The inaugural International Conference on Music Communication Science 5-7 December 2007, Sydney, Australia
http://marcs.uws.edu.au/links/ICoMusic

TOWARDS PHRASE STRUCTURE RECONSTRUCTION FROM EXPRESSIVE PERFORMANCE DATA

Maarten Grachten *Gerhard Widmer*

Department of Computational Perception

Johannes Kepler University, Linz, Austria

ABSTRACT

Using a simple pattern finding approach, we investigate to what degree patterns found in the tempo and loudness curves measured from piano performances of a classical piece coincide with repeated musical structures in the score that was performed. We show that high frequency content in such curves is more useful for finding repetitions of musical structures than low frequency content. In some cases removing low frequency content even improves the accuracy of pattern finding.

1. INTRODUCTION

It is commonly asserted that a primary function of musical expression is to clarify the structure of the music that is being played [3,5]. And indeed, countless studies of expressive music performance find that structural aspects of the musical score are in some way or another reflected in the expressive information that is extracted from performances. One such aspect is phrase structure, which is typically marked by a decrease of both tempo and loudness at phrase boundaries [9].

The observation that phrase structure is reflected in expressive tempo and loudness information as measured from performances, raises the question whether it would be possible to recognize the phrase structure by merely observing expressive information (and not, for example, pitch, or rhythmic information). This would form a complementary approach to studies that investigate regularities in expressive data in a score-driven way (for example [6]). From a practical point of view, performance-based phrase structure reconstruction could provide additional cues to systems that try to infer the structure of music pieces from, e.g., scores or MIDI files. Furthermore, applications such as score-following/automatic page turning could benefit from phrase-structure recognition.

An apparently discouraging argument against the endeavor of reconstructing phrase structure from expressive information is that a musician is by no means obliged to play repeated parts of the score in a similar way (a phenomenon termed `consistency' in [4]). It can even be argued that playing repeated parts in different ways is one of the aspects that make human performances intriguing. In practice however, there is often considerable agreement between the performance of repeated parts [6].

In this paper we investigate to what degree the phrase structure of a piece is reflected in the tempo and loudness information measured from performances of the piece. We do this by measuring how well patterns found in the tempo and loudness curves coincide with the phrase structure, more specifically melodic gestures, relatively small musical constructs (typically containing less than ten notes). Rather than determining the precise beginnings and endings of phrases and melodic gestures, our first goal is to establish which parts of the piece are repeated, and where. We measure accuracy in terms of how many of the instances of the pattern span repeated melodic gestures (precision), and how many repeated melodic gestures are identified as instances of the same pattern (recall). Obviously, a phrase structure reconstruction is not correct if the boundaries of the melodic gestures are not correct, but we believe that if repeated parts of the score are identified largely correctly, a useful step towards phrase structure reconstruction has been made.

2. RELATED WORK

It is undisputed that phrase structure is one of the factors that determine the expressive features of performance. Nevertheless, most of the work on automatic pattern finding and recognition of structure in music pieces has focused on score information. The pattern finding problem is often conceptually divided into a segmentation step, in which the boundaries of musical compounds are determined, and a clustering step, in which the delimited segments are grouped by identity, similarity or any other musically meaningful relation. Some approaches just deal with the segmentation problem [9,1], others deal with the clustering problem [2], or with both [8].

The strategy we present in this paper, as said before, deals with performance information rather than score information. Furthermore, no prior segmentation of the data is used. Instead, our algorithm considers all non-overlapping pairs of equally long subsequences as possible instances of a single pattern. In this sense, our approach is related to that of [4], in which patterns found in expressive information are used to characterize the degree to which performers play repeated parts similarly.

