MUSICAL MODELS FOR FOLK-SONG MELODY ALIGNMENT

Peter van Kranenburg, Anja Volk, Frans Wiering, Remco C. Veltkamp

Utrecht University, Department of Information and Computing Sciences

{petervk,volk,fransw,Remco.Veltkamp}@cs.uu.nl

ABSTRACT

In this paper we show that the modeling of musical knowledge within alignment algorithms results in a successful similarity approach to melodies. The score of the alignment of two melodies is taken as a measure of similarity. We introduce a number of scoring functions that model the influence of different musical parameters. The evaluation of their retrieval performance on a well-annotated set of 360 folk-song melodies with various kinds of melodic variation, shows that a combination of pitch, rhythm and segmentation-based scoring functions performs best, with a mean average precision of 0.83.

1. INTRODUCTION

In this paper we use alignment algorithms to measure the similarity of melodies. Alignment algorithms are widely used for comparison of sequences of symbols. Creating an alignment is a way to relate two sequences with each other by finding the best corresponding parts. Especially in the field of computational biology, where they are used to find corresponding patterns in protein or nucleotide sequences, many algorithms that align sequences have been developed. Sequence alignment is also suitable for assessing musical similarity for several reasons. Firstly, music unfolds in time, therefore, a model of music as a one-dimensional sequence of events seems appropriate. Secondly, manual alignments have extensively been used in folk-song research to evaluate relations between melodies. Thirdly, structural alignment is a prominent model in cognitive science for human perception of similarity [3].

Most alignment algorithms use a dynamic programming approach. One of the earliest variants is the Levenshtein distance [8], which is an edit distance: it computes how many operations are needed to transform one sequence into another. Needleman and Wunsch [9] proposed an algorithm that finds an optimal alignment of two complete sequences. The quality of an alignment is measured by the alignment score, which is the sum of the alignment scores of the individual symbols. If we consider two sequences of symbols $\mathbf{x} : x_1, \ldots, x_i, \ldots, x_n$, and $\mathbf{y} : y_1, \ldots, y_j, \ldots, y_m$, then symbol x_i can either be

aligned with a symbol from sequence **y** or with a gap. Both operations have a score, the substitution score and the gap score. The gap score is mostly expressed as penalty, i.e. a negative score. The optimal alignment and its score are found by filling a matrix D recursively according to:

$$D(i,j) = \max \begin{cases} D(i-1,j-1) + S(x_i, y_j) \\ D(i-1,j) - \gamma \\ D(i,j-1) - \gamma \end{cases}, \quad (1)$$

where $S(x_i, y_j)$ is the substitution scoring function, γ is the gap penalty, $D(0,0) = 0$, $D(i,0) = -i\gamma$, and $D(0,j) = -j\gamma$. $D(i,j)$ contains the score of the optimal alignment up to x_i and y_j and therefore, $D(m,n)$ contains the score of the optimal alignment of the complete sequences. We can obtain the alignment itself by tracing back from $D(m,n)$ to $D(0,0)$; the algorithm has both time and space complexity $O(nm)$. In our modeling, we use an extension of the algorithm proposed by Gotoh [5], which employs an affine gap penalty function without loss of efficiency. In this approach, the extension of a gap gets a lower penalty than its opening.

Mongeau and Sankoff [10] were among the first to adapt alignment algorithms to music. They used an extended version of the Needleman-Wunsch algorithm. Their scoring function takes both pitch and duration into account. Mongeau and Sankoff's approach has been quite influential, e.g. the search algorithm implemented in the search engine MELDEX [13] is based on this algorithm. Gómez et al. [4] successfully tested a modified version on a MIREX dataset. In general, alignment algorithms have often been used to match short melodic phrases against a larger database [1, 4, 7, 13, 14]. Typical tasks addressed with this approach are to find a tune in the database with QBH [1], different arrangements of a piece [14], or similar incipits given to the query [4]. We use the alignment between complete melodies in order to find melodies that belong to the same tune family. The similarity relations that have to be modeled originate in the oral transmission of folk-songs and differ from those in the previous tasks.

