Chapter 5
Computational Analysis of Musical Form

Mathieu Giraud, Richard Groult, and Florence Levé

Abstract Can a computer understand *musical forms*? Musical forms describe how a piece of music is structured. They explain how the sections work together through repetition, contrast, and variation: repetition brings unity, and variation brings interest. Learning how to hear, to analyse, to play, or even to write music in various forms is part of music education. In this chapter, we briefly review some theories of musical form, and discuss the challenges of computational analysis of musical form. We discuss two sets of problems, *segmentation* and *form analysis*. We present studies in music information retrieval (MIR) related to both problems. Thinking about codification and automatic analysis of musical forms will help the development of better MIR algorithms.

5.1 Introduction

Can a computer understand *musical forms*? Musical forms structure the musical discourse through repetitions and contrasts. The forms of the Western common practice era (e.g., binary, ternary, rondo, sonata form, fugue, variations) have been studied in depth by music theorists, and often formalized and codified centuries after their emergence in practice. Musical forms are used for pedagogical purposes, in composition as in music analysis, and some of these forms (such as variations or fugues) are also principles of composition that can be found inside large-scale works.

Mathieu Giraud
CNRS, France
Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL), Université de Lille 1, Villeneuve d'Ascq, France
e-mail: mathieu@algomus.fr

Richard Groult · Florence Levé
Laboratoire Modélisation, Information et Systèmes, Université Picardie Jules Verne, Amiens, France
e-mail: {richard, florence}@algomus.fr

For computer scientists working in music information retrieval (MIR) on symbolic scores, computational analysis of musical forms is challenging. On the one hand, very constrained music structures or forms can have rather consensual analyses, such as in an ABA form, and can therefore constitute good benchmarks for MIR algorithms. On the other hand, even if it seems difficult to automatically analyse more elaborate forms, where musicological or aesthetic considerations would seem to demand human expertise, an MIR analysis could provide a valuable time or computation gain towards systematic study and comparison of big musical data sets. More generally, symbolic MIR methods on musical forms are in their infancy and are a long way from the analysis done by performers, listeners, or music theorists: research into the codification and automatic analysis of musical forms will help the development of better MIR algorithms.

We begin this chapter by briefly reviewing some theories of musical form (Sect. 5.2). Two sets of MIR problems arise from these notions of musical structure: *segmentation*, which may be studied in all genres of music, and *form analysis*. We then present studies in MIR related to both problems (Sect. 5.3) as well as evaluation datasets and methods (Sect. 5.4). We conclude by discussing some perspectives as well as the usefulness of form analysis in MIR.

5.2 Musicological Motivation and MIR Challenges

5.2.1 Theories of Form

Musical forms describe how pieces of music are structured. Such forms explain how the sections work together through repetition, contrast, and variation: repetition brings unity, and variation brings interest. The study of musical forms is fundamental in musical education as, among other benefits, the comprehension of musical structures leads to a better knowledge of composition rules, and is the essential first approach for a good interpretation of musical pieces. Not every musical piece can be put into a formal restrictive scheme. However, forms do exist and are often an important aspect of what one expects when listening to music (Cook, 1987).

Even if musical forms have been discussed for several centuries, there are still debates between musicologists on their exact nature. The formalization of many musical forms was done in nineteenth century textbooks aimed at teaching composition to music students (Czerny, 1848; Marx, 1847; Reicha, 1824). Many theories of musical form and how to analyse form have arisen since (Nattiez, 1987; Ratner, 1980; Ratz, 1973; Schoenberg, 1967). Systematic methods for music analysis generally include some formalization of large-scale structure or form, such as the fundamental structure and more superficial structural levels (*Schichten*) of Schenkerian analysis (Schenker, 1935), the growth category of LaRue (1970), or the grouping structures and reductions of Lerdahl and Jackendoff's (1983) *Generative Theory of Tonal Music* (see also Chaps. 9 and 10, this volume).

Table 5.1 Musical forms from a thematic point of view. Letters designate contrasting sections

Form	Structure	Further details
Fugue	Contrapuntal form	
Variations	A A′ A″ A‴	
Binary	AA′, AB, AAB, ABB, AA′BB′	
Ternary	ABA, ABA′	
Rondo	ABACADA...	A: 'chorus', B, C, D: 'verse'
Sonata form	AB–Dev–AB	A: 'principal' theme/zone
Rondo-Sonata form	AB–AC–AB	B: 'secondary' theme/zone
Cyclic form		
Free forms		

Musical form is still an active field of research in musicology. Contemporary approaches to musical form include the "theory of formal functions", describing the role played by each section and formal processes inside and between them; "dialogic form", that is, form in dialogue with historically conditioned compositional options; and "multivalent analysis", which consists of form analysis along several independent elements (Caplin, 2000; Caplin et al., 2009).

5.2.2 Usual Forms

Table 5.1 lists the most common forms, taking the usual *segmentation* point of view: forms are here given as a succession of sections denoted by labels. Binary form consists of two sections, where one or both sections can be repeated, usually with some variation (as for the Bar form, AAB). The sections can use the same or different thematic material. They usually have a comparable length, but the second section can sometimes be longer. In ternary forms, a contrasting section comes between the two statements of the first section. A ternary form can be simple (as in the aria da capo or lied form), or compounded of binary forms (as in trio forms, such as Minuet/Scherzo) or even of ternary forms. Sonata form consists of an exposition, a development, and a recapitulation, and is generally built on two thematic areas.

One should note that the segmentation is only an element of the form analysis. The proper understanding of a ternary form often requires that one explains in which aspects (tonality, texture, rhythm, etc.) the contrasting section B differs from, or complements, the initial section A. Many forms also include high-level tension and dynamics that should be explained. Sonata forms combine these principles, including a tonal path: the piece goes to a secondary key, often the dominant, and then comes back to the tonic. A segmentation is thus a very partial view on a form.

More generally, there are many *free forms* that are more difficult to classify according to a section-based view. Baroque preludes and inventions are often viewed as free forms. However, this does not mean that they are not structured, but rather

that their organization depends on other components, such as texture or tension. An extreme case is some *open forms*, where the order of execution of the sections is not fixed by the composer, but where some components can still be conceived of in terms of form.

A full musical analysis that reveals a musical form should thus include consideration of *all organizational principles*, not just a simple segmentation. Note that there are typically many possible analyses, focusing on different aspects of a score. One may even start from different musicological hypotheses. For example, the conception of sonata form has evolved over time. Originally defined in two parts from a thematic perspective (Reicha, 1824), it was then theorized as a ternary form by A. B. Marx, who first named it, and C. Czerny (Czerny, 1848; Marx, 1847). Then Rosen returned to the view that this form derived from *binary* form, and that the most critical feature is the tonal and harmonic path followed (Rosen, 1980). The recent theory of Hepokoski and Darcy (2006) further emphasizes the importance of the medial caesura and rotations. Depending on the underlying musicological assumptions, an analysis of such a form will thus focus on different organizational principles.

5.2.3 Challenges for MIR

Based on these conceptions of musical form, we see two sets of challenges related to structure and form in MIR:

- *Segmentation* consists of taking a piece—either a score, or an audio file—and chunking it into sections, possibly labelled, as in Table 5.1. There are several MIR challenges involved in segmentation, such as identifying correct phrase or section boundaries and assessing the similarity between sections.
- *Form analysis* consists of predicting a segmentation together with *semantics* on this segmentation and on the global layout of the piece. The provided semantics can be both *internal*, indicating the formal function of a section and how the sections relate to each other and to the global layout ("B is a variation of A"); or *external*, explaining how the sections are organized with respect to some historical corpus or analytical practice ("A is in dialogic form", "A and B are the two thematic zones of a sonata form"). Finally, it has to build upon several elements (multivalent analysis): a thematic segmentation is one of these elements, but not the only one. For example, evolution of tension and tonal path are elements that should be included in a complete form analysis.