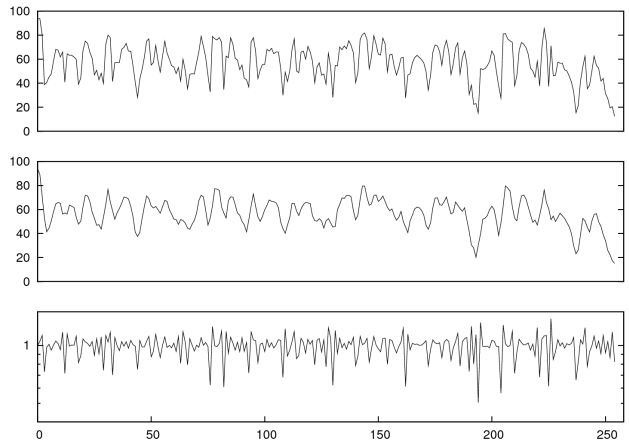

Figure 1: Decomposition of a tempo curve (from Schumann's Träumerei, performed by Horowitz, 1987) into slow and fast fluctuations. Top: Original tempo (in BPM); Middle: Smoothed Tempo (BPM); Bottom: Residual as a proportion of smoothed tempo (log scale)

towards the end of the phrase. Although this might facilitate finding the beginnings and endings of phrases, it possibly makes distinguishing phrases more difficult, since the tempo and loudness curves of distinct phrases have their overall parabolic form in common.

Assuming that identifying distinct phrases in a piece is hindered by the parabolic component they have in common, an obvious solution is to fit a set of second order polynomials to the tempo and loudness curves on the interval of each phrase and subtract these from the original curves (as in [10]). However, such an approach might introduce a bias towards the structure present in the score, a danger of the score-driven approach that we wish to avoid. As a simple non score-driven alternative, we apply a low-pass filter to the curves. The low-pass filtered curve contains only the lower frequency content. When this curve is subtracted from the original curve, the residual thus contains just the high frequency content. Most of the parabolic component will be contained in the low frequency curve. An example of a tempo curve and its low and high frequency components is shown in figure 1. A three point moving average filter is used as a low-pass filter both in the example and in the experiments. The peaks in the residual correspond to the sides of the parabolic forms, the points at which the original curve shows rapid changes.

3. METHOD

In this section we report the setup of an experiment in which we apply a pattern finding algorithm to expressive performance information in order to find repeated musical structures. We compare the results under three conditions: 1) using the original tempo and loudness curves, 2) using only the low-frequency content of the tempo and loudness curves, and 3) using only the high frequency content.

Data

The performance data used here stems from six performances of Schumann's piece ``Träumerei'' by renown pianists. The piece is played by Argerich, Kempff, Brendel, and Horowitz (of whom three different recordings of the piece are included). For each performance, instantaneous tempo and loudness information at half beat level is available (based on semi-automatic beat-tracking, cf. [12]). Thus, the expressive performance information extracted from an audio recording is represented as a chronological sequence of pairs of tempo and loudness values, where each pair corresponds to a half beat position in the score. The total sequence consists of 255 pairs.

Decomposition of Tempo and Loudness Curves

As stated in the introduction, typically both tempo and loudness curves convey phrase structure by a slowly evolving increase and decrease over the course of a musical phrase, roughly approximating a parabolic form. That is, the phrase is started relatively slow and soft, and after growing faster and louder towards the middle of the phrase, tempo and loudness decrease

Pattern Finding

We employ a simple pattern finding approach that is based on the correlation coefficient (r) between pairs of subsequences of tempo and loudness values. Tempo and loudness curves are treated in parallel, and we define the match score of a pair of subsequences as the average of the r values for tempo and loudness (we will refer to this average as the r value of the match). After a subsequence length l and a threshold α for the r values have been fixed, the pattern finding algorithm returns a graph where the vertices are subsequences, and edges represent a match ($r \geq \alpha$) between two subsequences. We define the patterns to be the connected components in the graph.

Although the instances of a single pattern do not overlap (overlapping subsequences are excluded from matching), the instances of different patterns may overlap. When the instances of two patterns overlap pairwise by a constant offset, the two patterns can be seen as parts of a larger pattern that covers both. In this case, the two patterns are fused, so that each instance of the new pattern spans an overlapping pair of instances of the old patterns. Especially for lower α values, this reduces the number of patterns considerably. Note that as a result of this fusing, patterns may have different lengths (although the instances of a single pattern of course *do* have the same length), and that the size l that is chosen acts a *minimum* size, rather than a fixed size.

