MELODIC SIMILARITY: LOOKING FOR A GOOD ABSTRACTION LEVEL

Maarten Grachten and *Josep-Lluís Arcos* and *Ramon López de Mántaras*
IIIA-CSIC - Artificial Intelligence Research Institute
CSIC - Spanish Council for Scientific Research
Campus UAB, 08193 Bellaterra, Catalonia, Spain.
Vox: +34-93-5809570, Fax: +34-93-5809661
Email: {maarten,arcos}@iiia.csic.es

ABSTRACT

Computing melodic similarity is a very general problem with diverse musical applications ranging from music analysis to content-based retrieval. Choosing the appropriate level of representation is a crucial issue and depends on the type of application. Our research interest concerns the development of a CBR system for expressive music processing. In that context, a well chosen distance measure for melodies is a crucial issue. In this paper we propose a new melodic similarity measure based on the I/R model for melodic structure and compare it with other existing measures. The experimentation shows that the proposed measure provides a good compromise between discriminatory power and the level of abstraction of melody representation.

1. INTRODUCTION

Computing melodic similarity is a very general problem with diverse musical applications ranging from music analysis to content-based retrieval. Choosing the appropriate level of representation is a crucial issue and depends on the type of application. For example, in applications such as pattern discovery in musical sequences [1], [4], or style recognition [4], it has been established that melodic comparison requires taking into account not only the individual notes but also the structural information based on music theory and music cognition [12].

Our research interest concerns the development of a CBR system for expressive music processing. In that context (e.g. for retrieval and reuse mechanisms), a well chosen distance measure for melodies is of importance. Some desirable features of such a measure are the ability to distinguish phrases from different musical styles and to recognize phrases that belong to the same song. We propose a new way of assessing melodic similarity, representing the melody as a sequence of I/R structures (conform Narmour's Implication/Realization (I/R) model for melodic structure [10]). The similarity is then assessed by calculating the edit-distance between I/R representations of melodies. We compared this assessment to assessments based on note representations [9], and melodic contour representations [2, 7].

We have found that the discriminatory power (using an entropy based definition) of the note level distance measure is much lower than that of the contour and I/R level measures. Also, taking into account interval durations within the contour level measures, tended to decrease the discriminatory power. We argue that the I/R level measure is an appropriate compromise that takes into account rhythmical/temporal information in an implicit way, without losing discriminatory power.

The paper is organized as follows: In Section 2 we briefly introduce the Narmour's Implication/Realization Model. In section 3 we describe the four distance measures we are comparing — the note-level distance proposed in [9], two variants of contour-level distance and the I/R-level distance we propose as an alternative. In section 4 we report the experiments performed using these four distance measures on a dataset that comprises musical phrases from a number of well known jazz songs. The paper ends with a discussion of the results, and the planned future work.

2. THE IMPLICATION/REALIZATION MODEL

Narmour [10, 11] has proposed a theory of perception and cognition of melodies, the Implication/Realization model, or I/R model. According to this theory, the perception of a melody continuously causes listeners to generate expectations of how the melody will continue. The sources of those expectations are two-fold: both innate and learned. The innate sources are 'hard-wired' into our brain and peripheral nervous system, according to Narmour, whereas learned factors are due to exposure to music as a cultural phenomenon, and familiarity with musical styles and pieces in particular. The innate expectation mechanism is closely related to the *gestalt theory* for visual perception [5, 6]. Gestalt theory states that perceptual elements

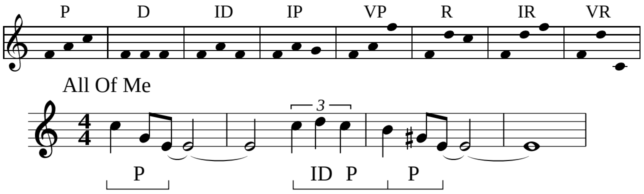

Figure 1. Top: Eight of the basic structures of the I/R model. Bottom: First measures of All of Me, annotated with I/R structures.

