Table of Contents

Viel Kleinkram
überarbeiten

+ ~~Deckblatt?~~

Symbolic Melodic Music Similarity

Ali Bektas and Paul Kröger
Advisor : Prof. Dr. Ulf Leser
Similarity Search
WS19-20

Humboldt Universität zu Berlin

15 April 2020

there does not exist

Woher?

○

[redacted]

[redacted]

Abstract. With an increasing need to compare [redacted] musical data in terms of their content the question of similarity between musical objects becomes a challenge. ~~While~~ is not a single method to ~~answer~~ all needs , various approaches have been introduced to the field in recent decades. ~~In~~ [redacted] , we introduce approaches which focus on symbolic melodic data , ~~that~~ is , a representation of the melody of a piece based on an alphabet of symbols. As the leading algorithms in this field , we examine J.Urbano's et al. algorithm which compares the shape of melodies and an algorithm by N.Orio et al. which constructs a tree out of melodies and uses the shortest path between them to determine similarity. Furthermore we will show how the MIREX competition handled the lack of an existing definition for similarity between pieces of music.

this was

V

1 Introduction

?
○

Similarity measures lie at the heart of Information Retrieval. This is also the case for [redacted] music similarity. For a subscriber of a music streaming application like Spotify , Last.fm etc. to be recommended latest albums or a [redacted] to find related ~~documents~~ in a database , determining similarity ~~between pieces of music~~ is crucial. [redacted] , Symbolic Music Similarity algorithms ~~use~~ symbols that represent the data as input , where in Audio Music Similarity one has pitch-time related values. ~~Within~~ Symbolic Music Similarity ~~we find~~ algorithms that deal with polyphonic data and ~~those~~ that deal with monophonic data. ~~In this~~ paper we will look at the field of Symbolic Melodic Music Similarity. ~~which only focuses~~ on monophonic data.

+Ref

[redacted]

[redacted]

[redacted]

Since we are mainly concerned with Western music , we can restrict ourselves to 12 tones in an octave , other cultures however may use more or less notes. In Middle Eastern countries , for instance . [redacted] there are 9 tones between two notes of a whole tone interval.

~~In our paper~~ we will follow the classification used by Velardo et al. [3]. According to Velardo et al. a ~~melodic~~ music similarity algorithm belongs to either of the four classes (1) Music Theory . (2) Cognition . (3) Mathematics or (4) Hybrid. Hybrid algorithms are usually formed by taking a linear combination of different similarity measures.

this was

[redacted]

[redacted]

+ 1 —

[redacted]

Vein Bland

V. commis

vgl

aber danach

~~In contrast, the~~

We will introduce two algorithms & ~~to~~ showcase how symbolic melodic data can be represented and compared ~~. The Systems~~ proposed by J. Urbano et al. [8] forms spline-sequence for a query melody and compares it with other spline-sequences using a sequence alignment method to determine similarity between pieces.

2 ~~bullets~~

~~We will also present an~~ approach by N.Orio et al. [4]. ~~It~~ incorporates routines to generalize the melody ~~, that heavily~~ depend on Music Theory, into simpler melodies while forming a tree structure to afterwards find the shortest path between given melodies within this tree.

? ~~Satz machen~~

After this we will introduce MIREX , the Music Information Retrieval Evaluation eXchange , that ~~arranges~~ annual competitions for researchers to test their algorithms against one another. We will explain how the subjectivity of music similarity ~~is in~~ these competitions reduced to a minimum by having experts establish a Ground Truth. We will explain how this process takes place based on the example of the 2005 competition. The resulting data , against which the result of an algorithms is to be compared , manifest ~~occasionally~~ a ~~non-total~~ order which makes it more difficult for precision and recall measures to come to a meaningful result. We will explain how this problem is solved with a novel measure called Average Dynamic Recall.

~~ing~~

~~against~~

~~Was~~?

