METHODOLOGICAL CONSIDERATIONS IN STUDIES OF MUSICAL SIMILARITY

Hamish Allan, Daniel Müllensiefen, Geraint Wiggins
Goldsmiths, University of London
New Cross, London SE14 6NW
{hamish.allan,d.mullensiefen,g.wiggins}@gold.ac.uk

ABSTRACT

There are many different aspects of musical similarity [7]. Some relate to acoustic properties, such as melodic [5], rhythmic [10], harmonic [9] and timbral [2]. Others are bound up in cultural aspects: artists involved in creation, year of first release, subject matter of lyrics, demographics of listeners, etc. In judgments about musical similarity, the relative importance of each of these aspects will change, not only for different listeners, but also for the same listener in different contexts [11]. Extra care must therefore be taken when designing studies in musical similarity to ensure that the context is an explicit variable. This paper describes the methodology behind our work in context-based musical similarity; introduces a novel system through which users can specify by example the context and focus of their retrieval needs; and details the design of a study to find parameters for our system which can also be adapted to test the system as a whole.

1 INTRODUCTION

Though musical similarity is multi-faceted, it is sometimes considered useful to distill it into a single measure; for example, to present a simple ordered list of results from a search-engine-style query. There are various measures for different aspects of similarity in the literature (for a full exploration of these, see [7]): melodic and timbral measures have generally received the most attention, but rhythmic and harmonic ones have also been considered, and metadata such as artist, lyrics, year of release, sales figures, chart position and label classification may also be examined. A single measure might combine several of these, but the relative weighting of the components of such a combination has a great impact of the utility of the measure. For a start, the perceptual prominence of aspects may vary across listeners; in addition, the context of the query exercise has an impact on which aspects are perceived as most salient [11].

We present a novel technique for allowing users to create example sets of pieces of music which are used to determine a weighting for various existing similarity measures. This allows for users lacking a musical vocabulary (or unwilling to constrict their query into one) to make queries like, "I think all these pieces of music are similar in a way that appeals to me. Find me music that is similar to them in the same way." The technique also incorporates a negative example set and a feedback loop through which the user can train the aspect more accurately (see Figure 1). We believe this dynamic approach addresses the subjectivity and context issues whilst retaining a good balance between expressive power and ease of use. It differs fundamentally from almost all existing approaches in the genre classification and similarity retrieval literature, in which parameters are fixed after an initial training phase rather than being iteratively tailored to the needs of individual users.

The technique only provides a relative weighting for the individual measures. Therefore, we also propose an experimental study to determine one or more initial weightings according to their perceptual salience for an average user (or several classes of user according to a clustering exercise). This is closely related to the task of weighting features to create a perceptually valid similarity measure [6]. However, when the separation of different aspects is considered, certain problematic assumptions are revealed which must be carefully examined. These assumptions, and a rigorous method which overcomes them, are the primary topic of this paper.

2 CONCEPTUAL OUTLINE OF ASPECT WEIGHTING

To demonstrate the concept of aspect weighting, we present an example using four distance measures and a corpus of 20 pieces of music. (Our full study will actually use a database of over 14,000 MIDI-encoded western pop songs and over 40 distance measures.) The selection and normalisation of these measures will be covered in detail in the next section, once we have explained our motivation.

Figure 2 shows two-dimensional projections of the example measures. For instance, these data might correspond to melodic (D_1), rhythmic (D_2), harmonic (D_3) and timbral (D_4) similarities. The projection itself is merely for demonstration: we assume only the pairwise distances between songs, rather than any of the underlying features. The shorter the distance, the greater the similar-

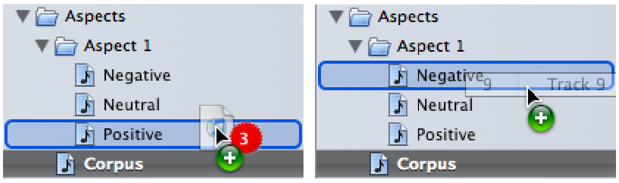

Figure 1. Choosing positive (left) and negative (right) examples by adding tracks to a playlist.

ity, with zero distance between identical pieces. Most of the measures will be metric spaces, though this does not necessarily have to be the case.