Contribution. In this paper we model various features of music as substitution scoring functions, which we incorporate in the Needleman-Wunsch-Gotoh algorithm. Using a set of melodies that are well-described regarding their different kinds of similarity relations, we evaluate the influence of these scoring functions on the retrieval performance. Our best scoring function combines several musical features and outperforms well-known approaches from literature.

2. DATA

2.1 The Annotated Corpus

The set of melodies studied in this paper is part of a larger collection of over 6000 encoded Dutch folk-songs hosted by the Meertens Institute in Amsterdam. In the ongoing project of digitization at this institute, the melodies are encoded both from ethnomusicological transcriptions of field recordings and from written sources of folk-songs, delivering several formats (humdrum **kern, MIDI, Lilypond). [1] For a subset of 360 melodies, detailed annotations have been created in order to describe similarity relations between melodies [16], resulting in a well-documented set of songs, the Annotated Corpus.

The melodies are grouped in so-called tune families. [2] All melodies in one group are considered to be historically related through the process of oral transmission. Since the history of each tune family is not fully documented, it is often not known whether two melodies are historically related. Instead, musicological experts decide whether melodies belong to the same tune family by assessing their melodic and textual similarity. In order to make the experts' musical intuition behind the similarity assessments explicit, we developed an annotation system (described in [16]). For the Annotated Corpus (consisting of 26 tune families) several dimensions of perceived similarity (contour, rhythm, motives, text) were numerically rated by the musicologists such that the similarity between the most typical melody of a tune family (the reference melody) and all other members of the tune family was described. The 26 tune families were chosen from the larger collection by an expert such that this set contains a representative diversity of similarity relations between members of a tune family. Comparing the annotations to the retrieval performance of alignment algorithms allows a detailed understanding of the success or failure of the models based on musicological insights.

2.2 Representation of melodies

For applying alignment algorithms, a melody has to be presented as a sequence of symbols. In our representation, each symbol represents a note. A symbol has a number of attributes, including: pitch (in base40 encoding), duration (rational number), score time (rational number), time in bar (rational number), onset (integer), current bar number (integer), current phrase number (integer), upbeat (boolean), current meter (rational number), free meter (boolean), accented (boolean), and time position within phrase (real number in $[0, 1]$). These attributes are used to compute substitution scores or other attributes. Figure 1 shows an example with some of the attributes.

Based on the encoded time signature, two levels of accents are distinguished: either accented or not accented. The first note of any group of two in a double meter and the first beat in any group of three beats in a triple meter is con-
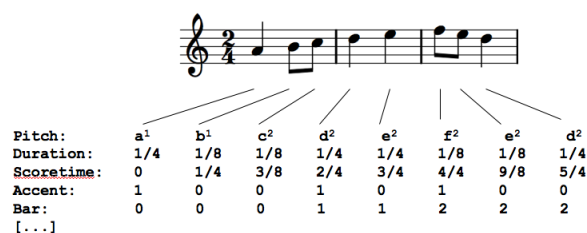

Figure 1. Representation of melodies.

sidered accented. All other notes are unaccented. Thus, in songs in free meter, or in songs with additive or asymmetrical meters, [3] which are very uncommon in this corpus, all notes are unaccented. Furthermore, phrase boundaries have been annotated by the encoders.

2.3 Rests

Most notated rests can be considered inessential. In particular at the end of phrases singers often take a breath, such that timing between the phrases is very variable. The exact encoding of rests as performed is therefore not reasonable. To make melodies more comparable, all rests have been replaced by a prolongation of the previous note.

2.4 Transposition Invariance

Since songs are notated in different keys, the similarity measure should be transposition invariant. To achieve this, a pitch histogram for both melodies is created that indicates for each pitch the total duration during the song. Then the shift at which the normalized histograms have maximal intersection is computed. Since the pitches are represented in base40 encoding, the shift of the histogram can be interpreted as the interval with which the one melody should be transposed in order to compare it to the other.

2.5 Normalization of Alignment Scores

Since the score of an alignment depends on the length of the sequences, normalization is needed to compare different alignment scores. Therefore, we divide the alignment score by the length of the shortest sequence.

