Royal Netherlands Academy of Arts and Sciences (KNAW)
KONINKLIJKE NEDERLANDSE AKADEMIE VAN WETENSCHAPPEN

A Comparison of Symbolic Similarity Measures for Finding Occurrences of Melodic Segments

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

published in
Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015
2015

document version
Publisher's PDF, also known as Version of record

document license
CC BY-ND

Link to publication in KNAW Research Portal

citation for published version (APA)
Janssen, B., van Kranenburg, P., & Volk, A. (2015). A Comparison of Symbolic Similarity Measures for Finding Occurrences of Melodic Segments. In M. Muller, & F. Wiering (Eds.), *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015, Málaga, Spain, October 26-30, 2015* International Society of Music Information Retrieval (ISMIR).

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the KNAW public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain.
• You may freely distribute the URL identifying the publication in the KNAW public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
pure@knaw.nl

Download date: 11. Feb. 2025

A COMPARISON OF SYMBOLIC SIMILARITY MEASURES FOR FINDING OCCURRENCES OF MELODIC SEGMENTS

Berit Janssen
Meertens Institute,
Amsterdam
berit.janssen
@meertens.knaw.nl

Peter van Kranenburg
Meertens Institute,
Amsterdam
peter.van.kranenburg
@meertens.knaw.nl

Anja Volk
Utrecht University,
the Netherlands
a.volk@uu.nl

ABSTRACT

To find occurrences of melodic segments, such as themes, phrases and motifs, in musical works, a well-performing similarity measure is needed to support human analysis of large music corpora. We evaluate the performance of a range of melodic similarity measures to find occurrences of phrases in folk song melodies. We compare the similarity measures correlation distance, city-block distance, Euclidean distance and alignment, proposed for melody comparison in computational ethnomusicology; furthermore Implication-Realization structure alignment and B-spline alignment, forming successful approaches in symbolic melodic similarity; moreover, wavelet transform and the geometric approach Structure Induction, having performed well in musical pattern discovery. We evaluate the success of the different similarity measures through observing retrieval success in relation to human annotations. Our results show that local alignment and SIAM perform on an almost equal level to human annotators.

1. INTRODUCTION

In many music analysis tasks, it is important to query a large database of music pieces for the occurrence of a specific melodic segment: which pieces by Rachmaninov quote *Dies Irae*? Which bebop jazz improvisers used a specific Charlie Parker lick in their solos? How many folk song singers perform a melodic phrase in a specific way?

In the present article, we compare a range of existing similarity measures with the goal of finding occurrences of melodic segments in a corpus of folk song melodies. This is a novel research question, evaluated on annotations which have been made specifically for this purpose. The insights gained from our research on the folk song genre can inform future research on occurrences in other genres.

We evaluate similarity measures on a set of folk songs, in which human experts annotated phrase similarity. We use these annotations as evidence for occurrences of melodic segments in related songs. If we know that a similarity measure is successful in finding the annotated occurrences in this set, we infer that the measures will be successful for finding correct occurrences of melodic segments of phrase length in a larger dataset of folk songs as well. We describe the dataset in more detail in Section 2.

In computational ethnomusicology various methods for comparing folk song melodies have been suggested: as such, correlation distance [12], city-block distance and Euclidean distance [14] have been considered promising. Research on melodic similarity in folk songs also showed that alignment measures reproduce human judgements on agreement between melodies well [16].

As this paper focusses 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 [7, 17] have shown success when 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.

Additionally, we apply the most successful similarity measures from the MIREX symbolic melodic similarity track in our research. The best measure of MIREX 2005 (Grachten et al. [4]), was evaluated on RISM incipits, which are short melodies or melodic segments, therefore relevant for our task. In recent MIREX editions the algorithm by Urbano et al. [15] has been shown to perform well on the EsAC folk song collection. [2]

We present an overview of the compared similarity measures in Table 1, listing the music representations to which these measures have been originally applied, and which we therefore also use in our comparisons. Moreover, we include information on the research fields from which the measures are taken, the database on which they were evaluated, if applicable, and a bibliographical reference to a relevant paper. We describe the measures in Section 3.