In our example, the user decides that tracks 2, 4 and 6 are all similar in her chosen aspect. A straightforward way for her to indicate this might be to drag them from a library of titles into a playlist marked “positive” (see Figure 1). Intuitively, this means that D_1 and D_3, in which those tracks are closer together than in D_2 and D_4 (see the solid thick lines in Figure 2) should be weighted more heavily. One way of achieving this is to divide the mean distance between all the tracks in D_k by the mean distance between only the tracks in D_k in the positive example set (i.e., dividing the mean of all the line lengths by the mean of only the three solid thick line lengths for each D_k in Figure 2). To avoid division by zero, a small constant should be added to both means. In this example, this gives an overall weighting for $D(+\{2,4,6\},-\phi)$ of $3.000 \cdot D_1 + 1.536 \cdot D_2 + 3.312 \cdot D_3 + 1.347 \cdot D_4$. According to this weighting, the nearest track on average to tracks 2, 4 and 6 is track 9. However, let us imagine that when we report this to the user, she puts track 9 into the negative example set. This means that our weighting does not reflect the aspect she had in mind. We need to incorporate the negative example set, for example by multiplying our existing weighting for D_k by the mean distance between all [positive, negative] pairs in D_k (i.e., the mean length of the three dashed lines for each D_k in Figure 2). Contrary to initial appearances, D_4 is now a much better measure for the aspect the user has in mind. In this example, the overall weighting $D(+\{2,4,6\},-\{9\})$ is now $0.462 \cdot D_1 + 0.270 \cdot D_2 + 0.492 \cdot D_3 + 1.030 \cdot D_4$. According to this weighting, the track nearest to tracks 2, 4 and 6 and furthest from 9 on average is now track 17 (the system can of course report more than one closest track). The user can continue to add positive and negative tracks to tune her description of the aspect she is interested in. We believe that this feedback cycle is a good way to determine the subjective and context-based similarity needs of an average user lacking a comprehensive agreed musical vocabulary. The authors are aware of one other piece of work [4] using relevance feedback in a similar way, but which requires both positive and negative initial examples and has a more complicated feedback cycle. There are, however, commercial recommendation systems based on collaborative filtering (e.g. Last.fm, Pandora) which use a thumbs-up / thumbs-down style of interface.

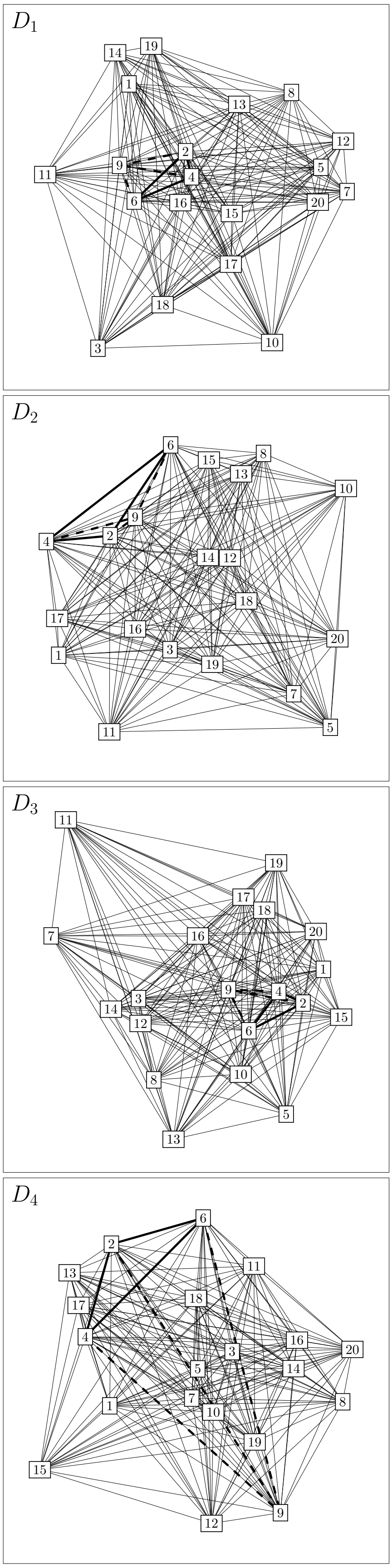

Figure 2. Two-dimensional projections of example distance measures.

3 DESIGN OF USER STUDY TO DETERMINE MUSICAL SIMILARITY

The example in Section 2 gives the distance measures equal weighting before the example sets are taken into account. However, there is no reason to assume that they all relate to aspects with equal perceptual salience. We have therefore designed an experimental study to determine an initial weighting for an average user. Although the balance of salient aspects may in fact be unique to each listener, an initial weighting derived from a study is likely to be better than a flat equal weighting. The results from the study may form clusters, in which case modelling more than one type of average user may be beneficial.