3. SCORING FUNCTIONS

3.1 Single substitution scoring functions

In this section we introduce a number of substitution scoring functions for different musical dimensions. They determine substitution scores that are based on musicological knowledge. Each function takes two symbols of the melodic sequence as input. The output of each scoring function is in the interval $[-1, 1]$.

First, we introduce scoring functions that are based on pitch-related features. The simplest scoring function determines whether two pitches are the same or not. The score

---

is either maximal or minimal:

$$S_{exactpitch}(x_i, y_j) = \begin{cases} 1 & \text{if } x_i = y_j \\ -1 & \text{if } x_i \neq y_j \end{cases} . \quad (2)$$

In oral transmission, slight changes of pitches are likely to occur, therefore, we allow substitution with pitches that are within a band with certain width:

$$S_{pitchb}(x_i, y_j) = \begin{cases} 1 - \frac{int(x_i, y_j)}{23} & \text{if } int(x_i, y_j) \leq 23 \\ -1 & \text{otherwise} \end{cases} . \quad (3)$$

We define $int(x_i, y_j) = |p(x_i) - p(y_j)| \bmod 40$, with $p(x)$ as the pitch of symbol x in base 40 encoding. A fifth is 23 in base 40 encoding. Thus, all intervals up to a fifth get a positive substitution score and all larger intervals are considered a bad match.

Another way to express the distance of two pitches is by their harmonic relation. The substitution of consonances gets a higher score than the substitution of dissonances:

$$S_{harm}(x_i, y_j) = \begin{cases} 1 & \text{prime} \\ 0.5 & \text{consonance} \\ 0.5 & \text{augmented prime} \\ -1 & \text{dissonance} \end{cases} . \quad (4)$$

The intervals are taken modulo octave. Consonances are minor and major third, perfect fourth, perfect fifth and minor and major sixth. The augmented prime gets a positive substitution score to favour alignments of songs that have a minor and a major variant.

Furthermore, we define two substitution functions that are based on melodic contour, taking either the contour of a phrase or of the entire melody into account:

$$S_{phrasecont}(x_i, y_j) = 1 - 2 * |p_{phr}(x_i) - p_{phr}(y_j)| . \quad (5)$$

$$S_{songcont}(x_i, y_j) = 1 - 2 * |p_{song}(x_i) - p_{song}(y_j)| . \quad (6)$$

Here $p_{phr}(x)\epsilon[0, 1]$ indicates the vertical position between the lowest and highest pitches of the phrase that x is part of, while $p_{song}(x)\epsilon[0, 1]$ indicates the vertical position between the lowest and highest pitches of the entire song. In determining the highest and lowest pitches, the notes in the upbeats of the phrases are disregarded, since these are very variable between variants of a song.

Next, we define three scoring schemes that are based on rhythmic features. In a simple approach based on note durations, the score is maximal if the durations are the same, and minimal otherwise:

$$S_{exactdur}(x_i, y_j) = \begin{cases} 1 & \text{if } d(x_i) = d(y_j) \\ -1 & \text{if } d(x_i) \neq d(y_j) \end{cases} , \quad (7)$$

in which $d(x)$ is the duration of the symbol x.

Metric accents derived from the notated time signature describe a further aspect of the rhythmic structure of melodies. We define a substitution function that uses these metric accents in the following way:

$$S_{accent}(x_i, y_j) = \begin{cases} 1 & \text{if } a(x_i) = a(y_j) \\ -1 & \text{if } a(x_i) \neq a(y_j) \end{cases} , \quad (8)$$

in which $a(x)$ indicates whether the symbol x is accented or not (for defining accents see section 2.2).

A more complex notion of metric accents based on the rhythmic structure of notes instead of the time signature is provided by Inner Metric Analysis (IMA) [15]. We define a scoring function that is determined by the metric weights of the notes, as computed by IMA:

$$S_{ima}(x_i, y_j) = 1 - 2 * |w(x_i) - w(y_j)| . \quad (9)$$

Here $w(x)$ denotes the metric weight of the symbol x scaled into the interval $[0, 1]$. For scaling, all weights were divided by the greatest weight in the song. The parameters for the IMA algorithm are the ones that are mostly used: $p = 2, l = 2$ (e.g., in [15]).

