TOWARDS STRUCTURAL ALIGNMENT OF FOLK SONGS

Jörg Garbers and Frans Wiering
Utrecht University
Department of Information and Computing Sciences
{garbers,frans.wiering}@cs.uu.nl

ABSTRACT

We describe an alignment-based similarity framework for folk song variation research. The framework makes use of phrase and meter information encoded in Humdrum scores. Local similarity measures are used to compute match scores, which are combined with gap scores to form increasingly larger alignments and higher-level similarity values. We discuss the effects of some similarity measures on the alignment of four groups of melodies that are variants of each other.

1 INTRODUCTION

In the process of oral transmission folk songs are reshaped in many different variants. Given a collection of tunes, recorded in a particular region or environment, folk song researchers try to reconstruct the genetic relation between folk songs. For this they study historical and musical relations of tunes to other tunes and to already established folk song prototypes.

It has often been claimed that their work could benefit from support by music information retrieval (MIR) similarity and alignment methods and systems. In practice however it turns out that existing systems do not work well enough out of the box [5]. Therefore the research context must be analyzed and existing methods must be adapted and non-trivially combined to deliver satisfying results.

1.1 Similarity and alignment

Similarity and alignment can be considered two sides of the same coin. In order to produce an automatic alignment we need a measure for the relatedness of musical units. Conversely, in order to compute the (local) similarity between two melodies we must know which parts of the melody should be compared.

Alignment can also be a prerequisite for higher-level similarity measures. In a previous paper we derived generalized queries from a group of folk song variants [3]. For a given group of musically related query melodies aligned by the user, we were able to retrieve melodies from a database that are good candidate members for this group.

Making a manual alignment is time-consuming and involves edit decisions, e.g. 'shall one *insert* a rest in one melody or *delete* a note in the other?'. When looking for good additional group members in a database, one should allow both options. However, keeping track of all options quickly becomes impracticable. In this paper we therefore look into automatic alignment of corresponding score positions and ways of controlling the alignment with basic similarity measures.

1.2 Overview and related work

In this paper we first discuss why automatic detection of (genetically related) folk song variants is very demanding and is a major research topic in its own. Next, to support research into similarity measures based on musically meaningful transformations, we develop a framework that helps to model the influence of local similarity measures on variation detection and alignment. Starting from the information encoded in our folk song collection, we motivate the use of available structural and metrical information within alignment directed similarity measures. Finally we compare automatically derived alignments with alignments annotated by an expert.

Generally, we follow a similar approach to Mongeau and Sankoff's [6], who tackled selected transformational aspects in a generic way. They set up a framework to handle pitch contour and rhythm in relation to an alignment-based dissimilarity (or quality) measure. They based their framework assumptions on musical

common sense and based their model parameter estimations on the discussion of examples.

We agree with them that musical time (as representated in common music notation) is a fundamental dimension in variant similarity. This distinguishes our approach from ones that deal with performance-oriented timing deviation problems. A shortcoming of Mongeau and Sankoff's global alignment algorithm is that it cannot handle aspects of musical form, such as repeats or reordering of parts. They are also criticized for sometimes ignoring barlines [2].

Contribution: In this paper we present an approach to tackle these shortcomings by proposing a phrase- and meter-based alignment framework. By using simple and out-of-the-box local similarity measures within the framework and studying the resulting alignments we prove that the framework is useful.

2 MUSICOLOGICAL MOTIVATION

For folk song researchers, an important question is to identify which songs or melodies are genetically related. A way for them to tackle this question is to order melodies into groups that share relevant musical features. We assume that in the process of incrementally building those melody groups, researchers construct a sort of mental alignment of related parts of the candidate melodies. The detection of relevant relationships between variants is affected by the perceived similarity of the material, the knowledge of common transformation phenomena and the discriminative power of the shared features with respect to the investigated corpus.

Researchers have identified transformations that range from a single note change to changes of global features like mood. A mood change can for example affect the tonality (major/minor), the number of notes (due to liveliness) and ambitus (due to excitation) [10].

2.1 Modeling musical similarity

To support folk song variation research, one could choose to model expert reasoning using rule systems. Such a system would consist of transformation rules and transformation sequences that taken together model the relation between melodic variants. A fundamental problem with this approach is that we are still a long way from sufficiently understanding music perception and cognition. Therefore it is impossible to fully formalize the necessary expert knowledge and model the rules of musicological discourse.