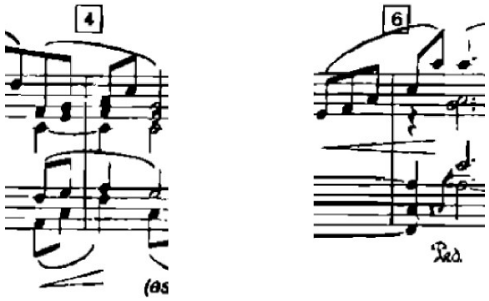

Figure 2: Two fragments from Schumann's Träumerei that were matched using tempo and loudness information

Evaluation

The patterns that are found are compared to the phrase analysis of the piece in terms of melodic gestures (MG) (that was adopted from [6]). We focus on the MG's in the soprano voice. For this voice, the piece has eight distinct MG's, most of which occur several times throughout the phrases. We evaluate repeated patterns found in the performance data by measuring how well they coincide with repetitions of the MG's.

We define the precision of a pattern as the degree of MG agreement among the instances of the pattern at each position. To this end, we define an *MGid* for each position, that is, a pair of (*MGlabel*,*offsetIntoMG)*. For example, the MGid of a position that is the first element of an instance of *MG2* would be (*MG2*,0). We define A to be the set of MGid's of all positions in the performance. The precision of a set of patterns is simply the average of the precisions per pattern, which is in turn the average of the precisions per position in the pattern. The precision at a position, finally, is the fraction of pattern instances with MGid a at that position, for the $a \in A$ that maximizes this fraction:

$$Prec = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L_n}\sum_{i=1}^{L_n}\max_{a\in A}\frac{|\{\, k \mid s_i^{k,n} = a \,\}|}{K_n},\quad 1\le k\le K_n$$

where N is the number of patterns, L_n is the length of the n-th pattern, K_n the number of instances the n-th pattern, and $s_i^{k,n}$ is the MGid corresponding to the i-th position of instance k of the n-th pattern.

Given a set of patterns, recall is defined as the average recall over all positions. Informally speaking, the recall at a position measures the largest fraction of related positions (in terms of MGid's) that is covered by a single pattern. More precisely, let a_j be the MGid of the j-th position, and let $B_j = \{\, x \mid a_{x_k,n} = a_j \,\}$ be the set of all instances of a_j. Furthermore, let $pos(s_i^{k,n})$ denote the (global) position of the i-th element of the k-th instance of the n-th pattern. The recall is then defined as:

$$Rec = \frac{1}{Q}\sum_{j=1}^{Q}\max_{\substack{1\le n\le N,\\ 1\le o\le L_n}}\frac{|B_j \cap \{\, pos(s_i^{k,n}) \mid i = o \,\}|}{|B_j|},\quad 1\le k\le K_n$$

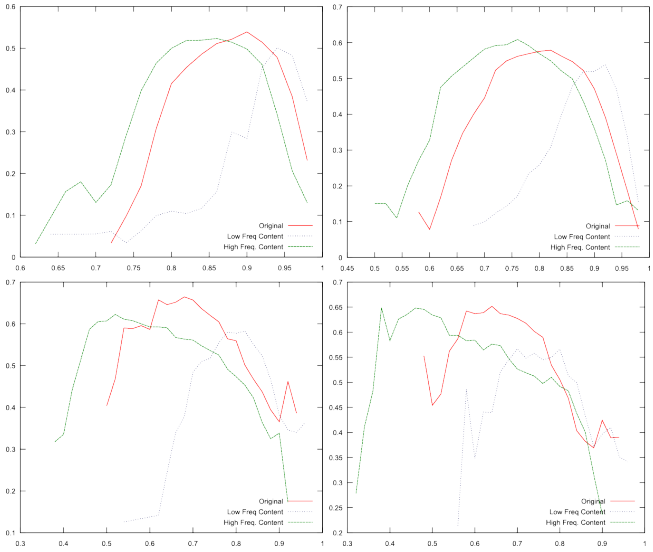

Figure 3: F-scores as a function of α, for segment sizes of 3 beats (upper left), 5 beats (upper right), 11 beats (lower left), and 15 beats (lower right)

where Q is the length of the sequence of tempo/loudness values. The interpretation of N, K_n, and L_n is as above.