We evaluate the different measures by comparison with human annotations of phrase occurrence, through quanti-

© Berit Janssen, Peter van Kranenburg, Anja Volk.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Berit Janssen, Peter van Kranenburg, Anja Volk. "A comparison of symbolic similarity measures for finding occurrences of melodic segments", 16th International Society for Music Information Retrieval Conference, 2015.

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

[2] http://www.esac-data.org

Similarity measure	Music representations	Research field	Dataset	Authors
Correlation distance (CD)	duration weighted pitch sequence	Ethnomusicology	-	[12]
City block distance (CBD)	pitch sequence	Ethnomusicology	-	[14]
Euclidean distance (ED)	pitch sequence	Ethnomusicology	-	[14]
Local alignment (LA)	pitch sequence	Ethnomusicology	MTC	[16]
Structure induction (SIAM)	pitch / onset	MIR	JKUPTD	[7]
Wavelet transform (WT)	duration weighted pitch sequence	MIR	JKUPTD	[17]
B-spline alignment (BSA)	pitch sequence	MIR	EsAC	[15]
I-R structure alignment (IRSA)	pitch, duration, metric weight	MIR	RISM	[4]

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, and which musical data the measures were tested on, if applicable.

fying the retrieval measures precision, recall and F1-score, and the area under the receiver-operating characteristic curve. The evaluation procedure is described in detail in Section 4.

The remainder of this paper is organised as follows: first, we describe our corpus of folk songs and the annotation procedure. Next, we give details on the compared similarity measures, and the methods used to implement the similarity measures. We describe our evaluation procedure before presenting the results, finally discussing the implications of our findings and concluding steps for future work.

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 [5]. MTC-ANN 2.0 contains 360 orally transmitted melodies, which have been transcribed from recordings and digitized in various formats. Various metadata have been added by domain experts, such as the tune family membership of a given melody: the melodies were categorized into groups of variants, or tune families. The variants belonging to a tune family are considered as being descended from the same ancestor melody [1]. 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 [2].

Even though MTC-ANN 2.0 comprises very well documented data, there are some difficulties to overcome when comparing the digitized melodies computationally. Most importantly, the transcription choices between variants can be different: where one melody is 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. This means that notes which are perceptually very similar might be hard to match based on the digitized information. Musical similarity measures might be sensitive to these differences, or they might be transposition or time dilation invariant, i.e. work equally well under different pitch transpositions or meters.

Of these 360 melodies categorized into 26 tune families, we asked three Dutch folk song experts to annotate similarity relationships between phrases within tune families. The annotators judged the similarity of phrases of 213 melodies belonging to 16 tune families, amounting to 1084 phrase annotations in total. The phrases contain, on average, nine notes, with a standard deviation of two notes. The dataset with its numerous annotations is publicly available. [3]

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. See an annotation example in Figure 1.

The three domain experts worked independently on the same data. To investigate the subjectivity of similarity judgements, we measured the agreement between the three annotators' similarity judgements using Fleiss' Kappa, which yielded $\kappa = 0.73$, constituting substantial agreement.

The annotation was organized in this way to guarantee that the task was feasible: judging the occurrences of hundreds of phrases in dozens of melodies (14714 comparisons) would have been much more time consuming than assigning labels to the 1084 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.

We focus on finding almost identical occurrences: if for a given query phrase q in one melody, at least one phrase r with exactly the same label (letter and number) appears in another melody s of the same tune family, we consider it an occurrence of melodic segment q in s. Conversely, if there is no phrase with exactly the same label as q in melody s, this constitutes a non-occurrence.

