full overview of all music representations and measures, with the resulting AUC as well as maximal ϕ with associated retrieval measures, can be found in Table A1 in the Appendix.

Figure 6 shows another comparison of ROC curves for the six similarity measures, with optimised music representations. We pick for each measure the music representation which results in the highest AUC. As we could not improve some measures (CD, SIAM) through other music representations, their curves are identical to those in Figure 4. We find that a number of measures (ED-DA, CBD-DA) perform much better than before as a result of the corrections for transposition and time dilation differences. Local alignment (LA-DA) and city block distance (CBD-DA) even outperform SIAM with these adjustments.

In Table 4, we report the area under the ROC curve for all measures with optimised music representations, as well as the maximised ϕ correlation coefficient with associated sensitivity, specificity, positive and negative predictive values.

With optimised music representation, local alignment and city block distance achieve values for ϕ close to that of structure induction (SIAM). The differences among these three measures can mainly be found in their sensitivity and positive predictive values, as SIAM and CBD-DA achieve lower sensitivity than LA-DA, but compensate by higher positive predictive values.

Euclidean distance is also improved considerably through duration and pitch adjustment; however, its ϕ is somewhat lower than that of the aforementioned measures. Correlation distance and wavelet transform could not be much improved through any of the tested music representations, and remain at relatively low ϕ values.

6.3 Discussion

The present section shows that transposition and time dilation differences have considerable influence on the results of several of the compared measures (CBD, ED, LA). We conclude that the relative success of local alignment in the previous section was caused by its pitch-adjusted music representation, and that city block distance and Euclidean distance perform much better on a pitch-adjusted representation. However, local alignment achieves slightly higher AUC than the distance measures for each representation, showing that it is the most effective overall.

As wavelet transform, correlation distance and structure induction (WT, CD, SIAM) are already defined as transposition invariant, they cannot be improved through pitch adjustment. Wavelet transform is improved through duration adjustment to some extent. The similarity threshold associated with maximal agreement ϕ is stricter for the duration-adjusted case, i.e. fewer matches are considered occurrences, accounting for higher positive predictive value but lower sensitivity (cf. Table A1). This leads us to the conclusion that the drawback of wavelet transform observed in the previous section, i.e. that it may miss phrase repetitions within a melody, cannot be resolved through our strategy for duration adjustment.

Correlation distance and structure induction perform slightly worse with duration adjustment as compared to their original music representation (cf. Table A1). For both measures, the similarity threshold associated with maximal agreement ϕ is not affected by duration adjustment. Duration adjustment increases the number of occurrences for both measures. As some of these occurrences are true positives, this leads to higher sensitivity. Inversely, we have seen that about a third of the automatic adjustments are incorrect, and these mis-

