Perception and modeling of segment boundaries in popular music

Perception and Modeling of Segment Boundaries in Popular Music

Ph.D. thesis

Michael J. Bruderer

The work described in this thesis was financially supported by Philips Research Laboratories Eindhoven and was carried out under the auspices of the J.F. Schouten School for User-System Interaction Research.

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Bruderer, Michael J.

Perception and modeling of segment boundaries in popular music / by Michael J. Bruderer.
- Eindhoven: Technische Universiteit Eindhoven, 2008. - Proefschrift. -
ISBN 978-90-386-1229-4
NUR 778
Keywords: Music perception / Music structure

Printing: Universiteitsdrukkerij Technische Universiteit Eindhoven

Perception and modeling of segment boundaries in popular music

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr.ir. C.J. van Duijn, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op donderdag 10 april 2008 om 16.00 uur

door

Michael Jonas Bruderer

geboren te Herisau, Zwitserland

Dit proefschrift is goedgekeurd door de promotor:

prof.Dr. A.G. Kohlrausch

Copromotoren:
dr. M.F. McKinney
en
dr. E. Cambouropoulos

Contents

1 Introduction

The present thesis deals with the perception and modeling of musical segmentation, i.e., how a musical piece can be partitioned into its constituent entities. The segmentation process is possible through the organization of the musical events in the piece. The listener perceives the organisation through the presence of cues in the music which can be used to segment the music.

Theories on music segmentation often make the distinction between changes in the temporal sequence of musical events and the analysis of similar segments or repetitions (e.g., Lerdahl & Jackendoff, 1983). These two principles are often related to the Gestalt rules of proximity and similarity, which were originally developed for the visual domain (Wertheimer, 1923) and have been shown to be valid for the segregation of an acoustic complex into perceptual streams (Bregman, 1990). The proximity rule states that elements close together temporally are grouped to an entity. The similarity rule states that two objects tend to be grouped together that appear similar. While proximity can be defined as to predict segment boundaries whenever there is a change in the properties of a sound event, the definition of similarity is more ambiguous (Cambouropoulos, 2001b). The difficulty arises from the fact that similar sound events often are not identical, but variations, which adds the additional issue of setting a criterion under which conditions two events are still considered similar and when not. Furthermore, repetition can independently occur along different dimensions of music (i.e., repetition of pitch intervals versus repetition of note duration).

1.1 Perception of music structure

The seminal work of Lerdahl and Jackendoff (1983), "A generative theory of tonal music" (GTTM), is a musicological theory for analysis of structure in music. In addition to being grounded in musicological theory it has also served as the basis for a number of studies on the *perception* of music structure (Deliège, 1987; Clarke & Krumhansl, 1990; Frankland & Cohen, 2004). Part of GTTM is the set of "grouping preference rules", which indicate where segment boundaries should be placed in the score (cf. Appendix C). The rules are based on musicological constructs, such as relative duration of consecutive notes, which serve as building blocks for higher order grouping and segmentation. Although these rules are musicological constructs, they have also

been tested for their perceptual relevance (Deliège, 1987; Frankland & Cohen, 2004). Deliège (1987) asked subjects to listen to and segment short excerpts of classical music, consisting of between 3 to 16 notes, by indicating segment boundaries on a series of dots, where each dot represented a note on the musical score. She found that subjects' preferred segments were generally in line with those predicted by grouping preference rules of GTTM. In a second study she placed two rules in conflicting places. Although not all possible combinations of the relative strengths of each cue were systematically tested, several rules were found to be more salient. These salient cues were timbre changes, register changes, and dynamic changes. In order to develop a computational model for music segmentation based on the qualitative rules of GTTM, Frankland and Cohen (2004) quantified four of the grouping preference rules and tested them against perceptual parsing of melodies. Subjects were asked to press a key on a computer keyboard whenever they heard a section start or end. The perceptual evaluation of the four rules showed that two were sufficient to predict the perceptual parsing. These two rules were attack-point (a long note in between two short notes) and to a lesser extent a rest. All in all these results support the perceptual validity of the grouping preference rules proposed in GTTM. It seems, thus, that certain rules, or cues, are more salient for conveying segment boundaries. If this relation between the weight of the different cues is the same for longer, more complex pieces, however, is still unclear. Furthermore, polyphonic pieces may include additional cues, which have so far not been identified.

Several considerations have to be taken into account when studying the perception of segment boundaries. One issue is related to the context in which musical stimuli are presented and to the length of the stimuli. If the musical material is very short, such as short excerpts consisting of a few notes (Deliège, 1987) or short melodies (Ferrand, 2004; Frankland & Cohen, 2004; Schaefer, Murre, & Bod, 2004), it is questionable whether the segmentation patterns can be related to longer and more complex musical pieces. The use of stimuli with a longer duration on the other hand, for instance complete tonal pieces (Krumhansl, 1996) or atonal pieces (Deliège & Ahmadi, 1990), makes it more difficult to identify and separate the individual cues responsible for the perception of segment boundaries.

A second methodological issue is how to register subjects' responses and how to define the precise temporal position of the boundary. Previous studies have asked subjects to indicate segment boundaries by pressing a foot pedal (Clarke & Krumhansl, 1990), a click with a computer mouse (Ferrand, 2004), or by pushing a key on the

computer keyboard (Deliège & Ahmadi, 1990). The difficulty with analyzing these indicated boundaries is that it is likely that given enough precision in the boundary recording, two subjects do not indicate the boundary precisely at the same time instance. If the score of the music piece is available one obvious method is to ask subjects to indicate the precise position of their perceptual boundary indications on the score notation, however, with the disadvantage of reducing participants to musically trained subjects (Clarke & Krumhansl, 1990). Another method is to ask subjects to describe for each boundary why it is salient, which can then be used to specify the precise boundary position in the score notation (Deliège, Mélen, Stammers, & Cross, 1996). For short stimuli, instead of the score notation also a line of dots has been used, where each dot represented a note (Deliège, 1987). If the score notation is not available or the stimuli consist of polyphonic pieces where the boundary indication can not easily be associated to the score properties, the method has to be adapted.

Moreover, once the boundary positions have been identified, it is not evident how to accumulate the boundary indications across different trials and subjects. If the boundaries have been indicated on a sequence of dots or on the score this is trivial. If, however, subjects were asked to indicate the segment boundaries by pressing, for instance a key, one way to accumulate boundary indications is to assign each boundary indication to the played note at that time instance (Frankland & Cohen, 2004), or to a set of notes (Schaefer et al., 2004). However, if the score constitutes several simultaneously played notes or subjects anticipate or delay their responses, the assigned note may not be the correct place of the boundary occurrence. A slightly different way is to use a temporal window, which can be related to the meter, i.e., the successive alternation of strong and weak beats, and integrate all boundary indications within, for instance, a one-beat (Spiro & Klebanov, 2006) or two-beat window (Krumhansl, 1996). The most versatile method, which is independent of metrical or note information, is to smoothen the boundary profile and take the maxima of the smoothed boundary profile as segment boundaries (Ferrand, 2004). One assumption of the studies which integrate the boundary indications across subjects is that the number of boundaries with which different subjects indicate a boundary can be taken as a measure of the boundary's salience, without giving empirical evidence.

Musicians often add an element of expression in music performance, which can influence the perception of structural elements. Musicians playing a piece often do not precisely adhere to the music notation as given in the score, but rather *interpret* the

music and shorten or lengthen certain notes and add micro-pauses. These expressive cues are often perceptually related to the structure of the piece. For instance, ritardandi and accelerandi (Todd, 1995) as well as crecendo and decrescendo (Todd, 1992) are often associated with the indication of musical phrases. Listeners expect these changes in the note duration and therefore have difficulties in perceiving lengthened notes at phrase boundaries (Repp, 1992). This observation was used in a recent study to identify segment boundaries by extracting lengthened notes (Chuan & Chew, 2007). In all, these results suggest that using expressively performed music aids listeners in perceiving the structural properties of the piece.

From previous studies on music structure perception it is unclear whether polyphony has an influence on the perception of segment boundaries. As we will see in the following, formal musicological models can currently only be applied to monophonic lines. It would be interesting to study how the segmentation of a monophonic version of a piece differs from the segmentation of polyphonic versions of the same piece.

1.2 Algorithms for music segmentation

There are many applications for the automatic analysis of music structure. Typically two different approaches can be distinguished. One approach is coming from musicology, where the goal is to extract segments using the score notation. The other approach stems from music information retrieval with the aim of automatically segmenting audio recordings.

1.2.1 Segmentation with formal musicological models

The common form to represent music used in musicology is the score notation (or the more modern MIDI format*). The score notation comprises a series of tone-heights and durations for each instrument as well as additional expressive information. Formal musicological segmentation algorithms generally reduce the input to a set of notes retaining only the properties of tone-height (pitch), the starting point and the duration of a note, as well as sometimes the level. This reduced representation is then analyzed

*MIDI is the abbreviation of Musical Instrument Digital Interface, a well-known protocol used to connect and control electronic instruments. Its representation of the music is similar to the score notation in that for each note the pitch, level, starting point, and duration are defined. The instrument with which the note is played is generally given for the whole track (a collection of notes), however, an instrument change event can also be given within a track.

for discontinuities. The discontinuities can be based on cues such as a pitch intervals or a level change (Tenney & Polansky, 1980; Lerdahl & Jackendoff, 1983; Cambouropoulos, 1998; Temperley, 2001). In the following a short overview over models of interest for the present thesis is given. An extended overview of the models used in this thesis is given in Appendix C.

In order to estimate these local discontinuities, a distance between two successive notes has to be defined. Instead of combining the different parameters into one single measure, often each cue is individually analyzed for changes and the sum of all differences is taken as the global segmentation profile. This process, however, introduces the problem of how to weight the individual cues, in order to indicate whether certain cues are more important for segmentation than others.

The grouping rules of "A generative theory of tonal music" (GTTM) (Lerdahl & Jackendoff, 1983) propose a set of rules for where a segment boundary can be set within a monophonic sequence of notes. The rules can be classified into three types (Clarke & Krumhansl, 1990, p. 215):

> 1) Preference rules based on the Gestalt principles of proximity and similarity, the hierarchical level to which the rule applies being determined by the strength of the Gestalt features. These are essentially responsive to surface features of the music.
> 2) A preference rule based on the grouping effects of pitch structure, based on the disposition of stable and unstable harmonic elements within the framework of the tonal system. This rule is responsive to relatively deeper structural features of the music.
> 3) Preference rules based on the more abstract principles of symmetry and motivic similarity, or "parallelism".

This set of rules is then applied to a monophonic line. Because the theory is not explicitly defined and can thus not be directly implemented, an expert has to apply these rules. To formalize the theory, Frankland and Cohen (2004) have quantified four of the rules and Hamanaka, Hirata, and Tojo (2006) have attempted to implement large parts of the theory into a computer algorithm.

Melisma by Temperley (2001) is a set of models that can extract basic kinds of musical information. The model of interest here, the "melodic phrase structure" extracts segment boundaries based on three rules. The first rule states that a boundary can be

introduced at places of large inter-onset intervals and large offset-to-onset intervals. The second rule prefers phrases having a length of eight notes. The third rule, finally, states that segment boundaries should be placed in parallel to the metrical structure. These three rules were then applied on monophonic melodies by using a technique called dynamic programming to predict phrase boundaries.

Instead of a large set of rules, the "local boundary detection model" (LBDM) by Cambouropoulos (1998) uses only two general rules to predict segment boundaries. The model examines the properties of three or four consecutive notes and applies two rules: 1) set a boundary whenever there is an interval between two successive notes and 2) choose the larger of two consecutive intervals. In principle an interval can be any property of a sound event, however, in his revised version the melody is analyzed for pitch changes, inter-onset intervals, and rests (Cambouropoulos, 2001a). The strength of the boundary is defined by the size of the interval, i.e., if the interval is larger, the strength of the boundary also increases.