There is also the question of which distance measures to select from the literature. Measures for which the task is more closely specified (e.g., "similarity in melodic contour" or "similarity in rhythmic complexity") are likely to have more perceptually reliable experimental validation, so we select those in preference to measures claiming 'generic' similarity. We select as many measures as we can, with the proviso that we must be able to find a sampling of the corpus such that every measure has a similar distribution of pairwise distances in the sample to that it has in the whole corpus. Because the sampling of a subset of the corpus is not independent from the whole set, we compare the distributions by taking the difference between the cumulative distribution functions for the whole set and the subset (see Figure 3; for more about using CDF to compare distribution, see [1]). The size of the sample is limited by the pragmatics of the user study as described below. We must also be able to normalise each measure so that the linear combination described above makes sense; as the pairwise distances are never negative, this can be achieved by dividing each measure's distances by their mean to give a range of values from 0 upwards with a mean of 1, giving each measure a similar scale whilst also allowing outliers to be incorporated in the example sets. Linear combinations of normalised measures may still be problematic if the percept to which the measure corresponds is non-linear or if the measures are not fully independent; further research needs to be done to determine the optimal combination. Also, we may have to restrict the number of measures selected, to avoid overfitting.

One way in which pieces of music can be similar is if they have the same large-scale structure; for example, quiet verse followed by loud chorus. However, we will be presenting short excerpts rather than whole pieces to subjects, and for the sake of simplicity the excerpts we choose will be largely self-similar (self-similarity being assessed using the same distance measures as those selected for the study).

We need from the user study an empirical distance measure to estimate each w_k:

$$D_{study} = \epsilon + \sum_{k=1}^{n} (w_k \cdot D_k) \tag{1}$$

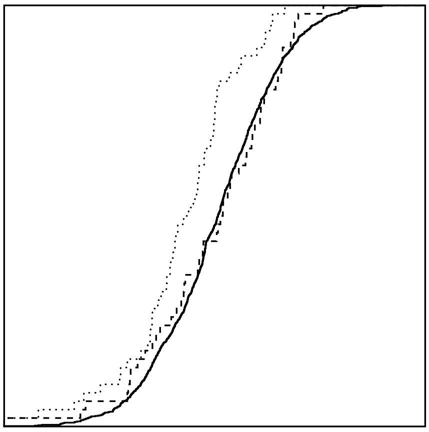

Figure 3. Comparing distributions using the cumulative density function. The solid line is the CDF of the parent distribution; the dashed line is the CDF of a much more representative sub-sample than that of the dotted line. The area between each curve can be taken to be the distance between each distribution.

for which the error term ϵ can be minimised. There are various empirical approaches that can be taken to obtain a complete pairwise distance matrix for D_{study}, but we choose the method of triadic comparisons [12] because it involves discrete decisions from subjects rather than use of a continuous scale, which means that subjects are less likely to have changed their calibrations as the experiment proceeds.

In this triadic comparison, excerpts from three pieces of music are played to the subject who is subsequently asked to judge which two are most similar. We have implemented a browser-based interface for this, with a graphical control designed to minimise visual biasing (see Figure 4). The control changes to reflect which piece the subject is currently listening to and also to indicate how to go about indicating their judgment. Subjects must listen to all three pieces in order; if they wish to listen again, they must listen to all three again, in the same order. The interface is connected to a back-end which records the subject's choices in a database; it also ensures that the triads are apportioned to subjects according to the methodological design principles described below.

For a complete block design (CBD) with n tracks, the number of different permutations is $n(n-1)(n-2)$. Let us suppose that a subject can listen to an excerpt from each of three tracks and make a judgement about their relative similarities within 30 seconds. This equates to 120 triads in an hour (with no time for breaks), which is the full permutations for $n = 6$ tracks. For $n = 8$, we have 336 triads or 5.6 hours. Clearly as the number of tracks rises, the number of triads quickly exceeds what a subject might reasonably be expected to tackle.