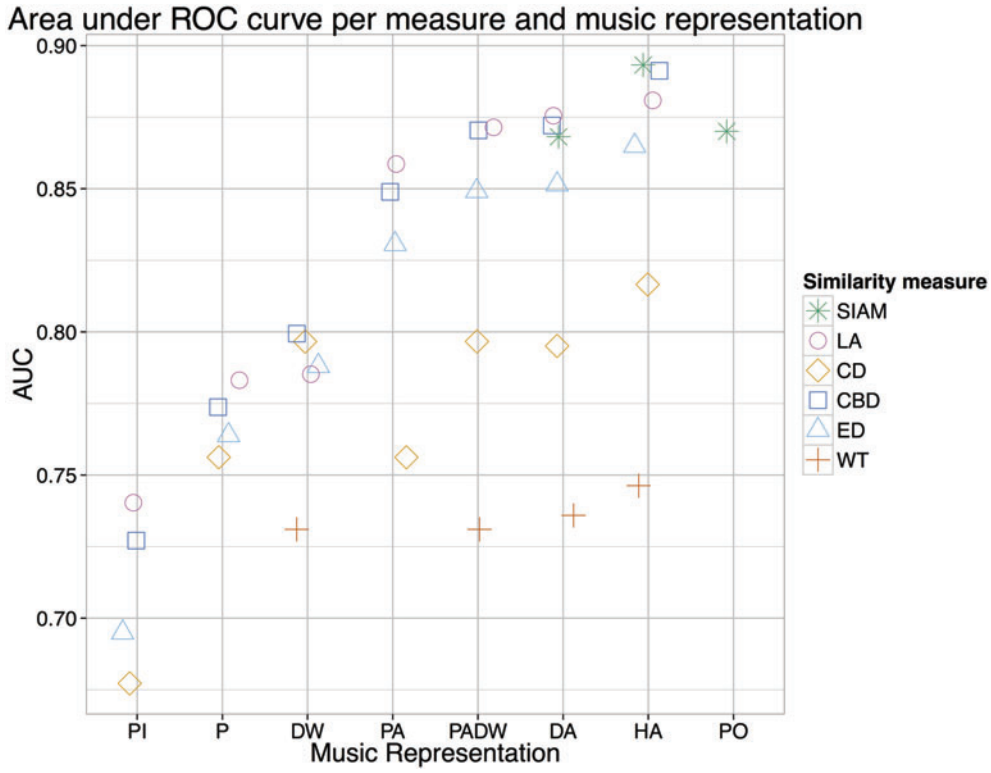

Fig. 5. The area under the ROC curves (AUC) of the similarity measures for different music representations: pitch interval (PI), pitch (P), duration weighted (DW), pitch adjusted (PA), pitch adjusted and duration weighted (PADW), metrically adjusted (DA), hand adjusted (HA), and pitch/onset (PO). For wavelet transform (WT) and structure induction (SIAM), not all music representations are applicable, and only SIAM uses the pitch/onset representation.

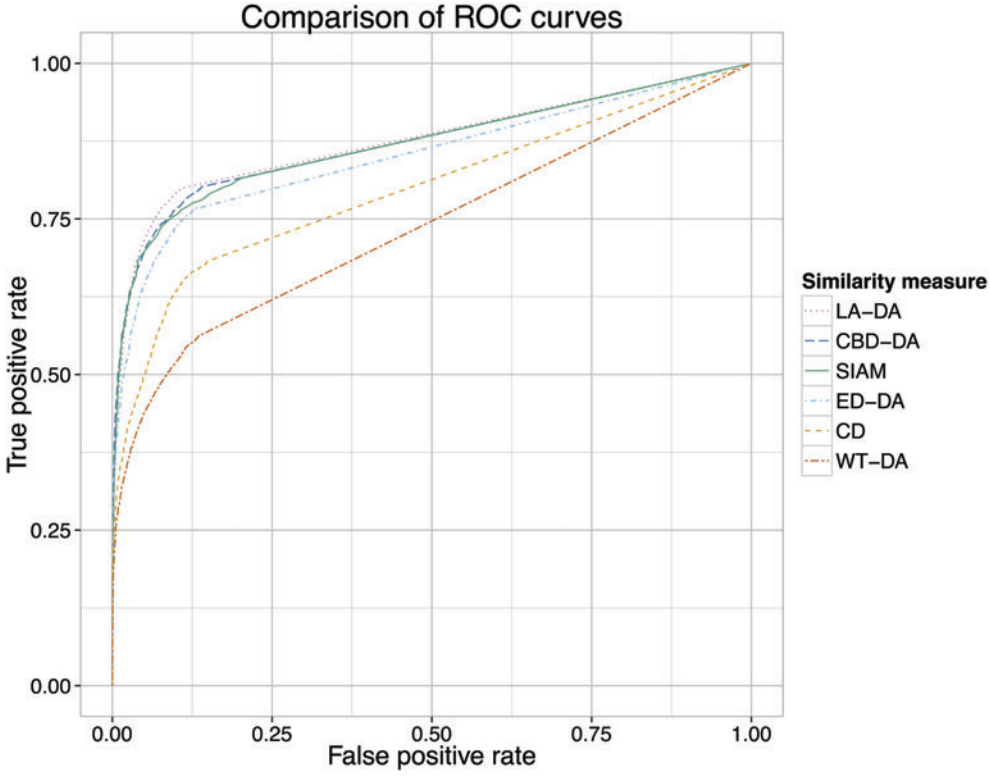

Fig. 6. The ROC curves for the various similarity measures with optimised music representations, showing the increase of false positive rate against the increase of the true positive rate, with the threshold as parameter.

Table 4. Results of the similarity measures with optimised music representations: area under the ROC curve (AUC), maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
WT-DA	0.736	0.454	0.320	0.985	0.772	0.903
CD	0.797	0.503	0.414	0.977	0.732	0.915
ED-DA	0.851	0.610	0.627	0.957	0.693	0.943
SIAM	0.870	0.665	0.632	0.973	0.787	0.945
CBD-DA	0.872	0.663	0.608	0.978	0.808	0.942
LA-DA	0.875	0.668	0.675	0.965	0.748	0.950

adjustments produce false positives, decreasing the positive predictive value.