Furthermore, we want to use the information of phrase boundaries given in our data-set. We introduce a scoring function based on the horizontal position within the phrase:

$$S_{phr}(x_i, y_j) = 1 - 2 * |phr(x_i) - phr(y_j)| , \quad (10)$$

in which $phr(x)\epsilon[0, 1]$ indicates for the symbol x the horizontal position in its phrase. This substitution function helps to keep phrases together in alignments.

3.2 Combination

The single substitution scoring functions defined in section 3.1 model isolated aspects of melodies. In order to model several aspects within one function to get closer to the multidimensionality of melodies, we combine substitution functions. We want alignments in which the aligned symbols are similar in all dimensions, therefore, we multiply the individual scores:

$$S_{combination}(x_i, y_j) = \prod_{k=1}^{n} S_k(x_i, y_j) , \quad (11)$$

in which each $S_k(x_i, y_j)$ is scaled into the interval $[0, 1]$, and the final score is scaled into $[-1, 1]$ back again.

3.3 Gap penalty function

We use an affine gap penalty function in which the penalty for opening a gap is 1, and the penalty for extending a gap is 0.1. Thus, variants of songs in which e.g. a phrase is repeated can be better aligned, since these penalties result in one long gap instead of many short gaps. Furthermore, the use of an affine gap penalty function prevents gaps from being scattered all over the alignment.

4. EVALUATION OF SCORING FUNCTIONS

The scoring functions are evaluated by their respective retrieval performance on our Annotated Corpus as described in section 2. To evaluate a scoring scheme each melody is taken once as query and the other melodies are sorted according to the normalized score of the alignment with the query melody. At all ranks the average recall and average precision for all ranking lists is computed. These values are plotted in a diagram. The criterion for relevance is the membership of the same tune family.

4.1 Evaluation of Single Substitution Functions

First, we study the performance of the single pitch-based substitution functions introduced in section 3.1. Variation in pitch is considered an important element of oral transmission (see e.g. [6]). Nevertheless, aligning melodies using the exact pitch information with the simplest function $S_{exactpitch}$ results in a relatively good performance (see Figure 2). Allowing pitch variation within a small range using the pitch band function improves this performance only slightly.
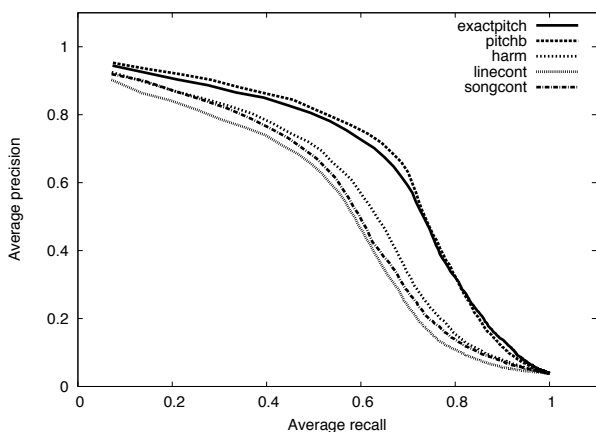
Figure 2. Retrieval performance of pitch-based substitution functions.

Both the harmonic and contour-based substitution functions perform worse than $S_{exactpitch}$. Considering the contour instead of the exact pitch sequence does not result in a better retrieval performance. Harmonic relations, which have otherwise successfully been used in models of melodic expectancy [11], do not improve the alignment of melodies of a tune family in comparison to exact pitch information.
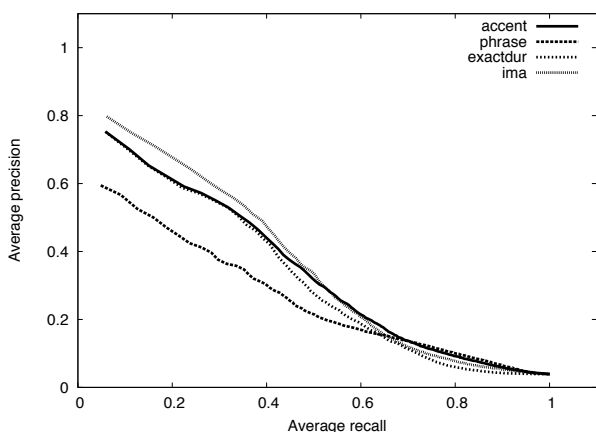
Figure 3. Retrieval performance of non-pitch-based substitution functions.

