AUTOMATED METHODS FOR ANALYZING MUSIC RECORDINGS IN SONATA FORM

Nanzhu Jiang
International Audio Laboratories Erlangen
nanzhu.jiang@audiolabs-erlangen.de

Meinard Müller
International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

ABSTRACT

The sonata form has been one of the most important large-scale musical structures used since the early Classical period. Typically, the first movements of symphonies and sonatas follow the sonata form, which (in its most basic form) starts with an exposition and a repetition thereof, continues with a development, and closes with a recapitulation. The recapitulation can be regarded as an altered repeat of the exposition, where certain substructures (first and second subject groups) appear in musically modified forms. In this paper, we introduce automated methods for analyzing music recordings in sonata form, where we proceed in two steps. In the first step, we derive the coarse structure by exploiting that the recapitulation is a kind of repetition of the exposition. This requires audio structure analysis tools that are invariant under local modulations. In the second step, we identify finer substructures by capturing relative modulations between the subject groups in exposition and recapitulation. We evaluate and discuss our results by means of the Beethoven piano sonatas. In particular, we introduce a novel visualization that not only indicates the benefits and limitations of our methods, but also yields some interesting musical insights into the data.

1. INTRODUCTION

The musical form refers to the overall structure of a piece of music by its repeating and contrasting parts, which stand in certain relations to each other [5]. For example, many songs follow a strophic form where the same melody is repeated over and over again, thus yielding the musical form $A_1A_2A_3A_4....$[1] Or for a composition written in rondo form, a recurring theme alternates with contrasting sections yielding the musical form $A_1BA_2CA_3D....$ One of the most important musical forms in Western classical music is known as *sonata form*, which consists of an *exposition* (E), a *development* (D), and a *recapitulation* (R), where the exposition is typically repeated once. Sometimes, one can find an additional introduction (I) and a closing coda (C), thus yielding the form IE_1E_2DRC. In particular, the exposition and the recapitulation stand in close relation to each other both containing two subsequent contrasting subject groups (often simply referred to as first and second theme) connected by some transition. However, in the recapitulation, these elements are musically altered compared to their occurrence in the exposition. In particular, the second subject group appears in a modulated form, see [4] for details. The sonata form gives a composition a specific identity and has been widely used for the first movements in symphonies, sonatas, concertos, string quartets, and so on.

In this paper, we introduce automated methods for analyzing and deriving the structure for a given audio recording of a piece of music in sonata form. This task is a specific case of the more general problem known as *audio structure analysis* with the objective to partition a given audio recording into temporal segments and of grouping these segments into musically meaningful categories [2, 10]. Because of different structure principles, the hierarchical nature of structure, and the presence of musical variations, general structure analysis is a difficult and sometimes a rather ill-defined problem [12]. Most of the previous approaches consider the case of popular music, where the task is to identify the intro, chorus, and verse sections of a given song [2, 9–11]. Other approaches focus on subproblems such as audio thumbnailing with the objective to extract only the most repetitive and characteristic segment of a given music recording [1, 3, 8].

In most previous work, the considered structural parts are often assumed to have a duration between 10 and 60 seconds, resulting in some kind of medium-grained analysis. Also, repeating parts are often assumed to be quite similar in tempo and harmony, where only differences in timbre and instrumentation are allowed. Furthermore, *global* modulations can be handled well by cyclic shifts of chroma-based audio features [3]. When dealing with the sonata form, certain aspects become more complex. First, the duration of musical parts are much longer often exceeding two minutes. Even though the recapitulation can be considered as some kind of repetition of the exposition, significant local differences that may last for a couple of seconds or even 20 seconds may exist between these parts. Furthermore, there may be additional or missing sub-structures as well as relative tempo differences be-

[1] To describe a musical from, one often uses the capital letters to refer to musical parts, where repeating parts are denoted by the same letter. The subscripts indicate the order of repeated occurrences.

tween the exposition and recapitulation. Finally, these two parts reveal differences in form of *local* modulations that cannot be handled by a global cyclic chroma shift.

The goal of this paper is to show how structure analysis methods can be adapted to deal with such challenges. In our approach, we proceed in two steps. In the first step, we describe how a recent audio thumbnailing procedure [8] can be applied to identify the exposition and the recapitulation (Section 2). To deal with local modulations, we use the concept of transposition-invariant self-similarity matrices [6]. In the second step, we reveal finer substructures in exposition and recapitulation by capturing relative modulation differences between the first and the second subject groups (Section 3). As for the evaluation of the two steps, we consider the first movements in sonata form of the piano sonatas by Ludwig van Beethoven, which constitutes a challenging and musically outstanding collection of works [13]. Besides some quantitative evaluation, we also contribute with a novel visualization that not only indicates the benefits and limitations of our methods, but also yields some interesting musical insights into the data.