In order to overcome this combinatorial explosion, Novello et al [6] suggest the use of a balanced incomplete block design (BIBD). Firstly, no combination of three pieces is reordered and presented in more than one per-

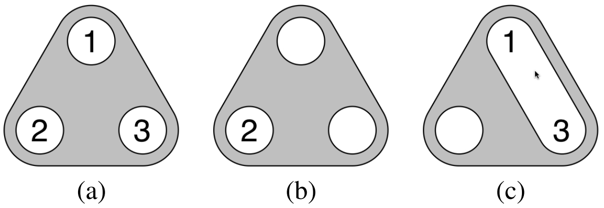

Figure 4. Main control from web interface to empirical study. (a) Before and after playback the control shows three numbers with equal emphasis. (b) During playback the control shows which piece is playing, e.g., piece 2 in this example. (c) After playback the user is asked to choose which two are most similar: here the mouse is hovering between tracks 1 and 3 and the control has changed to indicate that clicking here will select 1 and 3 as being most similar.

mutation. Secondly, only a subset of such combinations are presented: instead of presenting each pair of tracks $6(n-2)$ times (i.e., each pair alongside each other track), each pair is only presented λ times. Novello et al suggest $\lambda = 2$ which gives 102 triads for $n = 18$, as opposed to 4896 triads for the CBD. The tracks presented alongside each pair are balanced as far as possible, and the presentation order is varied to compensate for presenting only a single permutation of each triad: for a full description, see [12].

However, we believe that these steps are insufficient for the task at hand. Tversky [11] argues that presentation order alters the context in which similarity is judged. When a subject is asked to judge the similarity of tracks A, B and C, in that order, after hearing tracks A and B he has their common features in mind, and is likely to judge C along those same lines; whereas if he had heard B, C and A he would be judging A against the common features of B and C. Therefore it is not inconsistent for triads ABC and BCA to be given different judgments by the same subject. Although Novello et al try to balance the presentation order as far as possible, we believe that a full balancing for the incomplete design can only be achieved through knowledge of the relative perceptual salience of the aspects shared by the first two tracks in each triad, which is not known before the study (a circular dependency: Catch-22). Tversky also details other types of asymmetry, e.g., one piece may be considered a referent (such as an older or more well-known piece). Furthermore there is an inherent asymmetry in presenting a triad because the first and the third tracks are necessarily separated in time by the second, whereas the other two pairs run onto each other (even if the subject repeats listening to the triad immediately thereby running the third and first pieces together, there will still be a bias towards the other pairs). Therefore we propose that balancing can only really be achieved by presenting all the permutations of each combination of triad.

Similarly, the incomplete block design for $\lambda < n - 2$ also leaves gaps which cast doubt on the validity of complete pairwise distances being calculated from the study data. For example with $\lambda = 2$, if each of the two tracks presented alongside a given pair are similar in any aspect to at least one of the tracks in that pair, the pair will be judged completely dissimilar regardless of its actual similarity. That is, if the pair AB is presented in the contexts of C and D, if AC is judged the most similar of ABC and BD the most similar of BCD, A and B are taken to be have zero similarity, whereas if they had been presented in further contexts (E, F, etc.) they might have proved similar to some extent. Conversely, if AB and CD are very dissimilar, BC will receive the maximal similarity rating even if B and C are not particularly similar. Only a complete block design provides for enough granularity to construct a complete pairwise distance matrix.

As previously mentioned, for any number of pieces n more than just a handful, it is unreasonable to expect a single subject to tackle the complete block design. We therefore propose partitioning the block over multiple subjects. This is not ideal on account of the possibility of different listeners favouring different aspects, but as we are attempting to capture average listener characteristics, the compromise is justified. We should certainly be able to do better than a flat equal initial weighting.

The partitioning must also be balanced to maximise the return of information gathering (although redundancy for subject consistency testing must be re-incorporated later). The most obvious partitioning involves six subjects, each with a different permutation of each combination (see Figure 5). We can also partition each row further with the constraints that each piece is presented the same number of times to each user (see Figure 6). However, we can improve the balance by ensuring that there is no correlation between pieces and their positions in the triads (see Figures 7 and 8). We refer to this design as the Balanced Complete Block Partitioning (BCBP).

Our implementation of BCBP uses best-first search [8] to allocate a full combination of triads into groups as if for a single combination, according to the second balancing constraint described above (see Figure 9), then allocates permutations according to the third constraint.