Even if the term "musical form" is often applied to baroque, classical and romantic music analysed within traditional musicology, form analysis can be carried out on any genre of music, and many of the studies that will be discussed in this chapter focused on popular music.

Of course, segmentation and form analysis are related: segmentation is an essential component of an analysis of a musical form; and to assign labels to segments (A, A_1, A', B, etc.) requires knowledge about the relations between sections and their

semantics. Segmentation problems can (partially) be formalized and evaluated against reference analyses (see Sect. 5.4.2), and are an important set of challenges for the MIR community.

Nevertheless, we believe that the computational music analysis (CMA) community should aim at solving full "form analysis" problems, trying to have a better understanding of musical structure, beyond mere segmentation. Unfortunately, formalization of these tasks—and thus of their evaluation—is even more challenging than it is for segmentation. One may believe that some tasks inevitably resist computer formalization, as research in CMA should be linked to musicological practice, including subjectivity in the description of forms. However, this subjectivity might be modelled in some ways. Ideally, a model should be able to predict where there might be a disagreement about the form of a piece, and even predict several different formal analyses of the same piece where ambiguity exists.

5.3 Methods for Segmentation and Form Analysis

In this section, we review existing algorithmic methods for segmentation and form analysis.

5.3.1 Musical Structure and Segmentation

5.3.1.1 Phrase-Structure Analysis and Melodic Segmentation

Even if the aim of melodic segmentation is not to give a large-scale structure for a musical piece, it allows us to chunk music into phrases, which can provide clues about boundaries between structural sections.

A common way to handle *monophonic* data is to observe the differences between successive intervals or durations. It is not simply a large interval or a large duration that marks the end of a phrase, but the variation in these quantities. One reference algorithm here is Cambouropoulos' (2001, 2006) *local boundary detection model* (LBDM) which is a simple algorithm that achieves good results. This algorithm considers both melodic intervals, durations and rests, and detects points of segmentation as the maxima of some profile function. More recent work on melodic segmentation includes that of Muellensiefen et al. (2008), Wiering et al. (2009) and Rodríguez-López et al. (2014). Phrase segmentation can also be done on *polyphonic* data: for example, Rafailidis et al. (2008) proposed a method to infer *stream segments*, based on an n-nearest-neighbour clustering from grouping criteria.

Note that *pattern inference and matching* algorithms (see Chaps. 11–17, this volume) can also be used to detect repeating sections and predict their boundaries.

5.3.1.2 Global Segmentation

We focus here on algorithms that aim to provide a high-level segmentation of a musical piece. Structural music segmentation consists of dividing a musical piece into several parts or sections and then assigning to those parts identical or distinct labels according to their similarity. The founding principles of structural segmentation, whether it be from audio or symbolic sources, are homogeneity, novelty, and/or repetition.

Audio Sources Many methods for segmenting audio files are based on auto-correlation matrices (Foote, 1999). Some methods focus on the detection of repeated sections (Chai, 2006; Dannenberg and Goto, 2009; Dannenberg and Hu, 2002) possibly after the extraction of features from a signal (Peeters, 2007). Other methods are based on probabilistic approaches (Paulus and Klapuri, 2008), or on timbre characteristics (Levy et al., 2007). Maddage et al. (2009), Paulus et al. (2010) and Klapuri (2011) provide recent surveys. Some systems and algorithms can be applied to both audio and symbolic sources, such as that of Sargent et al. (2011), where a Viterbi algorithm predicts the segments, taking into account both a content-based cost, based on similarity between segments, and a "regularity cost" favouring segments of similar lengths. This regularity cost improves the segmentation on Western popular songs.

Symbolic Sources There is less work to date that focuses on segmentation of symbolic scores. On monophonic data, Chen et al. (2004) segments the musical piece into sections called "sentences". The phrases predicted by the LBDM algorithm are compared (using their first pitch and the subsequent contour) and clustered. The score is then processed another time to obtain a sequence of labels and to predict the actual starts of each section. For track-separated polyphonic data, Rafael and Oertl (2010) propose combining segmentations from different tracks into a global fragmentation. In each track, a set of repeated patterns is computed by a modified version of the algorithm proposed by Hsu et al. (1998) allowing transpositions. Then these sets are cleaned and further processed to obtain non-overlapping patterns. The segmentations are then clustered into a global segmentation by maximizing a score function that favours compatible segments from several tracks. However, besides some examples, the authors do not report any evaluation of this method.

Several authors have proposed systems for generating analyses at larger scales in accordance with well-established theories of tonal musical structure, such as Schenkerian analysis (Schenker, 1935) or Lerdahl and Jackendoff's (1983) *Generative Theory of Tonal Music* (GTTM) (e.g., Hamanaka and Tojo, 2009; Hirata and Matsuda, 2002; Kirlin and Jensen, 2011; Marsden, 2010; see also Chaps. 9 and 10, this volume). This is very challenging and still an open problem: in both Schenkerian analysis and GTTM theory, carrying out a musically relevant analysis requires the making of analytical choices that rely on a great deal of musical information. Other work has also tried to model specific large-scale features, such as tonal tension (Farbood, 2010; Lerdahl and Krumhansl, 2007), that may also result in a global segmentation of a piece.

5.3.1.3 Discussion

Are such methods able to analyse classical musical forms? As discussed earlier, segmentation is an important step towards the analysis of a musical structure, but to be satisfactory, a musical analysis cannot just provide a segmentation into similar or distinct parts. It is also necessary to identify the *formal function* of each segment of the piece, and ideally to indicate the evolution of the compositional material when it is used differently in several segments. Indeed, a segmentation ABABCAB could just as well correspond to a popular song with verses and chorus with a transition (a bridge), as to a classical sonata form where A and B are the principal and secondary themes, played twice, followed by a development and a recapitulation.

5.3.2 Systems for Analysing Specific Musical Forms

In this chapter, we present *computational systems tailored for the analysis of classical musical forms*, focusing on two particular cases: fugues and sonata form. In both cases, we start from a score in which all voices are separated, and we compute some local features using discrete tools (based on string comparison) and statistical approaches. These local features can be gathered into a global analysis.

Note that other methods have been designed to check whether a piece follows a given structure. For example, Weng and Chen (2005) built a tool to decide whether a piece is a fugue, a rondo, or none of these, with a method to find occurrences of thematic materials.

5.3.2.1 Fugue

Elements of Fugal Structure A fugue is a contrapuntal polyphonic piece built on several melodic themes, including a *subject* (S) and, in most cases, one or several *countersubjects* (CS1, CS2). A fugue is structured as a set of *voices*, where each voice is mostly a monophonic sequence of notes. In the first section, the *exposition*, the patterns are played by each voice in turn. First, the subject is stated in one voice until a second voice enters. The subject is then repeated in the second voice, generally transposed, while the first voice continues with the first countersubject, combining with the subject. Bruhn (1993, p. 43) states that

> the perfect little musical entity we call subject is in fact at the origin of the fugue. [...] The subject is responsible for the feelings of density and relaxation in the fugue, and it is the main force in creating structure.