2. COARSE STRUCTURE

In the first step, our goal is to split up a given music recording into segments that correspond to the large-scale musical structure of the sonata form. On this coarse level, we assume that the recapitulation is basically a repetition of the exposition, where the local deviations are to be neglected. Thus, the sonata form IE_1E_2DRC is dominated by the three repeating parts E_1, E_2, and R.

To find the most repetitive segment of a music recording, we apply and adjust the thumbnailing procedure proposed in [8]. To this end, the music recording is first converted into a sequence of chroma-based audio features [2], which relate to harmonic and melodic properties [7]. From this sequence, a suitably enhanced self-similarity matrix (SSM) is derived [8]. In our case, we apply in the SSM calculation a relatively long smoothing filter of 12 seconds, which allows us to better bridge local differences in repeating segments. Furthermore, to deal with local modulations, we use a *transposition-invariant* version of the SSM, see [6]. To compute such a matrix, one compares the chroma feature sequence with cyclically shifted versions of itself, see [3]. For each of the twelve possible chroma shifts, one obtains a similarity matrix. The transposition-invariant matrix is then obtained by taking the entry-wise maximum over the twelve matrices. Furthermore, storing the shift index which yields the maximum similarity for each entry results in another matrix referred to as *transposition index matrix*, which will be used in Section 3. Based on such transposition-invariant SSM, we apply the procedure of [8] to compute for each audio segment a fitness value that expresses how well the given segment explains other related segments (also called induced segments) in the music recording. These relations are expressed by a so-called path family over the given segment. The thumbnail is then defined as the segment that maximizes the fitness. Furthermore, a triangular scape plot representation is computed, which shows the fitness of all segments and yields a compact high-level view on the structural properties of the entire audio recording.

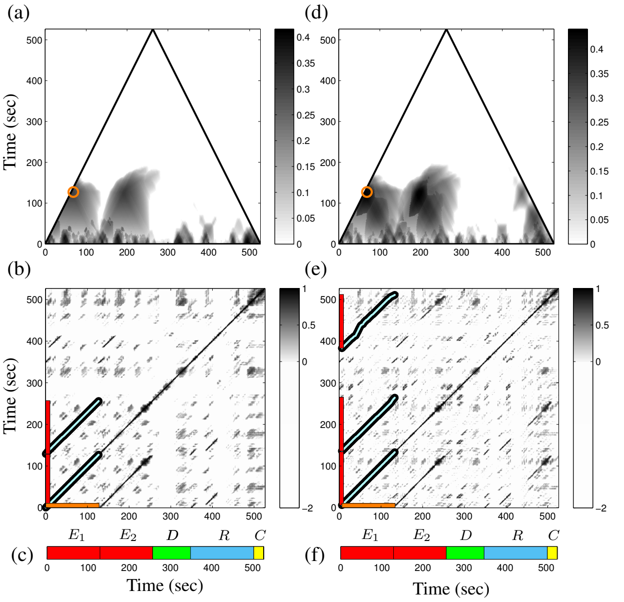

Figure 1: Thumbnailing procedure for `Op031No2-01` ("Tempest"). **(a)/(d)** Scape plot representation using an SSM without/with transposition invariance. **(b)/(e)** SSM without/with transposition invariance along with the optimizing path family (cyan), the thumbnail segment (indicated on horizontal axis) and induced segments (indicated on vertical axis). **(c)/(f)** Ground-truth segmentation.

We expect that the thumbnail segment, at least on the coarse level, should correspond to the exposition (E_1), while the induced segments should correspond to the repeating exposition (E_2) and the recapitulation (R). To illustrate this, we consider as our running example a Barenboim recording of the first movement of Beethoven's piano sonata Op. 31, No. 2 ("Tempest"), see Figure 1. In the following, we also use the identifier `Op031No2-01` to refer to this movement. Being in the sonata form, the coarse musical form of this movement is E_1E_2DRC. Even though R is some kind of repetition of E_1, there are significant musical differences. For example, the first subject group in R is modified and extended by an additional section not present in E_1, and the second subject group in R is transposed five semitones upwards (and later transposed seven semitones downwards) relative to the second subject group in E_1. In Figure 1, the scape plot representation (top) and SSM along with the ground truth segmentation (bottom) are shown for our example, where on the left an SSM without and on the right an SSM with transposition invariance has been used. In both cases, the thumbnail segment corresponds to part E_1. However, without using transposition-invariance, the recapitulation is not among the induced segments, thus not representing the complete sonata form, see Figure 1b. In contrast, using transposition-invariance, also the R-segment is identified by the procedure as a repetition

[2] In our scenario, we use a chroma variant referred to as `CENS` features, which are part of the Chroma Toolbox `http://www.mpi-inf.mpg.de/resources/MIR/chromatoolbox/`. Using a long smoothing window of four seconds and a coarse feature resolution of 1 Hz, we obtain features that show a high degree of robustness to smaller deviations, see [7] for details.