In summary, we can observe that transposition differences can be adequately resolved through pitch histogram intersection, while a better way of adjusting duration is needed, as the present approach of duration histogram intersection leads to many errors, and improves the performance only slightly or even not at all.

Based on our comparison of similarity measures with optimised music representations, city block distance and local alignment with pitch and duration adjustment, and structure induction (CBD-DA, LA-DA, SIAM) are the best approaches to finding occurrences of melodic segments in folk song melodies. None of them reach the level of agreement with the majority vote as the human annotators (cf. Table 2), however.

This leads to the question whether a combination of the best-performing measures might show better performance than the individual measures. This question will be investigated in the following section.

7. Combination of the best-performing measures

We combine the three best-performing measures (CBD-DA, LA-DA, SIAM), observing whether this combination improves performance, addressing Q3 from the introduction.

7.1 Method

For each measure, we retain only those matches which exceed the best similarity threshold, obtained from optimisation with respect to ϕ. For CBD-DA, matches with $dist(q, p) \leq 0.98$, for LA-DA, matches with $sim(q, s) \geq 0.55$, and for SIAM, matches with $sim(q, s) \geq 0.58$ are retained.

We combine the three best similarity measures in the same way as we combine the annotators' judgements to a majority vote. To this end, we redefine the notion of occurrence: we consider only those notes to constitute an occurrence which two or more measures detect as part of a match, given the respective measures' optimal similarity thresholds. We investigate how well this *combined measure* agrees with the annotators' majority vote.

Table 5. Results of a combined similarity measure from SIAM, CBD-DA and LA-DA, represented by the maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV).

ϕ	SEN	SPC	PPV	NPV
0.703	0.648	0.981	0.84	0.947

7.2 Results

Table 5 presents Matthews' correlation coefficient, sensitivity, specificity, positive and negative predictive values of the combined measure. The agreement $\phi = 0.703$ is higher than that of the individual measures, and outperforms the hand-adjusted music representations of all individual measures.

7.3 Discussion

The combined measure's increased performance is mainly caused by its positive predictive value ($PPV = 0.84$), which is considerably higher than the values achieved by any individual measure, and close to the values of two of the annotators. The sensitivity $SEN = 0.648$ is comparable to that of the individual measures, so it is still a lot lower than the annotators' sensitivity, meaning that the combined measure still misses more instances of melodic segments than human experts.

Based on our study, we find that the combined measure is the best currently achievable method for detecting occurrences of melodic segments automatically. However, we assume the same optimal threshold of the individual similarity measures over the whole data-set. This would be inappropriate if there were subgroups of the tested melodies which necessitate higher or lower thresholds to achieve optimal agreement with the annotations. Moreover, the agreement is also likely to vary in different subgroups of melodies, leading to different error rates, depending on the selection of melodies tested.

Therefore, in the next section, we proceed to test how leaving out tune families from the data-set affects the optimised similarity threshold of the three best-performing measures, and how much the agreement with the ground truth varies depending on the evaluated tune family.

8. Optimisation and performance of similarity measures for data subsets

In the present section, we investigate whether subgroups of our data-set affect the optimised threshold of the three best-performing similarity measures (LA-DA, CBD-DA and SIAM) to such an extent that it is inappropriate to assume one optimal threshold for the whole data-set. Moreover, we observe the variation of the agreement ϕ with the ground truth, depending on the evaluated subset. This analysis addresses research question Q4 from the introduction.

As the tune families form relatively homogenous subgroups of melodies within the Annotated Corpus, we use the 26 tune families as subsets. This has the disadvantage that the subsets have different sizes, but the advantage of knowing a priori that the subsets are different by human definition.

8.1 Method

For each of the 26 tune families, we optimise the similarity threshold for each measure, leaving that tune family out of the data-set. For this 'leave one tune family out' procedure, we remove the matches from the tune family under consideration from the data-set, and vary the similarity threshold in this reduced data-set, selecting the threshold that maximises Matthews' correlation coefficient ϕ with the ground truth.