Also, it is difficult to find out which approaches to folk song variation are the most promising, because there is little scholarly documentation about music transformations. An exception is Wiora's catalog of transformation phenomena [10]. It describes musical features and related transformations and is a good inspirational source for models of similarity and human reasoning about it. But it lacks descriptions of the contexts in which certain transformations are permissible. It also does not provide the means to estimate the plausibility of one transformation chain as compared to another when explaining a certain variation. It is common in folk song research to reason about the relatedness of *specific songs* rather than to provide comprehensive models. However, what is needed for MIR is a comprehensive theory about this kind of reasoning.

We chose a different approach to model variant relationships, namely to investigate local similarity. This closely follows musicological practice. Often some *striking similarities* between some parts are considered sufficient evidence for a variant relationship. This seems to leave MIR with the task to design local similarity measures for variation research. However we also need to model what is *striking* and what are *relevant parts* and we must find ways to turn local similarity values into overall relevance estimates.

2.2 Structural, textual and metrical alignment

In this section we mention some relations between the parts of a song and between parts of variant songs that support or inhibit the detection of variants. By doing so, we motivate the structure-based alignment framework that we describe in the next sections. Our assumptions stem from general experience with folk songs and from dealing with a collection of Dutch folk songs that we currently help making searchable. [9]

Strophes: In songs, music is tightly bound to the lyrics. When two different songs have similar lyrics, we can often simply use textual similarity to relate musical parts with high confidence. However, we cannot rely on text alone: if textually different variants or different strophes are encoded, we need to make use of musical similarities. This should be unproblematic in principle, since different strophes of a song typically

share the same basic melody.

Phrases: Musical phrases in a folk song are typically related to the verse lines of a poem. When dealing with automatic alignment we must be aware that the level of phrase indication might differ: one phrase in song A might match with two shorter phrases in song B, or the same rhythm can be notated with different note values.

Accents: Word accents in a verse typically follow a common scheme, such as dactyl and trochee. These accents often correspond to the metrical accents of the notes, which can be found by looking at the barlines and measure signatures. We cannot always assume the same meter across variants, since different durations can turn a 2-foot verse scheme into a 3-accent melody. Also, extra syllables in the lyrics may make it necessary to insert notes.

Accent (beat position) and phrase position may also be important for musical similarity of variants: in our previous work we found that pitches on stronger beat positions tend to be more stable across variants than less accented notes. There also seems to be a higher agreement between variants at the beginning and end of a strophe or a phrase (cadence notes), while inner variation is higher [4]. We will study these claims further in future research.

3 A STRUCTURE-BASED ALIGNMENT FRAMEWORK

In this section we describe three components for structure-based folk song alignment: hierarchical segmentation, phrase-level alignment and strophe alignment. The general idea of the proposed framework is to use alignments and alignment scores of smaller musical parts in the alignment process on higher levels. The more part A from the first melody resembles part B from the second melody according to a similarity measure, the more preferable is it to align A with B rather than with another part of the second melody with a lower similarity. However, constraints in higher level alignments can overrule this preference.

3.1 Hierarchical song structure

In accordance with the analysis given in the previous section, we split songs hierarchically into smaller parts. We work on manually encoded musical scores in which meter information and phrases that correspond to the lyrics are indicated. For each song, at least one strophe is encoded. For some songs two or more strophes are separately encoded. We convert our encodings to Humdrum **kern format with the usual = bar line markers and special `!!new phrase` comments.

We use the *Humextra* [7] toolkit to access the relevant information in *Humdrum* `**kern` files (one file per *strophe*). In our data model, each strophe contains a sequence of *phrases*. We currently do not deal with repeat structures, since they are not encoded in our data, but written out. Each phrase is split into *bar segments* using the given bar lines.

A bar in turn is recursively split into smaller *beat segments* delimited by metrically strong beat positions. These are not encoded and need not coincide with note events, but are inferred from the bar lines and the measure signature.

To each structural unit we attach both the corresponding Humdrum fragment, which can be used to retrieve notes and extra information such as the syllables, and a beat range, which identifies the start and duration of the segment measured in beats.