While some of these models assume that the metrical structure is independent of the grouping structure (Lerdahl & Jackendoff, 1983), the model proposed by Temperley (2001) first evaluates the metrical structure and uses the obtained meter to place phrase boundaries parallel to the metrical structure. There exists, however, also the idea to use the discontinuities in the musical surface to estimate the metrical structure (Cambouropoulos, 1998). It is, thus, not yet clear whether the metrical structure is the basis of the grouping structure, whether it follows from the grouping structure, or whether the two are completely independent.

One peculiarity in music theory is the avoidance of a formal definition of repetition, also called parallelism (Lerdahl & Jackendoff, 1983). Although repetition has been acknowledged to be important for music segmentation (Lerdahl & Jackendoff, 1983; Temperley, 2001), modeling repetition in computational models is a difficult task due to several constraints (Cambouropoulos, 2006a). Repeating patterns are often not identical, but ornamented or embody variations, which rises the issue of what patterns can still be considered repetitions. Furthermore, repetition in polyphonic music can occur in all voices as well as across different voices. From a perceptual point of view, however, it is unlikely that the listener does consciously perceive all repeating patterns equally, thus some patterns are more prominent than others. A further issue is the representation of the music and thus the coding of the representation. A sequence of notes can be defined, for instance, by the duration and pitch of each note. The

duration, however, can be coded as relative duration, i.e., quarter-note, or absolutely, i.e., in seconds. The extraction of repeating patterns may strongly be influenced by the way of representation. A final point concerns how to evaluate algorithms that extract repeating patterns. Presently little empirical data exists on the perception of repeating patterns. All these issues make it difficult to conceive an algorithm that can automatically extract repeating patterns.

A different way to music segmentation is to learn the phrasing structure of musical pieces from an annotated corpus. The underlying assumption is that subjects base their segmentation strategy on previously experienced phrase structures. In a study by Bod (2002) the frequency of occurrence of excerpts within a corpus with annotated segment boundaries predicted the presence of segment boundaries in unseen melodies. Using the Essen folk song collection, which contains approximately 5000 songs with manually identified segment boundaries (by musicologists, not perceptually), the author reported a correct phrase detection of more than 80%. A different approach is to look at music as a stream of information for which the entropy can be calculated for every note. Segment boundaries are then set at points with a change from high predictability to low predictability (Pearce & Wiggins, 2006). The advantage of these approaches is that the patterns where phrase boundaries occur are learned from the annotated corpus without the need of explicit rules. It is necessary, however, to have a large enough annotated corpus. So far, such a corpus only exists with annotations by musicologists and not for perceptually segmented pieces.

A limitation of segmentation models in musicology is that they are generally conceived for monophonic pieces. The underlying assumption is, thus, that the different voices in a polyphonic piece are perceptually independent (or can be separated) and can be segmented autonomously. However, recently it has been suggested that instead of using the voices as written in the score notation it is more meaningful to derive the perceptual voices from the polyphonic score notation first. Segmentation algorithms are then applied on the perceptual voices instead of the notated voices (Cambouropoulos, 2006b).

1.2.2 Segmentation using the audio recording

Audio recordings are the form in which music is most commonly represented. Furthermore, the audio recording often contains additional cues, mostly ignored by algorithms using the score notation, such as binaural properties and expressive

timing (e.g. deviations from the exact duration of the note and the introduction of micro-pauses). The audio recording of songs also often contains the recorded voice, which adds the semantics of the lyrics to the music. Thus, the audio recording adds additional parameters that may influence listeners' perception of boundaries, which is not necessarily present in the score notation or a MIDI representation.

In recent years a number of studies on automatic segmentation of Western popular music have been published. The basic approach of these algorithms is to first extract features from the signal and then to look for repetitions in the features. The features used previously have been:

- Timbral, such as the mel-frequency cepstral coefficients (MFCC) (Foote, 1999; Logan & Chu, 2000; Peeters, Burthe, & Rodet, 2002),
- Based on the harmonic content like constant cue transform (Lu, Want, & Zhang, 2004) or chromograms (Bartsch & Wakefield, 2001; Goto, 2006),
- Based on the rhythmic content (Rauber, Pampalk, & Merkl, 2002).

The features are extracted over the course of the piece of music using short time windows, resulting in a feature vector. Often, the distance between each two elements of the vector is calculated, resulting in a self-similarity matrix (Foote, 1999), which can then be further processed, for instance, by finding repeating patterns via dynamic time warping (Chai, 2005) or singular value decomposition (Foote & Cooper, 2003). However, all these algorithms are based on basic signal properties that have little relation to musicological aspects, and have not been verified musicologically or perceptually.

Recently, there has been an increasing interest in extracting also musicologically inspired features. For instance, Maddage (2006) extracted the rhythmical structure by detecting notes onset, the chord patterns, as well as parts with singing-voice. These features were then used to find repeating patterns. Although these features take more musically specific data into account, the extracted features are still relatively simple compared to the musicological information available in the score notation.

1.3 Goals of the thesis

In the previous sections a number of aspects in the perception and modeling of musical structure have been discussed. The present thesis attempts to extend the scientific

knowledge in several ways: 1) Previous studies that asked subjects to segment pieces of music (Krumhansl, 1996; Ferrand, 2004; Frankland & Cohen, 2004; Spiro & Klebanov, 2006) assumed that the number of subjects indicating a boundary within a certain time window can be taken as a measure for the salience of the boundary, without explicit testing this assumption. Only one study, Clarke and Krumhansl (1990), asked subjects also to rate the salience of their own indicated boundaries. In the present thesis we extend this method by asking subjects to rate a number of boundaries taken from the boundary indications across all subjects and compare the salience rating with the number of indications of these boundaries. 2) Previous studies either used monophonic melodies (Ferrand, 2004; Frankland & Cohen, 2004; Schaefer et al., 2004) or used complete pieces (Deliège & Ahmadi, 1990; Krumhansl, 1996; Ferrand, 2004; Spiro & Klebanov, 2006), but none compared the perception of segment boundaries for different versions of the same piece of music, i.e., monophonic versus polyphonic representations. 3) Formal musicological models that evaluate the musical surface for discontinuities have so far only been tested on short, monophonic melodies (Cambouropoulos, 2001a; Temperley, 2001; Frankland & Cohen, 2004). It is interesting to study whether these models are also apt to predict segment boundaries in complete pieces of music and to analyze which combination of cues is best for predicting perceptual segment boundaries.

One assumption taken for granted is that music can be segmented. As we have seen above, previous studies found that subjects are able to segment very different types of music if asked to do so (Deliège, 1987; Clarke & Krumhansl, 1990; Krumhansl, 1996; Frankland & Cohen, 2004; Spiro, 2007). Nevertheless, it may be possible that some kind of music is difficult to segment, or at least very ambiguous in its structural properties. As the focus of the present thesis is on music taken from Western popular music, which has generally a clear structure, it was expected that subjects are able to segment our selection of musical pieces.

In summary, the main questions to be addressed in this thesis are the following.

1. Is there a correlation between the number of subjects that indicate a segment boundary (within a time window) and an explicit salience rating of that boundary? Are there other measures for the salience of segment boundaries?

2. Is there an influence of polyphony on segmentation, i.e., do subjects segment pieces differently if only the melodic line is presented, compared to a polyphonic version or to a polyphonic recording?

3. When using the underlying cues of existing formal musicological models, which combination of cues leads to the best prediction of the perceptual boundary profiles?

1.4 Applications of music segmentation

Knowledge about the perception of structural boundaries can be used in many ways. However, in order to be useful, the identified segments often also need to be annotated, such as chorus, verse, or first theme. Assuming such annotations as given, several applications are imaginable.

One application is music browsing. By knowing the structure of musical pieces, one could, for example, not only skip from song to song, but also within a song, like skipping directly to the beginning of the chorus. Automatic structure analysis could also improve automatic DJ applications by identifying compatible sections of songs for mixing and sequencing. And the structural knowledge could also lead to music summarization, by for instance, extracting the chorus of each song of a large collection and then providing an overview of the content of the music collection.

Other possible applications are compression algorithms. Commonly-found compression algorithms, such as the zip format, use repetition as the attribute for reducing the size of files. Compression algorithms used for audio, on the other hand, rely on short-term perceptual properties to reduce the size of the audio signal. Music, and in particular popular music, is often highly repetitive, like the chorus-verse pattern. This repetition could be exploited to decrease the size of the compressed audio file.

Finally, automatic music structure analysis could be used to improve a variety of other music information retrieval applications, such as music transcription, classification, and comparison. Knowing the structure could help to extract additional cues, or help improving higher level algorithms, such as music transcription. Moreover, once the structure of a piece of music is known, more complex features of the repeated patterns could be extracted with the advantage that they only need to be extracted once and thus increase efficiency of complex feature extraction.

1.5 Outline of the thesis

The core of the thesis is organized as follows: The second chapter describes an experimental design used to investigate the perception of structural boundaries. Two experiments were conducted: In the first, subjects were asked to segment monophonic melodies taken from Western popular music; in the second, subjects were asked to rate the salience of a selection of segment boundaries taken from the first experiment and give a description of cues responsible for boundaries. The segmentation pattern obtained in the first experiment was correlated with the mean salience ratings from the second experiment. Finally, the boundary indications were compared with the segmentation profiles predicted by three musicological models.

The third chapter extends the investigation to polyphonic music. The experimental method from Chapter 2 was used to obtain segmentation and salience data from two different polyphonic representations of the same songs. Finally, the data from these experiments were compared with those obtained using monophonic stimuli. The goal was to examine the influence of the representation on the perception of segment boundaries.

The fourth chapter dissects two formal models taken from musicology into their underlying cues. The cues are then evaluated individually as well as combined for their ability to predict the perceptual segment boundary data obtained in the previous chapters. The most salient predicted boundaries were also compared with the most salient perceptual boundaries, to investigate how well the most salient predicted boundaries correspond to them. Finally, the optimal combination of cues for the two polyphonic representations was evaluated. The results of the experiments were then compared across the three different representations. The goal was to examine the influence of the representation on the perception of segment boundaries.

The thesis concludes with a summary of the main findings and suggestions for further research.

2 The perception of structural boundaries in melody lines of Western popular music*

Two experiments were conducted to investigate the perception of structural boundaries in six popular music songs. In the segmentation experiment, subjects were asked to indicate perceived segment boundaries in monophonic representations of the songs, synthesized from the MIDI score. In the salience rating experiment, subjects were asked to rate the salience of a number of boundaries selected from the outcome of the segmentation experiment, and to describe the perceptual cues for each boundary. The segmentation experiment showed that there is a wide variety in the number and temporal positions of perceived boundaries across subjects. However, certain boundaries in the music are indicated by nearly all subjects. The salience rating experiment showed a moderate correlation between subjects' boundary salience ratings. Comparing the outcome of the two experiments, we found a significant correlation between the frequency of boundary indications and the corresponding salience rating of that boundary. These findings suggest that both methods can be used equally well for evaluating the perceptual boundaries. The perceptual boundaries were also compared to boundaries predicted by three musicological models. The comparison of the perceptual boundaries with the predicted boundaries showed a moderate correlation between the perceptual and predicted boundaries.

*This chapter is based on Bruderer, M.J., McKinney M.F., and Kohlrausch, A. "The perception of structural boundaries in melody lines of Western popular music", accepted for publication in Musicae Scientiae with the ESCOM young researcher award 2006.

2.1 Introduction

One of the spontaneous processes in listening to music is the perception of segmentation structure. Although several theories (Tenney & Polansky, 1980; Lerdahl & Jackendoff, 1983; Cambouropoulos, 2001a; Temperley, 2001) have addressed the topic, the relation between theory of musical structure and the perception of structural boundaries is still not clear.

Gestalt psychologists postulated a variety of rules governing perceptual organization (e.g., Wertheimer, 1923), which define how to group objects into larger entities. Although originally conceived for the visual domain, these rules were also applied to and tested for auditory perception (i.e., Bregman, 1990; Deutsch, 1999). In music theory, these rules were applied in, for instance, "A Generative Theory of Tonal Music" (GTTM) by Lerdahl and Jackendoff (1983). The component in GTTM that applies the Gestalt rules to segmentation of monophonic lines is called the "grouping structure". For our purpose, the most applicable portion of GTTM is the set of "Grouping Preference Rules" (GPR), which are seven rules that can be classified into three types (Clarke & Krumhansl, 1990). The first type of rules are based on the two Gestalt principles of *proximity* and *similarity* (changes). The second type of rules are based on the *deep level harmonic structure* (time-span reduction and prolongation stability). The third type of rules are based on *repetition*, called "parallelism".