Next, we use this 'leave one tune family out' optimised threshold to detect occurrences in the considered tune family, and observe the resulting agreement (ϕ_{tf}) with the ground truth of this tune family. This gives us a different value ϕ_{tf} for the 26 tune families. Ideally, we would like ϕ to be high on average, and show small variance.

For comparison of the optimised thresholds *thres* after leaving out one tune family, we standardise them, using the arithmetic mean and standard deviation of all similarity scores produced by a given measure.

$$thres_{std} = \frac{thres - \overline{sim}}{\mathbf{SD}(sim)} \tag{20}$$

As a result, the standardised threshold $thres_{std}$ is mapped into a space centred on 0, representing the average similarity score, and in which each unit represents one standard deviation of the similarity scores $\mathbf{SD}(sim)$. As city block distance has similarity values ranging from $0 \leq dist \leq 5.29$, while local alignment and structure induction are bounded by the interval $sim = (0, 1]$, the standardisation allows better interpretation of the differences between optimised thresholds.

To compare the variation in agreement ϕ_{tf} of the individual measures, the combined measure and the annotators with the ground truth, we use a Tukey box and whiskers plot (Tukey, 1977), in which the median is indicated by a horizontal line, and the first (25%) and third (75%) quartiles of the data by the horizontal edges of the box. All data exceeding the first and third quartiles by no more than 1.5 times the inter-quartile range are represented by vertical lines. All data outside this range are considered outliers and plotted as individual dots.

8.2 Similarity thresholds

The thresholds vary very little when specific tune families are left out of the optimisation procedure: most 'leave one tune family out' optimisations result in the same optimal threshold as the optimisations on the full data-set in the previous section, indicated by black stripes in Figure 7. There are some minor deviations, but none larger than 0.3 standard deviations, noticeable in SIAM's thresholds.

8.3 Agreement with ground truth

The agreement with the ground truth, measured in the tune family-dependent Matthews' correlation coefficient ϕ_{tf}, depends greatly on the considered tune family, as can be seen in Figure 8. This is true especially for the similarity measures SIAM and CBD-DA, which result in a wide range of values for ϕ_{tf}, while LA-DA shows less variation in ϕ_{tf}.

The combined measure (COMB) achieves consistently higher agreement with the ground truth than the measures of which it is composed, as can be seen in its higher mean (indicated by a horizontal line in the box plot), though its variation between $0.45 < \phi_{tf} < 0.83$, depending on the evaluated tune family, is considerable.

The annotators are more consistent than the similarity measures overall, but there are some remarkable outliers for all of them, some as low as $\phi_{tf} = 0.47$, which is comparable to some of poorest algorithmic results.

8.4 Discussion

The thresholds vary little when leaving out tune families from the optimisation procedure (cf. Figure 7), indicating that it is reasonable to assume the same optimal similarity threshold throughout the whole data-set. This means that the combined measure can also be applied with one similarity threshold per measure to the whole data-set.

The variation in agreement when evaluated against the tune families separately (cf. Figure 8) indicates that SIAM and CBD-DA are less robust towards differences between tune families than LA-DA and the combined measure.

Less variation in ϕ_{tf} means that a measure is more consistent with respect to the number of errors it produces, regardless of the tune family under consideration. Neither any of the individual measures, nor the combined measure shows enough consistency that a computational folk song study using them should consider the error constant over all subsets of a folk song corpus.

As the annotators also show considerable variation in their agreement with the majority vote, it is unlikely that a computational method can find occurrences in this folk song corpus without producing variable amounts of errors, depending on the evaluated tune family.

9. Conclusion

We have investigated the success of six similarity measures at finding occurrences of melodic segments in folk songs. We