The fugue alternates between other instances of the subject and the countersubjects (either in their initial form, altered or transposed), and developments on these patterns called *episodes* (E). The episodes can contain *cadential passages*, and are often composed with *harmonic sequences*, which are passages where a pattern is consecutively repeated starting on a different pitch, possibly modulating from one tonality

to another. At the end of the fugue, one often finds *stretti* (shorter and narrower statements of the head of the subject) and *bass pedals*.

Algorithms for Fugue Analysis Many MIR studies take examples from corpora of fugues, for pattern extraction or grammar inference (e.g., Sidorov et al., 2014), but few studies have been specifically devoted to fugue analysis. Browles (2005) proposed several heuristics to aid in the selection of candidate fugue subjects using the repeated pattern extraction algorithms of Hsu et al. (1998). These algorithms maximize the number of occurrences of repeating patterns.

We have proposed several tools in a system for automated fugue analysis (Giraud et al., 2015, 2012). Starting from a symbolic score that is separated into voices, we compute the end of the subject and its occurrences (including augmentations or inversions), and then retrieve CS1 and possibly CS2. Harmonic sequences, cadences, and pedals are detected. Some of these elements are then combined to sketch the global structure of the fugue. Figure 5.1 shows an example of the output of the system on a fugue in B♭ major by J. S. Bach.

S/CS1/CS2 Patterns All the occurrences of S/CS1/CS2, as well as the patterns that form harmonic sequences in episodes, are not necessarily identical, but can be transposed or altered. The similarity score between a pattern and the rest of the fugue can be computed via dynamic programming using equations similar to the ones proposed by Mongeau and Sankoff (1990), and the alignment is retrieved through backtracking in the dynamic programming table. A threshold determines whether an occurrence is retained or discarded. Various methods have been proposed for measuring similarity between pitch intervals. These include directly computing intervals between *diatonic* pitches, strict pitch equality, "up/down" classes (Ghias et al., 1995), "step/leap intervals" (Cambouropoulos et al., 2005) and "quantized partially overlapping intervals" (QPI) (Lemström and Laine, 1998). We use very conservative substitution functions: repeated patterns are detected using a substitution function by considering the diatonic similarity of pitch intervals, and forcing the duration of all except the first and the last notes to take the same value.

The subject is first heard alone by a unique voice until the second voice enters, but the end of the subject is generally not exactly at the start of the second voice—in all but one of 36 fugues by Bach or Shostakovich, the opening statement of the subject ends between eight notes before and six notes after the start of the second entry. In the first voice, we thus test patterns that finish between eight notes before and six notes after the start of the second voice: each of these candidates is matched against all the voices, and the candidate having the best total score on all its occurrences is then selected. The subject can appear augmented (all durations are doubled) or inverted (all pitch intervals are reversed), and once the subject is known, the same matching algorithm retrieves the inversions or the augmentations of the subject.

The countersubject starts right after the subject (or very infrequently, after some additional notes), and ends (in the majority of cases) exactly at the same position as the subject. The second countersubject (when it exists) starts and ends approximately at the start and end, respectively, of the third occurrence of the subject. We use the

same matching algorithm to retrieve CS1/CS2. We allow a CS1 only when there is a concurrent S, and a CS2 only when there is a CS1 in a "compatible" position.

Other Analytical Elements and Global Structure Using the substitution function, we look for partial harmonic sequences—that is, consecutive transposed occurrences of the same pattern across at least two voices (Giraud et al., 2012). Every bass note that lasts strictly more than four beats (for binary metres) is labelled "bass pedal" (or six beats for ternary metres). We try to identify a perfect authentic cadence (V–I in root position) by detecting a dominant chord followed by a complete or partial tonic chord with the root in the bass.

Finally, we use a hidden Markov model to combine previous elements to discover the sequence of states that structure the fugue (exposition, codetta, further exposition, further episodes). The observed symbols are the elements output by the preceding analysis steps (entries of subjects and countersubjects, harmonic sequences).

Results Our aim with this system was to provide *algorithms with high precision* rather than high recall. As all the content of a fugue is somewhat derived from the base patterns, what is interesting is not to locate as many approximate occurrences as possible or to infer very short patterns, but to provide an analysis with some *semantics*: the occurrences of the patterns and of other analytical elements should be organized into a meaningful analysis. The system has been tested on 36 fugues of Bach and Shostakovich, and was successful in finding the main patterns and the global structure in at least half of the fugues (Giraud et al., 2015). In both the Bach and Shostakovich corpora, the precision of pattern detection is more than 85% for S, CS1, and CS2. For example, false positive subjects were found in only three of Bach's fugues and four of Shostakovich's fugues. Most of the time, false positive subjects detected by the system are in fact relevant to the analysis, because they correspond to incomplete thematic patterns. Finally, the analysis of some fugues yielded poor results, especially when the subject detection failed. Some noteworthy cases were the double and triple fugues, where additional subjects have their own expositions in the middle of a fugue, as in Bach's Fugue in C♯ minor (BWV 849). Because the detection of subjects is done at the beginning of a fugue, our system was not able to detect these other subjects and their occurrences.

5.3.2.2 Sonata Form

A movement in sonata form is built on a *tonal progression* and on two thematic zones (primary, P, and secondary, S) (Hepokoski and Darcy, 2006). A first step towards computing an analysis of such a movement is to be able to detect the global structure (exposition–development–recapitulation).

A study on audio signals managed to find these approximate boundaries (Jiang and Müller, 2013). We presented preliminary results on identifying such global structure in symbolic signals (David et al., 2014). After detecting putative phrase endings, we try to locate a P zone, from the start to a phrase ending, that is repeated in the recapitulation. Then we try to locate, after P, an S zone, under the constraint

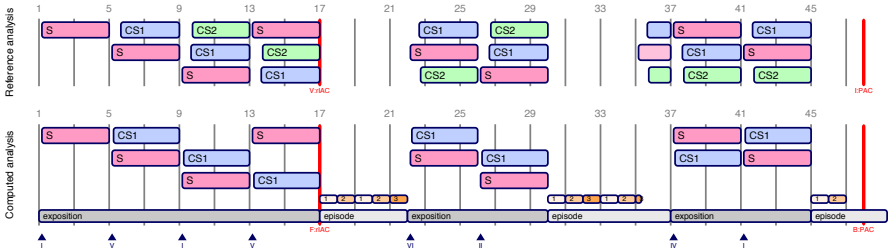

Fig. 5.1 Reference annotations on Fugue in B♭ major (BWV 866) from the first book of the Well-Tempered Clavier by J. S. Bach (top), and computed analysis (bottom). The computed analysis detects thematic patterns (S/CS1) with scale degrees, harmonic sequences (1/2/3), and the two cadences (rIAC and PAC). Here the algorithms find a good structure, even if the second countersubject (CS2) and some incomplete patterns at measures 35 and 36 are undetected

of a transposition from the dominant in the recapitulation. In a corpus of eleven first movements from string quartets by Mozart, Haydn or Schubert, this method allows for the retrieval of the majority of exposition/recapitulation pairs. The starting boundary of P in the recapitulation is often precise, but not the P/S separation. The predicted S zone can be detected before the start of the actual S, because the end of the transition is often already transposed in the recapitulation. An objective could be to precisely locate the *medial caesura*, when it exists, marking the end of the P zone and the beginning of the S zone (Hepokoski and Darcy, 2006).