are (in the process of perception) grouped together to form a single perceived whole (a 'gestalt'). This grouping follows certain principles (*gestalt principles*). The most important principles are *proximity* (two elements are perceived as a whole when they are perceptually close), *similarity* (two elements are perceived as a whole when they have similar perceptual features, e.g. color or form, in visual perception), and *good continuation* (two elements are perceived as a whole if one is a 'good' or 'natural' continuation of the other). Narmour claims that similar principles hold for the perception of melodic sequences. In his theory, these principles take the form of *implications*: Any two consecutively perceived notes constitute a melodic interval, and if this interval is not conceived as complete, or closed, it is an *implicative interval*, an interval that implies a subsequent interval with certain characteristics. In other words, some notes are more likely to follow the two heard notes than others. Two main principles concern *registral direction* and *intervallic difference*. The principle of registral direction states that small intervals imply an interval in the same registral direction (a small upward interval implies another upward interval, and analogous for downward intervals), and large intervals imply a change in registral direction (a large upward interval implies another upward interval and analogous for downward intervals). The principle of intervallic difference states that a small (five semitones or less) interval implies a similarly-sized interval (plus or minus 2 semitones), and a large intervals (seven semitones or more) implies a smaller interval.

Based on these two principles, melodic patterns can be identified that either satisfy or violate the implication as predicted by the principles. Such patterns are called *structures* and labeled to denote characteristics in terms of registral direction and intervallic difference. Eight such structures are shown in figure 1(top). For example, the P structure ('Process') is a small interval followed by another small interval (of similar size), thus satisfying both the registral direction principle and the intervallic difference principle. Similarly the IP ('Intervallic Process') structure satisfies intervallic difference, but violates registral direction.

Additional principles are assumed to hold, one of which concerns *closure*, which states that the implication of an interval is inhibited when a melody changes in direction, or when a small interval is followed by a large interval. Other factors also determine closure, like metrical position (strong metrical positions contribute to closure, rhythm (notes with a long duration contribute to closure), and harmony (resolution of dissonance into consonance contributes to closure).

We have designed an algorithm to automate the annotation of melodies with their corresponding I/R analyses. The algorithm implements most of the 'innate' processes mentioned before. The learned processes, being less well-defined by the I/R model, are currently not included. Nevertheless, we believe that the resulting analysis have a reasonable degree of validity. An example analysis is shown in figure 1(bottom).

3. MEASURING MELODIC DISTANCES

For the comparison of the musical material on different levels, we used a measure for distance that is based on the concept of *edit-distance* (also known as Levenshtein distance [8]). In general, the edit-distance between two sequences is defined as the minimum total cost of transforming one sequence (the source sequence) into the other (the target sequence), given a set of allowed edit operations and a cost function that defines the cost of each edit operation. The most common set of edit operations contains insertion, deletion, and replacement. Insertion is the operation of adding an element at some point in the target sequence; deletion refers to the removal of an element from the source sequence; replacement is the substitution of an element from the target sequence for an element of the source sequence.

Because the edit-distance is a measure for comparing sequences in general, it enables one to compare melodies not only as note sequences, but in principle any sequential representation can be compared. In addition to comparing note-sequences, we have investigated the distances between melodies by representing them as sequences of directional intervals, directions, and I/R structures, respectively.

These four kinds of representation can be said to have different levels of abstraction, in the sense that some representations convey more concrete data about the melody than others. Obviously, the note representation is the most concrete, conveying absolute pitch, and duration information. The interval representation is more abstract, since it conveys only the pitch intervals between consecutive notes. The direction representation abstracts from the size of the intervals, maintaining only their sign. The I/R representation captures pitch interval relationships by distinguishing categories of intervals (small vs. large) and it characterizes consecutive intervals as similar or dissimilar. The scope of this characterization (not all interval-pairs are necessarily characterized), depends on metrical and rhythmical information.