Our main emphasis is that the field suffers from the subjectivity of music similarity with no agreed upon definition existing of what constitutes similarity between pieces of music. Many algorithms lack empirical evaluation and with no Ground Truths existing, the data that does exist is very hard to compare.

In die Conclusion schieben

~~is~~

2 Algorithms

2.1 A Graph Based Approach

first

transform

Get

Orio et al. [4] introduce ~~in their paper~~ a series of operations ~~to reduce~~ melodies into a single large tree. Melodies are segmented and then these segments are added ~~into the~~ tree as terminal nodes. The intermediate nodes represent generalization of the segments. In a single step of generalization a segment is transformed into a simpler segment by deleting less important notes in the given segment. Which notes are less important is decided by three weight coefficients : (1) its underlying harmonic function. (2) its metric position and (3) the interval between the tone and the root of the underlying chord, that is a harmonic grouping of notes that are played together forming a unified musical structure.

Warum andere Bücher (dep?)

Sätze machen

In order to determine the harmonic function of a note, harmonic analysis must be applied. Harmonic analysis is the process of making statements about in which way notes of a sequence are related to each other. These three coefficents must then be determined in such a way that it reflects the priorities of the human ear when considering two songs to be similar. ~~In their paper~~ Orio et al. state that they chose to manually annotate the functions of notes in order to prevent any problems that could have otherwise arisen during automated annotation , which could cause wrong simplification steps.

~~leaves~~

Which (x) von welchen Songs?

thus

+2

Chapter 4+5 auch ankündigen

Da fehlt was. Wie kommt man von einem Tree zum andern? Der Match fehlt

The distance between two documents is expressed as the median of shortest distance between all segments. The similarity function $s(c_i, q)$ between a document c_i from a collection of N documents and a query document q is calculated as :

$$s(c_i, q) = 1 + \frac{d(c_i, q)}{\sum_{j=1}^{N} \frac{d(c_i, c_j)}{N-1}} \tag{1}$$

The similarity function is normalized and authors mention that 'the normalization factors can be computed off-line to speed up retrieval'. Authors also mention that the tree representation can offer novel ways to view a collection and see visually to what extent two songs were considered to be similar. cf ?

Kurz aber OK

✓ Ref

2.2 Urbano Melody-Shape

In 2010 J.Urbano et al. proposed a ~~new~~ method to calculate similarity between symbolic melodic pieces, comparing the shapes of melodies created by ~~looking at~~ considering notes as points on a pitch-time plane and interpolating a curve through those points, [8]. Based on this ~~method~~ principle, Urbano developed two algorithms for the

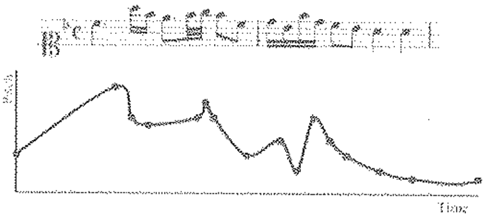


Fig. 1. Notes represented as a curve on a pitch-time plane [2]

MIREX Competition, Shape (in three variants: ShapeL, ShapeG and ShapeH) which only uses the pitch dimension, and Time, which uses both the pitch and the time dimension [2]. Both algorithms separate melodies into sequences of notes, and use Uniform B-Splines to ~~get~~ obtain a spline sequence representation. Two spline sequences are then compared using sequence alignment. Based upon the results of the first Mirex competitions Urbano entered he also developed the ShapeTime system, which collects the n most relevant documents out of the data set, against which the query is run, using the ShapeH system and then ranks those documents using the time system. This was done because the ShapeH system performed better in rank unaware measures, whilst the time system performed better in rank aware measures in the competition. The ShapeTime system usually achieved the best overall results, as we will show ~~later in the evaluation section~~ in chapter XY.