5.4 Benchmark Data and Evaluation

In this section, we discuss how to evaluate and compare algorithms for segmentation and form analysis. As computer scientists, we would like to have computer-readable reference datasets that may be used as a "ground truth" to evaluate algorithms. But as music theorists, we know that there is rarely one correct analysis of a given piece: listeners, players, or analysts often disagree or at least propose several points of view.

We now discuss the possibility of having reference datasets, present datasets for segmentation and form analysis, and finally discuss evaluation methods.

5.4.1 Can We Have a Reference Segmentation or a Reference Music Form Analysis?

A music theorist performing an analysis focuses on particular aspects of a score with respect to what she wants to reveal. The relevant observations are not necessarily the same as the ones perceived by a listener or by a player. When one wants to evaluate computational analysis methods, one thus needs to ask *which specific task one needs*

to evaluate, and one has to use an adapted ground truth. Which analytical points to consider is not so straightforward and depends on the studied piece and on some musicological assumptions.

Different opinions exist regarding the definition of basic analytical elements, the placement of section boundaries, and the functional labelling of sections. In the following, we will see how different datasets encode the same popular song (Figs. 5.3 and 5.4). In classical music, there are also many points where different people may have different answers. Considering different analyses of fugues from the first book of Bach's Well-Tempered Clavier, for 8 out of 24 fugues, at least two musicological sources disagree on defining the *end* of the subject, that is the main theme of a fugue (Fig. 5.2). Indeed, in fugues, whereas the starting points of subjects are notable cognitive events, their endings are not always important and may be constructs arising from music theory. More radically, considering that the ambiguity can be part of the music, Tovey (1924) writes about J. S. Bach's Fugue in E major (BWV 854) from the first book of the Well-Tempered Clavier: "It is not worthwhile settling where the subject ends and where the countersubject begins".

On the other hand, there is consensus among many music theorists, players and listeners about some analytical elements. The fact that reaching consensus may be difficult on some points should not prevent us from trying to formalize some elements. Ideally we should have a collection of reference analyses highlighting different points of view on the same piece by different musicologists. This could be achieved either by having several annotators agree on a reference annotation, or, even better, by having analyses or alternative solutions by several annotators. Finally, dealing with ambiguities does not exclude the possibility of having an "exact" symbolic approach: one can indicate in the dataset that one may consider two or more boundaries for a section, but the precise position of each boundary can be specified.

5.4.2 Datasets on Music Segmentation

Several reference datasets exist for segmentation/structuration purposes, trying to take into account multiple dimensions of music (Peeters and Deruty, 2009). The four datasets presented below are plain-text files describing the structure of audio files as annotations at some offsets (measured in seconds). These datasets consist mostly of popular music, although some "classical" works can be found in them too. Annotations have been produced manually, often using a platform such as Sonic Visualiser.[1]

SALAMI The SALAMI dataset (Structural Analysis of Large Amounts of Music Information) (Smith et al., 2011), version 2.0 (March 2015), contains the structure of 1164 audio files, each one analysed by *one or two listeners*, totalling 1933 annotation files, and 75191 segments.[2] The annotation format can describe "musical similar-

[1] http://www.sonicvisualiser.org/

[2] http://ddmal.music.mcgill.ca/research/salami/annotations

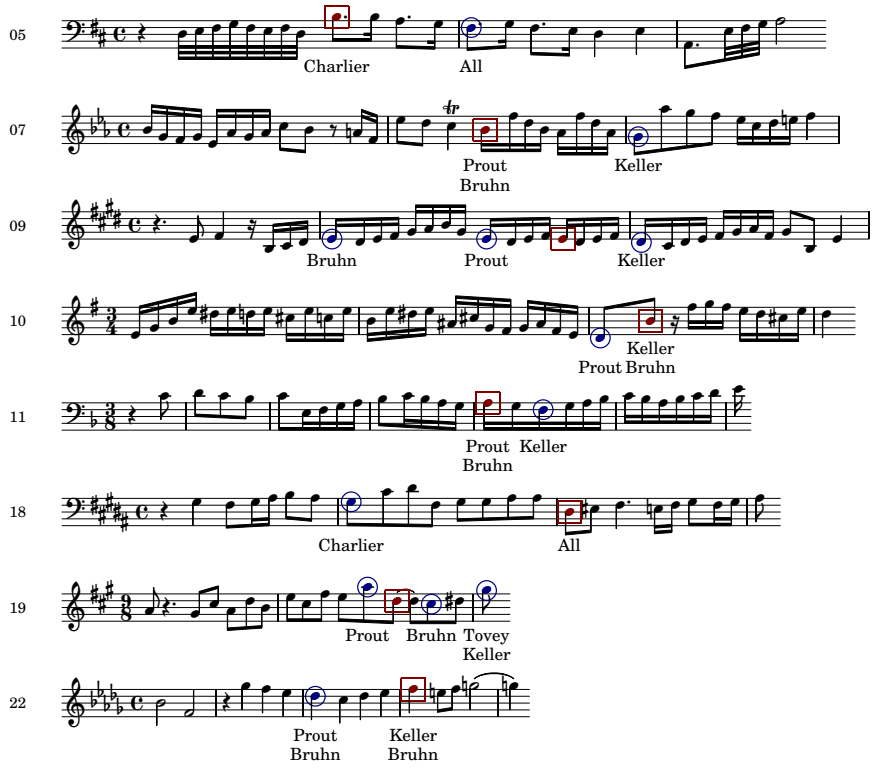

Fig. 5.2 The eight subjects in the first book of Bach's *Well-Tempered Clavier* where at least two sources disagree on the end of the subject (Giraud et al., 2015). Circled notes show the proposed endings of the subjects. For example, in Fugue number 7, Prout (1910) and Bruhn (1993) state that the subject ends on the B♭ (tonic), with a cadential move that is strengthened by the trill. Keller (1965) states that the subject ends after the sixteenth notes, favouring a longer subject that ends on the start of the countersubject. Charlier (2009) has a motivic approach, resulting in shorter subjects in some fugues

ity, function, and instrumentation". Musical similarity is described on two levels: lowercase letters indicate small-scale similar segments, whereas uppercase letters indicate sections. Functions are described by a controlled vocabulary of 20 labels ("intro", "verse", etc.). Some of these labels are specific to a genre or a form ("head", "main theme", "exposition", "development", etc.). The leading instruments are also described—instruments are delimited by opening and closing parentheses (Fig. 5.3, right, and Fig. 5.4, right). The fact that some files are encoded by *two* listeners allows for some ambiguities to be included, coming from different annotators.

IRISA Semiotic Annotations The Semiotic Annotations dataset (Bimbot et al., 2012, 2014), version 2012, describes the structure of 383 audio files in 6725 seg-

ments, produced and cross-checked by two annotators.[3] Authors identified *segments* respecting a "System-and-Contrast" model, a segment being usually a sequel of four elements where the last element has some contrast within a logical progression. This enables traditional antecedent/consequent *abac* periods to be modelled. Further extensions can model shorter or longer periods such as *ab* or *abaac*. Segments are clustered according to the similarity of their logical progression and are labelled according to some functional properties (for instance I for an introduction and C for a central part, often the chorus). Composite labels are also used, such as A_B for an overlap between two segments. The method usually requires a limited number of symbols to model a piece, providing a high-level view of its structure (Fig. 5.3, left). These annotations rely on a methodology that is well-defined in the guidelines for this dataset—ideally, they should be independent of the annotator.