Among the rules of GPR, two rules (GPR 2 and 3) are based on local details in the musical surface: The proximity rule sets a group boundary if there is a slur, rest, or a long note between two short notes. The change rule sets a boundary when a change in register, dynamics, articulation, or length occurs. The remaining rules of GPR state how to combine the outcome of these two rules or relate the boundaries to other features such as reduction, symmetry, or repetition of the monophonic line.

While the grouping rules were developed from musicological theory, their *perceptual* validity has also been assessed by various studies. Frankland and Cohen (2004) quantified four of the rules postulated by GTTM, analyzed a number of nursery tunes and classical melodies, and asked subjects to segment the same pieces. They then compared the perceptual profiles with the predictions of the quantified rules. From the four rules, only two, *attack-point* (long note in between two short notes) and *rest*, were needed to account for the perceptual segmentations.

Deliège (1987) asked subjects to segment excerpts consisting of a few notes taken

from the classical music repertoire, where each excerpt contained one of the grouping rules of GTTM. The obtained segmentations largely supported the rules postulated in GTTM. Based on the segmentation patterns she also suggested two additional rules to complement the rules of GTTM; *change in timbre*, and *change in harmony*.

Clarke and Krumhansl (1990) asked musicians to segment two complete pieces of classical music and to describe and rate the salience of selected boundaries. The obtained boundaries and their descriptions were found to be closely predicted by the rules of GTTM. In general, these findings validate the perceptual relevance of the GTTM grouping rules.

Apart from GTTM, other theoretical models have been proposed for the segmentation of melodies, including the "Local Boundary Detection Model" (LBDM) by Cambouropoulos (2001a) and the "Melisma" model by Temperley (2001). Both models are based on local discontinuities in melodies, for instance, a long note in between two short notes. The main difference between GTTM and these two models is that the placement of the segment boundaries in GTTM requires an expert listener having high level musicological knowledge and, therefore, cannot be automated. LBDM and Melisma, on the other hand, are defined with explicit rules allowing quantitative implementation in a computer program. Although the LBDM and Melisma have been tested for their validity before, their evaluation was based on segmentations provided by music analysts, which do not necessarily correspond to perceptual boundaries.

A common property of all three models is that they evaluate the very local context, usually a few consecutive notes and often referred to as musical "surface", to estimate the presence of a segment boundary. Higher level processes, such as repetition (i.e., Cambouropoulos, 2006a) or reduction (i.e., Lerdahl & Jackendoff, 1983) are not taken into account. This focus on the local context is of relevance here, because it is generally assumed that structure in music is perceived in a hierarchical manner (Meyer, 1956; Schenker, 1935; Deutsch & Feroe, 1981; Lerdahl & Jackendoff, 1983). The assumption is that global hierarchical structure has a strong influence on the perception of segment boundaries, i.e., the perception of segment boundaries is not only influenced by the local context, but also by perceiving global structural properties. This theoretical view has recently been challenged by a series of experiments that found that the local context perceptually prevails over global processing of musical structure (Tillmann & Bigand, 1998; Tillmann, Bigand, & Madurell, 1998). Another perceptual study with a series of experiments on the process of cue abstraction found a strong influence of surface

cues on segmentation (Deliège et al., 1996). It seems, thus, that the importance of hierarchical structure is rather of theoretical nature and not necessarily perceived.

An important factor in a music perception experiment is musical training, as it may bias the outcome of the experiment. It has been shown, for instance, that musical training influenced the judgment of octave equivalence or note scale functions of different tones (see Smith, 1997). A more recent study, however, suggests that through exposure to music, nonmusicians can accomplish many musicological tasks, although with lower performance than musicians (Bigand & Poulin-Charronnat, 2006). There were, for instance, only small differences between musicians and nonmusicians in abstracting the trajectory of melodies, in expectations of chords within a consonant/dissonant context, or in learning of a new musical grammar. Further support of a low influence of musical training is given by an experiment in which subjects were asked to solve musical puzzles (Tillmann et al., 1998). A piece of music was segmented into smaller pieces, randomly arranged, and subjects were asked to place the segments into the correct order. For simple as well as more difficult puzzles, both musicians and nonmusicians had difficulties in solving the problem. Furthermore, both groups of subjects seemed to perceive the same kind of harmonic structure, suggesting that nonmusicians and musicians use the same underlying principles, but musicians do this more efficiently.

The perceptual segmentation of a piece seems to be little affected by musical training. Deliège (1987) found no differences in segmentation patterns from musicians and nonmusicians in her study on the segmentation of short excerpts. The only difference she found between the two groups was that musicians tend to place fewer boundaries and group more items into a longer segment. This result is in direct contrast to the findings of Ferrand (2004), who reported that musicians indicate a significantly higher number of segment boundaries than nonmusicians. The accumulated boundary profiles, however, were significantly correlated between musicians and nonmusicians. A study by Krumhansl (1996) reported a considerable consensus of boundary indications between musicians and nonmusicians. Deliège and Ahmadi (1990) asked subjects to segment two atonal musical pieces and could not find general differences between the segmentation patterns of musicians and nonmusicians. These studies seem to indicate that musical training has little influence on the perception of structural boundaries in music, at least for the accumulated boundary profiles.

In the following, two experiments are described that extend previous research on

boundary perception in three ways. First, instead of Western classical music that has been used in previous studies (Deliège & Ahmadi, 1990; Clarke & Krumhansl, 1990; Krumhansl, 1996; Spiro, 2007), the present study uses melodies extracted from Western popular music. The stimuli were monophonic representations, which can also be segmented with recent models based on the monophonic musical score. Second, the obtained perceptual boundaries were compared with boundaries predicted by the LBDM, the Melisma model, and the quantified rules of GTTM (Cambouropoulos, 1997; Temperley, 2001; Frankland & Cohen, 2004). These models have been tested on short melodies but, to the authors' knowledge, not on full-length pieces of music with a duration of several minutes. Third, the stimuli used contain changes in timbre, a cue so far rarely tested empirically. It must be mentioned, however, that the changes in timbre were primarily placed at the end of the melodic lines and therefore often coincided with structural boundaries. Nevertheless, for simplicity, we will use the term "change in timbre" for this cue.

The method developed in the present study extends a method used by Clarke and Krumhansl (1990) and make it applicable also to stimuli for which the score is not available. In a first experiment subjects were asked to indicate segment boundaries in three trials. However, instead of using a temporal window related to a specific metrical level or notes to integrate the inherent dispersion of the boundary indications a different method is proposed to obtain an accumulated boundary profile. As we found out only very recently, similar ideas had been described independently in an unpublished doctoral thesis by Ferrand (2004). In a second experiment, subjects were asked to rate the salience of a set of boundaries and to describe the boundary cues. Instead of rating the salience of their own indicated boundaries, as in the study by Clarke and Krumhansl (1990), the selection of boundaries for the second experiment was based on the accumulated boundary profile across all subjects. It was, thus, possible that subjects had to rate the salience of boundaries not indicated by themselves. Furthermore, subjects could also listen *across* the boundary, thus the music did not stop at each boundary indication as in Clarke and Krumhansl (1990). We also assigned the given salience rating to the boundary cues, to estimate which cues were associated mainly with salient boundaries. The assignment of the two resulted in a novel measure for the importance of a boundary cue.

The issues addressed in the present study were: 1) How consistent are subjects in their segmentation pattern across repeated trials? And what is the consistency

between subjects? 2) Can the outcome of the first experiment, the number of boundary indications, be compared to the salience ratings of the second experiment and what is the relationship between the two results? 3) How well can musicological models predict the boundaries obtained in the perceptual experiment?

2.2 Experiment 1: Segmentation experiment

The aim of the first experiment was to evaluate how subjects segment melodies. The points of focus were first, whether subjects segment consistently, i.e., if different subjects perceive the same boundaries, and second, whether the boundary indications within the three trials were similar or whether there was a systematic change. It was expected that at certain time instances clear perceptual cues for boundaries are present, and that these cues trigger boundary perception in all subjects. Other time instances with fewer or less strong cues may trigger boundary perception for fewer subjects, so that different boundaries are indicated with a different number of boundary indications within a given piece of music.

2.2.1 Method

Stimuli

For this study only the monophonic MIDI representation of the melody lines were used and rendered to audio with a MIDI synthesizer. Twenty songs were first selected having the following criteria: they had to cover different subgenres of popular music, some of the songs should contain strong tempo changes, one of the songs should have lyrics that our subjects could not understand, and we wanted two different audio versions of one song, one being vocal and the other being instrumental. Finally, it was necessary to have the audio as well as the corresponding MIDI file available. Twenty songs adhering to these criteria were selected and then analyzed for temporal dispersion of theoretical boundary cues. The idea was to select those songs, for which the potential boundary cues were not highly correlated over time, because only in that case there is a chance to find the contribution of individual cues for boundary perception. The songs were evaluated for their segment boundaries with the help of the "Local Boundary Detection Model" (LBDM) by Cambouropoulos (2001a) and the quantified rules of GTTM by Frankland and Cohen (2004). Both models were applied to the melody of each song.

Table 2.1: The selected songs used in the two experiments.

Song title	Artist	Duration	Album	Publisher
"Heart to hurt"	Kousuke Morimoto	3:17	RWC-MDB-P-2001 Nr. 72 RWC Database	Goto, Hashiguchi, Nishimura, and Oka (2002)
"Live and let die"	Paul McCartney	3:10	All the best	MPL communication Ltd.
"Moondance"	Van Morrison	4:33	Moondance	Warner Bros Records
"And when I die"	Blood, Sweat, and Tears	4:01	Super Hits: Blood, Sweat, and Tears	Sony
"Body and soul" (vocal)	Billy Holiday	2:57	Best of the best Gold: Billie Holiday	Sony
"Body and soul" (instrumental)	Coleman Hawkins	3:00	A retrospective 1929-1963	BMG

In addition to the cues of the two monophonic models, three polyphonic cues were extracted: the beginning of an instrument voice, the ending of an instrument voice, and the ending of harmonic cycles (V-I). From the twenty songs, the six songs that had the lowest mean correlation between these different types of boundary cues were selected. The selected songs are shown in Table 2.1.

MIDI representations of the songs were manually time-aligned to an audio version to provide a more realistic representation of the timing. The MIDI files were obtained from publicly available sources, see Appendix A. From each song we extracted manually a monophonic MIDI representation, which consisted of the melody line. Where the melody line was not present, the most salient accompaniment was chosen. The melody line was synthesized with the Pad 3 (polysynth) MIDI instrument, as this was distinguishably different from natural instruments. The accompaniment was synthesized with the timbres as in the polyphonic MIDI file. Across the six songs these timbres were piano, cello, trumpet, rin tink, and guitar. This procedure resulted in between 3 and 15 timbre changes per song. The two MIDI versions of "Body and soul" were essentially the same, except that one was aligned to the vocal version and the melody was synthesized with the Pad 3 (polysynth) and the other to the instrumental version of "Body and soul" with the melody having the timbre of a saxophone.

Procedure

Subjects were asked to listen to and segment six melodic lines into smaller pieces by pressing the space bar on the computer keyboard each time they perceived a boundary. The exact task description was: "Please press the space bar when you hear a segment boundary (phrase, section, passage)". When subjects did not understand the task based on this description, they were asked to divide the song into smaller meaningful pieces. The six songs were presented to the subjects in random order and they listened to each song four times in a row. Initially, subjects listened to a given song without pressing any key to get familiar with the piece, and in the following three presentations they were asked to indicate segment boundaries by pressing a key on the computer keyboard. Subjects could not pause the playback of the song during listening and there was no visual information about the song (e.g., no score or timeline) provided to the subjects. At the end of the four presentations, subjects were asked to rate their familiarity with the song before the experiment on a seven-point scale, from unknown to very well known. After the experiment, subjects were asked for any comments about the experiment.