Figure 3 shows retrieval performance for the scoring functions that do not involve pitch information. Although rhythmic features have been considered quite stable within oral transmission (see [6]), all rhythm-related substitution functions perform worse than pitch-related functions.

S_{ima} performs at the top of the ranking slightly better than S_{accent}, however S_{accent} performs slightly better in the low range. In general the difference between the two models is quite small, indicating that the accents of the notated bars are synchronous to the accents based on notes onsets.

4.2 Evaluation of Combinations of Single Substitution Functions

In a next step, we combine rhythmical, metrical and segmentation data. First, we combine the best of the pitch-related functions (S_{pitchb}) with rhythmical and phrase functions. Figures 2 and 3 show that the individual substitution functions perform worse than S_{pitchb}, but from the curves of $S_{pitchb-accent}$, $S_{pitchb-phrase}$, $S_{pitchb-exactdur}$, and $S_{pitchb-ima}$ in Figure 4 it appears that combinations yield better retrieval performance for all combinations but $S_{exactdur}$. Since $S_{exactdur}$ is binary and the combination is by multiplication, the pitch similarity for symbols with no exact correspondence in duration is lost.

Combination with the other two rhythmic functions (S_{ima} and S_{accent}) show equal improvement. The rather modest improvement when considering metric accents in comparison to the single substitution function S_{pitchb} contradicts the hypothesis that pitches among melodies of the same tune family are more stable on metrically accented notes than on metrically weak positions as assumed in [2]: obviously pitches on metrically weak positions also vary to only a small extent. The phrase information yields the greatest improvement.

Finally, we evaluate the retrieval performance of the combination of the best substitution functions. We choose S_{ima} as the metric scoring function, S_{pitchb} is the best pitch based scoring function. S_{phr} improved retrieval results by stimulating phrase boundaries to be aligned. The retrieval performance of the combination $S_{pitchb-ima-phr}$ shown in Figure 4 shows even better performance results than the combinations of two single substitution functions. If we average the precision of all relevant items for all queries, we get a mean average precision of 0.83 for this combination. Choosing S_{accent} instead of S_{ima} gives nearly the same retrieval performance.
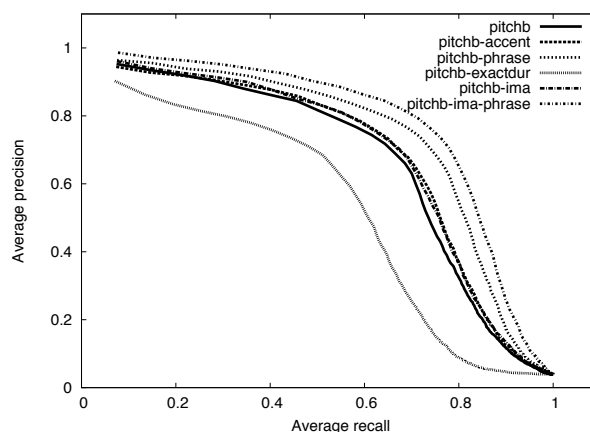
Figure 4. Retrieval performance of combinations of substitution functions.

To evaluate the scalability, we performed the same test with a data-set containing all 4863 classified songs and with the same 360 queries. The results yield a mean average precision of 0.67.