AIST Annotations AIST Annotations (Goto, 2006) (National Institute of Advanced Industrial Science and Technology) describe the structure of 276 files of the RWC database (Goto, 2004),[4] in 4485 segments (Fig. 5.3, middle). Some annotations also include other elements such as beats or melody line. Synchronized MIDI files, obtained by the procedure described by Ewert et al. (2009), are available for the Popular Music database and Classical Music database, allowing this dataset to be used for symbolic MIR studies.

Isophonics The C4DM (Centre for Digital Music) provides segmentation for 301 files (Mauch et al., 2010) (Fig. 5.4, left). Some of these files also have other annotations (chords, keys, and beats).[5]

Do These Datasets Agree? Figures 5.3 and 5.4 show examples of these annotations on the same files. Staying at the *segmentation* level, one can observe that the boundaries of these manual annotations generally coincide (with slight differences in the offsets) and that the clustering of labels is also generally done in the same way. However, several differences are found when one goes to functional labelling, getting closer to what could be a form analysis. Even in popular songs, identifying the formal function of a chorus, a verse or a bridge is sometimes ambiguous.

5.4.3 Datasets with Reference Analyses of Musical Forms

The datasets described above are well adapted to popular music. On the other hand, large musical forms, such as sonata form, feature complex relationships between parts that cannot be satisfactorily modelled simply by labelled sections. Moreover, the musical score, when it exists, contains many details that can be analysed beyond a raw segmentation. We argue for the need for *specific reference datasets for musical forms* adapted to computational music analysis.

[3] http://metissannotation.irisa.fr/

[4] https://staff.aist.go.jp/m.goto/RWC-MDB/AIST-Annotation/

[5] http://www.isophonics.net/content/reference-annotations

```
IRISA                                 | AIST                                  | SALAMI
0.0000000   0.0625000   #             |                                       | 0.000000000  Silence
0.0625000   22.3629358  I-|I/C        | 0     2233  ''intro''                 | 0.081269841  a, A, Intro, (synthesizer
22.3629358  44.6824407  C             | 2233  4466  ''bridge A''              | 5.665668934  b, synthesizer)
44.6824407  55.8520639  D/2           | 4466  5582  ''bridge B''              | 11.276757369 a, (guitar
55.8520639  78.1794653  A             | 5582  7815  ''verse A''               | 16.867596371 b, guitar)
78.1794653  100.5029185 B             | 7815  10047 ''verse B''               | 22.437596371 a, B, Chorus, (voice)
100.5029185 122.8303200 C             | 10047 12280 ''chorus A''              | 27.972766439 b, (synthesizer)
122.8303200 134.0038914 D/2           | 12280 13396 ''bridge B''              | 33.568571428 a, (voice
134.0038914 156.3273446 A             | 13396 15629 ''verse A''               | 39.137959183 b, voice)
156.3273446 178.6507978 B             | 15629 17861 ''verse B''               | 44.715374149 c, C, Solo, (synthesizer
178.6507978 200.9703027 C             | 17861 20094 ''chorus A''              | 50.311360544 c, synthesizer)
200.9703027 212.1359776 D/2           | 20094 21210 ''bridge B''              | 55.902448979 d, D, Verse, (voice
212.1359776 234.4752238 C0            | 21210 23443 ''bridge A''              | 61.462131519 d'
234.4752238 239.2091857 -             | 23443 24046 ''ending''                | 67.063356009 d
239.2091857 242.5652248 #             |                                       | ...
```

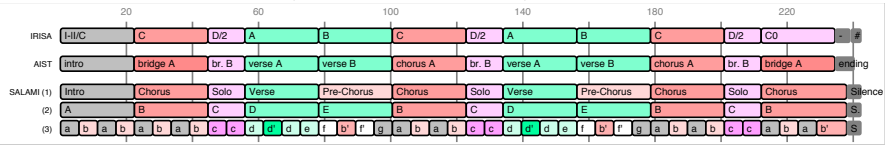

Fig. 5.3 Reference annotations from IRISA, AIST and SALAMI for "Spice of Life", sung by Hisayoshi Kazato (track 4 of RWC-MDB-P-2001). The three annotations generally agree on the segmentation even though their onsets differ slightly. Note that the four sections labelled `B, Chorus` by SALAMI are labelled `C`, `C`, `C`, and `C0` by IRISA and `bridge A`, `chorus A`, `chorus A`, and `bridge A` by AIST. The SALAMI annotation has several levels, giving both functions (1), sections (2) and small-scale segments (3)

```
                Isophonics                                SALAMI

0.000   0.980           silence                 |  0.000000000    Silence
0.980   18.709          verse                   |  0.505034013    A, a, Verse, (voice
18.709  35.988          verse                   |  9.912834467    a
35.988  53.174          refrain                 |  18.702879818   a
53.174  70.518          verse                   |  27.498321995   a
70.518  87.667          refrain                 |  36.109433106   B, b, Chorus
87.667  105.349         verse_(instrumental)    |  44.602471655   b
105.349 122.596         verse                   |  53.247551020   A, a, Verse
122.596 155.139         refrain_with_ending     |  61.883242630   a', voice)
155.139 160.400         silence                 |  66.161360544   c
                                                |  70.657346938   B, b, Chorus, (voice
                                                |  79.133197278   b, voice)
                                                |  87.652335600   C, c, Interlude
                                                |  96.556575963   c
                                                |  105.326258503  A, a, Verse, (voice
                                                |  113.987437641  a
                                                |  123.241791383  B, b, Chorus
                                                |  131.276349206  b
                                                |  139.812698412  b, Fade-out, voice)
                                                |  160.474172335  End
```

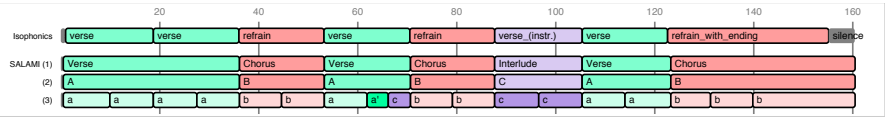

Fig. 5.4 Reference annotations from Isophonics and SALAMI on the Beatles song "Yellow Submarine". The annotations mainly agree on the segmentation even if the onsets differ slightly, and the segments are different: the first two segments labelled `verse` by Isophonics correspond to only one segment, labelled `A` (`aaaa`) by SALAMI: `verse` corresponds to the sub-segments `aa` or `aa′c`, with `a′` and `c` having half the length of the other sub-segments. The last `Chorus` (`refrain_with_ending` for Isophonics) has three small-scale segments, `bbb`, instead of only two

```
== S [length 4 start +1/8]
     S  1, 13, 37
     A  5, 22, 41
     T  9, 26

== S-inc [base S length 2]
     A  35


== cadences
     *  17 (V:rIAC)
     *  48 (I:PAC)
```

```
== CS1 [length 3 start -1/4]
       S  6, 23, 42
       A  10, 27, 38
       T  14

== CS1-inc [base CS1 length 1]
       S  36


== CS2 [length 3 start -3/16]
       S  10, 27
       A  14
       T  23, 38, 42

== CS2-inc [base CS2 length 1]
       T  36
```

Fig. 5.5 Reference annotations from the Algomus group on Fugue in B♭ major (BWV 866) by J. S. Bach, obtained from several sources including the books by Bruhn (1993) and Prout (1910). These annotations show the entries of the subject (S), countersubjects (CS1 and CS2) in soprano (S), alto (A), and tenor (T) voices, as well as incomplete entries (S-inc, CS1-inc, CS2-inc). The position of each occurrence is denoted by measure numbers (1 is the first beat of measure 1), but the actual start is shifted from this logical position (one eighth note forward for S, one quarter note backward for CS1, three sixteenth notes backward for CS2). This fugue also contains perfect and root position imperfect authentic cadences (PAC and rIAC) in the tonic (I) and dominant (V) keys. A graphical representation of this annotation can be found at the top of Fig. 5.1

As we saw in the introduction, there are different conceptions of musical forms. So the nature of a specific "musical form analysis" depends on the musicological assumptions on which it is based. There are many "correct" or at least "pertinent" ways to analyse musical forms. The points of view will not be the same when one focuses on a particular piece, on the evolution of the techniques used by a composer, or on a corpus including works of other composers.