An example may illustrate how the interval, direction and I/R measures assess musical material. In figure 2, three musical fragments are displayed. The direction measure rates A – B and A – C as equally distant, which is not surprising since A differs by one direction from both B and C. The interval measure rates A as closer to B than

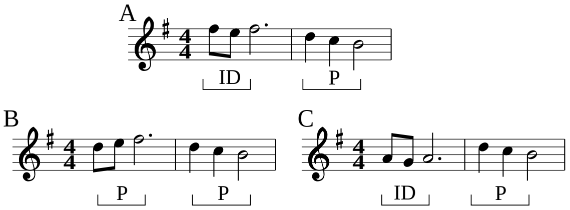

Figure 2. An example illustrating differences of similarity assessments by the interval, direction and I/R measures.

to C. The most prominent difference between A and C in terms of intervals is the jump between the last note of the first measure and the first note of the second. In fragment A this jump is a minor third down, and for C it is a perfect fourth up. It can be argued that this interval is not really relevant, since the first three and the last three notes of the fragments form separate perceptual groups. The I/R distance assessment does take this separation into account, as can be seen from the I/R groupings and rates fragment A closer to C than to fragment B.

The next subsections briefly describe our decisions regarding the choice of edit-operations and weights of operations for each type of sequence. We do not claim these are the only right choices. In fact, this issue deserves further discussion and might benefit also from empirical data conveying human similarity ratings of musical material.

3.1. An edit-distance for note sequences

In the case of note sequences, we have followed Mongeau and Sankoff's approach [9]. They propose to extend the set of basic operations (insertion, deletion, replacement) by two other operations that are more domain specific: *fragmentation* and *consolidation*. Fragmentation is the substitution of a number of (contiguous) elements from the target sequence for one element of the source sequence; conversely, consolidation is the substitution of one element from the target-sequence for a number of (contiguous) elements of the source sequence. In musical variations of a melody for example, it is not uncommon for a long note to be fragmented into several shorter ones, whose durations add up to the length of the original long note.

The weights of the operations are all linear combinations of the durations and pitches of the notes involved in the operation. The weights of insertion and deletion of a note are equal to the duration of the note. The weight of a replacement of a note by another note is defined as the sum of the absolute difference of the pitches and the absolute difference of the durations of the notes. Additionally, there is a weight factor for the duration difference, in order to control the relative importance of pitch and duration attributes. Fragmentation and consolidation weights are calculated similarly: the weight of fragmenting a note n_1 into a sequence of notes $n_2, n_3, ..., n_N$ is again composed of a pitch part and a duration part. The pitch part is defined by the sum of the absolute pitch differences between n_1 and n_2, n_1 and n_3, etc. The duration part is defined by the absolute difference between the duration of n_1, and the summed durations of $n_2, n_3, ..., n_N$. Just like the replacement weight the fragmentation weight is a weighted sum of the pitch and duration parts. The weight of consolidation is exactly the converse of the weight of fragmentation.

3.2. An edit-distance for contour sequences

One way to conceive of the contour of a melody is as comprising the intervallic relationships between consecutive notes. In this case, the contour is represented by a sequence of signed intervals. Another idea of contour is that it just refers to the melodic direction (up/down/repeat) pattern of the melody, discarding the sizes of intervals (the directions are represented as 1,0,-1, respectively). In our experiment, we have computed distances for both kinds of contour sequences.

We have restricted the set of edit operations for both kinds of contour sequences to the basic set of insertion, deletion and replacement, thus leaving out fragmentation and consolidation, since there is no correspondence to fragmentation/consolidation as musical phenomena (as there is trivially in the case of note-sequences). The weights for replacement of two contour elements (intervals or directions) is defined as the absolute difference between the elements, and the weight of insertion and deletion is defined as the absolute value of the element to be inserted/deleted (conform Lemström and Perttu [7]).