For all phrases and all melodies, within their respective tune families, we observe whether the annotators agree on occurrence or non-occurrence of query phrases q in melodies s. The agreement for these judgements, 14714 in total, was analyzed with Fleiss' Kappa, with the result $\kappa = 0.51$ denoting moderate agreement. This highlights the ambigu-

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

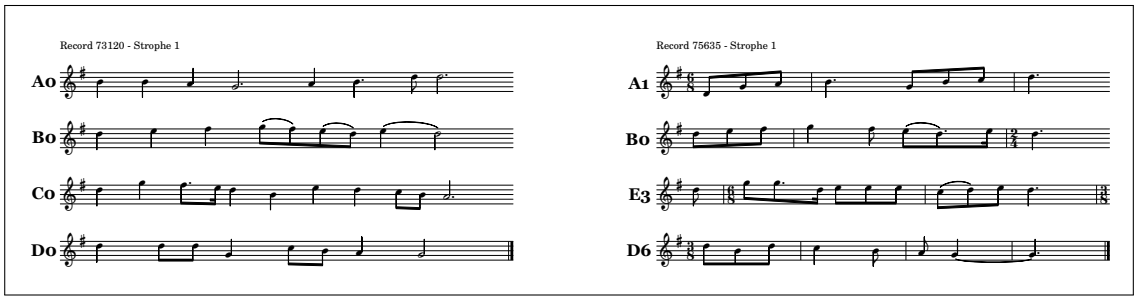

Figure 1. An example for two melodies from the same tune family with annotations.

Annotators	Precision	Recall	F1-score
1 and 2	0.745	0.763	0.754
1 and 3	0.803	0.75	0.776
2 and 3	0.788	0.719	0.752

Table 2. The retrieval scores between annotators. For instance, annotator 2 agrees to 75% with the occurrences detected by annotator 1. The scores are symmetric.

ity involved in finding occurrences of melodic segments.

To compare the annotators' agreement with the performance of the similarity measures in the most meaningful way, we also compute the precision, recall and F1-score of each annotator in reproducing the occurrences detected by another annotator. Table 2 gives an overview of these retrieval scores. A higher retrieval score for a given similarity measure would indicate overfitting to the judgements of one individual annotator.

3. COMPARED SIMILARITY MEASURES

In this section, we present the eight compared similarity measures. We describe the measures in three subgroups: first, measures comparing fixed-length note sequences; second, measures comparing variable-length note sequences; third, measures comparing more abstract representations of the melody.

For our corpus, as melodies are of similar length, we can transpose all melodies to the same key using pitch histogram intersection. For each melody, a pitch histogram is computed with MIDI note numbers as bins, with the count of each note number weighed by its total duration in a melody. The pitch histogram intersection of two histograms h_q and h_r, with shift σ is defined as

$$PHI(h_q, h_r, \sigma) = \sum_{k=1}^{l} \min(h_{q,k+\sigma}, h_{r,k}), \qquad (1)$$

where k denotes the index of the bin, and l the total number of bins. We define a non-existing bin to have value zero. For each tune family, we randomly pick one 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.

Some similarity measures use note duration to increase precision of the comparisons, others discard the note duration, which is an easy 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 quarter note by 16 pitch values, an eighth note by 8 pitch values, and so on. Onsets of small duration units, especially triplets, may fall between these sampling points, which shifts their onset slightly in the representation. Besides, a few similarity measures require music representation as *onset, pitch* pairs, or additional information on metric weight.

3.1 Similarity Measures Comparing Fixed-Length Note Sequences

To formalize the following three measures, we refer to two melodic segments q and r of length n, which have elements q_i and r_i. The measures described in this section are distance measures, such that lower values of $dist(q, r)$ 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 segment which is most similar to the query segment is retained as a match. The implementation of the fixed-length similarity measures in Python is available online. [4] It uses the *spatial.distance* library of scipy [10].

Scherrer and Scherrer [12] suggest correlation distance to compare folk song melodies, represented as duration weighed 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.

$$dist(q, r) = 1 - \frac{\sum_{i=1}^{n}(q_i - \bar{q}) \cdot \sum_{i=1}^{n}(r_i - \bar{r})}{\sqrt{\sum_{i=1}^{n}(q_i - \bar{q}) \cdot \sum_{i=1}^{n}(r_i - \bar{r})}} \qquad (2)$$

Steinbeck [14] proposes two similarity metrics for the classification of folk song melodies: city-block distance and Euclidean distance (p.251f.). He suggests to compare pitch sequences, next to various other features of melodies such as their range, or the number of notes in a melody. As we are interested in finding occurrences of segments

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

rather than comparing whole melodies, we analyze pitch sequences.