Apparatus

The stimuli were synthesized offline with Steinberg's Cubase CS internal MIDI synthesizer. They were presented over Beyer Dynamics DT 990 Pro headphones in a sound-insulated listening room at the Philips Research Laboratories Eindhoven, The Netherlands, at a comfortable listening level. The playback of the stimuli and the recordings of the subjects' responses were controlled by a computer using a custom-made program, which played the synthesized renderings of the pieces. The delay between an indication by pressing a key and the recording of the indication by the computer was measured to have a mean value of 25 ms (σ = 6 ms). The delay was measured in recording the audio playback through the headphones with the sound generated by the keypress on the keyboard. The recordings as used as stimuli and the recorded audio containing the sound of the keypresses were then manually aligned to estimate the delay between the sound of the keypress and the boundary indication. In the experiment, no additional auditory (e.g. computer generated clicks) or visual feedback was provided during and after the playback.

Subjects

The experiment was performed by 21 subjects, 19 male and 2 female. Subject age ranged from 23 to 38 years with a mean of 26.2 years. None of them were professional musicians. Musical training varied from none to 21 years of practical musical training (μ=5.5 years, $\sigma = 7.0$) and from none to ten years of theoretical training (μ=2.3 years, σ=3.2).

2.2.2 Results

The total number of boundary indications ranged from 1 to 84 across subjects and songs. The mean number of boundary indications was 28.3 for the first trial, 27.8 for the second trial, and 27.9 for the third trial, with an overall mean of 28.0. The numbers of boundary indications between the three trials were tested with an ANOVA which showed no significant difference in the number of boundary indications across trials ($F_{2,340} = 0.12$, p = 0.89). No interaction between song and repeated trials was found ($F_{10,340} = 0.35$, p = 0.97). There was a significant effect of the song ($F_{5,340} = 28.2$, $p < 0.001$) and subject ($F_{20,340} = 62.5$, $p < 0.001$). These results suggest that subjects performed similarly in terms of the number of boundary indications across the three trials.

To see whether the boundary indications were also placed at similar time instances, we calculated a perceptual segmentation profile by converting the indicated boundary time stamps for each trial to a pulse train, and then smoothing the pulse train by convolving it with a Gaussian window resulting in a smoothed boundary profile. The procedure of smoothing is shown in Figure 2.1. The temporal positions and heights of the maxima in the smoothed boundary profile were taken as the accumulated boundary indications.

An important parameter in the smoothing process is the size of the temporal window used for the Gaussian. The optimal window size was defined with the following criteria: minimizing the frequency of windows containing more than one boundary indication within one trial while maximizing the inclusion of boundary indications across all three trials. To have a better estimation of the influence of the window size on the smoothing process, we first collapsed the boundary indications across the three trials for each individual subject. We then analyzed this pattern with a rectangular window of specific size. The onset of the window was initially placed at the first indication, and was

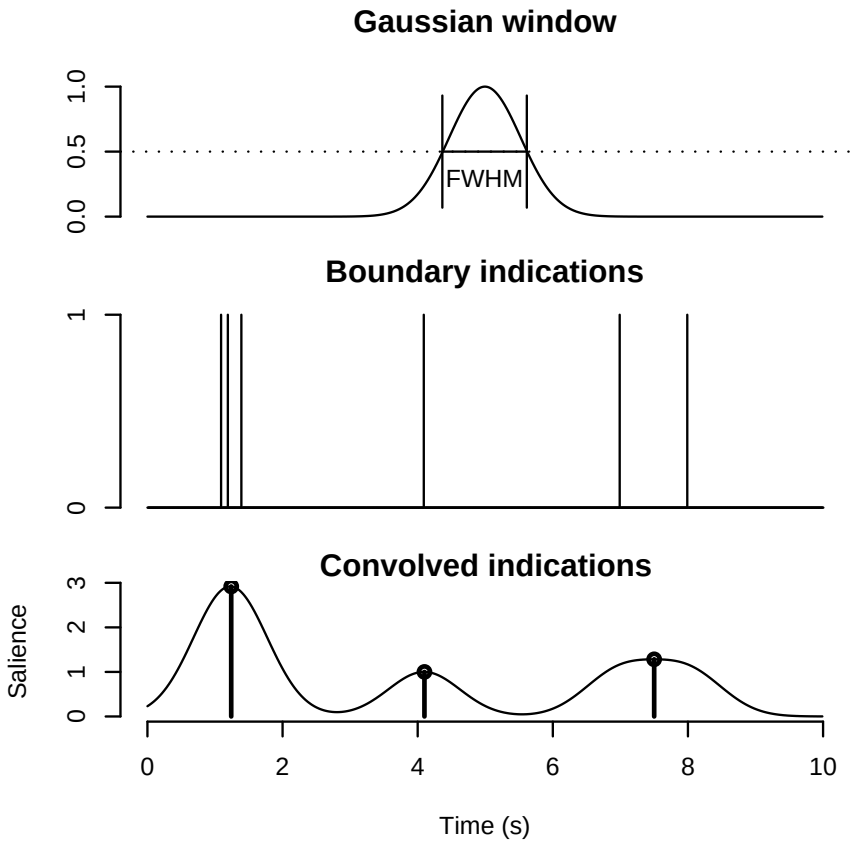


Figure 2.1: Visual representation of the Gaussian smoothing. The top row shows the Gaussian function with the full width at half maximum of 1.25 s, the second row some examples of indicated boundaries, and the third row shows the indicated boundaries convolved with the Gaussian function resulting in the smoothed boundary profile with the relative maxima, which are taken as the time points of a boundary.

then shifted to the next indication. We counted how many of these windows contained exactly one boundary indication from each trial, i.e., one boundary indication in the first trial, one indication in the second trial, and one indication in the third trial. We also counted how many windows contained at least two boundary indications from the same trial. The number of windows containing boundaries from each trial and the number of windows containing at least one double indication within one trial are plotted in Figure 2.2. In order to estimate the optimal window size, the difference between the number of windows containing an indication in all trials and the number of windows containing at least one double indication was calculated and the maximum was selected as the optimal window size. The figure shows that across all six songs there is some variation in the value of the optimal window size. The optimal window size is around 1.25 s for all songs. This variation of the window size could not be attributed to any obvious properties of the songs (see, for instance, the two versions of "Body and soul", which should, in theory, lead to the same window size as they were very similar). We decided therefore to use a fixed window size of 1.25 s for all songs.

To derive a measure of the within-subject consistency, the smoothed pulse train of each trial for each individual subject was correlated (using Pearson's correlation) with each of the two other trains for the same song, which resulted in three correlation values per subject and song (with the mean across subjects shown per song in the first row of Table 2.2). As the two versions of the song "Body and soul" were closer to each other than any other two songs, we attributed the difference of the within-subject consistency to response variability. Based on the difference between the two versions of "Body and soul" we thus interpreted the within-subject consistency as similar across the six songs.

Another measure of within-subject consistency was generated by first collapsing the boundary indications across the three trials. At each occurrence of a boundary indication the number of boundaries was counted within the range of a window starting at the boundary indications. The number of windows containing a boundary indication from each of the three trials over the total number of boundary indications was our second measure of consistency. Table 2.2, second row, shows the consistency measures for each song. The results showed no strong influence of the song on within-subject consistency. In general, subjects were consistent over the three trials in about half to two-thirds of the boundary indications. This result is consistent with subjects commenting that they sometimes missed a boundary or added an additional one over the course of the three trials.

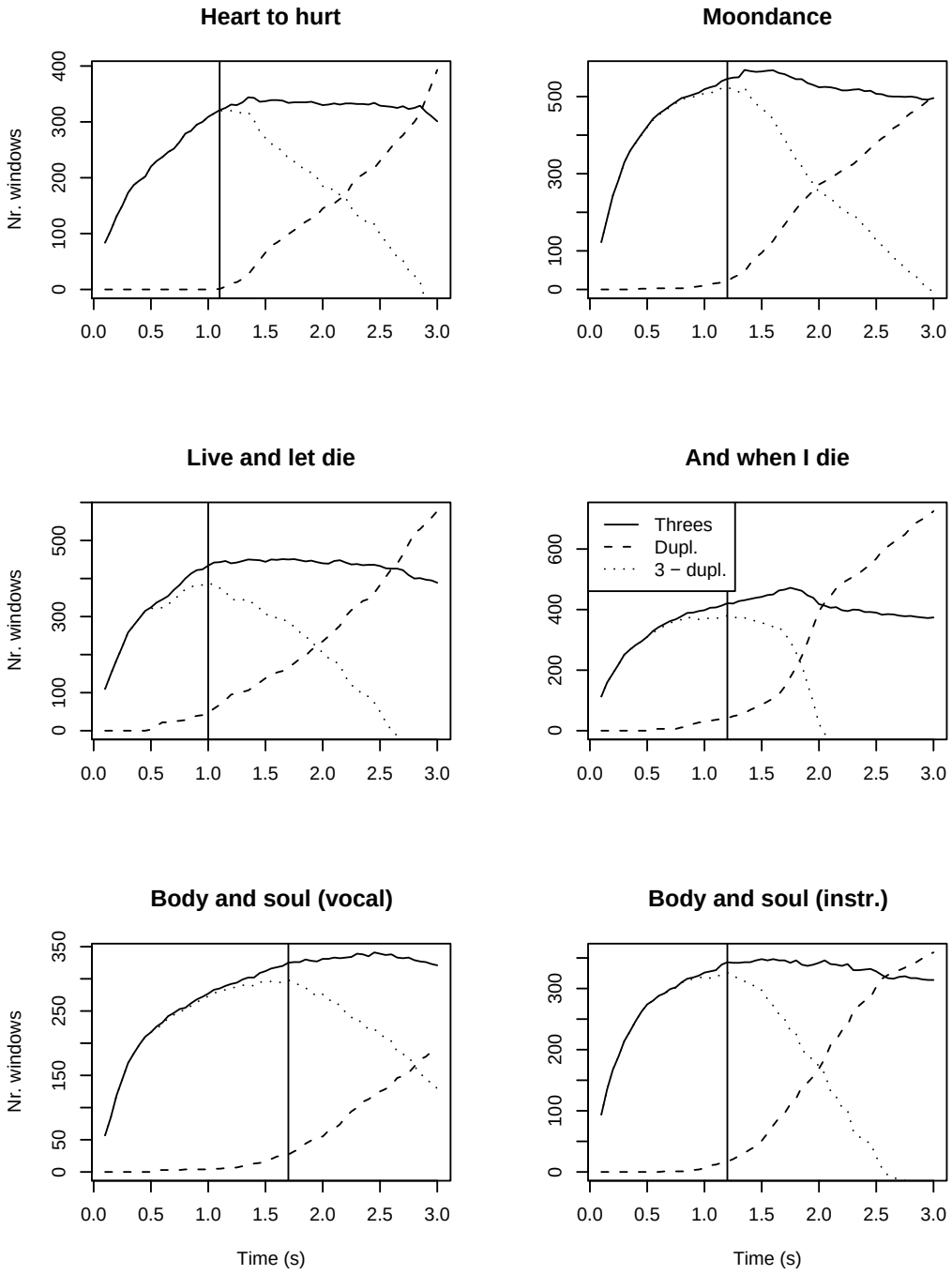


Figure 2.2: The estimation of the optimal temporal window size for all six songs. The plot shows the number of windows containing exactly one boundary indication from each trial (continuous curve), the number of windows containing more than one boundary indication from one trial (dashed curve), and the difference of the two (dotted curve) as a function of the window size. The vertical line indicates the maximum of the dotted curve.

Table 2.2: Measures for within- and across-subject consistency. The first two rows show two different within-subject consistency measures. (A) Mean correlation between the three pairs of trials for the same song and subject. (B) Percentage of boundary indications that also contain boundary indications from the other two trials within the optimal temporal window as a fraction of the total number of boundary indications. The bottom row shows the mean correlation across subjects, obtained similar to the first method of the within-subject correlation. The columns indicate the songs: (1): Heart to hurt, (2): Moondance, (3): Live and let die, (4): And when I die, (5): Body and soul (vocal), (6): Body and soul (instrumental).