When retrieving the notes for a structural unit, special care is taken to handle boundaries: incomplete bars can occur not only at the beginning or end of a strophe but also at the beginning or end of a phrase. A bar segment will only return those notes of the bar that are part of the parent phrase segment, and likewise for a beat segment.

3.2 Phrase level alignment

The user of the framework chooses an elementary similarity measure sim that is defined on either bar segments or beat segments. The framework computes a similarity value for any pair of segments (one segment from melody A, the second from melody B).

To combine segments of two phrases into a **phrase level** alignment, we use standard global string alignment techniques with match scores and deletion and insertion related gap scores [11]. We define the match score of two segments a and b as being equal to $sim(a, b)$. The scaling of gap scores with respect to sim is left to the user.

To cope with the common phenomenon of different durations of the upbeat bar and the last bar in the same phrase in two variants, we support different gap scores for *inner* and (left/right) *outer* gaps. Also, we could

use local instead of global alignment methods to look for motivic similarity instead of phrase similarity [6].

Future improvements will support *augmentation*, *diminution*, *fragmentation* and *consolidation* as described in [6] in combination with segment constraints. We will also look into inner-phrase form deviations, such repeats of a bar or beat segment. (See section 5.)

3.3 Strophe alignment

For the **strophe level** alignment the framework employs phrase alignments: from the alignment scores similarity values are calculated for all possible pairs of phrases (one phrase from melody 1, one from melody 2). Different phrase-level similarity values from alternative elementary similarity measures and from non-alignment similarity measures (e.g. on cadence tones) can be consolidated into one similarity value. Then a string alignment technique can be used again to find the best alignments of the phrase sequences based on these similarity values. This handles transformations from ABCD to ABD.

To assume sequential similarity and to use phrases only once on the strophe level would sometimes be misleading. Consider simple transformations from AABB to AAB or BBAA. Therefore the framework supports the creation of alignments where one strophe is fixed and each phrase p of it can be matched against any of the phrases q of the phrase set S variant strophe: $MatchScore(p, S) = max_{q \epsilon S}\{similarity(p, q)\}$

To cover cases where strophes of one song differ significantly, the framework simply performs all pairwise comparisons between all strophes from one variant with all strophes from the other variant.

4 EVALUATION

To study the usefulness of our framework, we compared alignments produced by framework-based models with manual alignments (annotations). One of the authors selected sets of similar phrases from a variant group and produced for each set a multiple alignment of their segments in a matrix format (one line per phrase, one column per set of corresponding segments). Segments are identified by their starting metrical position (e.g. 3.2 for bar 3, second part) and all segments in a bar must be of the same size.

From each multiple alignment annotation of N phrases we derived N(N-1)/2 pairwise alignments. We compared these to automatic alignments derived from specific framework setups. Each setup consists of:

A basic distance measure (*seed*) acting on segments defined by the expert annotation. The segments usually stem from the first subdivision of the bar (one half of a bar in 6/8, one third of a bar in 9/8 measure). Exception: a 4/4 variant in a 6/8 melody group is split into four segments per bar.

A normalization to turn the segment distance values into match scores between 0 and 1. We employ $e^{-distance}$ as the match score for this experiment.

Gap penalty scores for inner and outer gaps (between 0 and 1.5 for this experiment). Note: gap scores are subtracted and match scores are added in an alignment.

For each setup we generated a log file. For overall performance comparisons we produced summary fitness values per variant group and across all tested variant groups. For the fitness of a setup for a particular variant group (*group fitness*), we counted all pairwise alignments in which the automatic alignment has the same gap positions as the annotation and divided this number by N(N-1)/2. For the *overall fitness* of a setup, we took the average of the group fitnesses.

4.1 Results

Four (not necessarily representative) groups of variant phrases were manually aligned and used for evaluation. We only present a summary about the lessons learned from studying the log files [1], which contain annotations, links to the musical scores, alignment results, failure analysis information and summaries.

The overall performance of selected setups is shown in table 4.1 and discussed in the next sections.

4.2 Discussion of distance seeds

In this section we discuss the performance of increasingly complex elementary distance measures (seeds). Baselines are provided by **trivial** and **random**.