City-block distance and Euclidean distance are not transposition invariant, but as they are applied to pitch sequences, they are time dilation invariant. All the fixed-length measures in this section will be influenced by small variations affecting the number of notes in a melodic segment, such as 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 formalize the following three 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 lower values of $sim(q, s)$ indicate higher similarity. The implementation of these methods in Python is available online. [4]

Mongeau and Sankoff [8] suggest the use of alignment methods for measuring music similarity, and they have been proven to work well for folk songs [16]. We apply local alignment [13], which returns the similarity of a segment within a melody which matches the query best.

To compute the optimal local alignment, a matrix $A(i, j)$ is recursively filled according to equation 3. The matrix is initialized 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.

$$A(i,j) = \max \begin{cases} A(i-1, j-1) + subs(q_i, s_j) \\ A(i, j-1) + W_{insertion} \\ A(i-1, j) + W_{deletion} \\ 0 \end{cases} \quad (3)$$

We apply local alignment to pitch sequences. In this representation, local alignment is not transposition invariant, but 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

$$subs(q_i, s_j) = \begin{cases} 1 & \text{if} \quad q_i = s_j \\ -1 & \text{otherwise} \end{cases} . \quad (4)$$

The local alignment score is the maximum value in the alignment matrix, normalized by the number of notes n in the query segment.

$$sim(q, s) = \frac{1}{n} \max_{i,j}(A(i, j)) \quad (5)$$

Structure Induction Algorithms [7] formalize a melody as a set of points in a space defined by note onset and pitch, and perform well for musical pattern discovery [6]. 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.

$$\mathbf{T} = \begin{pmatrix} q_{i,onset} \\ q_{i,pitch} \end{pmatrix} - \begin{pmatrix} s_{j,onset} \\ s_{j,pitch} \end{pmatrix} \quad (6)$$

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 $\mathbf{T}$.

$$MTP(q, s, \mathbf{T}) = \{q_i | q_i \in q \wedge q_i + \mathbf{T} \in s\} \quad (7)$$

We analyze the pattern matching method SIAM, defining the similarity of two melodies as the length of the longest maximally translatable pattern, normalized by the length n of the query melody:

$$sim(q, s) = \frac{1}{n} \max_{\mathbf{T}} |MTP(q, s, \mathbf{T})| \quad (8)$$

3.3 Similarity Measures Comparing Abstract Representations

The following three methods transform the melodic contour into a more abstract representation prior to comparison.

Velarde et al. [18] use wavelet coefficients to compare melodies: melodic segments are transformed with the Haar wavelet. 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 [17]. We use the authors' Matlab implementation to compute wavelet coefficients of duration weighed pitch sequences, and compute city-block distance between the coefficients of query segment and match candidates.

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

Urbano et al. [15] transform note trigrams to a series of B-spline interpolations, which are curves fitted to the contours of the note trigrams. The resulting series of B-splines of two melodies are then compared through alignment. Different B-spline alignment approaches have performed well in various editions of MIREX for symbolic melodic similarity. [5]

We apply the ULMS2-ShapeL algorithm, [6] using the most recent version, different from its original publication [15]. This algorithm discards the durations of the notes and returns the local alignment score of query segments and melodies. The score is normalized by the length

[5] http://www.music-ir.org/mirex/wiki/2012:Symbolic_Melodic_Similarity_Results

[6] https://github.com/julian-urbano/MelodyShape

n of the query segment. This similarity measure is of variable length, sensitive to time dilation, but transposition invariant.

Grachten's method [4] relies on Implication-Realization (IR) structures, as introduced by Narmour [9] as basic units of melodic expectation. Grachten et al. transform melodies into IR structures using a specially developed parser. The similarity of melodies is then determined based on the alignment of the IR structures. This method was successful in the MIREX challenge for symbolic melodic similarity of 2005. [7]

In preparation of IR-structure alignment, we use Grachten's [4] IR-parser, which takes the onset, pitch, duration and metric weight of a melody and infers the corresponding IR structures. To this end, we exclude all melodies which do not have annotated meter ($n = 65$), needed for the computation of metric weight, from the corpus. We align the IR-structures with the same insertion and deletion weights and the same substitution function as Grachten's publication, but as we are interested in finding occurrences, we use local alignment rather than the original global alignment approach. Through the transformation of the note sequences to IR-structure sequences, this similarity measure is transposition invariant, but it is sensitive to time dilation and ornamentation, which might affect the detected IR-structures.