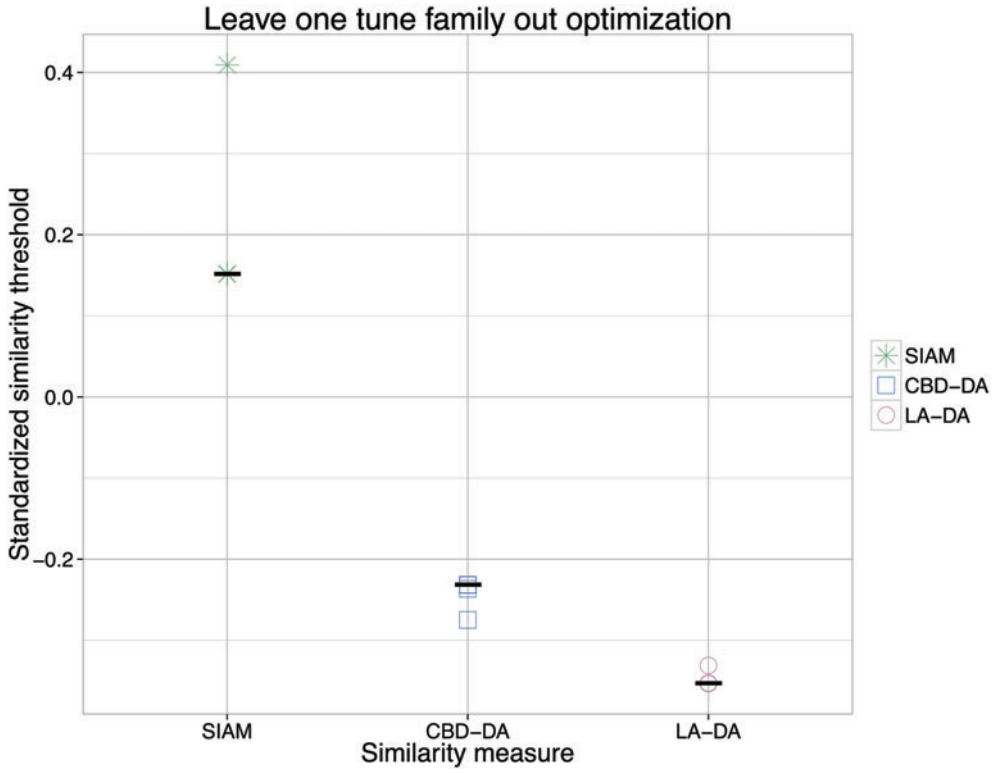

Fig. 7. The thresholds resulting from 'leave one tune family out' optimisation. The black stripes indicate the threshold of the optimisation of the full data-set. All of the measures' thresholds are close to each other.

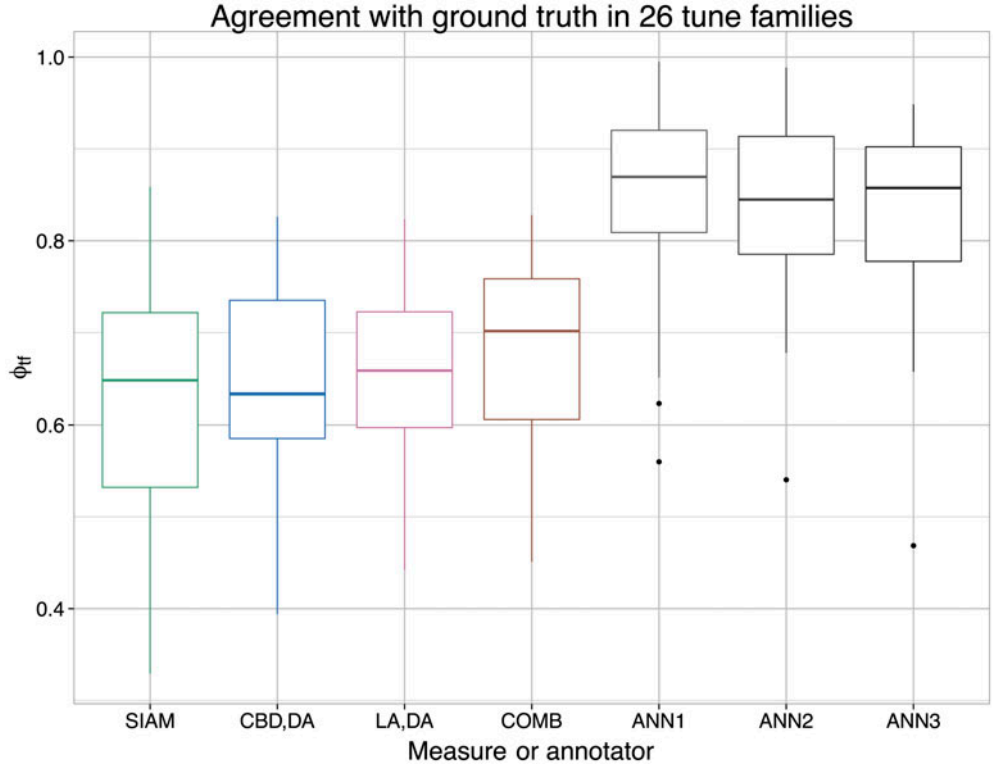

Fig. 8. The agreement (in ϕ) of the three similarity measures and the annotators with the majority vote, evaluated separately for each tune family. The similarity measures show more variation than the annotators, even though there are also some remarkable low outliers for the annotators.