The trivial distance 0 turns into a match score of 1 to any pair of segments. As a consequence the actual alignment depends on the gap scores and the algorithm execution order only. In our test setup this always means that the algorithm always chooses left

Seed	IG	OG	G1	G2	G3	G4	A1	A2
trivial	0.0	0.0	33	50	20	10	28	34
random	0.5	0.3	50	16	30	20	29	32
random	1.5	1.0	83	50	20	10	40	51
beatPos	0.0	0.0	33	16	20	10	19	23
beatPos	0.5	0.3	33	50	20	10	28	34
beatPos	1.5	1.0	100	50	20	10	45	56
events	0.5	0.3	100	50	40	30	55	63
events	1.5	0.5	100	50	20	30	50	56
ptdAbs	0.5	0.3	66	83	50	30	57	66
ptdAbs	1.5	1.0	100	83	40	20	60	74
ptdRel	0.5	0.3	66	50	50	20	46	55
ptdRel	1.5	1.0	100	50	50	20	55	66

Table 1. Alignment fitness. IG/OG: inner/outer gap scores. G1-4: percentage of well-aligned phrases per group. A1: average of G1-4; A2: average of G1-3.

(outer) gaps to compensate for the beat segment count difference of the variants.

A random distance between 0 and 1 leads to a more even distribution of gaps. When outer gaps are cheaper, there is a preference for outer gaps. Interestingly in our examples *random* performs better than *trivial*, because the manual alignments contain more right than left outer gaps. As a consequence we should consider to lower the right outer gap penalty with respect to the left in future experiments.

To study the performance using phrase and meter information only, we defined the **beatPos** distance as the difference of the segment number relative to the previous barline. The second segment of a bar thus has distance 1 to the first segment. The algorithm should prefer to align barlines this way. Surprisingly it performed worse than *trivial*. However, we found that too many (relatively cheap) gaps were introduced to match as many segments as possible. We compensated this in another test run with gap penalties greater than 1 and achieved much higher fitness than *trivial*. In general there were only few examples where both phrases were supposed to have inner gaps at different positions.

The next distance measure **events** measures the difference of the number of notes in a segment. Tied notes that begin before the segment starts count for both the current and the preceding segment. The effect of this measure is that regions of same event density (related to tempo or rhythm) are preferred matches. *Events* performs overall better in the alignment than *beatPos*.

To take both onset and pitch into account at the same time, we used proportional transportation distance (PTD) [8]: we scaled the beat segment time into [0..1] and weighted the note events according to their scaled duration. As the ground distance we used $(4\Delta pitch^2 + \Delta onset^2)^{-2}$ with pitches measured as MIDI note numbers modulo 12. Our distance measure **ptdAbsolute** takes advantage of the fact that the given melodies are all in the same key. It performs best in comparison with the previous measures.

If we do not assume the same key, it does not make sense to employ absolute pitch, but instead one can compare only the contours. One approach to tackle this is **ptdRelative**, which takes the minimum over 12 *ptdAbsolute* distances. However, it performs much worse. The reason for this is that, given this distance measure, two segments that each contain a single note always have a distance 0. One should therefore apply this measure only on larger segments or model the tonal center in the state of the alignment process (see section 5).

4.3 Discussion of annotation groups

The four variant groups were chosen to display different kind of alignment problems.

The manual alignment of the group G1 (*Wat hoor ik hier...*) does not contain any inner gaps. There is little variation in absolute pitch, so *ptdAbsolute* achieves 100% the same alignments. The framework proves to be useful and handles different kind of measures (6/8 and 9/8) correctly.

Group G2 (*Ik ben er ...*) contains one inner gap. According to the annotation, the final notes *d,c,b* of one variant (*72564*) are diminished (*d,c* lasts one beat instead of two beats). Because the framework does not handle such transformations yet, this was annotated as a gap. For the similarity measures this gap is hard to find, probably because the neighboring segments provide good alternative matches and often a right outer gap is preferred. Lowering the gap penalties leads to the introduction of unnecessary extra gaps. However, *ptdAbsolute* with high gap penalties achieves 83% success and misses only one pair (*72564* and *72629*), because it matches *d* with *e*. The framework deals well with aligning 4/4 with 6/8 measures.