4. EVALUATION

We evaluate the potential success of a similarity measure through comparing the retrieved occurrences to the annotators' judgements, separately for each annotator. Different thresholds on the similarity measures determine which matches are accepted as occurrences, or rejected as non-occurrences. For the distance measures (CD, CBD, ED, WT), matches with similarity values below the threshold, for the other measures, matches with similarity values above the threshold are considered occurrences.

The relationship between true positives and false positives for each measure is summarized in a receiver-operating characteristic (ROC) curve with the threshold as parameter. The area under the ROC curve (AUC) determines whether a similarity measure overall performs better than another, for which we calculate confidence intervals and statistical significance using DeLong's method for paired ROC curves, based on U statistics [3, 11]. Furthermore, we report the maximally achievable retrieval measures precision, recall and F1-score with relation to the ground truth.

5. RESULTS

We have analyzed the results with respect to all annotators, resulting in the same ranking of the similarity measures. Due to space constraints, we report and discuss our results in relation to annotator 1. We show the ROC curves of the eight different measures in Figure 2, which display the true positive rate against the false positive rate at different thresholds. The more of the higher left area a ROC curve covers in a graph, the better; this indicates that the two classes are better separable.

[7] http://www.music-ir.org/mirex/wiki/2005:Symbolic_Melodic_Similarity_Results

Measure	F1-score	Precision	Recall	AUC
Baseline	0.68	0.52	1	n/a
CD	0.68	0.51	1	0.549
CBD	0.68	0.51	1	0.574
ED	0.68	0.51	1	0.568
LA	0.73	0.7	0.78	0.790
SIAM	0.73	0.75	0.71	0.787
WT	0.69	0.57	0.87	0.732
BSA	0.72	0.65	0.81	0.776
IRSA	0.69	0.54	0.95	0.683

Table 3. Results of the compared similarity measures for different music representations: the maximal F1-score, the associated precision and recall, and the area under the ROC curve (AUC).

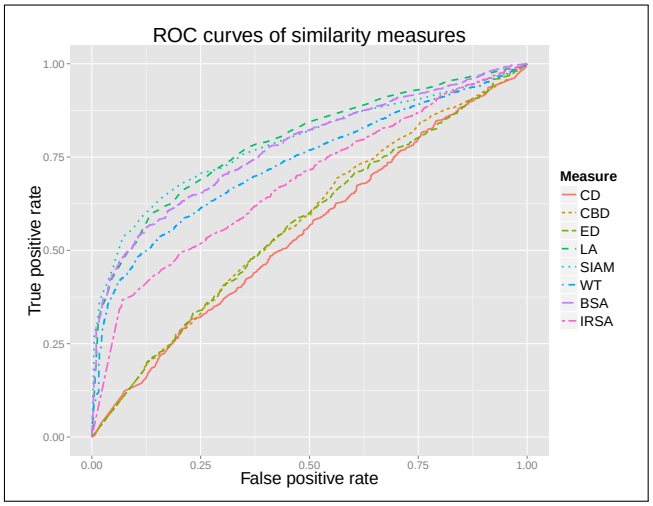

Figure 2. The ROC curves for the various similarity measures, showing the increase of false positive rate against the increase of the true positive rate, as a parameter of the threshold.

From Figure 2 it can be seen that the similarity measures suggested in computational ethnomusicology (CD, CBD, ED) perform only marginally above chance. IR-structure alignment and wavelet transform obtain better results, and B-spline alignment, local alignment and SIAM perform best.

We summarize the area under the ROC curve (AUC), the maximally achieved F1-score, as well as the associated precision and recall in Table 3. We include a baseline in this table which assumes that every compared melody contains an occurrence of the query segment, which leads to perfect recall, but poor precision, as the chance for a segment to occur in a given melody are only about 50%.