tested how well the similarity measures would find occurrences of phrases, evaluating their results against the majority vote of three annotators' judgements of phrase occurrences in Dutch folk songs. We summarise the answers to the four research questions posed in the introduction, and conclude with some steps for future work.

Regarding the question of which similarity measure is best suited for finding occurrences (Q1), our results of Section 5 indicate that structure induction and local alignment are the most successful approaches for this task given the music representation for which they were defined. However, when duration as well as pitch information is supplied, and time dilation and transposition are corrected, city block distance performs even slightly better than structure induction, and the results of local alignment can be improved, as shown in Section 6.

We show that the performance of all similarity measures can be improved when time dilation and transposition differences between folk songs are adjusted (Q2, Section 6). The best way to adjust pitch differences automatically is histogram intersection, leading to much improved results. Providing information on the duration as well as pitch of compared notes improves the success of all measures considerably, but time dilation differences remain a problem. Our approach to adjust durations automatically through histogram intersection led to slight improvement for some measures, but no improvement for others.

A combination of the best-performing measures (SIAM, CBD-DA, LA-DA) does indeed perform better than each measure individually (Q3), and is the best measure arising from our comparison. It produces about 16% spurious results, close to the values of human annotators. However, the combined measure misses about a third of the relevant instances of query segments, whereas the annotators miss only around 10%. In consequence, the combined measure is not a replacement for human judgements on melodic occurrences, but to our knowledge produces the best results with the current similarity measures and music representations.

In Section 8, we show that the agreement of the three best-performing similarity measures with the ground truth differs depending on the evaluated tune family (Q4). However, we also show that human annotators show almost as much variation. Our optimisation of the similarity threshold on subsets of the full data-set also leads to almost no change in the selected similarity thresholds of SIAM, CBD-DA and LA-DA, meaning that it is appropriate to assume the same threshold for the whole data-set. Yet, in statistical analyses of occurrences detected by these measures or the combined measure, it would be inappropriate to assume the same error rate throughout the whole data-set. When categories within a music collection are defined, as is the case with tune families in the Meertens Tune Collections, it is therefore advisable to make use of these categories and to assume different error terms for each of them.

Further research into alternative similarity measures and better ways of representing musical information is needed to improve the success of computational detection of melodic occurrences. Our research on music representation indicates that better methods to adjust time dilation differences will lead to much improved results. Moreover, other weighting schemes for local alignment still need to be explored. Another area of improvement is the combination of the judgements from different similarity measures into one combined measure, for which more successful ways than the currently used majority vote approach may be found.

The annotations used in this study distinguish between two levels of instances, those which are 'related but varied' and those which are 'almost identical'. We have focused on the latter category in the current study; it would be interesting to see whether the best-performing similarity measures of this study and their combination would also work best for the 'related but varied' category, and if so, in how much the optimal similarity thresholds would be affected.

It is also important to validate our findings in different folk song corpora, and in different genres. Unfortunately, no comparable annotations on occurrences in folk songs exist to our knowledge. Annotations in works of Classical music, used as validation sets for pattern discovery, might be an interesting ground of comparison. More annotation data and comparative research are needed to overcome some of the challenges we have presented in finding occurrences of melodic segments in folk songs, and in melodies from other genres, and to ascertain the robustness of computational methods.

Acknowledgements

We thank Gissel Velarde for kindly providing her code, and the Music Cognition Group of the University of Amsterdam for valuable feedback at several stages of the research. We also thank Sanneke van der Ouw, Jorn Janssen and Ellen van der Grijn for their annotations. Last but not least, many thanks to Alan Marsden and two anonymous reviewers for their insightful comments on the manuscript.

Funding

Berit Janssen and Peter van Kranenburg are supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences under the auspices of the Tunes & Tales project. For further information, see http://ehumanities.nl. Anja Volk is supported by the Netherlands Organisation for Scientific Research through an NWO-VIDI [grant number 276-35-001].