Additionally, one could argue that when comparing two intervals, it is also relevant how far the two notes that constitute each interval are apart in time. This quantity is measured as the time interval between the starting positions of the two notes, also called the Inter Onset Interval (IOI). We incorporated the IOI into the weight functions by adding it as a weighted component. For example, let P_1 and IOI_1 respectively be the pitch interval and the IOI between two notes in sequence 1 and P_2 and IOI_2 the pitch interval and IOI between to notes in sequence 2, then the weight of replacing the first interval by the second, would be $|P_2 - P_1| + k \cdot |IOI_2 - IOI_1|$. The weight of deletion of the first interval would be $1 + k \cdot IOI_1$.

3.3. An edit-distance for I/R sequences

The sequences of (possibly overlapping) I/R structures (I/R sequences, for short) that the I/R parser generated for the musical phrases, were also compared to each other. Just as with the contour sequences, it is not obvious which kinds of edit operations could be justified beyond insertion, deletion and replacement. It is possible that research investigating the I/R sequences of melodies that are musical variations of each other, will point out common transformations of music at the level of I/R sequences. In that case, edit operations may be introduced to allow for such common transformations. Presently however, we know of no such common transformations, so we allowed only insertion, deletion and replacement.

As for the estimation of weights for edit operations upon I/R structures, note that unlike the replacement operation, the insertion and deletion operations do not involve any comparison between I/R structures. It seems reasonable to make the weights of insertion/deletion somehow proportional to the 'importance' or 'significance' of the I/R structure to be inserted/deleted. Ideally the (unformalized) notion of significance of an I/R structure would depend on the context of the structure. However, this would not make sense in the case of editing sequences, as this would create a cyclic dependence among the weights of edit operations. Therefore we propose to take the size of an I/R structure, referring to the number of notes the structure spans, as a more practical indicator of the significance of an I/R structure. The weight of an insertion/deletion of an I/R structure can then simply be the size of the structure.

The weight of a replacement of two I/R structures should assign high weights to replacements that involve two very different I/R structures and low weights to replacements of an I/R structure by a similar one. The rating of distances between different I/R structures (which to our knowledge has as yet remained unaddressed) is an open issue. Distance judgments can be judged on class attributes of the I/R structures, for example whether the structure captures a realized or rather a violated expectation. Alternatively, or in addition, the distance judgment of two instances of I/R structures can be based on instance attributes, such as the number of notes that the structure spans (which is usually but not necessarily three), the registral direction of the structure, and whether or not the structure is chained with neighboring structures.

Aiming at a straight-forward definition of replacement weights for I/R structures, we decided to take into account four attributes. The first term in the weight expression is the difference in size (i.e. number of notes) of the I/R structures. Secondly, a cost is added if the direction of the structures is different (where the direction of an I/R structure is defined as the direction of the interval between the first and the last note of the structure). Thirdly, a cost is added if one I/R structure is chained with its successor and the other is not (this depends metrical and rhythmical information). Lastly, a cost is added if the two I/R structures are not of the same kind (e.g. *P* and *VP*). A special case occurs when one of the I/R structures is the *retrospective* counterpart of the other (a retrospective structure generally has the same up/down contour as it's prospective counterpart, but different interval sizes; for instance, a retrospective P structure typically consists of two large intervals in the same direction, see [10] for details). In this case, a reduced cost is added, representing the idea that a pair of retrospective/prospective counterparts of the same kind of I/R structure is more similar than a pair of structures of different kinds.