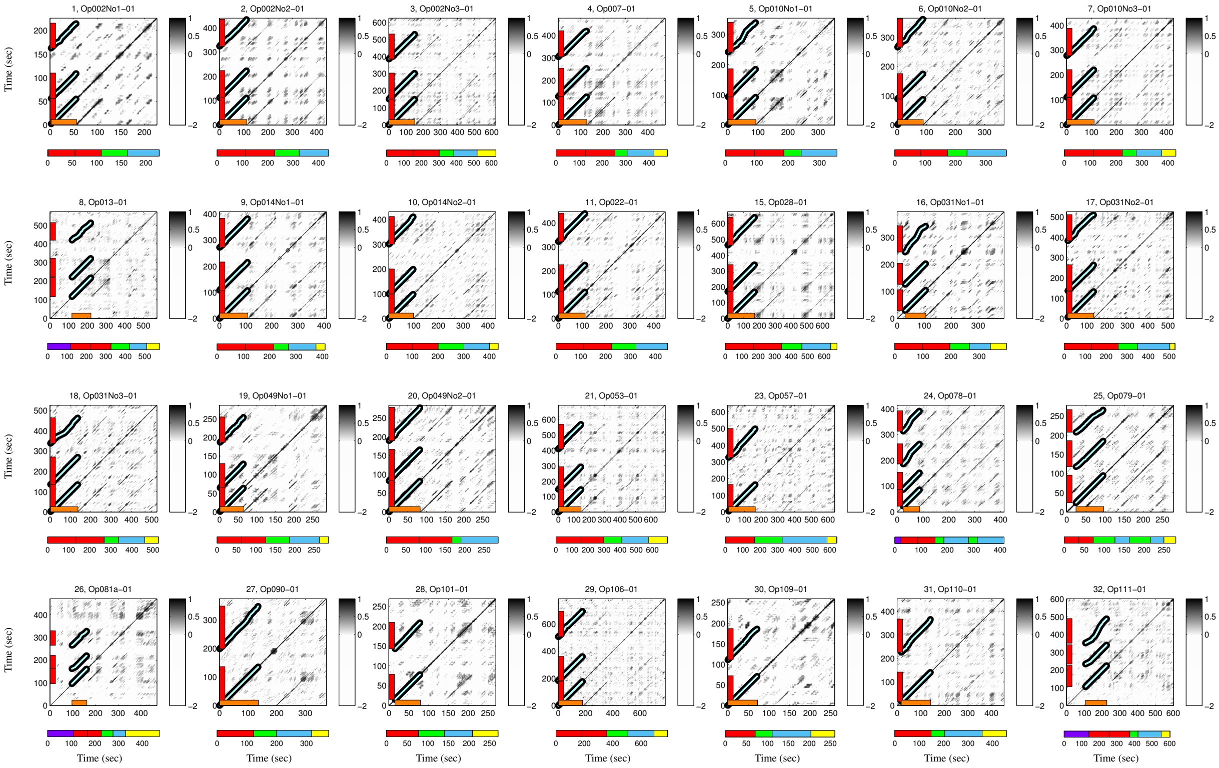

Figure 2: Results of the thumbnailing procedure for the 28 first movements in sonata form. The figure shows for each recording the underlying SSM along with the optimizing path family (cyan), the thumbnail segment (indicated on horizontal axis) and the induced segments (indicated on vertical axis). Furthermore, the corresponding GT segmentation is indicated below each SSM.

of the E_1-segment, see Figure 1e.

At this point, we want to emphasize that only the usage of various smoothing and enhancement strategies in combination with a robust thumbnailing procedure makes it possible to identify the recapitulation. The procedure described in [8] is suitably adjusted by using smoothed chroma features having a low resolution as well as applying a long smoothing length and transposition-invariance in the SSM computation. Additionally, when deriving the thumbnail, we apply a lower bound constraint for the minimal possible segment length of the thumbnail. This lower bound is set to one sixth of the duration of the music recording, where we make the musically informed assumption that the exposition typically covers at least one sixth of the entire movement.