In any case, one needs to annotate both features occurring in all parts or voices (cadences, sequences, texture, global segmentation) and at the voice or instrument level (e.g., melodic themes). This latter level is still valid when the voices are not explicitly separated, such as in piano or guitar music. The annotations should rely on one or several musicological reference analyses indicating the locations and the durations of precise elements (e.g., for sonata form, the thematic zones in the exposition and recapitulation or thematic elements in the development). Offsets can be exact symbolic values (measure numbers and positions within the measure).

For annotators, a problem is that even a detailed musicological analysis does not always provide the level of precision required in a formalized dataset. For example, the books of Bruhn (1993) on Bach's fugues detail the list of occurrences of the S and CS patterns, including complementary details for the strongly varied occurrences. However, very slight variations are not always described, and an annotator transcribing this analysis has to encode the precise location to report the exact offset value in the reference dataset. Moreover, since human language itself can be ambiguous, the

encoder of the reference analysis may sometimes have to make some interpretations, even if the intended meaning is obvious most of the time. Such interpretations should be recorded in the formalized dataset.

Finally, to provide subtle analyses of musical pieces closer to human interpretations, algorithms should also model ambiguities, ideally by predicting several solutions with their associated likelihoods.

Algomus Fugue Reference Analysis We released a reference analysis for the 24 fugues of the first book of Bach's Well-Tempered Clavier and the first 12 fugues of Shostakovich (Op. 87) (Giraud et al., 2015), totalling about 1,000 segments.[6] The dataset is built on several musicological references, such as the analysis of Bruhn (1993), and was produced and cross-checked by several annotators. The annotations give the symbolic positions of the occurrences of subjects and countersubjects, as well as cadences and pedals (Fig. 5.5). Slight modifications of the thematic patterns, such as varied start or ending or delayed resolutions, are also reported. Our dataset differs from the ones presented in Sect. 5.4.2: the length and the position of each pattern in each voice are given, in terms of the number of measures and beats, and several analytical elements beyond patterns are specified.

The purpose of this dataset is to give "correct" analytical elements for evaluable tasks, that should be part of a more complete analysis. In most of the fugues, there is consensus between theorists on some very technical points. Indeed, considering again Fig. 5.2, all our sources perfectly agree on the definition of the subject in 16 out of the 24 fugues. There may be also agreements on modulations, some cadences, and so on. The algorithms can thus be evaluated on all these points. We also reported ambiguous definitions of the subject. Further work should be done to encode even more ambiguities concerning the exact boundaries of other elements constituting the analysis.

5.4.4 Evaluating Segmentation and Form Analysis Algorithms

Even once a reference dataset is established, there are several ways to assess the efficiency of a segmentation algorithm or of a form analysis system against this dataset taken as a "ground truth": one can focus on a segmentation, on some precise borders, on a more semantic segmentation with relations between the sections, or on a subjective assessment of the quality of the algorithm.

Frame-Based Evaluation—How Often Does the Algorithm Predict the Right Section? On audio files, segmentation can be seen as a prediction of a label for every audio "frame", for example with a prediction every 40 milliseconds. On symbolic data, algorithms may also predict sections at the resolution of the smallest symbolic duration, or at another resolution, such as one quarter note or one measure. The most simple evaluation is then to consider frames one by one and to compute the ratio

[6] http://www.algomus.fr/datasets

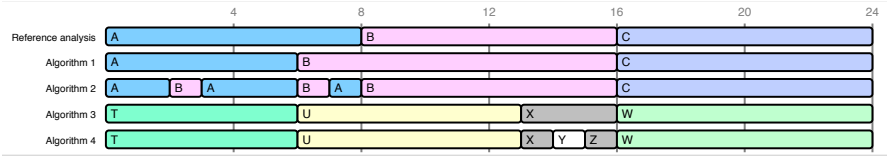

Fig. 5.6 Comparing segmentation labels from a reference analysis (top) against several algorithm predictions on a piece with 24 measures

of correctly predicted frames over the total number of frames. This is equivalent to comparing the *length* of the predicted sections in the computed analysis and in the ground truth: what is the proportion of section A that is found by the algorithm? For example, in Fig. 5.6, considering the measures as frames, 22 out of 24 frames are "correctly" predicted by both algorithms 1 and 2.

However, these evaluations need to assert which label in one analysis is equivalent to another label in the other analysis. When the predicted labels are different from those of the ground truth (such as for algorithms 3 and 4 in Fig. 5.6), one solution can be to link each computed segment with the largest overlapping segment in the ground truth (T with A, U with B and W with C), giving here 21 out of 24 frames for both algorithms 3 and 4. These evaluations may not behave correctly when a ground truth section is mapped to several predicted sections. Here algorithm 3 can be considered to have a better result than algorithm 4, because section B is segmented into only two sections, U and X. A solution is to measure the *mutual information* between the algorithm output and the ground truth. Extending the ideas of Abdallah et al. (2005), Lukashevich (2008) defines scores reporting how an algorithm may over-segment (S_o) or under-segment (S_u) a piece compared to the ground truth. These scores are based on normalized entropies.

On algorithms segmenting audio signals, all these evaluation measures, and other ones, were compared in a meta-analysis of the MIREX 2012 structural segmentation task results (Smith and Chew, 2013). On symbolic data, one can also consider the *number of notes* in each section—"How many notes are correctly put into section A?"—thus preventing biases that might arise from assigning greater weight to long notes than to short notes.

Boundary-Based Evaluation—How Many Section Boundaries Are Correctly Predicted? Frame-based evaluations ignore the actual boundaries of the segmentation. In particular, they do not take into account the fact that an algorithm may make too many transitions between sections. In Fig. 5.6, all the methods presented above give the same evaluation for algorithms 1 and 2, but algorithm 1 may be preferable because it predicts longer sections.

An evaluation can thus focus on *section boundaries*. One can compute the usual *precision* (ratio of correct predictions to all predictions) and *recall* (ratio of correct predictions to all ground-truth annotations). For example, in Fig. 5.6, there are 2 boundaries in the ground truth (A/B, B/C). Algorithm 1 successfully predicts the B/C boundary; it thus has a recall of 1/2 and a precision 1/2 (2 predicted boundaries,

including the false boundary A/B). Algorithm 2 successfully predicts the two ground truth boundaries: it has a perfect 2/2 recall. However, as it predicts 4 other false boundaries, its precision is only 2/6.