4.3 Comparison with Related Methods

Figure 5 shows comparisons of our best scoring scheme with alignment methods from literature. For the method of Mongeau and Sankoff [10] the parameters were taken as given by Mongeau and Sankoff. The normalization was done by dividing the alignment score by the sum of the durations of both sequences. DiffEd and rawEd were taken from the Simile alignment toolbox without change [12]. It appears that our $S_{pitchb-ima-phr}$ performs best.
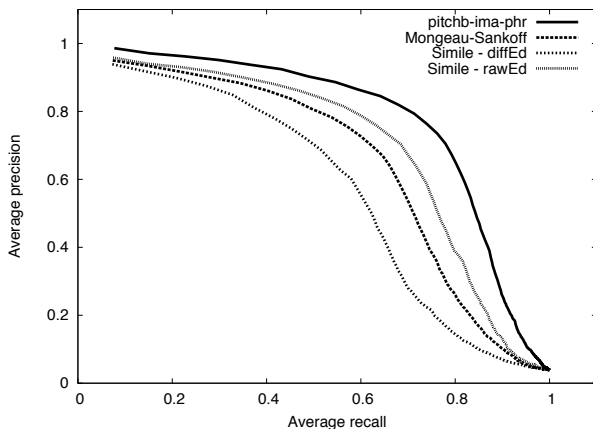

Figure 5. Comparison with related methods.

5. RETRIEVAL PERFORMANCE PER TUNE FAMILY

The classification of the melodies into tune families by musicological experts is based on a number of musical dimensions. The importance of the different dimensions and the form of interaction between them varies to a great extent among tune families. Therefore, finding a similarity model that performs well on all tune families is a challenging task. The retrieval performance of the scoring function $S_{pitchb-ima-phr}$ shows per tune family a considerably stable success, with average precision values ranging from 1 to 0.8 for 23 out of 26 tune families. Hence, this function works reasonably well on the majority of tune families. For three tune families the function shows only moderate retrieval performance, which are *Een lindeboom stond in het dal 1* (short: *Lindeboom*), *Daar reed een jonkheer 1* (short: *Jonkheer*) and *Heer Halewijn 4* (short: *Halewijn*), with respective average precision of 0.71, 0.67 and 0.65. Table 1 gives an overview of low ranking results extracted from rankings in which the reference melody of the tune family (see section 2.1) was used as the query. For *Lindeboom*, Figure 6 shows the first line of the query along along with a good match and the two melodies with ranks 77 and 140. The melodies on these low ranks are quite different from the query. This is reflected by the experts' annotations: for the two melodies on ranks 77 and 140 all global musical

Tune family	Relevant melodies at rank
Lindeboom	77 and 140
Jonkheer	78, 98 and 167
Halewijn	19, 40, 74 and 101

Table 1. Overview over low ranking results.

Figure 6. From top to bottom: incipits of reference melody (query), melodies on ranks 1, 77 and 140..

dimensions (rhythm, contour and importance of motives) are rated *somewhat similar*, while for most of the other melodies of *Lindeboom* at least two of these dimensions are rated *very similar*.

In the tune family *Jonkheer* the melody on rank 78 is notated with a different phrase structure than the query, such that two phrases correspond to one phrase of the query. Since phrase information is used for the alignment, this inconsistent phrase assignment introduces lower scores. The melody on rank 98 has a very different formal structure than the query: while the query has the form ABA', this melody has the form AAA'A". As a consequence, notes in phrases that are aligned with each other differ to a great extent. The melody on rank 167 is quite different from the query (see Figure 7), which is reflected by low ratings of the experts in both local and global musical dimensions.

The tune family *Halewijn* is according to the experts most of all characterized by a similar rhythmic organization. Melodies of this group differ considerably regarding pitch, such that the contour is mostly rated as only *somewhat similar*. The averaged annotation values for all musical dimensions in this tune family show a significant ($p = 0.02$) linear correlation with the distances obtained using the $S_{pitchb-ima-phr}$ rater — hence the lowest ranked melodies tend to receive low similarity scores in the annotations. For the melody on rank 74 the expert commented that it is possible that this melody does not belong to the tune family.

6. CONCLUSION AND FUTURE WORK

We have shown that the inclusion of musical knowledge in alignment algorithms improves the assessment of similarity among folk song melodies. By evaluating different substitution scoring functions, we found that our pitch-related functions lead to better recognition than rhythm-related functions. The use of phrase information improved

Record 73426 - Strophe 1

Er was eens een heer, hij was wel - le - ge-moed

En hij droeg er een gou - wen krans on - ner zijn hoed

Hij droeg er een gouwe en een gou-wen krans on - ner zijn hoed

Record 73929 - Strophe 1

Daar reed eens een heer - - die

was wel - di - di - don

was wel don da ron di di da ri don

was wel - der - ge - moed

Figure 7. Reference melody of *Jonkheer* (top) and melody on rank 167 (bottom).

the retrieval results considerably. The best combination of functions, combining a pitch-based, a rhythm-based and a segmentation-based scoring function, outperforms related methods from literature.