No.	Piece ID	GT Musical Form	P	R	F
1	Op002No1-01	E_1E_2DR	0.99	0.90	0.90
2	Op002No2-01	E_1E_2DR	0.99	0.96	0.96
3	Op002No3-01	E_1E_2DRC	0.95	0.97	0.97
4	Op007-01	E_1E_2DRC	1.00	0.99	0.99
5	Op010No1-01	E_1E_2DR	0.99	0.93	0.93
6	Op010No2-01	E_1E_2DR	0.95	0.86	0.86
7	Op010No3-01	E_1E_2DRC	0.93	0.94	0.94
8	Op013-01	IE_1E_2DRC	0.96	0.95	0.95
9	Op014No1-01	E_1E_2DRC	0.97	0.97	0.97
10	Op014No2-01	E_1E_2DRC	0.94	0.96	0.96
11	Op022-01	E_1E_2DR	1.00	0.97	0.97
12	Op026-01	-	-	-	-
13	Op027No1-01	-	-	-	-
14	Op027No2-01	-	-	-	-
15	Op028-01	E_1E_2DRC	1.00	0.99	0.99
16	Op031No1-01	E_1E_2DRC	0.83	0.74	0.74
17	Op031No2-01	E_1E_2DRC	0.90	0.85	0.85
18	Op031No3-01	E_1E_2DRC	0.99	0.98	0.98
19	Op049No1-01	E_1E_2DRC	0.96	0.91	0.91
20	Op049No2-01	E_1E_2DR	1.00	0.96	0.96
21	Op053-01	E_1E_2DRC	0.99	0.97	0.97
22	Op054-01	-	-	-	-
23	Op057-01	$EDRC$	0.92	0.78	0.78
24	Op078-01	$IE_1E_2D_1R_1D_2R_2$	0.98	0.84	0.84
25	Op079-01	$E_1E_2D_1R_1D_2R_2C$	0.50	0.55	0.55
26	Op081a-01	IE_1E_2DRC	0.86	0.88	0.88
27	Op090-01	$EDRC$	0.76	0.85	0.85
28	Op101-01	$EDRC$	0.97	0.89	0.89
29	Op106-01	E_1E_2DRC	0.99	0.98	0.98
30	Op109-01	$EDRC$	0.92	0.86	0.86
31	Op110-01	$EDRC$	0.91	0.81	0.81
32	Op111-01	IE_1E_2DRC	0.65	0.64	0.64
	Average		0.92	0.86	0.89

Table 1: Ground truth annotation and evaluation results (pairwise P/R/F values) for the thumbnailing procedure using Barenboim recordings for the first movements in sonata form of the Beethoven piano sonatas.

To evaluate our procedure, we use the complete Barenboim recordings of the 32 piano sonatas by Ludwig van Beethoven. Among the first movements, we only consider the 28 movements that are actually composed in sonata form. For each of these recording, we manually annotated the large-scale musical structure also referred to as ground-truth (GT) segmentation, see Table 1 for an overview. Then, using our thumbnailing approach, we computed the thumbnail and the induced segmentation (resulting in two to four segments) for each of the 28 recordings. Using pairwise P/R/F-values [3], we compared the computed segments with the E- and R-segments specified by the GT annotation, see Table 1. As can be seen, one obtains high P/R/F-values for most recordings, thus indicating a good performance of the procedure. This is also reflected by Figure 2, which shows the SSMs along with the path families and ground truth segmentation for all 28 recordings. However, there are also a number of exceptional cases where our procedure seems to fail. For example, for `Op079-01` (No. 25), one obtains an F-measure of only 0.55. Actually, it turns out that for this recording

[3] These values are standard evaluation measures used in audio structure analysis, see, e. g. [10].

the D-part as well as R-part are also repeated resulting in the form $E_1E_2D_1R_1D_2R_2C$. As a result, our minimum length assumption that the exposition covers at least one sixth of the entire movement is violated. However, by reducing the bound to one eighth, one obtains for this recording the correct thumbnail and an F-measure of 0.85. In particular, for the later Beethoven sonatas, the results tend to become poorer compared to the earlier sonatas. From a musical point of view, this is not surprising since the later sonatas are characterized by the release of common rules for musical structures and the increase of compositional complexity [13]. For example, for some of the sonatas, the exposition is no longer repeated, while the coda takes over the role of a part of equal importance.