3.4. Computing the Distances

The minimum cost of transforming a source sequence into a target sequence can be calculated using the following recurrence equation for the distance d_{ij} between two sequences $a_1, a_2, ..., a_i$ and $b_1, b_2, ..., b_j$:

$$d_{ij} = min \begin{cases} d_{i-1,j} + w(a_i, \emptyset) & \text{(a)} \\ d_{i,j-1} + w(\emptyset, b_j) & \text{(b)} \\ d_{i-1,j-1} + w(a_i, b_j) & \text{(c)} \\ d_{i-1,j-k} + w(a_i, b_{j-k+1}, ..., b_j), 2 \leq k \leq j & \text{(d)} \\ d_{i-k,j-1} + w(a_{i-k+1}, ..., a_i, b_j), 2 \leq k \leq i & \text{(e)} \end{cases}$$

for all $1 \leq i \leq m$ and $1 \leq j \leq n$, where m is the length of the source sequence and n is the length of the target sequence. The terms on the right side respectively represent the cases of (a) deletion, (b) insertion, (c) replacement, (d) fragmentation and (e) consolidation. Additionally, the initial conditions for the recurrence equation are are:

$$\begin{aligned} d_{i0} &= d_{i-1,j} + w(a_i, \emptyset) && \text{deletion} \\ d_{0j} &= d_{i,j-1} + w(\emptyset, b_j) && \text{insertion} \\ d_{00} &= 0 \end{aligned}$$

For two sequences a and b, consisting of m and n elements respectively, we take d_{mn} as the distance between a and b. The weight function w, defines the cost of operations (which we discussed in the previous subsections). For computing the distances between the contour and I/R sequences respectively, the terms corresponding to the cost of fragmentation and consolidation are simply left out of the recurrence equation.

4. EXPERIMENTATION

A crucial question is how the behavior of each distance measure can be evaluated. One possible approach could be to gather information about human similarity ratings of musical material, and then see how close each distance measure is to the human ratings. Although this approach would certainly be very interesting, it has the practical disadvantage that it may be hard to obtain the necessary empirical data. For instance, it may be beyond the listener's capabilities to confidently judge the similarity of musical fragments longer than a few notes, or to consistently judge hundreds of fragments. Related to this is the more fundamental question of whether there is any consistent 'ground truth' concerning the question of musical similarity (see [3] for a discussion of this regarding musical artist similarity). Leaving these issues aside, we have chosen a more pragmatic approach, in which we compared the ratings of the various distance measures, and investigate possible differences in features like discriminating power. Another criterion to judge the behavior of the measures is to see how they assess distances between phrases from the same song versus phrases from different songs. This criterion is not ideal, since it is not universally true that phrases from the same song are more similar than phrases from different songs, but nevertheless we believe this assumption is reasonably valid.

The comparison of the different distance measures was performed using 124 different musical phrases from 40

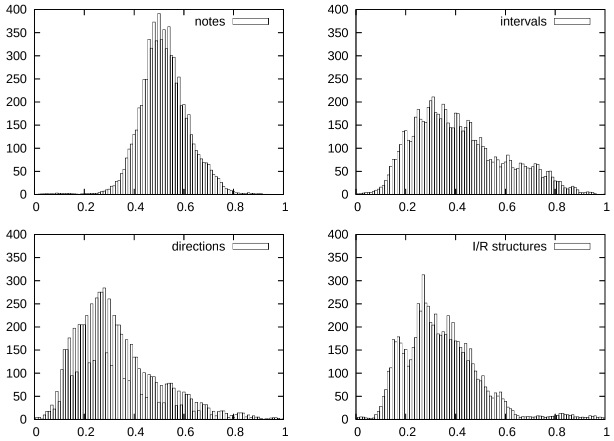

Figure 3. Distribution of distances for four melodic similarity measures. The x axis represents the normalized values for the distances between pairs of phrases. The y axis represents the number of pairs that have the distance shown on the x axis.

different jazz songs from the Real Book. The musical phrases have a mean duration of eight bars. Among them are jazz ballads like 'How High the Moon' with around 20 notes, many of them with long duration, and Bebop themes like 'Donna Lee' with around 55 notes of short duration. Jazz standards typically contain some phrases that are slight variations of each other (e.g. only different beginning or ending) and some that are more distinct. This is why the structure of the song is often denoted by a sequence of labels such as A1, A2 and B, where labels with the same letters denote phrases that are similar.