	(1)	(2)	(3)	(4)	(5)	(6)	Mean
Within-subject analysis, method (A)	0.62	0.67	0.65	0.54	0.61	0.66	0.62
Within-subject analysis, method (B)	0.53	0.63	0.60	0.52	0.55	0.61	0.58
Across-subject analysis	0.37	0.34	0.38	0.32	0.41	0.36	0.37

The analysis of the boundary indications across the three trials can also be used to see how precise subjects were in their notations. We, therefore, collapsed the boundary indications across the three trials and calculated the standard deviation of the boundary indications within a temporal window centered at each boundary indication, with the window size set to the optimal 1.25 s. If a subject indicated exactly one boundary indication for each trial within the optimal window, thus when the subjects were consistent across the three trials, the standard deviation of the time instance of the boundary indication was calculated. The standard deviation of the boundary indications across the three trials, averaged across subjects, and songs, was 193 ms with a window size of 1.25 s. It seems, thus, that subjects were relatively precise and consistent in their boundary indications if they indicated a boundary consistently in all three trials.

We also correlated the within-subject consistency with subjects' familiarity with the song and with the musical training of the subjects. The mean within-subject consistency across the three trials was not significantly correlated with the familiarity with the song before the experiment ($r=0.04$, $p=0.67$). The mean within-subject consistency had a low correlation with the musical training, both with the number of years of theoretical music training ($r=0.25$, $p<0.005$) and with the number of years of musical practice ($r=0.27$, $p<0.005$), suggesting that there seems to be no linear relation between song familiarity or musical training and within-subject correlation. To analyze this further, we selected the seven subjects having the lowest amount of practical musical training ($\mu = 0$ years) and compared the within-subject correlation with the within-subject correlation of the

seven subjects having the highest amount of practical musical training ($\mu = 14.4$ years, $\sigma = 5.4$). The musically most experienced subjects had a mean within-subject correlation of 0.69, while the least experienced subjects had a mean within-subject correlation of 0.56, for which a Welch t-test showed that they were significantly different from each other across the six songs ($t = -5.32$, $df = 81.7$, $p < 0.01$). This result suggests that subjects with more musical experience are more consistent in their boundary indications across the three trials.

To analyze across-subject consistency, the boundary profiles from the three trials were averaged for each subject and correlated pair-wise between subjects. The mean correlation across all songs and subjects was 0.37, ranging from -0.15 to 0.87. Multidimensional scaling analyses of the pair-wise across-subject correlation values showed no obvious clustering of subjects for individual songs, nor across all songs suggesting that, for our stimuli and pool of subjects, there are no obvious groups of subjects that segment the songs in a similar manner. The across-subject consistency across the six songs is shown in the last row of Table 2.2. The similarity of the consistency across the six songs is represented in the similar consistency of the two versions of the song "Body and soul", which were the same except for a different timbre for the melody and a different time alignment. It was therefore expected that both songs would have the same across-subject consistency. The variability of the across-subject consistency across all songs is similar to the variability between the two versions of "Body and soul".

It is possible that subjects with higher musical training are more consistent among themselves than subjects without musical training. To analyze this we calculated the across-subject consistency for the seven subjects with the most practical musical training and compared the consistency with the across-subject consistency of the seven subjects without practical musical training. For all six songs, a t-test showed no significant difference between the two groups, not for individual songs ($0.61 < p < 0.79$) and not across all songs ($p = 0.77$), suggesting that across subjects, the resulting boundary profile is little influenced by musical training.

We were also interested in the influence of the window size on the within- and across-subject consistency. We calculated the mean of the pair-wise correlations of the smoothed boundary profiles for individual trials (within) and of the mean boundary profile for different subjects (across) for window sizes between 0.5 and 2.0 s, with the results shown in Table 2.3. The table shows that the relation between within-

Table 2.3: The within- and across-subject consistency depending on the window size used to construct the accumulated boundary profiles. The within-subject consistency was calculated by pair-wise correlating the smoothed boundary indications of each trial. The across-subject consistency was calculated in correlating the mean boundary profile pair-wise across subjects.

Window size (s)	0.5	1.0	1.5	2.0
Mean within-subject consistency	0.51	0.61	0.64	0.65
Mean across-subject consistency	0.29	0.36	0.38	0.39

and across-subject consistency is similar, independent of the chosen window size in that the within-subject consistency is much higher than the across-subject consistency. The table also shows that except for a window size of less than 1.0 s, there is little influence of the window size on the within- and across-subject consistency. These results suggest that a window size between one to two seconds does not critically influence the consistency measure.

To obtain a per-song profile of the boundary indications, we summed the smoothed pulse trains across trials and across subjects. An example of such a summed smoothed boundary profile with the corresponding music notation is given in the first two rows of Figure 2.3. The example shows a section of about 20 s duration from the song "Heart to hurt". The perceptual profile contains two strong boundaries at 149 s and 161 s.

The maxima of this summed profile were extracted and are plotted for each song in Figure 2.4. The profiles show that certain boundaries were perceived by nearly all subjects (the theoretical maximum is 63, 3 trials $\times$ 21 subjects) and others only by a few subjects – a pattern that was observed across all songs. For some songs in Figure 2.4 the boundaries receiving many indications have a relatively regular interval, as most clearly seen in "Heart to hurt" with the four bar structure or in "Moondance" with the two bar structure at the beginning of the song. These regular intervals could be an indication that subjects perceive typical compositional structures.

2.2.3 Discussion

The goal of the segmentation experiment was to explore how subjects segment pieces of music taken from Western popular music. The main findings of this experiment were that there is a variety in the number of times a boundary was indicated across subjects. Some boundaries were indicated by nearly all subjects and others only once by few subjects. This pattern was found consistently for all six songs. Previous studies have

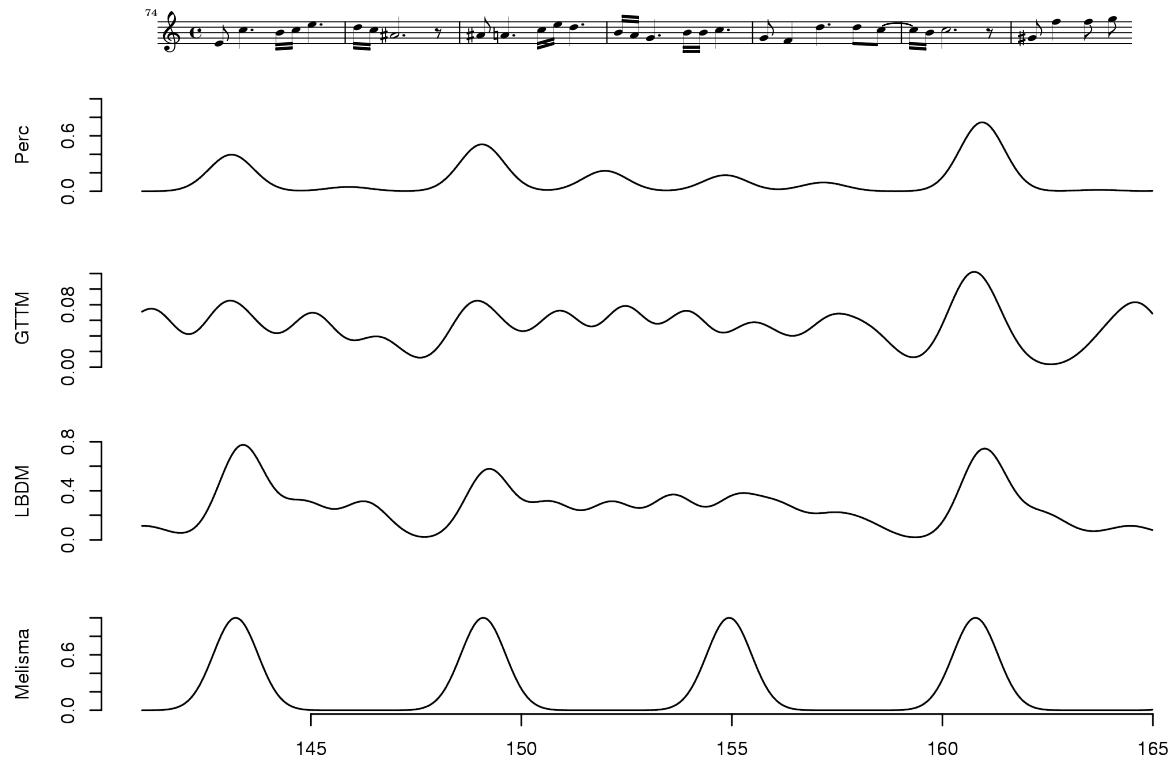


Figure 2.3: An example of the smoothed boundary profiles with the music notation in the top row for the song "Heart to hurt". The first smoothed boundary profile shows the perceptual boundaries with the height normalized by the theoretical maximum. The following three rows show the smoothed boundary profiles of the predicted boundaries of the three models GTTM, LBDM, and Melisma.

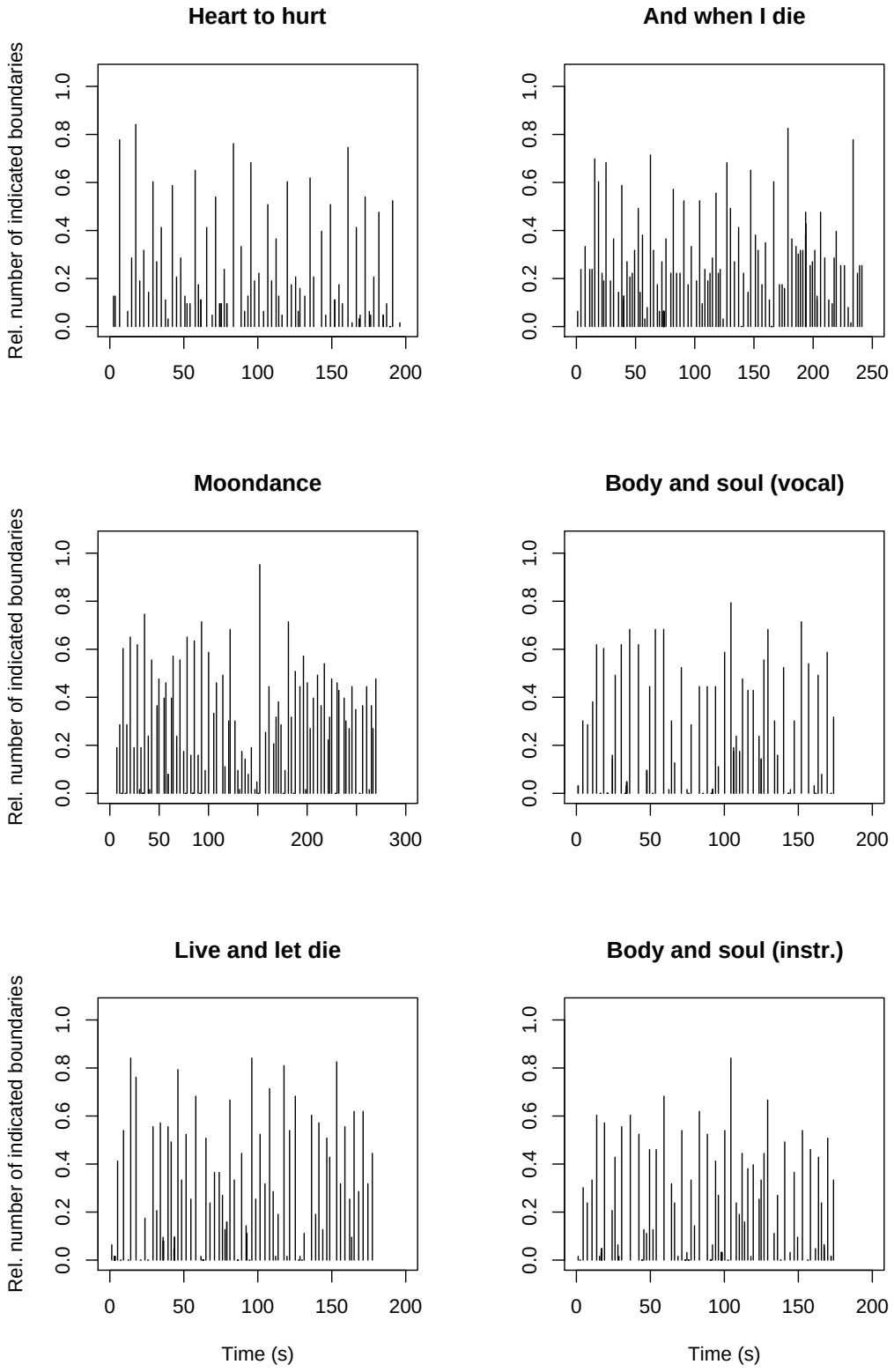


Figure 2.4: The boundary profiles showing the relative number of boundary indications per time window. Each vertical line indicates the point in time of the boundary shown with the relative number of subjects that indicated a boundary within the optimal window size of 1.25 s given by the line height. The theoretical maximum of possible boundary indications was 63 (21 subjects with 3 trials each), thus a relative number of boundary indications of 1.0 corresponds to 63 boundary indications.