3. FINE STRUCTURE

In the second step, our goal is to find substructures within the exposition and recapitulation by exploiting the relative harmonic relations that typically exist between these two parts. Generally, the exposition presents the main thematic material of the movement that is contained in two contrasting *subject groups*. Here, in the first subject group ($G1$) the music is in the *tonic* (the home key) of the movement, whereas in the second subject group ($G2$) it is in the *dominant* (for major sonatas) or in the *tonic parallel* (for minor sonatas). Furthermore, the two subject groups are typically combined by a *modulating transition* (T) between them, and at the end of the exposition there is often an additional closing theme or *codetta* (C). The recapitulation contains similar sub-parts as the exposition, however it includes some important harmonic changes. In the following discussion, we denote the four sub-parts in the exposition by E-$G1$, E-T, E-$G2$, and E-C. Also, in the recapitulation by R-$G1$, R-T, R-$G2$, and R-C. The first subject groups E-$G1$ and R-$G1$ are typically repeated in more or less the same way both appearing in the tonic. However, in contrast to E-$G2$ appearing in the dominant or tonic parallel, the second subject group R-$G2$ appears in the tonic. Furthermore, compared to E-T, the transition R-T is often extended, sometimes even presenting new material and local modulations, see [4] for details. Note that the described structure indicates a tendency rather then being a strict rule. Actually, there are many exceptions and modifications as the following examples demonstrate.

To illustrate the harmonic relations between the subject groups, let us assume that the movement is written in C major. Then, in the exposition, E-$G1$ would also be in C major, and E-$G2$ would be in G major. In the recapitulation, however, both R-$G1$ and R-$G2$ would be in C major. Therefore, while E-$G1$ and R-$G1$ are in the same key, R-$G2$ is a modulated version of E-$G2$, shifted five semitones upwards (or seven semitones downwards). In terms of the maximizing shift index as introduced in Section 2, one can expect this index to be $i = 5$ in the transposition index matrix when comparing E-$G2$ with R-$G2$.[4] Similarly, for minor sonatas, this index is typically $i = 9$, which corresponds to shifting three semitones downwards from the tonic parallel to the tonic.

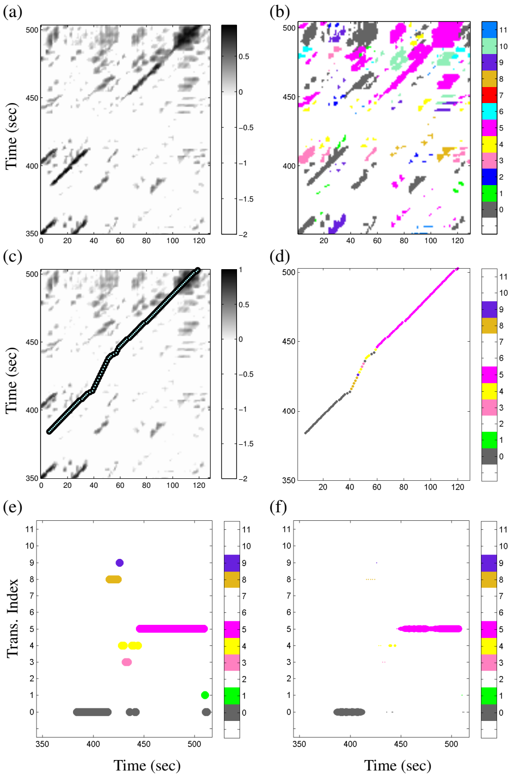

Figure 3: Illustration for deriving the WRTI (weighted relative transposition index) representation using `Op031No2-01` as example. **(a)** Enlarged part of the SSM shown in Figure 1e, where the horizontal axis corresponds to the E_1-segment and the vertical axis to the R-segment. **(b)** Corresponding part of the transposition index matrix. **(c)** Path component of the optimizing path family as shown in Figure 1e. **(d)** Transposition index restricted to the path component. **(e)** Transposition index plotted over time axis of R-segment. **(f)** Final WRTI representation.

Based on this observation, we now describe a procedure for detecting and measuring the relative differences in harmony between the exposition and the recapitulation. To illustrate this procedure, we continue our example `Op031No2-01` from Section 2, where we have already identified the coarse sonata form segmentation, see Figure 1e. Recall that when computing the transposition-invariant SSM, one also obtains the *transposition index matrix*, which indicates the maximizing chroma shift index [6]. Figure 3a shows an enlarged part of the enhanced and thresholded SSM as used in the thumbnailing procedure, where the horizontal axis corresponds to the exposition E_1 and the vertical axis to the recapitulation R. Figure 3b shows the corresponding part of the transposition index matrix, where the chroma shift indices are displayed in a color-coded form.[5] As revealed by Figure 3b, the shift indices corresponding to E-$G1$ and R-$G1$ are zero (gray color), whereas the shift indices corresponding to E-$G2$ and R-$G2$ are five (pink color). To further emphasize these relations, we focus on the path that encodes the sim-

[4] We assume that the index encodes shifts in upwards direction. Note that the shifts are cyclic, so that shifting five semitones upwards is the same as shifting seven semitones downwards.