Splitting a complete block design for $n = 18$, giving each subject 102 triads (51 minutes) each requires 48 subjects. Note that a fully balanced BCBP partitioned over 48 subjects is equivalent to a BIBD with $n = 18$ and $\lambda = 2$ given to 48 subjects if the BIBD is balanced between subjects as well as within them. (Novello et al use 36 subjects which gives a maximum coverage of 0.75)

Only some group sizes are suitable for balancing. The constraints are as follows: the combinations must be divisible into groups of equal size (Equation 2), and each group must contain an equal number of occurrences of each piece (Equation 3).

$$\binom{n}{3} \bmod g = 0 \qquad (2)$$

$$\frac{\binom{n}{3}}{g} \bmod n = 0 \qquad (3)$$

```
ABC ABD ABE ACD ACE ADE BCD BCE BDE CDE
ACB ADB AEB ADC AEC AED BDC BEC BED CED
BAC BAD BAE CAD CAE DAE CBD CBE DBE DCE
BCA BDA BEA CDA CEA DEA CDB CEB DEB DEC
CAB DAB EAB DAC EAC EAD DBC EBC EBD ECD
CBA DBA EBA DCA ECA EDA DCB ECB EDB EDC
```

Figure 5. Full permutations for five pieces. Partitioning for six subjects is straightforward: one row for each subject, ten pieces each.

```
ACD ABE ACE BCD BDE   ABD ABC ADE BCE CDE
ADC AEB AEC BDC BED   ADB ACB AED BEC CED
CAD BAE CAE CBD DBE   BAD BAC DAE CBE DCE
CDA BEA CEA CDB DEB   BDA BCA DEA CEB DEC
DAC EAB EAC DBC EBD   DAB CAB EAD EBC ECD
DCA EBA ECA DCB EDB   DBA CBA EDA ECB EDC
```

Figure 6. Full permutations for five pieces, partitioned for twelve subjects, balanced so that each piece appears three times for each subject.

```
ABC DAE CEA BCD EDB   ABD CDA EAB BCE DEC
ACB DEA CAE BDC EBD   ADB CAD EBA BEC DCE
BAC ADE ECA CBD DEB   BAD DCA AEB CBE EDC
BCA AED EAC CDB DBE   BDA DAC ABE CEB ECD
CAB EDA ACE DBC BED   DAB ACD BEA EBC CDE
CBA EAD AEC DCB BDE   DBA ADC BAE ECB CED
```

Figure 7. Balanced Complete Block Partitioning (BCBP). Full permutations for five pieces, partitioned for twelve subjects, balanced so that each piece not only appears three times for each subject, but also appears once in each position.

```
ADF EAC DCB BFE   ABD FAC CEB DFE   AEB DAE CBF FCD
AFD ECA DBC BEF   ADB FCA CBE DEF   ABE DEA CFB FDC
DAF AEC CDB FBE   BAD AFC ECB FDE   EAB ADE BCF CFD
DFA ACE CBD FEB   BDA ACF EBC FED   EBA AED BFC CDF
FAD CEA BDC EBF   DAB CFA BCE EDF   BAE EDA FCB DFC
FDA CAE BCD EFB   DBA CAF BEC EFD   BEA EAD FBC DCF

          ACD BAF DEB EFC   ABC EAF BFD CDE
          ADC BFA DBE ECF   ACB EFA BDF CED
          CAD ABF EDB FEC   BAC AEF FBD DCE
          CDA AFB EBD FCE   BCA AFE FDB DEC
          DAC FBA BDE CEF   CAB FEA DBF ECD
          DCA FAB BED CFE   CBA FAE DFB EDC
```

Figure 8. Balanced Complete Block Partitioning (BCBP). Full permutations for six pieces, partitioned into thirty groups, balanced so that each piece appears twice in each group. Note that although each piece cannot appear once in each position as per Figure 7, the positioning is nonetheless balanced; e.g., in the bottom right hand group, pieces D and F both appear in positions 1 and 2, pieces A and B in positions 2 and 3, and pieces C and E in positions 1 and 3.

```
1. Distribute all combinations of $n$ pieces over $g$ groups by adding one to each group in turn from an alphabetically ordered list of all combinations until the list is exhausted. This produces a partitioning closer to being balanced than if the first $n$ pieces had gone to the first group, the second $n$ to the second group, and so on. Create an initial node for a priority queue, with this grouping as the node's state and with the average variance of the number of times each piece appears in each group as its score.

2. Remove a node from the head of the queue and perform the following on it:
       For each group $G_i$ $(i = 1 \dots g)$:
           Count the number of each piece in the group
           For each triad $T_a$ containing each piece appearing more than the average number of times:
               For each other group $G_j$ $(j = 1 \dots g, i \neq j)$:
                   Count the number of each piece in the group
                   For each triad $T_b$ containing each piece appearing fewer than the average number of times:
                       Create a copy of the node with $T_a$ and $T_b$ exchanged, and add it to the priority queue with its score as described in the step above.
   Repeat this step until the score of the node at the head of the queue is zero (i.e., its grouping is fully balanced).
```