found a similar pattern in the accumulated boundary profile. Krumhansl (1996) asked 15 subjects with a range of musical training to segment the Piano Sonata K. 282 by Mozart. She accumulated the indicated section boundaries within two beat temporal windows. Three section boundaries were indicated by all subjects while many more boundaries were indicated by only a few subjects. In a study by Frankland and Cohen (2004), 123 subjects were asked to indicate section boundaries in short monophonic melodies. Indications were accumulated by assigning them to the note being played at that moment. In each of the six melodies, a few (two to seven) boundaries reached proportions above 0.8, while the great majority of indicated boundaries had proportions below 0.2. These findings suggest that there is no binary boundary profile in the sense that there is a boundary or not, but rather a collection of boundaries with associated saliences.

To accumulate the boundary indications, which have intrinsic temporal scattering, across trials and subjects several methods have been used in previous studies. Schaefer et al. (2004) and Frankland and Cohen (2004), for instance, assigned each boundary indication to the note being played during the indication. A quarter note had a duration of between around 250 to 500 ms for the stimuli used in Frankland and Cohen (2004) and 600 ms for the study of Schaefer et al. (2004). Given that indications in Frankland and Cohen (2004) were aligned with note durations between one quarter and three quarter notes the "de facto" integration time covered the range between 250 and 1500 ms. Schaefer et al. (2004) defined "critical points" where more than half of the subjects indicated a boundary within three successive notes, which would cover a duration of 1800 ms for three quarter notes. Krumhansl (1990) and Spiro and Klebanov (2006) used a temporal window with a size related to the duration of a measure, the former used a two beat window, the latter a one beat window. To obtain an accumulated boundary profile across subjects (and trials), all boundary indications falling in this window were integrated and a histogram was calculated. The experiments reported by Ferrand (2004) used a Gaussian window with a window size (FWHM) of 360 ms to obtain an accumulated boundary profile. This relatively short window size was derived from an analysis of the typical time difference between the boundary indications and the closest on- or offset of a MIDI note. As the author discusses his window size is too small to capture all indications given by the different subjects, certainly for prominent boundaries. Our method for obtaining the optimal window size, which is based on the variability of intra-subject boundary indications, leads to a larger window size.

Summarizing, the obtained "optimal" window size of 1.25 s used in our study is in the same range as the one used in the majority of previous studies.

Besides choosing an appropriate duration for the integration window one also has to decide *where* to place the window. Previous studies have mainly either used a rectangular window aligned to the meter of the piece (one or two beats, Krumhansl, 1996; Spiro, 2007) or aligned to the note being played at the indicated time instance (Frankland & Cohen, 2004; Schaefer et al., 2004). The difficulty of these approaches is that it is unclear how to treat boundary indications at the border of a temporal windows or the note, for instance if subjects anticipate the boundary or delay their indication. The method presented here, where the pulse train of the boundary indications is convolved with a Gaussian window does not define a priori where the window can be placed. Boundaries further away in time can be included, but with a lower weight. We would like to add that many of the arguments given for the choice of a Gaussian window have also been derived independently by Ferrand (2004).

Schaefer et al. (2004) reported an influence of musical training on the segmentation of 10 children's songs. They found that highly trained subjects had a lower mean variance in the identified boundaries for boundaries indicated by more than half of the subjects. Furthermore, they found that the segmentation pattern of musicians violated the boundaries predicted by Gestalt rules more often than patterns of musically less experienced subjects. Our results support the influence of musical training on the within-subject correlation – subjects with more musical training had a higher mean within-subject correlation. Across subjects, however, no clear pattern between the segmentation profiles of musically trained subjects and subjects without musical training could be found, which is in agreement with other studies (Deliège & Ahmadi, 1990; Krumhansl, 1996; Spiro, 2007). It seems, thus, that subjects with musical training have a clearer idea on when to segment than musically less trained subjects and therefore segment more consistently across the three trials. Across subjects, however, the segmentation profile seems to be less influenced by musical training.

The method used in the segmentation experiment does not give any insights *why* subjects indicate a boundary in that specific place. In addition, it leaves open how the total number of boundary indications is related to the boundary salience of a boundary. To investigate these two issues, a second experiment was performed in which subjects were asked to explicitly *rate* the salience of selected boundaries and to *describe* the cues contributing to the perception of the boundary.

2.3 Experiment 2: Salience rating experiment

In previous studies, it was assumed that the number of subjects indicating a specific boundary represents a measure of the *salience* of that boundary (Deliège, 1987; Deliège & Ahmadi, 1990; Krumhansl, 1996; Frankland & Cohen, 2004). This implicit measure of a boundary salience, however, is not necessarily correlated with the perceived boundary salience.

For this reason subjects are asked in the second experiment to explicitly rate the salience of a subset of boundaries. The resulting boundary salience ratings were then correlated with the boundary profiles of the segmentation experiment to see how well the implicit measure of salience corresponds to the explicit measure.

2.3.1 Method

Stimuli

The same monophonic representations of six popular songs were used as in the segmentation experiment. These representations included timbre changes. The selection of the boundaries for the salience rating experiment was based on the boundary profiles from the segmentation experiment. The selection included all of the strong boundaries, i.e., the 10% of all indicated boundaries with the highest number of indications, and two to three moderate as well as two to three lower boundaries. The lower boundaries were chosen such that they did not have a strong boundary in their vicinity to avoid confusion with more salient boundaries. Between 16 and 19 boundaries per song were selected, in total 103 across all songs.

Procedure

The task description of the salience rating experiment was: “In the second experiment you are asked to rate the salience of the given boundaries and write down the cues that make it a boundary.” Subjects rated boundary salience on a seven point scale labeled from “no boundary” (0) to “very strong boundary” (6). For each boundary, subjects were asked to write down the cues in the musical signal responsible for the boundary. If a subject did not understand the terms salience or cue, salience was explained as the “strength of a boundary” and cue as “what in the music lets you hear the boundary”. Subjects were also advised to give a rating of zero and not to describe the boundary

if they thought there was no boundary. The interface allowed subjects to skip to any place in the song and listen to parts of it as many times as necessary to do the task. Subjects could also switch back and forth between the different boundaries and adjust the rating if needed. When they were content with their ratings and the descriptions of the boundaries, they saved the results and went on to the next song. After the experiment subjects were asked to comment on the experiment.

Apparatus

The salience rating experiment used the same apparatus as the segmentation experiment except for the interface, which showed the timeline of the song as a horizontal line with vertical bars indicating the boundaries to be rated. A moving pointer indicated the momentary position of the playback. Each vertical boundary line had a slider on which subjects could indicate the boundary salience on a scale from 0 to 6. In addition, there was a text field, where subjects wrote down, for each boundary, the cues that made them perceive the boundary. All boundary positions and their ratings of one song remained visible until subjects finished judging a specific song.

Subjects

The same subjects from the segmentation experiment participated in the salience rating experiment.

2.3.2 Results

In a first analysis, the scale used for the salience rating of the subjects was calculated. Across the six songs, the minimum mean-boundary rating was 0.3, the maximum was 5.7, indicating that subjects used the whole range for rating the salience of the boundaries.

To see whether subjects gave similar ratings to boundaries, we calculated the correlation between the salience ratings across all subject pairs. The mean of the correlation across all songs and subject pairs was 0.55, the minimum 0.26 and maximum 0.78 (all significant, $p<0.01$). For the individual songs, the mean correlations across subjects' salience ratings were: "Heart to hurt": 0.64, "Moondance": 0.59, "Live and let die": 0.40, "And when I die": 0.59, "Body and soul (vocal)": 0.49, "Body and

Table 2.4: The "cue classes" used to classify the descriptions given by subjects for the boundary cues.

Group	Cue class
Vertical/horizontal	Harmonic progression and tonality change
	Melody change
Rhythm	Tempo change
	Rhythm change
Timbre	Timbre change
Dynamics	Level change
	Break
	Global structure
	Repetitions

soul (instrumental)": 0.64. Thus, overall subjects gave moderately consistent salience ratings.

To analyze the descriptions subjects gave for the boundary cues, we classified their descriptions into "cue classes", shown in Table 2.4. The first group of cues are descriptions about the horizontal or vertical development of the stimuli at the boundary point. The second group are changes in the rhythm or in the tempo. The third group contains all descriptions about timbre changes. The fourth group comprises changes in dynamics and breaks. The last group are descriptions that are not basic cues, but rather complex summaries of the song structure. Descriptions of the last group, thus, can in principle also comprise cues from the other four groups but were not used in the wording as written by our subjects. The precise wording of the boundary cue descriptions and how they were classified is given in Appendix B.1.

All terms were classified into these "cue classes" and the number of times a certain term was mentioned was counted (shown in parenthesis in Table 2.6). In total there were 2163 boundary descriptions (21 subjects times 103 boundaries). The "cue classes" most often mentioned were change in timbre (530 times), global structure (450 times), and break (450 times).

As an example of how subjects described boundaries, the two boundaries at 149 s and 161 s in Figure 2.3 will be analyzed. Subjects described these boundaries using the "cue classes" as shown in Table 2.5. The first column indicates the "cue class", the second and third column the number of used terms for the boundary at 149 s and at 161 s, respectively.

For the boundary at 149 s, the main descriptions were melody-changes and breaks.

Table 2.5: The cue descriptions and the number of times subjects used the term for the boundary at 149 s and 161 s for the song "Heart to hurt".

Cue class	149 s	161 s
Harmonic progression	0	2
Melody change	6	7
Tempo change	0	1
Level change	4	3
Break	6	3
Global structure	4	7
Repetition	0	8

Subjects also varied in their perception of the boundary salience, some wrote that there was no boundary and rated it correspondingly low, others thought that it was a salient boundary. For the boundary at 161 s, the main descriptions were repetition, melody changes, and global structure. Subjects in particular notated that the music returned to the chorus after the boundary, but in a transposed form. Subjects with musicological knowledge also noted the perfect cadenza and the change in harmony through the modulation.

The "cue classes" also reveal some peculiarities. For instance, the cue level change was often used as a boundary description, despite the fact that all notes were played at the same level. A further investigation showed that this cue was mainly mentioned when there was a short rest after the boundary, when other subjects did not perceive a boundary, or when there was a change in instrumentation. Another cue of interest is harmonic progression, as one would assume that harmonic progressions are mainly conveyed through chord progressions and not easily perceived in monophonic music. Here, however, the class harmonic progression was mainly used when there was a clear return to the tonica, thus, when the melody ended with a cadenza or when there was a change in tonality. Furthermore, the description of harmonic progression was used by subjects with extended musical training only, which were used to pay attention to harmonic progressions. The class break was mainly used if the music stopped for a moment, i.e., when there was a gap between the ending of one note and the beginning of the following note.