References

Bade, K., Nürnberger, A., Stober, S., Garbers, J., & Wiering, F. (2009). *Supporting folk-song research by automatic metric learning and ranking*. Proceedings of the 10th International Society for Music Information Retrieval Conference, Kobe, Japan (pp. 741–746).

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

Boot, P., Volk, A., & de Haas, W. B. (2016). Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research, 45*, 223–238.

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

Conklin, D. & Anagnostopoulou, C. (2011). Comparative pattern analysis of cretan folk songs. *Journal of New Music Research, 40*, 119–125. doi:10.1080/09298215.2011.573562

Cuthbert, M. S., & Ariza, C. (2010). *music21: A toolkit for computer-aided musicology and symbolic music data*. Procedings of the 11th International Society for Music Information Retrieval Conference, Utrecht, Netherlands (pp. 637–642).

Eerola, T., Jäärvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception: An Interdisciplinary Journal, 18*, 275–296.

Flexer, A., & Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research, 45*, 239–251. doi:10.1080/09298215.2016.1200631

Garbers, J., Volk, A., van Kranenburg, P., Wiering, F., Grijp, L. P., & Veltkamp, R. C. (2007). *On pitch and chord stability in folk song variation retrieval*. International Conference on Mathematics and Computation in Music, Berlin, Germany (pp. 97–106).

Grachten, M., Arcos, J. L., & López de Mántaras, R. (2005). *Melody retrieval using the implication/realization model*. Music Information Retrieval Evaluation eXchange, London.

Hillewaere, R., Manderick, B., & Conklin, D. (2009). *Global feature versus event models for folk song classification*. Proceedings of the 10th International Society for Music Information Retrieval Conference, Kobe, Japan (pp. 729–733).

Janssen, B., van Kranenburg, P., & Volk, A. (2015). *A comparison of symbolic similarity measures for finding occurrences of melodic segments*. Proceedings of the 16th International Society for Music Information Retrieval Conference, Málaga, Spain (pp. 659–665).

Juhász, Z. (2006). A systematic comparison of different European folk music traditions using self-organizing maps. *Journal of New Music Research, 35*, 95–112. doi:10.1080/09298210600834912

van Kranenburg, P., de Bruin, M., & Grijp, L.P. (2014). *The Meertens Tune Collections* (Tech. Rep.). Amsterdam: Meertens Online Reports No. 2014-1.

van Kranenburg, P., Janssen, B., & Volk, A. (2016). *The Meertens Tune Collections: the Annotated Corpus (MTC-ANN) versions 1.1 and 2.0* (Tech. Rep.). Amsterdam: Meertens Online Reports No. 2016-1.

van Kranenburg, P., Volk, A., & Wiering, F. (2013). A comparison between global and local features for computational classification of folk song melodies. *Journal of New Music Research, 42*, 1–18. doi:10.1080/09298215.2012.718790

Louhivuori, J. (1990). Computer aided analysis of Finnish spiritual folk melodies. In H. Braun (Ed.), *Probleme der Volksmusikforschung* (pp. 312–323). Bern: Peter Lang.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA - Protein Structure, 405*, 442–451. doi:10.1016/0005-2795(75)90109-9

Meredith, D. (2006). Point-set algorithms for pattern discovery and pattern matching in music. In Crawford, T., & Veltkamp, R.C. eds. *Content-Based Retrieval. Dagstuhl Seminar Proceedings 06171*. Dagstuhl, Germany.

Meredith, D. (2014). *COSIATEC and SIATEC Compress: Pattern discovery by geometric compression*. Music Information Retrieval Evaluation Exchange, Taipei.

Mongeau, M., & Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities, 24*, 161–175.

Müllensiefen, D., & Frieler, K. (2007). Modelling experts notions of melodic similarity. *Musicae Scientiae, 11*, 183–210. doi:10.1177/102986490701100108

Oliphant, T. E. (2007). Python for scientific computing. *Computing in Science and Engineering, 9*, 10–20. doi:10.1109/MCSE.2007.58

Olthof, M., Janssen, B., & Honing, H. (2015). The role of absolute pitch memory in the oral transmission of folksongs. *Empirical Musicology Review, 10*(3), 161–174.

Powers, D. M. W. (2007). Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *Journal of Machine Learning Technologies, 2*, 24. doi:10.1.1.214.9232

Scherrer, D. K., & Scherrer, P. H. (1971). An experiment in the computer measurement of melodic variation in folksong. *The Journal of American Folklore, 84*, 230–241.