With the 124 jazz phrases we performed all the possible pair-wise comparisons (7626) using the four different measures. The resulting distance values were normalized per measure. Figure 3 shows the distribution of distance values for each measure. The results for the direction and interval measures were obtained by leaving IOI information out of the weight function (i.e. setting the k parameter to 0, see section 3.2).

The first thing to notice from figure 3 is the difference in similarity assessments at the note-level on the one hand, and the interval, direction and I/R-levels on the other hand. Whereas the distance distributions of the last three measures are more spread across the spectrum with several peaks, the note level measure has its values concentrated around one value. This suggests that the note-level measure has a low discriminatory power. We can validate this by computing the entropy as a measure of discriminatory power: Let $p(x)$, $x \in [0,1]$ be the normalized distribution of a distance measure D on a set of phrases S, discretized into k bins, then the entropy of D on S is

$$H(D) = -\sum_{0}^{1} p(k) \ln p(k)$$

where $p(k)$, is the probability that the distance between a pair of phrases is in bin k. The discriminatory power for note, interval, direction and I/R measures are then 4.41, 5.27, 5.12 and 4.91, respectively.

An interesting detail of the note measure distribution is a very small peak between 0 and 0.2 (hard to see in the plot). More detailed investigation revealed that the data points in this region were comparisons between 'partner' phrases of the same song (e.g. the A1 and A2 variants). This peak is also observable in the I/R measure, in the range $0 - .05$, In the interval and direction measure the peak is 'overshadowed' by a much larger neighboring peak. This suggests that the note and I/R measures are better at separating *very much resembling* phrases from *not much resembling* phrases than the interval and direction measures. However, the note measure lacks a subtle assessment necessary for separation of the *not-much-resembling* category into sub-categories.

The interval, direction and I/R measures seem to have higher discriminatory power for phrases that are not near-identical. In particular, the various peaks in their distribution are evidence that these measures cluster the phrases in some way, since the within-cluster comparisons produce an accumulation of low-distance values, and the between-cluster comparisons of the various clusters produce peaks at higher distance values, depending on how close the clusters are to each other.

The distributions of the interval and direction measures in figure 3 show the assessments that did not include any kind of rhythmical/temporal information. Contour representations that ignore rhythmical information are sometimes regarded as too abstract, since this information may be regarded as an essential aspect of melody [13, 14]. Therefore, we tested the effect of weighing the inter-onset time intervals (IOI) on the behavior of the interval and distance measures. Increasing the weights of IOI did improve the ability to separate *within-song* comparisons (comparing phrases from the same song) from the *between-song* comparisons (comparing phrases from different songs). However, it decreased the discriminatory power of the measures. In figure 4, the distance distributions of the direction measure are shown for different weights of IOI. Note that, as the IOI weight increases, the form of the distribution smoothly transforms from a multi-peak form (like those of the interval, direction and I/R measures in figure 3), to a single-peak form (like the note-level measure in figure 3). That is, the direction level assessments with IOI tend to resemble the more concrete note level assessment, with a degradation in discriminatory power from 5.12 for $k = 0$ to 4.81 for $k = 2$. A similar effect was observed for the interval measure.

This shows that taking into account rhythmic information in a straight-forward manner (by weighing the IOI's in calculating the edit-distance), decreases the discriminating power of the direction and interval measures for the set of musical phrases under consideration. In this respect, the I/R measure is an interesting alternative, since it does abstract from the literal musical surface, but at the same time rhythmical information is not completely ignored.