Those next

ShapeH This system ignores the time dimension and only focuses on the shape of a melody. It uses spline-span sequences 3 notes long which results in a polynomial of degree 2 for each spline. ~~They~~ are then differentiated, resulting in polynomials of degree 1. A dynamic programming table is then filled using a global alignment algorithm. The score of a cell (i,j) is computed by :

$$H(i,j) = max \begin{Bmatrix} H(i-1,j-1) + s(a_i,b_j) \\ H(i-1,j) + s(a_i,-) \\ H(i,j-1) + s(-,b_j) \end{Bmatrix}$$

and

where H is the dynamic programming table~~, a and~~ b the compared spline-span sequences. The highest score from the table is ~~then~~ used as the similarity score between the melodies. This ~~hybrid~~ alignment approach is used in favor of just the global alignment algorithm because Urbano argues that listeners put more emphasis on the beginning of a melody rather than the end when determining similarity.
The operations are defined as follows :

Also better Präfix?

prefix-based

- Insertion : $s(-,n) = -(1 - f(n))$.
- Deletion : $s(n,-) = -(1 - f(n))$
- Match : $s(n,n) = 1 - f(n)$

relative?

Where f(n) denotes the frequency of a spline in the spline-span sequence. Urbano argues that the more often a spline occurs in the spline-sequence, the less important it is for the comparison.
The substitution or mismatch score is calculated based on the shape of the spans where two melodies with a similar shape only get a small penalization. There are three possible scenarios :

- If the derivative sign at the start and at the end of the splines are the same, they are considered to have a similar shape and there is only a small penalization.
- If the derivative sign only matches at the start or the end of the splines, they are considered to be less similar and there is a medium penalization.
- If the derivative sign at the start and the end of the spline does not match, they are considered to be not similar and there is a large penalization.

✓

SB

Due to looking at polynomials of degree 2 it is sufficient to only regard the start and end points of the splines, because they can only change their direction once within the span. It should be noted here that Urbano also submitted the ShapeG and ShapeL systems to Mirex competition until 2013, which only differ from the ShapeH system in that they use a strictly global and a strictly local alignment approach, respectively, instead of the hybrid alignment approach. However the ShapeH system on average produced the best results in the Mirex competitions ~~as we will show~~ later. (see Chapter XY) also

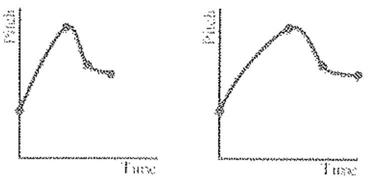


Fig. 2. Normalization of spline [2]

Time This system uses the time dimension ~~as well~~ and calculates the area between splines to determine similarity. It uses spline-span sequences 4 notes long which results in a polynomial of degree 3 for each spline. They are ~~then~~ differentiated, resulting in polynomials of degree 2. Each span duration is normalized to the same length. (Fig 2) It uses the same hybrid alignment approach as the ShapeH system. The operations are defined as follows :

- Insertion : $s(-,n) = -diff_p(n,\phi(n)) - \lambda k_t * diff_t(n,\phi(n))$.
- Deletion : $s(n,-) = -diff_p(n,\phi(n)) - \lambda k_t * diff_t(n,\phi(n))$.
- substitution : $s(n,m) = -diff_p(n,m) - \lambda k_t * diff_t(n,m)$
- Match : $s(n,n) = 2\mu_p + 2\lambda k_t \mu_t = 2\mu_p(1 + k_t)$.

φ

Unclear? Das Integral?

$diff_p$ and $diff_t$ are defined as the area of the compared first derivatives from the splines pitch and time function. $\phi(n)$ describes the ~~area~~ between n and the x-axis. μ_p, μ_t as well as k_t are weighting constants where μ_p and μ_t being the mean scores returned by $diff_p$ and $diff_t$ respectively, and k_t weighing time dissimilarity corresponding to pitch dissimilarity. $\lambda = \mu_p/\mu_t$ is a constant used to normalize time dissimilarity scores with respect to pitch dissimilarity scores. 6

distance

3 MIREX

in organises

MIREX is a platform for researchers ~~of~~ Music IR. It ~~arranges~~ annual competition where researchers present their algorithms. The subbranch "Symbolic Melodic Music Similarity" was active during ~~year~~ 2005-2015 and doesn't take place since ~~2016.~~ then

any more

As we mentioned in the Introduction, one of the main problems of Symbolic Melodic Music Similarity is that there is no consensus over a universal measure of similarity. In order to circumvent this issue, MIREX ~~consu~~lts ratings given by experts of the field. The results of the competitors were ~~then~~ compared against this so-called Ground Truth.