These evaluations are well adapted to melodic segmentation. Again, one has to decide which boundary in one analysis is equivalent to another boundary in the other analysis. In symbolic data, one can require that the boundaries *exactly* coincide: most of the time, a phrase, a section, or even some events such as a modulation, starts on a precise note or chord, and the goal of segmentation could be to retrieve this exact point. One can also allow some flexibility with a tolerance window, either in audio time (0.5 seconds, 3 seconds) (Sargent et al., 2011), or, for symbolic data, in terms of notated duration ("two quarter notes", "one measure") or in number of notes in a monophonic sequence ("one note", "two notes").

Section-Based Evaluation—How Many Sections Are Correctly Predicted? One can also choose to focus on whole sections, evaluating the proportion of sections correctly predicted. Again, thresholds can be added to this detection; should the start and the end exactly coincide? In Fig. 5.6, depending on the threshold, the performance of algorithm 1 will be evaluated to either 1/3 (only section C) or 3/3 (all sections).

Evaluating Form Analysis—Is My Computer a Good Music Analyst? Implementing frame-based, boundary-based and section-based evaluations allow us to quantify different aspects of the *segmentation* task. But these procedures do not evaluate other elements such as structural labelling. As we mentioned in Sect. 5.2.3, the challenge of *form analysis* needs to go beyond segmentation. In Fig. 5.6, are the A/B/C and T/U/W labels only symbols, or do they have some semantics? To evaluate an algorithm aiming to analyse a pop song, a fugue, or a sonata form movement, an evaluation should be conducted on the segmentation elements but also on the global output, including the semantics of the proposed analysis.

The results given by the computer should be compared to a reference analysis, when it exists, or, better, evaluated by musicologists, asserting how close a particular analysis is to a "musically pertinent" analysis, keeping in mind that there may be several pertinent analyses. Along with specific comments, some subjective notations such as "bad", "correct" or "good" can be used. Such an evaluation can pinpoint the strengths and the weaknesses of an algorithm on different pieces in the same corpus. Ideally, such an evaluation should be done by several experts.

Of course, as there is no absolute "musicologically correct" analysis, even this expert evaluation cannot be considered final and only evaluates what is expected according to a particular analysis model. Nevertheless, students in music analysis are evaluated on their homework, both on technical formal points (harmonic progressions, cadences, segmentations, etc.) and even on the aesthetic aspects of their work. Algorithms need to be evaluated too, and the difficulty of developing sound methodologies for carrying out such evaluation should not deter us from attempting to do so.

5.5 Discussion and Perspectives

There are many MIR and CMA challenges in segmentation and form analysis, in designing new algorithms, establishing reference datasets and conducting evaluations. Segmentation has been studied more extensively on audio signals than it has on symbolic representations, and some evaluation datasets are already available. Some studies have now begun to specifically address form analysis challenges, but more work is needed in this field of research.

One should remember that musical forms evolved (and perhaps were even designed) for *pedagogical purposes* in composition, as well as in music analysis and performance. The student who follows a lecture on music analysis learns to take a score and to recognize and analyse a fugue, a variation, or a sonata form. The student in harmony or composition learns how to write something that should sound like a fugue by Bach, respecting some rules or not. This is also true of performers: the artist playing a fugue has in some way to play with this special form, choosing between several plausible interpretations.

All these forms are also *building blocks* or *techniques* that one can find inside large-scale works. If you know how to hear, how to analyse, how to perform, or perhaps how to write a fugue, you will be able to hear, to analyse, to perform, or to write a fugato passage in some larger work. A further MIR/CMA challenge is to detect fragments of forms inside large works, or, more generally, to describe an unknown score with more elements than a segmentation. A solution is to test several known forms and take the best matched form, but further research should be done to propose better pipelines that predict the musical form on-the-fly.

Returning to the original question, "Can a computer understand musical forms?", we are not even sure that, as casual or experienced listeners, or even as music theorists, we ourselves hear and *understand* a fugue or a sonata form correctly. However, we all know that there is not a unique way to hear or to understand any musical piece—and it is not the relation to the form that makes a piece a pleasure to listen to. What we do know is that the process of learning to hear, to analyse, to play, and even to write such forms is important in musical education. Knowledge of these forms is one of the key ingredients for a better understanding of repertoires and genres.

We believe that algorithms in computational musical analysis are in their infancy: they are like students in a first-year music analysis classroom. They learn or they infer rules. They need to be evaluated—even if there are many different analyses that can be done, some are definitely more correct than others. Perhaps these student algorithms do not understand the big picture with respect to musical form, but they are learning how to handle some analytical concepts—thematic analysis, segmentation, and other musical parameters—that can also be useful for other MIR applications. One day, algorithms may perhaps manage to do more complete analyses, breaking the rules and including aesthetic and comparative elements.

Acknowledgements This work was supported by a grant from the French Research Agency (ANR-11-EQPX-0023 IRDIVE).

References

Abdallah, S., Noland, K., and Sandler, M. (2005). Theory and evaluation of a Bayesian music structure extractor. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 420–425, London, UK.

Bimbot, F., Deruty, E., Sargent, G., and Vincent, E. (2012). Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, pages 235–240, Porto, Portugal.

Bimbot, F., Sargent, G., Deruty, E., Guichaoua, C., and Vincent, E. (2014). Semiotic description of music structure: an introduction to the quaero/metiss structural annotations. In *Proceedings of the AES 53rd International Conference on Semantic Audio*, pages P1–1, London, UK.

Browles, L. (2005). Creating a tool to analyse contrapuntal music. Bachelor Dissertation, University of Bristol, UK.

Bruhn, S. (1993). *J. S. Bach's Well-Tempered Clavier. In-Depth Analysis and Interpretation*. Mainer International.

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)*, La Habana, Cuba.

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation. *Music Perception*, 23(3):249–268.

Cambouropoulos, E., Crochemore, M., Iliopoulos, C. S., Mohamed, M., and Sagot, M.-F. (2005). A pattern extraction algorithm for abstract melodic representations that allow partial overlapping of intervallic categories. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, pages 167–174, London, UK.

Caplin, W. E. (2000). *Classical Form: A Theory of Formal Functions for the Instrumental Music of Haydn, Mozart, and Beethoven.* Oxford University Press.

Caplin, W. E., Hepokoski, J., and Webster, J. (2009). *Musical Form, Forms & Formenlehre—Three Methodological Reflections*. Leuven University Press.

Chai, W. (2006). Semantic segmentation and summarization of music: Methods based on tonality and recurrent structure. *IEEE Signal Processing Magazine*, 23(2):124–132.

Charlier, C. (2009). *Pour une lecture alternative du Clavier Bien Tempéré*. Jacquart.

Chen, H.-C., Lin, C.-H., and Chen, A. L. P. (2004). Music segmentation by rhythmic features and melodic shapes. In *IEEE International Conference on Multimedia and Expo (ICME 2004)*, pages 1643–1646.

Cook, N. (1987). *A Guide to Musical Analysis*. Oxford University Press.

Czerny, C. (1848). *School of Practical Composition*. R. Cocks & Co.

Dannenberg, R. B. and Goto, M. (2009). Music structure analysis from acoustic signals. In Havelock, D., Kuwano, S., and Vorländer, M., editors, *Handbook of Signal Processing in Acoustics*, pages 305–331. Springer.

Dannenberg, R. B. and Hu, N. (2002). Pattern discovery techniques for music audio. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, pages 63–70, Paris, France.

David, L., Giraud, M., Groult, R., Louboutin, C., and Levé, F. (2014). Vers une analyse automatique des formes sonates. In *Journées d'Informatique Musicale (JIM 2014)*.