[5] For the sake of clarity, only those shift indices are shown that correspond to the relevant entries (having a value above zero) of the SSM shown in Figure 3a.

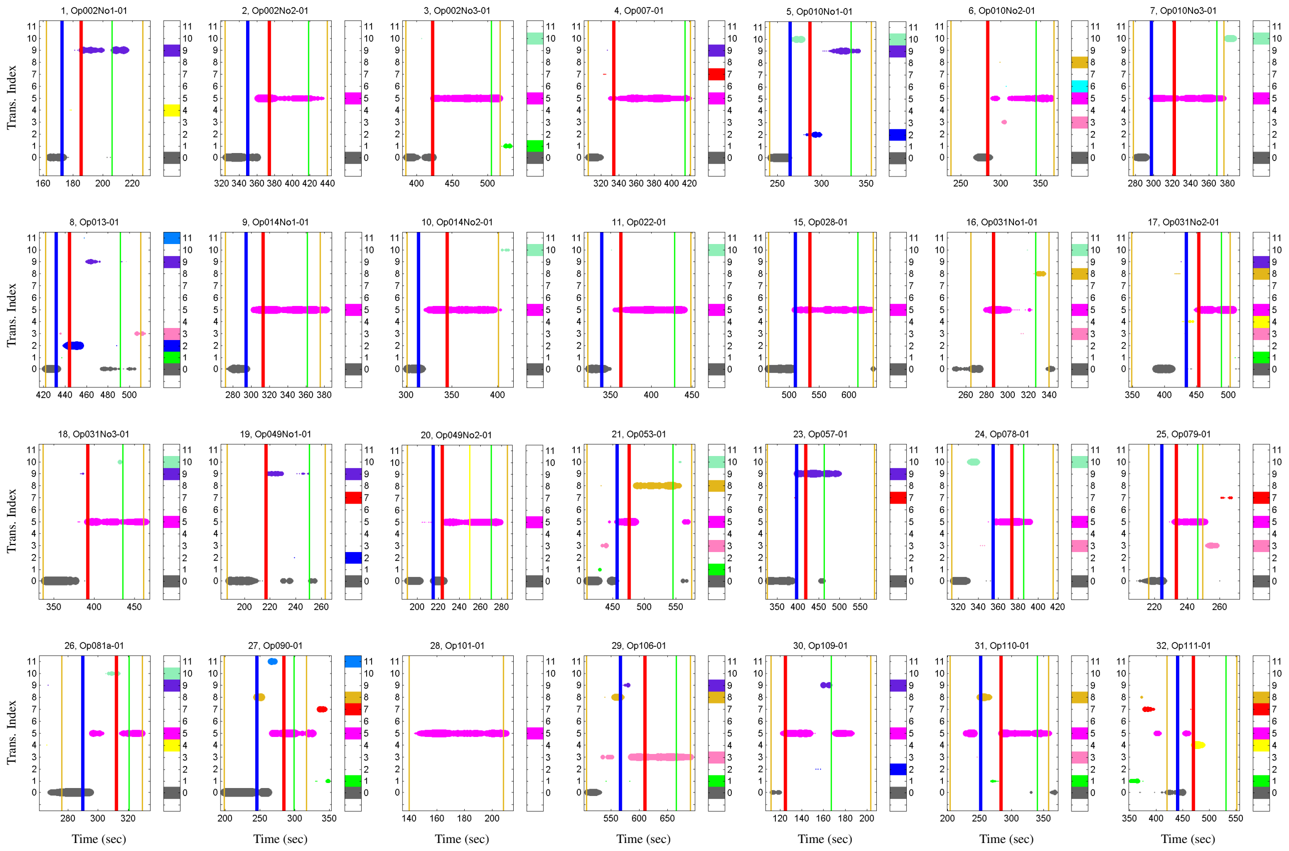

Figure 4: WRTI representations for all 28 recordings. The manual annotations of the segment boundaries between R-$G1$, R-T, R-$G2$, and R-C are indicated by vertical lines. In particular, the blue line indicates the end of R-$G1$ and the red line as the beginning of R-$G2$.

ilarity between E_1 and R, see Figure 3c. This path is a component of the optimizing path family computed in the thumbnailing procedure, see Figure 1e. We then consider only the shift indices that lie on this path, see Figure 3d. Next, we convert the vertical time axis of Figure 3d, which corresponds to the R-segment, into a horizontal time axis. Over this horizontal axis, we plot the corresponding shift index, where the index value determines the position on the vertical index axis, see Figure 3e. In this way, one obtains a function that expresses for each position in the recapitulation the harmonic difference (in terms of chroma shifts) relative to musically corresponding positions in the exposition. We refine this representation by weighting the shift indices according to the SSM values underlying the path component. In the visualization of Figure 3f, these weights are represented by the thickness of the plotted dots. In the following, for short, we refer to this representation as the WRTI (weighted relative transposition index) representation of the recapitulation.