3.1 The Ground Truth

The RISM A/II collection which was used as the collection in the competition of 2005 contains 476.000 documents. Experts were asked to rate similarity of ~~docu-~~

melodies

was

Komma nach

melodies

the duration

ments in this collection to given queries. The experts were not asked to evaluate similarity to all documents, since that would take too much time. A series of filtering processes were applied to reduce the number of relevant documents.

?

Among several of them documents were filtered based on:

- The interval between the highest and lowest note within the melody
- The largest interval between subsequent notes.
- The editing distance between the query and the document. In order to find the editing distance, a document is projected onto a string, that contains the alphabet U("up") , D("down") , R("repeat").

the

Interval in music means ~~a degree~~ between two notes , that specifies ~~their~~ relation to one another in terms of the number of half-steps between them ~~and~~ if ~~they~~ are ordered in an ascending order or descending.

between 2 notes

?

Different filtering steps are used based on characteristic features of the documents. In order to prevent relevant documents from being filtered out , a limit of 300 documents was set. To come to a convenient number of documents, residual documents were then manually reduced to a collection of 50 documents.

The relevant documents were ~~then~~ given to the experts to be ordered. Experts were given the freedom to choose which documents were to include at all. The rankings were then grouped together for each document , ordered by their by median rank and then by mean rank. Every document is then compared against higher ranked documents by Wilcoxon rank sum test , that measures the probability of the null-hypothesis , that is , the probability that the ratings of a relatively small group of experts reflect those of a larger group. When there is no compelling evidence that documents actually differ in terms of median ranks , they are ~~grouped together~~. As a consequence of this there is often no total order among ranked documents.

(x) Logic: Grenze which doc within recall

be

ranked equally

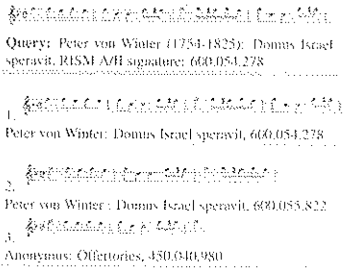

Query: Peter von Winter (1754-1825): Domus Israel speravit, RISM A/II signature: 600.054.278
1.
Peter von Winter: Domus Israel speravit, 600.054.278
2.
Peter von Winter : Domus Israel speravit, 600.055.822
3.
Anonymus: Offertories, 450.040.980

Fig. 3. Ground Truth for Winter: "Domus Israel speravit" [7]

In [7] 31 experts were asked to order relevant documents to the given query in Fig. 4. The second and the third documents in the resulting list differ from the given query in such small ways , that it was hard to conclude a total order among results.

3

w.r.t a query

To emphasize the group boundaries better , Typke, Veltkamp. Wiering [5] introduces a new measure called Average Dynamic Recall.

3.2 Average Dynamic Recall

The Authors mention nine criteria that they considered when they introduced ADR ~~.~~ among which ~~we want to list the following.~~: cer

- The measure doesn't need the ground truth to be completely ordered
- Violations of the correct order should be punished if they happen across group boundaries.