Using the above definitions of precision and recall, we evaluate the overall accuracy of a set of patterns with the F-score:

$$F = 2\cdot Prec \cdot Rec \,/\, (Prec + Rec)$$

Results and Discussion

As an illustrative example, figure 2 displays two score fragments that were matched based on the tempo and loudness of the performance. Although the fragments are not instances of the same melodic gesture according to the phrase analysis, there are several interesting similarities. For example, the position in the metrical grid is the same, both fragments end in a chord, and are not immediately continued in the soprano voice, and the soprano voices in both cases are largely ascending. Also, both fragments contain a crescendo. This however was not a necessary nor a sufficient condition for the match, since other instances of the same pattern did not contain a crescendo, nor did the pattern contain all crescendos.

For each of the six recordings, we applied the pattern finding algorithm to the original tempo and loudness curves (OR), the low-frequency components (LF), and the high-frequency components (HF) respectively, using various segment sizes and α values. Figure 3 shows the F-scores (averaged over the six recordings) for each of the three curve types as a function of α, for four different segment sizes.

Unsurprisingly, the F-scores for LF and OR peak at higher *r*-thresholds than HF (regardless of segment size). This is in accordance with our hypothesis that the low frequency components make it harder to discriminate different MG's, and thus need a higher *r*-threshold. A more interesting result is that pattern finding on LF gives systematically lower performance than on the others, implying that the high frequency components of the tempo and loudness curves contain essential information for telling MG's apart. Moreover, pattern finding on HF gives results that are comparable to the results for OR, and in some cases (for example, segment size 10) even better.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have described a first step towards phrase reconstruction from expressive performance data, by detecting patterns in tempo and loudness curves. Although we have not addressed the question of finding the exact boundaries of musical phrases, we have found that repetitions of musical structures can be identified with modest success.

Moreover, our experiments show that removing the low frequency content from the tempo and dynamics curves hardly decreases, and sometimes even increases the ability to find expressive patterns that coincide with musical structures. It must be noted however that, even if six different performances were used, the current experiment covers just one musical piece. Further experiments are needed to investigate to what extent the results generalize to pieces that are performed at very regular tempos.

Lastly, although the F-score that was used for evaluation is a good indicator of the accuracy of the individual patterns that are found, it does not fully describe the accuracy of a set of patterns when interpreted as a hypothetical phrase structure. For example, it does not explicitly measure redundancy between patterns, nor incorrect phrase boundaries. More elaborate evaluation will be required to address such issues.

Acknowledgments

This work is funded by the Austrian National Science Fund (FWF) under project number P19349-N15. Sebastian Flossmann commented on draft versions of this paper.

5. REFERENCES

1. Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In Proceedings of the International Computer Music Conference (ICMC'2001), Havana, Cuba.

2. Cambouropoulos, E. and Widmer, G. (2000). Melodic clustering: Motivic analysis of Schumann's Träumerei. In Proceedings of the III Journeés d'Informatique Musicale, Bordeaux, France.

3. Clarke, E. F. (1991). Expression and communication in musical performance. In Sundberg, J., Nord, L., and Carlson, R., editors, Music, Language, Speech and Brain. MacMillan Academic and Professional Ltd.

4. Madsen, S. T. and Widmer, G. (2006). Exploring pianist performance styles with evolutionary string matching. International Journal on Artificial Intelligence Tools, 15(4):495–513.

5. Palmer, C. (1997). Music performance. Review of Psychology, 48:115–138.

6. Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. Journal of the Acoustical Society of America, 88:622–641.

7. Repp, B. H. (1995). Diversity and commonality in music performance - An analysis of timing microstructure in Schumann's "Träumerei". Journal of the Acoustical Society of America, 92(5):2546–2568.

8. Rolland, P. (1999). Discovering patterns in musical sequences. Journal of New Music Research, 28 (4):334–350.

9. Temperley, D. (2001). Cognition of Basic Musical Structures. MIT Press, Cambridge, Mass.

10. Tobudic, A. and Widmer, G. (2003). Playing Mozart phrase by phrase. In Proceedings of the Fifth International Conference on Case-Based Reasoning (ICCBR-03), nr. 2689 in Lecture Notes in Artificial Intelligence, Springer-Verlag.

11. Todd, N. (1989). A computational model of rubato. Contemporary Music Review, 3 (1).

12. Widmer, G. (2005). Studying a creative act with computers: Music performance studies with automated discovery methods. Musicae Scientiae, IX(1):11–30.