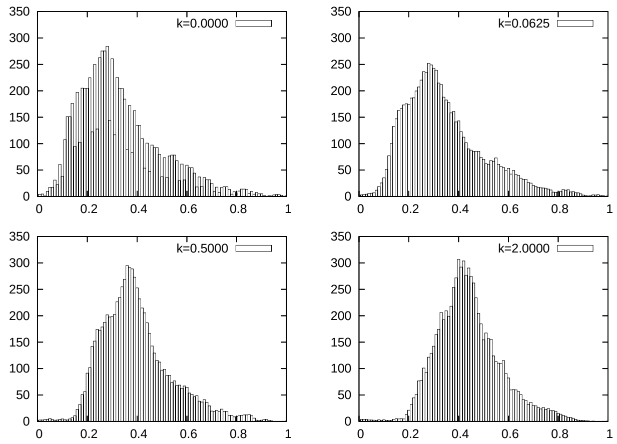

Figure 4. Distributions of distances of the direction measure for various weights of inter-onset intervals.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a new way of assessing melodic similarity and compared it with existing methods for melodic similarity assessment, using a dataset of 124 jazz phrases from well known jazz songs.

The discriminatory power (using an entropy based definition) on the whole dataset was highest for the (most abstract) contour and I/R level measures and lowest for the note level measure. This suggests that abstract melodic representations serve better to differentiate between phrases that are not near-identical (e.g. phrases belonging to different musical styles) than very concrete representations. It is conceivable that the note-level distance measure is too fine-grained for complete musical phrases and would be more appropriate to assess similarities between smaller musical units (e.g. musical motifs).

The experimentation also showed that the note and I/R level measures were better at clustering phrases from the same song than the contour (i.e. interval and direction) level measures. This might be due to the fact that rhythmical information is missing in the contour level measures. Taking into account this information (by weighting the IOI values in the edit operations) in the contour level measures improved their ability separate *within-song* comparisons from *between-song* comparisons, at the cost of discriminatory power on the whole dataset.

It may be concluded that the distance measure based on I/R representations is a good compromise between very concrete and very abstract melodic representations. It incorporates rhythmic information in an implicit way, allowing the measure to separate *within-song* comparisons from *between-song* comparisons, while maintaining its discriminative power on assessments that involve more diverse musical phrases.

In the future, we wish to investigate the usefulness of the similarity measures to cluster phrases from the same musical style. Some initial tests indicated that in particular the contour and I/R measures separated bebop style phrases from ballads. Possibly, further categorizations can also be made. However, for definitive conclusions in this direction, more research (with explicitly labeled data) is needed.

6. REFERENCES

[1] David Cope. *Computers and Musical Style*. Oxford University Press, 1991.

[2] W. J. Dowling. Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, 85(4):341–354, 1978.

[3] D. P. W. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence. The quest for ground truth in musical artist similarity. In *Prooceedings of the 3rd International Conference on Music Information Retrieval*. ISMIR, 2002.

[4] D. Hörnel and W. Menzel. Learning musical structure and style with neural networks. *Computer Music Journal*, 22 (4):44–62, 1998.

[5] K. Koffka. *Principles of Gestalt Psychology*. Routledge & Kegan Paul, London, 1935.

[6] W. Köhler. *Gestalt psychology: An introduction to new concepts of modern psychology*. Liveright, New York, 1947.

[7] Kjell Lemström and Sami Perttu. Semex - an efficient music retrieval prototype. In *First International Symposium on Music Information Retrieval (ISMIR'2000)*, Plymouth, Massachusetts, October 23-25 2000.

[8] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.

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

[10] E. Narmour. *The Analysis and cognition of basic melodic structures : the implication-realization model*. University of Chicago Press, 1990.

[11] E. Narmour. *The Analysis and cognition of melodic complexity: the implication-realization model*. University of Chicago Press, 1992.

[12] P.Y. Rolland. Discovering patterns in musical sequences. *Journal of New Music Research*, 28 (4):334–350, 1999.

[13] J. Schlichte. Der automatische vergleich von 83.243 musikincipits aus der rism-datenbank: Ergebnisse - nutzen - perspektiven. *Fontes Artis Musicae*, 37:35–46, 1990.

[14] R. Typke, P. Giannopoulos, R. C. Veltkamp, F. Wiering, and R van Oostrum. Using transportation distances for measuring melodic similarity. In *Prooceedings of the 4th International Conference on Music Information Retrieval*. ISMIR, 2003.