We compare the AUC values of the different measures in Figure 3, showing confidence intervals and significance of the pairwise differences between adjoining measures, indicated by stars ($*p < .5$, $**p < .01$, $***p < .001$).

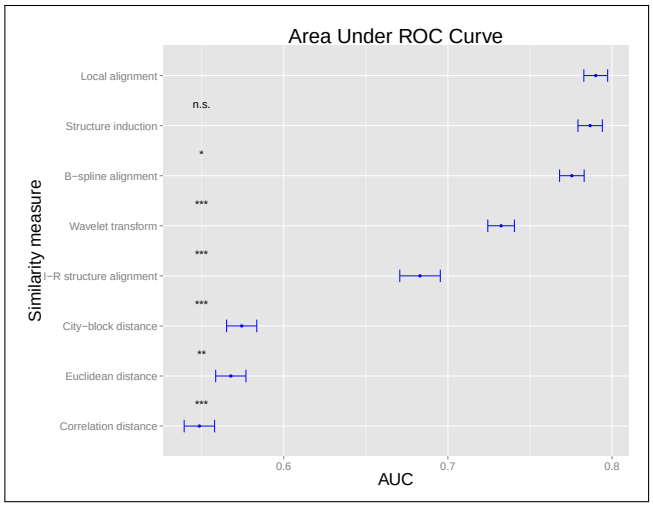

Figure 3. The area under the ROC curve of all similarity measures, ordered by the most successful to the least successful methods. The error bars indicate the confidence intervals, and significant difference between adjoining measures is indicated by stars (*p< .5, **p< .01, ***$p < .001$).

6. DISCUSSION

Our results indicate that the distance measures (CD, CBD, ED) do not work very well, which contradicts the intuitions of the computational ethnomusicologists who propose them. This suggests that variations on pitch height and contour, which mostly affect these measures, are not the most informative aspect for human judgements on musical similarity. Embellishments of a note sequence through extra notes, for instance to accommodate slightly varied lyrics, on the other hand, would cause considerable decrease of measured similarity, while they will be perceived as minor variation, if at all by human listeners.

Measures from symbolic melodic similarity (BSA, IRSA) and pattern discovery (WT) perform better overall. Among these, I-R structure alignment performs least well. This performance might be improved by optimising the alignment scores for our dataset; the alignment weights were trained on RISM incipits and might therefore not fit the folk songs optimally.

Wavelet transform seems to capture some essential notions of music similarity for finding correct occurrences, showing that essentially the same technique - fixed-length comparison with city-block distance - can be much more successful if it is applied to a different abstraction level than pitch sequences. Possibly a variable-length comparison step would yield even better results.

As expected from its success in symbolic melodic similarity MIREX tracks, B-spline alignment successfully retrieves a large portion of relevant occurrences annotated by human experts. However, it does not perform as well as some of the other measures in our comparison.

Confirming earlier research on melodic similarity in folk songs, alignment performs well in our task. We show that local alignment is very successful in correctly identifying occurrences, even with a very simple substitution score, which only rewards equal pitches. Even better results might be achieved with different weights and substitution scores.

SIAM, to our knowledge, has not been evaluated for detecting phrase occurrences in folk song melodies yet, but performs on the same level as local alignment. This implies that SIAM is a good candidate for finding occurrences of melodic segments successfully, especially in corpora where transposition differences cannot be resolved through pitch histogram intersection, for instance in classical music and jazz, where key changes might make the estimation of transposition more difficult.

With maximal F1-scores of 0.73, the results of local alignment and SIAM come close to the between-annotator F1-scores between 0.75 and 0.78. This shows that we cannot do much better for our problem on this dataset without overfitting.

7. CONCLUSION

We conclude that both local alignment and SIAM seem adequate methods for finding occurrences of melodic segments in folk songs. Based on the retrieval scores, they find almost the same amount of relevant occurrences as human annotators among each other.