The importance of each individual cue was estimated by relating the salience rating of a boundary with the descriptions given to that boundary. We defined the *mean term rating* of a class as the mean of all individual salience ratings subjects gave when they

described a boundary with the term of the class. If, for example, the term "harmonic progression" was used as a description twice across the six songs, once by subject A, who gave a salience rating of 4, and once by subject B, who gave a salience rating of 6, the mean term rating of "harmonic progression" would be 5. One criticism of the mean term rating could be that a high mean term rating may not say much if it is obtained by rarely mentioned terms. It may be necessary, therefore, to include the number of times a certain class was used into the mean term rating. However, such an incorporation of the number of times a cue was mentioned is not straightforward. Here, we therefore used the mean term rating as defined above.

Table 2.6 shows the mean term ratings for each class in our "cue classes". The classes harmonic progression, although mentioned only a few times, and timbre change had the highest overall mean term rating. Another cue with a relative high mean term rating was rhythm change. The classes melody change, global structure, and repetition had a mean term rating that was neither high nor low, but all of them were mentioned often. The mean term rating of the class tempo change seems to represent the properties of the songs. For two songs which had strong tempo changes, "And when I die" and "Live and let die", the cue tempo changes had a high mean term rating, while for the other songs the cue tempo changes had a low mean term rating and the cue was much less often used. The two classes describing a change in intensity of sound, level change and break, were often used as descriptions — however, these classes had the lowest mean term rating, despite being mentioned often, and were thus mainly used for boundaries with a low salience.

Clarke and Krumhansl (1990) used a similar experimental setup as described here. They also reported the terms used for describing a boundary and the number of times each term was mentioned. They asked, in contrast to us, to describe the indicated boundary cues of each subject's individual boundaries instead of boundaries taken from all subjects. The stimuli used were two pieces taken from Western Classical music, Stockhausen's Klavierstück IX and Mozart's Fantasie in C minor (K. 475). To compare their results to the current findings, we classified their descriptions into our "cue classes" (see Appendix B.2 for how we classified their descriptions into our "cue classes"). We then calculated the mean term rating for each cue, also shown in Table 2.6. From these results, it can be seen that several cues have a similar mean term rating in the two studies, such as harmonic progressions and rhythm change. Some cues, however, have a different mean term rating for our songs and the two classical pieces. For instance,

Table 2.6: Mean term ratings of the classified descriptions. The number of times a term was mentioned is indicated in parenthesis. Results from Clarke and Krumhansl (1990) are shown in the rows below the current results (Stockhausen and Mozart).

	Harmonic prog.	Melody	Tempo change	Rhythm change
Heart to hurt	4.71 (7)	4.03 (63)	1.50 (2)	4.85 (13)
Moondance	5.33 (3)	4.20 (50)	1.00 (2)	4.45 (33)
Live and let die	5.60 (5)	5.00 (60)	5.26 (31)	5.62 (24)
And when I die	6.00 (2)	4.10 (63)	4.60 (15)	5.12 (24)
Body and soul (v)	5.19 (16)	3.92 (39)	4.00 (2)	4.40 (10)
Body and soul (i)	5.62 (8)	3.94 (50)	2.00 (2)	5.89 (9)
Overall	5.29 (41)	4.22 (325)	4.61 (54)	5.00 (113)
Stockhausen	6.00 (3)	4.90 (21)	– (0)	5.39 (7)
Mozart	5.07 (20)	4.86 (13)	5.22 (12)	3.79 (2)
Overall classical	5.19 (23)	4.89 (34)	5.22 (12)	5.03 (9)

	Timbre change	Level change	Break	Global	Repetition
Heart to hurt	5.41 (75)	3.45 (47)	3.56 (59)	4.60 (70)	4.21 (28)
Moondance	5.28 (39)	3.65 (40)	3.24 (89)	4.56 (84)	3.73 (92)
Live and let die	5.19 (159)	4.55 (51)	4.55 (20)	4.71 (72)	4.02 (53)
And when I die	5.24 (153)	3.41 (44)	4.36 (75)	4.46 (70)	3.94 (32)
Body and soul (v)	5.33 (43)	3.12 (43)	3.03 (126)	4.05 (77)	4.22 (46)
Body and soul (i)	5.39 (61)	2.64 (36)	2.30 (81)	4.40 (77)	4.32 (50)
Overall	5.28 (530)	3.52 (261)	3.30 (450)	4.46 (450)	4.02 (301)
Stockhausen	3.52 (6)	4.03 (10)	3.59 (9)	5.14 (13)	2.78 (26)
Mozart	4.81 (13)	4.95 (13)	– (0)	5.17 (21)	5.83 (6)
Overall classical	4.40 (19)	4.55 (23)	3.59 (9)	5.16 (34)	3.35 (32)

change in timbre has a higher mean term rating in the current study, while change in level has a higher mean term rating for the two pieces taken from classical music. These differences are likely to be the result of the stimuli used – our stimuli contained changes in timbre while the two classical pieces were mono-timbral and our stimuli had normalized levels while the two classical pieces were performed by a pianist. Aside from these differences caused by the acoustic signals, it seems that the mean term rating is rather similar for our six songs and the two classical music pieces, suggesting that certain cues are stronger associated with more salient cues.

Musical training seems to be of low influence on the frequency of terms with which subjects described the boundary cues. There was no significant correlation between musical training and the frequency of terms used to describe the boundary: Neither the number of years of music theory studies ($r=0.12$, $p=0.6$), nor the number of years practicing an instrument ($r=0.10$, $p=0.66$) were significantly correlated with the frequency of terms. To further analyze whether there was an influence of musical training on the boundary cue descriptions, the boundary descriptions of the seven subjects with the highest musical training were compared with seven subjects without musical training. A Welch t-test showed no significant difference in the total frequency of terms used by the two groups ($t(10.65) = -0.34$, $p = 0.74$). However, there was a significant difference in the frequency of terms used for the "cue class" global ($t(10.61) = -2.33$ $p < 0.05$), thus subjects with more musical training do also use more often terms describing the general structure of a song, such as chorus-verse patterns.

2.3.3 Comparison of different measures of boundary salience

The two experiments have revealed a number of possible measures for the estimation of boundary salience. Apart from the salience rating, four additional salience measures were extracted to investigate which of them best correlate with the perceptual salience ratings. Two measures were derived from the segmentation experiment and two measures from the descriptions obtained in the salience rating experiment.

1. The number of boundary indications within a time window of optimal size (1.25 s).
2. The standard deviation of the temporal position of the boundary indications within a time window of optimal size (1.25 s).
3. The frequency of terms used by the subjects to describe the boundary.
4. The number of different classes used in the description of the boundary.

The most interesting comparison is between the number of boundary indications and the salience rating because if these two measures are correlated, the number of indications within a time window can be directly used as a measure of the boundary salience. Figure 2.5 shows the relation between these two measures for all boundaries included in the salience rating experiment. The different symbols indicate the six different songs. The figure shows that the number of boundaries within the optimal window size is indeed correlated with the mean boundary salience rating. The overall correlation between these two measures was r=0.68 ($p<0.001$), and for individual songs: "Heart to hurt": 0.86, "Moondance": 0.37, "Live and let die": 0.76, "And when I die": 0.53, "Body and soul (vocal)": 0.63, "Body and soul (instrumental)": 0.85. Thus, even though the song "Moondance" had a lower correlation, the overall correlation between the two measures is moderately high.

We also calculated the correlation between the explicit salience rating and the three other measures. The frequency of terms mentioned, i.e., with how many terms subjects described the boundaries, had the highest correlation with the boundary salience rating (r=0.91, $p<0.001$). The higher correlation of the salience ratings with the frequency of terms compared to the number of *different* classes (r=0.63, $p<0.001$) shows that subjects describe a salient boundary often with more terms, but not with more classes. This might be an indication that stronger boundaries do not necessarily stem from more *different* cues.

The standard deviation of the boundary indications had a moderate correlation with the salience rating (r=0.57, $p<0.001$). The motivation to compute the standard deviation was that subjects may be more precise in indicating a strong boundary than a weak boundary. In such a case, a negative correlation between the standard deviation and the other salience measures was expected. However, the result found suggests that a higher rated boundary also has a higher deviation of the boundary indications within a temporal window.

2.3.4 Discussion

The goals of the salience rating experiment were (1) to evaluate the salience of boundaries in a more explicit manner and (2) to verify the assumption made in previous studies that the number of boundary indications across subjects is a measure of boundary salience (Clarke & Krumhansl, 1990; Frankland & Cohen, 2004; Krumhansl, 1996). The significant correlation between the two measures, the number of boundary

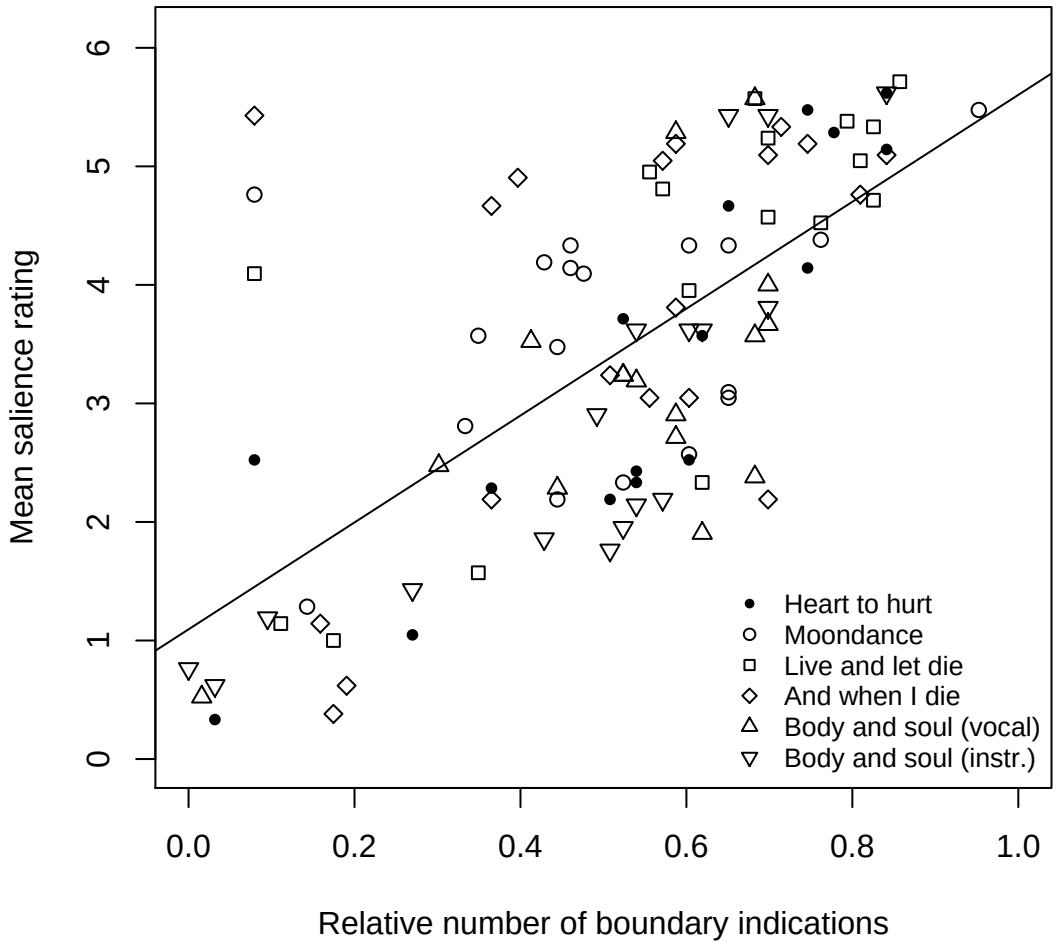


Figure 2.5: The relative number of boundary indications from the segmentation experiment on the abscissa and the mean salience rating of the selected boundaries from the salience rating experiment on the ordinate. The correlation between the two measures is 0.68 ($p < 0.001$). The line is the linear regression of the mean salience rating and the relative number of boundary indications across all songs.

indications within a time window obtained in the segmentation experiment and the salience rating obtained in the salience rating experiment, are consistent with the assumption of the previous studies.