Group G3 (*Moeder ik kom ...*) contains a repeated segment. Variants that have no repeat are annotated with a gap at the first occurrence of the segment. However, there is no compelling reason why this gap cannot

be assigned to the second occurrence. This ambiguity accounts for many "failures" in the test logs.

Group G4 (*Heer Halewijn*) was chosen because of its complexity. Only when looking at the annotation for a while the chosen alignment becomes understandable. It is mainly based on tonal function of pitches and contains many inner gaps. For pairs of phrases, other alignments are plausible as well, but in the multiple alignment several small hints together make the given annotation convincing. Therefore there are only few correct automatic alignments. Interestingly, however, the algorithm manages to align a subgroup (*72256*, *74003*, and *74216*) without failure.

5 CONCLUSION

We have presented a structure-based alignment and similarity framework for folk song melodies represented as scores. We have done initial tests that show both the usefulness and limitations of our segmentation, alignment and evaluation approach. We see two continuations. First, we should use the framework to study similarity seeds that take the observed stability of beginnings and endings into account (see section 2.2).

Second, the alignment framework needs to be developed further into several directions. 1) We did not so far pay any attention to the relationship between the statistical properties of the distance measure, its normalization and the value of the gap penalties. 2) We should support the modeling of *states* and non-linear gap costs. 3) *Multiple alignment* strategies should be incorporated in order to relate more than two melodies. The need for this became apparent in the last alignment group. Multiple alignments are particularly needed for group queries [3]. Therefore we will not only evaluate the quality of the alignments but also the performance of melody retrieval using these alignments.

Acknowledgments. This work was supported by the Netherlands Organization for Scientific Research within the WITCHCRAFT project NWO 640-003-501, which is part of the CATCH-program.

6 REFERENCES

[1] Log files for this paper. `http://pierement.zoo.cs.uu.nl/misc/ISMIR2008/`.

[2] University Bonn Arbeitsgruppe Multimedia-Signalverarbeitung. Modified Mongeau-Sankoff algorithm. `http://www-mmdb.iai.uni-bonn.de/forschungprojekte/midilib/english/saddemo.html`.

[3] J. Garbers, P. van Kranenburg, A. Volk, F. Wiering, L. Grijp, and R. C. Veltkamp. Using pitch stability among a group of aligned query melodies to retrieve unidentified variant melodies. In Simon Dixon, David Bainbridge, and Rainer Typke, editors, *Proceedings of the eighth International Conference on Music Information Retrieval*, pages 451–456. Austrian Computer Society, 2007.

[4] J. Garbers, A. Volk, P. van Kranenburg, F. Wiering, L. Grijp, and R. C. Veltkamp. On pitch and chord stability in folk song variation retrieval. In *Proceedings of the first International Conference of the Society for Mathematics and Computation in Music*, 2007. (`http://www.mcm2007.info/pdf/fri3a-garbers.pdf`).

[5] P. van Kranenburg, J. Garbers, A. Volk, F. Wiering, L.P. Grijp, and R. C. Veltkamp. Towards integration of MIR and folk song research. In *ISMIR 2007 proceedings*, pages 505–508, 2007.

[6] M. Mongeau and D. Sankoff. Comparison of musical sequences. In *Computers and the Humanities*, volume 24, Number 3. Springer Netherlands, 1990.

[7] C. Sapp. Humdrum extras (source code). `http://extras.humdrum.net/download/`.

[8] Rainer Typke. *Music retrieval based on melodic similarity*. PhD thesis, Utrecht University, 2007.

[9] A. Volk, P. van Kranenburg, J. Garbers, F. Wiering, R. C. Veltkamp, and L.P. Grijp. A manual annotation method for melodic similarity and the study of melody feature sets. In *ISMIR 2008 proceedings*.

[10] Walter Wiora. Systematik der musikalischen Erscheinungen des Umsingens. In *Jahrbuch für Volksliedforschung 7*, pages 128–195. Deutsches Volksliedarchiv, 1941.

[11] D. Yaary and A. Peled. Algorithms for molecular biology, lecture 2. `http://www.cs.tau.ac.il/~rshamir/algmb/01/scribe02/lec02.pdf`, 2001.