Ewert, S., Müller, M., and Grosche, P. (2009). High resolution audio synchronization using chroma onset features. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009)*, pages 1869–1872.

Farbood, M. (2010). A global model of musical tension. In *Proceedings of the 11th International Conference on Music Perception and Cognition (ICMPC 11)*.

Foote, J. (1999). Visualizing music and audio using self-similarity. In *Proceedings of the 7th ACM International Conference on Multimedia (MM99)*, pages 77–80, Orlando, FL.

Ghias, A., Logan, J., Chamberlin, D., and Smith, B. C. (1995). Query by humming: Musical information retrieval in an audio database. In *Proceedings of the 3rd ACM International Conference on Multimedia*, pages 231–236, San Francisco, CA.

Giraud, M., Groult, R., Leguy, E., and Levé, F. (2015). Computational fugue analysis. *Computer Music Journal*, 39(2):77–96.

Giraud, M., Groult, R., and Levé, F. (2012). Detecting episodes with harmonic sequences for fugue analysis. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, Porto, Portugal.

Goto, M. (2004). Development of the RWC music database. In *International Congress on Acoustics (ICA 2004)*, pages I–553–556.

Goto, M. (2006). AIST annotation for the RWC music database. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)*, pages 359–360, Victoria, Canada.

Hamanaka, M. and Tojo, S. (2009). Interactive GTTM analyzer. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, pages 291–296, Kobe, Japan.

Hepokoski, J. and Darcy, W. (2006). *Elements of Sonata Theory: Norms, Types, and Deformations in the Late-Eighteenth-Century Sonata*. Oxford University Press.

Hirata, K. and Matsuda, S. (2002). Interactive music summarization based on GTTM. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)*, Paris, France.

Hsu, J. L., Liu, C. C., and Chen, A. (1998). Efficient repeating pattern finding in music databases. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM 1998)*, pages 281–288, Bethesda, MD.

Jiang, N. and Müller, M. (2013). Automated methods for analyzing music recordings in sonata form. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, Curitiba, Brazil.

Keller, H. (1965). *Das Wohltemperierte Klavier von Johann Sebastian Bach*. Bärenreiter.

Kirlin, P. B. and Jensen, D. D. (2011). Probabilistic modeling of hierarchical music analysis. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, pages 393–398, Miami, FL.

Klapuri, A. (2011). Pattern induction and matching in music signals. In Ystad, S., Aramaki, M., Kronland-Martinet, R., and Jensen, K., editors, *Exploring Music Contents: 7th International Symposium, CMMR 2010, Málaga, Spain, June 21–24, 2010. Revised Papers*, volume 6684 of *Lecture Notes in Computer Science*, pages 188–204. Springer.

LaRue, J. (1970). *Guidelines for Style Analysis*. Harmonie Park Press.

Lemström, K. and Laine, P. (1998). Musical information retrieval using musical parameters. In *Proceedings of the International Computer Music Conference (ICMC 1998)*, pages 341–348, Ann Arbor, MI.

Lerdahl, F. and Jackendoff, R. S. (1983). *A Generative Theory of Tonal Music*. MIT Press.

Lerdahl, F. and Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception*, 24(4):329–366.

Levy, M., Noland, K., and Sandler, M. (2007). A comparison of timbral and harmonic music segmentation algorithms. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, pages 1433–1436, Honolulu, HI.

Lukashevich, H. M. (2008). Towards quantitative measures of evaluating song segmentation. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pages 375–380, Philadelphia, PA.

Maddage, N., Li, H., and Kankanhalli, M. (2009). A survey of music structure analysis techniques for music applications. In Grgic, M., Delac, K., and Ghanbari, M., editors, *Recent Advances in Multimedia Signal Processing and Communications*, volume 231 of *Studies in Computational Intelligence*, pages 551–577. Springer.

Marsden, A. (2010). Schenkerian analysis by computer. *Journal of New Music Research*, 39(3):269–289.

Marx, A. B. (1837–1847). *Die Lehre von der musikalischen Komposition, praktisch theoretisch*. Breitkopf und Härtel.

Mauch, M., Cannam, C., Davies, M., Dixon, S., Harte, C., Kolozali, S., Tidhar, D., and Sandler, M. (2010). OMRAS2 metadata project 2009. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, Utrecht, The Netherlands.

Mongeau, M. and Sankoff, D. (1990). Comparison of musical sequences. *Computers and the Humanities*, 24(3):161–175.

Muellensiefen, D., Pearce, M., and Wiggins, G. (2008). A comparison of statistical and rule-based models of melodic segmentation. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, pages 89–94, Philadelphia, PA.

Nattiez, J.-J. (1987). *Musicologie générale et sémiologie*. Christian Bourgeois.

Paulus, J. and Klapuri, A. (2008). Music structure analysis using a probabilistic fitness measure and an integrated musicological model. In *International Conference on Music Information Retrieval (ISMIR 2008)*, pages 369–374.

Paulus, J., Müller, M., and Klapuri, A. (2010). Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, pages 625–636, Utrecht, The Netherlands.

Peeters, G. (2007). Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria.

Peeters, G. and Deruty, E. (2009). Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *International Workshop on Learning the Semantics of Audio Signals*, pages 75–90.

Prout, E. (1910). *Analysis of J. S. Bach's Forty-Eight Fugues (Das Wohltemperirte Clavier)*. E. Ashdown.

Rafael, B. and Oertl, S. M. (2010). MTSSM – A framework for multi-track segmentation of symbolic music. In *World Academy of Science, Engineering and Technology Conference*, Turkey.

Rafailidis, D., Nanopoulos, A., Manolopoulos, Y., and Cambouropoulos, E. (2008). Detection of stream segments in symbolic musical data. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR 2008)*, Philadelphia, PA.

Ratner, L. (1980). *Classical Music: Expression, Form, and Style*. Schirmer.

Ratz, E. (1973). *Einführung in die musikalische Formenlehre*. Universal Edition. Second edition.

Reicha, A. (1824). *Traité de haute composition musicale*. A. Diabelli.

Rodríguez-López, M. E., Volk, A., and Bountouridis, D. (2014). Multi-strategy segmentation of melodies. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 207–212, Taipei, Taiwan.

Rosen, C. (1980). *Sonata Forms*. W. W. Norton.

Sargent, G., Bimbot, F., and Vincent, E. (2011). A regularity-constrained Viterbi algorithm and its application to the structural segmentation of songs. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, FL.

Schenker, H. (1935). *Der freie Satz*. Universal Edition.

Schoenberg, A. (1967). *Fundamentals of Musical Composition*. Faber & Faber.

Sidorov, K., Jones, A., and Marshall, D. (2014). Music analysis as a smallest grammar problem. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*, pages 301–3016, Taipei, Taiwan.

Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., and Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, Miami, FL.

Smith, J. B. L. and Chew, E. (2013). A meta-analysis of the MIREX structural segmentation task. In *Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pages 251–6, Curitiba, Brazil.

Tovey, D. F., editor (1924). *Forty-Eight Preludes and Fugues by J. S. Bach*. Associated Board of the Royal Schools of Music.

Weng, P.-H. and Chen, A. L. P. (2005). Automatic musical form analysis. In *Proceedings of the International Conference on Digital Archive Technologies (ICDAT 2005)*, Taipei, Taiwan.

Wiering, F., Nooijer, J. D., Volk, A., and Tabachneck-Schijf, H. J. M. (2009). Cognition-based segmentation for music information retrieval systems. *Journal of New Music Research*, 38(2):139–154.