Next, we will develop scoring functions that reflect more advanced musicological models. We will use the annotations to evaluate the results by means of the quality of alignments. Since the occurrence of similar motives in related melodies was considered important by the musicological experts, we will investigate how corresponding motives can be aligned. For testing the suitability of our approach to model similarity within oral transmission in general, we will evaluate its performance on different collections. These steps contribute to the development of a similarity rater adequate for oral transmission that is based on musically advanced models of melodic similarity.

7. REFERENCES

[1] N.H. Adams, M.A. Bartsch, J.B. Shifrin and G.H. Wakefield: "Time Series Alignment for Music Information Retrieval", *Proceedings of the 5th International Conference on Music Information Retrieval*, Barcelona, 2004.

[2] 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", *Proceedings of the First International Conference of the Society for Mathematics and Computation in Music*, 2007.

[3] R.L. Goldstone: "Similarity, interactive activation, and mapping", *Journal of Experimental Psychology: Learning, Memory, and Cognition*, Vol. 20, pp. 3–28, 1994.

[4] C. Gómez, S. Abad-Mota, and E. Ruckhaus: "An Analysis of the Mongeau-Sankoff Algorithm for Music Information Retrieval", *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, pp. 109–110, 2007.

[5] O. Gotoh: "An Improved Algorithm for Matching Biological Sequences", *Journal of Molecular Biology*, Vol. 162, pp. 705–708, 1982.

[6] E. Klusen, H. Moog, and W. Piel: "Experimente zur mündlichen Tradition von Melodien", *Jahrbuch für Volksliedforschung*, Vol. 23, pp. 11–32, 1978.

[7] K. Lemström and E. Ukkonen: "Including Interval Encoding into Edit Distance Based Music Comparison and Retrieval", *Proceedings of the AISB'00 Symposium on Creative & Cultural Aspects and Applications of AI & Cognitive Science*, Birmingham, pp. 53–60, 2000.

[8] V. Levenshtein: "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", *Soviet Physics Doklady*, Vol. 10, No. 8, pp. 707–710, 1966.

[9] S.B. Needleman and C.D. Wunsch: "A general method applicable to the search for similarities in the amino acid sequence of two proteins", *Journal of Molecular Biology*, Vol. 48, No. 3, pp. 443-53, 1970.

[10] M. Mongeau and D. Sankoff: "Comparison of musical sequences", *Computers and the Humanities*, Vol. 24, pp. 161–175, 1990.

[11] E.H. Margulis: "A Model of Melodic Expectation". *Music Perception*, Vol. 22, No. 4, pp. 663–714, 2005.

[12] D. Müllensiefen and K. Frieler: "Cognitive Adequacy in the Measurement of Melodic Similarity: Algorithmic vs. Human Judgements," *Computing in Musicology*, Vol. 13, pp. 147–177, 2004.

[13] L.A. Smith, R.J. McNab, and I.H. Witten: "Sequence-Based Melodic Comparison: A Dynamic-Programming Approach", *Computing in Musicology*, Vol. 11, pp. 101–117, 1998.

[14] A. Uitdenbogerd and J. Zobel: "Melodic Matching Techniques for Large Music Databases", *The Seventh International Multimedia Conference*, Orlando, Florida, 1999.

[15] A. Volk: "Persistence and Change: Local and Global Components of Metre Induction using Inner Metric Analysis", *Journal of Mathematics and Computation in Music*, Vol. 2, No. 2, pp. 99–115, 2008.

[16] A. Volk, P. van Kranenburg, J. Garbers, F. Wiering, R.C. Veltkamp, L.P. Grijp: "A manual annotation method for melodic similarity and the study of melody feature sets", *Proceedings of the Ninth International Conference on Music Information Retrieval*, pp. 101–106, 2008.