Figure 9. Algorithm for balancing a complete block partitioning, based on a priority queue [3]. This algorithm balances the partitioning of the unordered combination of n pieces over g groups (i.e., a single row in Figure 6). A similar algorithm (not detailed here) is employed to balance the ordering of pieces within each triad (as per Figures 7 and 8).

A third constraint may be added to ensure perfect balancing: that for each group, each piece appears exactly the same number of times in each position (Equation 4).

$$\frac{3\binom{n}{3}}{n \cdot g} \bmod 3 = 0 \tag{4}$$

However, a partitioning can still be considered balanced if each position is occupied an equal number of times by an equal number of pieces (see Figure 8).

To test for consistency, each subject should also be given a number of triads (e.g. 5) more than once (within-subject consistency) and a number of common triads (e.g. 5) should be given to every subject (between-subjects consistency). These courses of action will violate the balancing constraints, but are important for determining the admissibility of results. The extra triads will be injected into each subject's partitions roughly equally spaced throughout.

The BCBP method is by no means limited to the study for our initial weighting; it is suitable for any task to which a BIBD might be applied, such as attempting to determine a ground truth. We are also intending to use a triadic design based on the same web interface for the evaluation of the aspect weighting system described in Section 2.

4 CONCLUSION

We have presented an examination of the methodology of studies in musical similarity, and outlined why incomplete block designs for user studies, even if balanced, may not capture enough information for determining a complete pairwise distance measure. We have described the Balanced Complete Block Partitioning which is designed to address these issues and described how a study based on this design might be implemented. We have also presented a novel system for context-based specification of musical similarity and a user-centric interface for such a system. We are currently undergoing a study to parameterise this system using pieces from our database of more than 14,000 MIDI-encoded pieces of western popular music, the results of which will be presented at ISMIR 2007 and published in a later paper.

ACKNOWLEDGEMENTS

The first author is supported by a Doctoral Bursary from the Department of Computing at Goldsmiths, University of London. The second author is supported by EPSRC grant EP/D038855/1.

5 REFERENCES

[1] Hamish Allan and Geraint Wiggins. Further aspects of similarity. In *Proceedings of the 2nd Digital Music Research Network Summer Conference*, 2006.

[2] Jean-Julien Aucouturier. *Ten Experiments on the Modelling of Polyphonic Timbre*. PhD thesis, University of Paris 6, Paris, France, May 2006.

[3] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*, chapter 6.5, pages 138–142. MIT Press and McGraw-Hill, 2nd edition, 2001.

[4] Michael I. Mandel, Graham E. Poliner, and Daniel P.W. Ellis. Support vector machine active learning for music retrieval, 2006.

[5] Daniel Müllensiefen and Klaus Frieler. Optimizing measures of melodic similarity for the exploration of a large folk song database. In *Proceedings of the Fifth Annual International Symposium on Music Information Retrieval: ISMIR 2004*. Universitat Pompeu Fabra, 2004.

[6] Alberto Novello, Martin F. McKinney, and Armin Kohlrausch. Perceptual evaluation of music similarity. In *Proceedings of the Seventh Annual International Symposium on Music Information Retrieval: ISMIR 2006*. University of Victoria, British Columbia Canada, 2006.

[7] Elias Pampalk. *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Vienna University of Technology, Vienna, Austria, March 2006.

[8] J Pearl. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, 1984.

[9] Jeremy Pickens. *Harmonic Modeling for Polyphonic Music Retrieval*. PhD thesis, University of Massachusetts Amherst, MA, USA, May 2004.

[10] Godfried Toussaint. A comparison of rhythmic similarity measures. In *Proceedings of the Fifth Annual International Symposium on Music Information Retrieval: ISMIR 2004*. Universitat Pompeu Fabra, 2004.

[11] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[12] Susan C. Weller and A. Kimball Romney. *Qualitative Research Methods: Systematic Data Collection*. 10. Sage Publications, California, 1988.