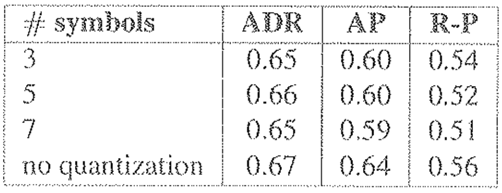
In comparison to standard measures such as recall and precision , ADR is specifically tailored for partially ordered lists. The ADR is calculated as the sum of recall ~~success~~ over number of steps . The relevant documents are at the beginning the documents within the first group. When this group of elements is completely retrieved , the number of relevant documents will get larger by adding the next group about which it is known that there is evidence that documents do differ in terms of median ranks. In each step the ~~success~~ of recall is calculated as the ratio of found currently relevant documents over the number of currently relevant documents.

Was ist das?

4 Evaluation

4.1 The graph based algorithm of Orio et al.

For the evaluation authors used the RISM A/II collection with Ground Truth and queries from MIREX 2005.



# symbols	ADR	AP	R-P
3	0.65	0.60	0.54
5	0.66	0.60	0.52
7	0.65	0.59	0.51
no quantization	0.67	0.64	0.56

+Tab

Alle Spalten erklären

Fig. 1. Manual segmentation using projections with ranges of different sizes [4]

Wo?

Authors experimented with the input data to see what ~~kind of an~~ impact it would have ~~in the success of the algorithm~~ if the melodies were to be projected on to simpler melodies , which still reflect basic elements of the starting melody. Figure 1 shows the averaged results of the algorithm when using quantization with alphabets of various sizes. Quantization can be seen as a function that

shown in Fig

is

strange

projects a given melody onto a sequence of symbols. As an example to a quantization with 3 symbols ~~can~~ be a function that projects a melody onto the alphabet {'up','down','repeat'}. We observe that reducing a melody into a simpler melody , results in loss of result quality , though often not significant enough for success of retrieval. Where memory size plays a significant role , quantization can be a preferable process.

weighting scheme	ADR	AP	R-P
$3H, MH, 3M$	0.67	0.64	0.56
$4H, MH, 3M$	0.65	0.63	0.56
$7H, MH, 3M$	0.65	0.64	0.56
$3H, MH, 4M$	0.67	0.64	0.55
$3H, MH, 7M$	0.61	0.60	0.51
$3H, MS, 3M$	0.66	0.63	0.52

Fig. 5. Different segmentation techniques with no quantization [4]

for

Figure 2 shows use of different weight measures. As mentioned in section 2.1 weight measures are important ~~when~~ choosing the more important note among other notes in a measure. For harmonic weight the authors experimented with different alphabets size 3,4 and 7. An example to an alphabet of size 3 would consist in grouping the first , fourth and sixth degrees to a group, grouping second and seventh to another and rest to the third group. How these groupings are formed is based on Music Theory. The projection of a single note in the starting melody is based upon its harmonic meaning within the musical context. While reducing the different harmonic functionalities into one larger group, it is important to make assumptions, that ~~result~~ in a grouping with the least amount of information loss. With the same consideration, melodic weights have been grouped together in terms of their intervallic function. As to metric weight, there are two different schemes ~~that are proposed~~ : (1) A simple subdivision in terms of strong and weak beats and (2) *"a hierarchical organization depending on the ~~position in the measure~~"*. It can be observed that, even though the differences are not significant ~~enough~~, the better results are obtained when weighting is based on more generalizing schemes.

✓

A

The generalization processes that graph based algorithm ~~in Section 2~~.1 use are highly dependent upon annotations of harmonic functions of melodic data , which are rarely included in symbolic melodic data. Authors see this as a drawback since a false assumption of a harmonic function of a note in input could easily result in a false segmentation of a piece within. The tree that is formed as musical data are added to it has a tendency to grow ~~in~~ sublinear fashion ~~as argues by the authors.~~ We see this sublinear tendency as an advantage for memory consumption.

?

Sagt wer? Wie genau?

lead to

?

Höhe oder Größe?

Jedes Segment ist [illegible]

Selbstständigkeitserklärung

April 15, 2020

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und nur unter Zuhilfenahme der angegebenen Quellen und Hilfsmittel verfasst habe.

Name	Datum	Unterschrift
Ali Bektas	15.04.2020	