Figure 4 shows the WRTI representations for the 28 recordings discussed in Section 2. Closely following [13], we manually annotated the segments corresponding to $G1$, T, $G2$, and C within the expositions and recapitulations of these recordings [6], see Table 2. In Figure 4, the segment corresponding to R-T is indicated by a blue vertical line (end of R-$G1$) and a red vertical line (beginning of R-$G2$). Note that for some sonatas (e. g., `Op002No3-01` or `Op007-01`) there is no such transition, so that only the red vertical line is visible. For many of the 28 recordings, as the theory suggests, the WRTI representation indeed indicates the location of the transition segment by a switch from the shift index $i = 0$ to the shift index $i = 5$ (for sonatas in major) or to $i = 9$ (for sonatas in minor). For example, for the movement `Op002No1-01` (No. 1) in F minor, the switch from $i = 0$ to $i = 9$ occurs in the transition segment. Or for our running example `Op031No2-01` (No. 17), there is a clearly visible switch from $i = 0$ to $i = 5$ with some further local modulations in between. Actually, this sonata already constitutes an interesting exception, since the shift of the second subject group is from the dominant (exposition) to the tonic (recapitulation) even though the sonata is in minor (D minor). Another more complex example is `Op013-01` (No. 8, “Pathétique”) in C minor, where E-$G1$ starts with $E^\flat$ minor, whereas R-$G1$ starts with F minor (shift index $i = 2$) before it reaches the tonic C minor (shift index $i = 9$). Actually, our WRTI representation reveals these harmonic relations.

To obtain a more quantitative evaluation, we located the transition segment R-T by determining the time position (or region) where the shift index $i = 0$ (typically corresponding to R-$G1$) changes to the most prominent non-zero shift index within the R-segment (typically corresponding to R-$G2$ and usually $i = 5$ or $i = 9$), where we neglect all other shift indices. This position (or region) was computed by a simple sweep algorithm to find the optimal position that separates the weighted zero-indices (which should be on the left side of the optimal sweep line) and the weighted indices of the prominent index (which should

[6] As far as this is possible due to many deviations and variations in the actual musical forms.

No.	Piece ID	$G1$	T	$G2$	C	$\Delta(G1)$	In(T)	$\Delta(G2)$
1	Op002No1-01	10.6	12.6	20.8	20.4		y	
2	Op002No2-01	26.0	24.4	44.2	21.1		y	
3	Op002No3-01	37.9	-	82.9	12.3	-0.6	n	
4	Op007-01	29.0	-	80.7	5.7	-11.5	n	
5	Op010No1-01	23.2	22.4	45.9	22.4		y	
6	Op010No2-01	46.2	-	60.3	22.2		n	2.0
7	Op010No3-01	20.1	24.7	46.2	7.5	-5.6	n	
8	Op013-01	10.1	12.1	47.2	18.8		y	
9	Op014No1-01	22.8	18.6	48.4	13.9		y	
10	Op014No2-01	13.0	31.4	55.7	-		y	
11	Op022-01	17.5	23.5	65.7	19.8		y	
12	Op026-01	-	-	-	-	-	-	-
13	Op027No1-01	-	-	-	-	-	-	-
14	Op027No2-01	-	-	-	-	-	-	-
15	Op028-01	45.2	24.7	80.3	25.4	-4.0	n	
16	Op031No1-01	21.6	-	40.2	12.6	-12.5	n	
17	Op031No2-01	85.7	19.6	34.9	13.6	-5.4	n	
18	Op031No3-01	55.4	-	42.9	25.7	-10.3	n	
19	Op049No1-01	30.5	-	33.5	12.5	-6.0	n	
20	Op049No2-01	24.6	8.6	26.2	15.2		n	8.9
21	Op053-01	47.6	19.3	69.2	29.1		y	
22	Op054-01	-	-	-	-	-	-	-
23	Op057-01	70.3	22.7	43.7	120.8	-7.3	n	
24	Op078-01	41.7	18.9	11.7	29.5	-15.9	n	
25	Op079-01	8.0	8.9	13.2	2.9		y	
26	Op081a-01	13.9	22.3	8.3	8.8		y	
27	Op090-01	47.1	38.9	14.1	18.2		y	
28	Op101-01	-	-	-	-	-	-	-
29	Op106-01	60.0	43.4	55.5	24.9	-36.7	n	
30	Op109-01	13.7	-	41.9	36.6	-6.1	n	
31	Op110-01	47.8	32.0	56.0	17.3	-26.0	n	
32	Op111-01	20.3	29.9	61.0	20.4		y	