Smith, T., & Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology, 147*, 195–197.

Steinbeck, W. (1982). *Struktur und Ähnlichkeit. Methoden automatisierter Melodienanalyse*. Kassel: Bärenreiter.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Urbano, J., Lloréns, J., Morato, J. & Sánchez-Cuadrado, S. (2011). Melodic similarity through shape similarity. In Ystad, S.I., Aramaki, M., Kronland-Martinet, R., & Jensen, K.. eds. *Exploring Music Contents*: 7th International Symposium, CMMR 2010, Málaga, Spain, June 21–24, 2010. Revised Papers (LNCS 6684). (pp. 338–355).

Velarde, G., & Meredith, D. (2014). *A wavelet-based approach to the discovery of themes and sections in monophonic melodies*. Music Information Retrieval Evaluation Exchange, Taipei, Taiwan.

Velarde, G., Weyde, T., & Meredith, D. (2013). An approach to melodic segmentation and classification based on filtering with the Haar-wavelet. *Journal of New Music Research, 42*, 325–345. doi:10.1080/09298215.2013.841713

Volk, A., & van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae, 16*, 317–339. doi:10.1177/1029864912448329

Appendix A.

Table A1 shows the influence of music representations on all the compared measures.

Table A1. Area under ROC curve, maximal ϕ correlation coefficient with associated sensitivity (SEN), specificity (SPC), positive and negative predictive values (PPV, NPV) for all similarity measures in all applicable music representations.

Measure	AUC	ϕ	SEN	SPC	PPV	NPV
WT-DW	0.731	0.459	0.368	0.976	0.703	0.909
WT-PADW	0.731	0.459	0.368	0.976	0.703	0.909
WT-DA	0.736	0.454	0.320	0.985	0.772	0.903
WT-HA	0.746	0.460	0.324	0.986	0.778	0.904
ED-PI	0.695	0.373	0.243	0.985	0.718	0.894
ED-P	0.764	0.468	0.482	0.948	0.589	0.922
ED-DW	0.788	0.540	0.441	0.980	0.774	0.919
ED-PA	0.831	0.554	0.616	0.940	0.612	0.940
ED-PADW	0.849	0.618	0.619	0.962	0.716	0.942
ED-DA	0.851	0.610	0.627	0.957	0.693	0.943
ED-HA	0.865	0.612	0.610	0.962	0.714	0.941
CBD-PI	0.727	0.424	0.365	0.967	0.634	0.908
CBD-P	0.774	0.499	0.425	0.973	0.708	0.916
CBD-DW	0.799	0.581	0.468	0.985	0.824	0.923
CBD-PA	0.849	0.589	0.564	0.966	0.720	0.935
CBD-PADW	0.870	0.663	0.601	0.979	0.818	0.941
CBD-DA	0.872	0.663	0.608	0.978	0.808	0.942
CBD-HA	0.891	0.696	0.651	0.978	0.822	0.948
CD-PI	0.677	0.313	0.214	0.979	0.617	0.889
CD-P	0.756	0.426	0.266	0.990	0.810	0.897
CD-PA	0.849	0.589	0.564	0.966	0.720	0.935
CD-DW	0.797	0.503	0.414	0.977	0.732	0.915
CD-PADW	0.797	0.503	0.414	0.977	0.733	0.915
CD-DA	0.795	0.501	0.420	0.975	0.720	0.916
CD-HA	0.817	0.525	0.448	0.975	0.734	0.919
LA-PI	0.740	0.470	0.416	0.967	0.662	0.915
LA-P	0.783	0.533	0.491	0.967	0.695	0.925
LA-DW	0.785	0.573	0.510	0.974	0.750	0.928
LA-PA	0.859	0.621	0.646	0.956	0.695	0.946
LA-PADW	0.871	0.665	0.658	0.968	0.759	0.948
LA-DA	0.875	0.668	0.675	0.965	0.748	0.950
LA-HA	0.881	0.682	0.695	0.965	0.753	0.953
SIAM-PO	0.870	0.665	0.632	0.973	0.787	0.945
SIAM-DA	0.868	0.663	0.641	0.971	0.772	0.946
SIAM-HA	0.893	0.696	0.688	0.970	0.783	0.953