The measures investigated in this paper were applied to specific music representations. A wider range of music representations will be compared in future work. Moreover, the results will need to be analyzed in more detail with special attention to the cases where the similarity measures err, i.e. are false positives and false negatives more frequent for a specific tune family? And if so, do the annotators also disagree most on these same tune families? Besides, it is important to investigate the true positives as well, and ascertain that they are found in the correct positions in a melody.

The similarity measures compared in this article can be applied to other music corpora, which will give even deeper insights into relationships between melodies based on melodic segments that are shared between them. We can learn much about melodic identity and music similarity from both the confirmation and refutation of our findings in other music genres.

8. ACKNOWLEDGEMENTS

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 (276-35-001). We thank Gissel Velarde, Maarten Grachten and Julián Urbano for kindly providing their code and helpful comments, Sanneke van der Ouw, Jorn Janssen and Ellen van der Grijn for their annotations, and the anonymous reviewers for their detailed suggestions.

9. REFERENCES

[1] Samuel P. Bayard. Prolegomena to a Study of the Principal Melodic Families of British-American Folk Song. *The Journal of American Folklore*, 63(247):1–44, 1950.

[2] Michael Scott Cuthbert and Christopher Ariza. music21 : A Toolkit for Computer-Aided Musicology and Symbolic Music Data. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, number Ismir, pages 637–642, 2010.

[3] Elizabeth R. Delong, David M. Delong, and Daniel L. Clarke-Pearson. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44(3):837–845, 1988.

[4] Maarten Grachten, Josep Lluís Arcos, and Ramon López de Mántaras. Melody Retrieval using the Implication / Realization Model. In *MIREX-ISMIR 2005: 6th International Conference on Music Information retrieval*, 2005.

[5] Peter Van Kranenburg, Martine De Bruin, Louis P Grijp, and Frans Wiering. The Meertens Tune Collections. Technical report, Meertens Online Reports, Amsterdam, 2014.

[6] David Meredith. COSIATEC and SIATECCompress: Pattern Discovery by Geometric Compression. In *Music Information Retrieval Evaluation eXchange*, 2014.

[7] David Meredith, Kjell Lemström, and Geraint A. Wiggins. Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4):321–345, 2002.

[8] Marcel Mongeau and David Sankoff. Comparison of Musical Sequences. *Computers and the Humanities*, 24:161–175, 1990.

[9] Eugene Narmour. *The Analysis and Cognition of Basic Melodic Structures. The Implication-Realization Model.* University of Chicago Press, Chicago, 1990.

[10] Travis E. Oliphant. Python for Scientific Computing. *Computing in Science and Engineering*, 9(3):10–20, 2007.

[11] Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-charles Sanchez, and Markus Müller. pROC : an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinformatics*, 12(1):77, 2011.

[12] Deborah K. Scherrer and Philip H. Scherrer. An Experiment in the Computer Measurement of Melodic Variation in Folksong. *The Journal of American Folklore*, 84(332):230–241, 1971.

[13] T.F. Smith and M.S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195 – 197, 1981.

[14] Wolfram Steinbeck. *Struktur und Ähnlichkeit. Methoden automatisierter Melodienanalyse*. Bärenreiter, Kassel, 1982.

[15] Julián Urbano, Juan Lloréns, Jorge Morato, and Sonia Sánchez-Cuadrado. MIREX 2012 Symbolic Melodic Similarity: Hybrid Sequence Alignment with Geometric Representations. In *Music Information Retrieval Evaluation eXchange*, pages 3–6, 2012.

[16] Peter van Kranenburg, Anja Volk, and Frans Wiering. A Comparison between Global and Local Features for Computational Classification of Folk Song Melodies. *Journal of New Music Research*, 42(1):1–18, 2012.

[17] Gissel Velarde and David Meredith. A Wavelet-Based Approach to the Discovery of Themes and Sections in Monophonic Melodies. In *Music Information Retrieval Evaluation eXchange*, 2014.

[18] Gissel Velarde, Tillman Weyde, and David Meredith. An approach to melodic segmentation and classification based on filtering with the Haar-wavelet. *Journal of New Music Research*, 42(4):325–345, December 2013.