Table 2: Ground truth annotation and evaluation results for finer-grained structure. The columns indicate the number of the sonata (No.), the identifier, as well as the duration (in seconds) of the annotated segments corresponding to R-$G1$, R-T, R-$G2$, and R-C. The last three columns indicate the position of the computed transition center (CTC), see text for explanations.

be on the right side of the optimal sweep line). In the case that there is an entire region of optimal sweep line positions, we took the center of this region. In the following, we call this time position the *computed transition center* (CTC). In our evaluation, we then investigated whether the CTC lies within the annotated transition R-T or not. In the case that the CTC is not in R-T, it may be located in R-$G1$ or in R-$G2$. In the first case, we computed a negative number indicating the directed distance given in seconds between the CTC and the end of R-$G1$, and in the second case a positive number indicating the directed distance between the CTC and the beginning of R-$G2$. Table 2 shows the results of this evaluation, which demonstrates that for most recordings the CTC is a good indicator for R-T. The poorer values are in most case due to the deviations in the composition from the music theory. Often, the modulation differences between exposition and recapitulation already start within the final section of the first subject group, which explains many of the negative numbers in Table 2. As for the late sonatas such as `Op106-01` (No. 29) or `Op110-01` (No. 31), Beethoven has already radically broken with conventions, so that our automated approach (being naive from a musical point of view) is deemed to fail for locating the transition.

4. CONCLUSIONS

In this paper, we have introduced automated methods for analyzing and segmenting music recordings in sonata form. We adapted a thumbnailing approach for detecting the coarse structure and introduced a rule-based approach measuring local harmonic relations for analyzing the finer substructure. As our experiments showed, we achieved meaningful results for sonatas that roughly follow the musical conventions. However, (not only) automated methods reach their limits in the case of complex movements, where the rules are broken up. We hope that even for such complex cases, automatically computed visualizations such as our introduced WRTI (weighted relative transposition index) representation may still yield some musically interesting and intuitive insights into the data, which may be helpful for musicological studies.

Acknowledgments: This work has been supported by the German Research Foundation (DFG MU 2682/5-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer IIS.

5. REFERENCES

[1] Mark A. Bartsch and Gregory H. Wakefield. Audio thumbnailing of popular music using chroma-based representations. *IEEE Transactions on Multimedia*, 7(1):96–104, 2005.

[2] Roger B. Dannenberg and Masataka Goto. Music structure analysis from acoustic signals. In David Havelock, Sonoko Kuwano, and Michael Vorländer, editors, *Handbook of Signal Processing in Acoustics*, volume 1, pages 305–331. Springer, New York, NY, USA, 2008.

[3] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1783–1794, 2006.

[4] Hugo Leichtentritt. *Musikalische Formenlehre*. Breitkopf und Härtel, 12. Auflage, Wiesbaden, Germany, 1987.

[5] Richard Middleton. Form. In Bruce Horner and Thomas Swiss, editors, *Key terms in popular music and culture*, pages 141–155. Wiley-Blackwell, 1999.

[6] Meinard Müller and Michael Clausen. Transposition-invariant self-similarity matrices. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 47–50, Vienna, Austria, 2007.

[7] Meinard Müller and Sebastian Ewert. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 215–220, Miami, FL, USA, 2011.

[8] Meinard Müller, Nanzhu Jiang, and Peter Grosche. A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *IEEE Transactions on Audio, Speech & Language Processing*, 21(3):531–543, 2013.

[9] Meinard Müller and Frank Kurth. Towards structural analysis of audio recordings in the presence of musical variations. *EURASIP Journal on Advances in Signal Processing*, 2007(1), 2007.

[10] Jouni Paulus, Meinard Müller, and Anssi P. Klapuri. Audio-based music structure analysis. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)*, pages 625–636, Utrecht, The Netherlands, 2010.

[11] Geoffroy Peeters. Deriving musical structure from signal analysis for music audio summary generation: “sequence” and “state” approach. In *Computer Music Modeling and Retrieval*, volume 2771 of *Lecture Notes in Computer Science*, pages 143–166. Springer Berlin / Heidelberg, 2004.

[12] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pages 555–560, Miami, FL, USA, 2011.

[13] Donald Francis Tovey. *A Companion to Beethoven's Pianoforte Sonatas*. The Associated Board of the Royal Schools of Music, 1998.