The "cue classes" provide a means to evaluate which of the cues are more often used and may thus be more prominent than others. The "cue classes" were empirically established and can therefore be compared to theoretical cues. As reported by Clarke and Krumhansl (1990), the boundary cue descriptions are closely related to the GTTM rules. Furthermore, the "cue classes" allow the extraction of additional cues, which have not yet been incorporated into theoretical models, such as, for example, cues related to rhythm (rhythm change or tempo change). There are also other cue descriptions that are of interest. For instance, level change is an often-mentioned cue. However, there were no changes in the intensity of the played notes, thus level change rather was a description of a change in instrument or of a break. An additional cue that deserves discussion is the cue harmonic progression. In music theory it is often assumed that harmonic progressions or changes in harmony are mainly conveyed through chord progressions. In melodies, however, there is only one note at the time, thus the stimuli used in this study had no chord progressions. The harmonic progressions, therefore, must be implied. Perceptually, the implied harmony of a melody has been observed, for instance, in completeness rating experiments (Krumhansl & Kessler, 1982). All in all, the "cue classes" allow thus to further investigate which cues subjects consciously recognize in the perception of segment boundaries.

The descriptions of the boundaries indicate in two ways that a change in instrument is an important cue. The first indication is that the term "change in timbre" is often mentioned. The second indication is that the mean term rating of change in timbre is high. The result may be biased by the fact that the stimuli were monophonic representations, artificially extracted from the song. There were changes in timbre whenever the melody started or stopped playing, so the timbre changes could quite well coincide with section beginnings or ends, that may have influenced boundary perception. Nevertheless, the two experiments underline the importance of timbre change and thus support the results of Deliège (1987), who suggested the adding of timbre change to the GTTM rules.

An important caveat with our application of "cue classes" is that it is limited by the listeners' musical knowledge. Although we could only find a low correlation between the frequency of terms used for describing the boundary cues and musical

training for certain cues, the description of a boundary depends on the knowledge and awareness of the subject. The frequency of terms mentioned for a certain class, therefore, is an indication of what boundary cues in the music are *consciously* perceived and not necessarily what cues are present in the music. Therefore, theoretical models are important because they give a more objective analysis of the different cues present in the music. In the next section we correlate the perceptual boundaries with boundaries predicted by theoretical models.

2.4 Comparison of perceptual boundaries and theoretical model predictions

We compared the perceptual boundary profiles to predictions from three models: The combination of four rules taken from "A Generative Theory of Tonal Music" (GTTM) quantified by Frankland and Cohen (2004); the "Local Boundary Detection Model" (LBDM) by Cambouropoulos (2001a); and "Melisma" by Temperley (2001). All three models operate on monophonic pieces of music and can thus be directly applied to the current stimuli. The aim of the present evaluation was to see how well these theoretical models can predict perceptual boundaries.

Although the preference rules in GTTM define where to place segment boundaries, they do not specify how to combine the outcome of the different rules. This is of particular relevance in cases where one rule predicts a boundary that another rule does not. Frankland and Cohen (2004), therefore, quantified four of the rules from GTTM and attributed a value between 0 and 1 to each rule, indicating the strength of the rule's contribution to a particular boundary. The quantification then allowed the perceptual testing of the different rules. In addition, the results of the different rules can be summed to provide a global boundary profile. The four quantified rules were: Rest, attack-point, register-change, and length-change. The rest rule predicts the perception of a boundary whenever a rest occurs in the music, attack-point indicates whenever there is a long note in between two short notes, and register- and length-change predict when there is either a change in consecutive pitch interval sizes or a change of the durations between successive notes. All four rules take four consecutive notes to evaluate the boundary strength between the second and third note. The window of four notes is then slid over all notes to obtain a boundary profile of the whole piece. Although two of the rules, attack point and rest, have shown to be much better predictors for perceptual

boundaries (cf. Frankland & Cohen, 2004), the implementation does not give explicit weights for the combination of the rules. In the present study we therefore concatenated the perceptual and the predicted boundary profiles of all six songs and calculated the optimal weight combination. The optimal weight combination across all six songs were: 0.40 × rest + 0.04 × attack-point + 0.05 × register + 0.14 × length-change. As the predicted boundary profiles are not orthogonal to each other it is not straight-forward to interpret the weights of each cue. It seems, however, that the most salient cue of the quantified GTTM is the rest rule.

The LBDM of Cambouropoulos (1997) uses two general rules for segmentation. The first rule, called identity-change rule, indicates where in the musical signal a boundary is to be expected: "Amongst three successive objects, boundaries may be introduced on either of the consecutive intervals formed by the objects if these intervals are different. If both intervals are identical, no boundary is suggested." The second rule, called proximity rule, then specifies on which of the two intervals the boundary is more likely to occur: "Amongst three successive objects that form different intervals between them, a boundary may be introduced on the larger interval, i.e., those two objects will tend to form a group that are closer together (or more similar to each other)" (Cambouropoulos, 1997, p. 282). As in the GTTM model quantified by Frankland and Cohen (2004), a weight can be assigned to each cue. In his refined version (Cambouropoulos, 2001a), the author suggests the weights of 0.25 for pitch intervals, 0.50 for inter-onset intervals, and 0.25 for rests, thus a change in the duration of notes is given a higher weight than pitch intervals and rests. The total sum of the weighted predicted boundary strengths then represents the global boundary profile.

The Melisma model of Temperley (2001) uses an approach similar to the GTTM that describes the structure of music. The model is a combination of different components, each representing an aspect of musical structure: Metrical structure, melodic phrase structure, contrapuntal structure, pitch spelling (i.e., whether a note is G# or Ab), harmonic structure, and key structure. Some of the components can be enhanced by adding the output of one component as an input for another component. The theory defines, similarly to GTTM, two types of rules, the well-formedness rules, defining legal structures, and the preference rules, defining preferred of all possible structures. The part of the model used here, the melodic phrase structure component, consists of one well-formedness rule, which defines that groups cannot be overlapping and three preference rules: 1) Phrase boundaries are preferred at large inter-onset intervals or

large offset-to-onset intervals. 2) Groups of approximately eight notes are preferred. 3) It is preferable to place a boundary at parallel points in the metrical structure (Temperley, 2001, pp. 68–70).

Temperley's model is not only precisely described in his book, but the author also provides an implementation. For the first rule, the gap rule, the inter-onset and offset-to-onset intervals are summed up as a general boundary profile, called "gap score", where a boundary is introduced if the "gap score" passes a certain threshold. The other two preference rules are implemented as penalties: The second rule, the phrase length rule, assigns a penalty to groups having a different size than eight notes, and the third rule, the metrical parallelism rule, assigns a penalty for groups not beginning on the metrical structure. The model estimates phrase boundaries and it is not clear how to apply the three rules to obtain a global boundary profile. We therefore took the output of the default implementation as the global boundary profile. In contrast to the other two models, Temperley's implementation is context sensitive in the sense that the boundary prediction does not only depend on discontinuities in the musical surface (gap rule), but also on the location of the previous boundary (phrase length rule).

Timbre change has also been proposed as a possible cue for music segmentation (Lerdahl & Jackendoff, 1983) and has been assessed perceptually (Deliège, 1987). The models above do not take timbre change into account, probably because melodies often do not contain any changes in instrumentation. Since the melodies used in the present study did contain changes in timbre, although coinciding with the melodic structure, we added change in timbre to the present evaluation as an additional cue to all three models. Change in timbre was quantified to a value of 1 when a change in timbre occurred and zero otherwise.

For each of the three models a single predicted boundary profile was calculated for each song (see also Appendix C for precise definitions of the models). Each boundary profile was smoothed by convolving it with a Gaussian window with a FWHM of 1.25 s (cf. Figure 2.1). An example of the smoothed boundary profiles of the predicted boundaries for the three models is shown in Figure 2.3. The figure shows that especially the boundary at 161 seconds is predicted by all three models. To evaluate the performance of the models, the predicted boundary profiles were correlated with the perceptual boundaries (using the Pearson correlation). The models were tested individually, as well as in combination with the additional cue of change in timbre. Parameters of all models were set to their default values. These implementations

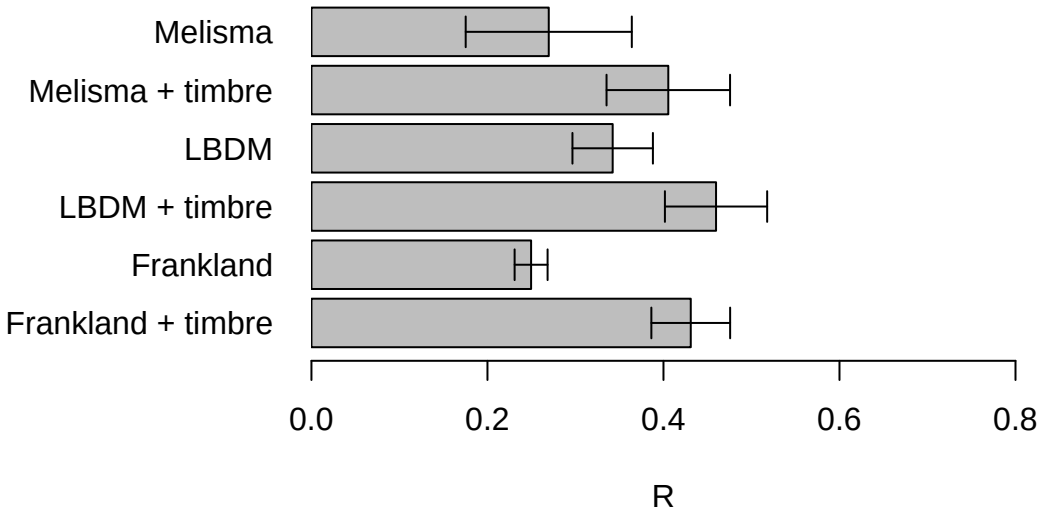


Figure 2.6: The correlation between the three models and the perceptual boundaries. The three models were tested as defined and with the additional cue of timbre change. The error bars indicate the standard error of the mean of the correlation across the six songs.

resulted in six different evaluations: GTTM, GTTM with timbre, LBDM, LBDM with timbre, Melisma (in the 2003 version) and Melisma with timbre.

Figure 2.6 shows the correlation between the boundaries predicted by the linear regression of theoretical cues and the perceptual boundaries. Both, the theoretical boundary placements, as well as the perceptual boundary indications, used the smoothed boundary profile instead of the precise points in time to counterbalance the impreciseness of the perceptual boundary indications (see the results section of the segmentation experiment for details about the smoothing process). The performance of the LBDM seems to be best and the GTTM rules as quantified by Frankland and Cohen (2004) worst. The performance of the quantified GTTM rules, however, could be improved if the weights of each cue are optimized for each individual songs instead of taking fixed weights for all songs. The additional timbre-change cue improved the performance of all models. Overall, however, there was only a moderate correlation between the model predictions and the perceptual boundaries.

For the different songs, the mean correlation across all models are 0.58 and 0.39 for “Heart to hurt” with and without timbre-change, respectively, 0.24 and 0.18 for “Moondance”, 0.49 and 0.18 for “Live and let die”, 0.40 and 0.29 for “And when I

die", 0.42 and 0.35 for "Body and soul (instrumental)", and 0.45 and 0.34 for "Body and soul (vocal)". Thus, for the song "Moondance", the models predict the perceptual boundaries considerably worse then for the other songs. One possible explanation for the low correlation for "Moondance" is the highly repetitive structure of the song, which is not taken into account in the tested models as the model cues do not extract repeating patterns.

The aim of the present evaluation was to estimate the performance of musicological models to predict our perceptual boundaries obtained in segmenting six popular music songs. The results show that a moderate correlation between the predicted boundary profiles can be obtained with these models that use simple cues extracted from the musical score. For a better understanding, however, a much more extensive evaluation has to be done, which would exceed the present study. We, therefore, refer to Chapter 4 where the models are evaluated in much more detail. Here, we would like to emphasize that by using a linear combination of the cues change in pitch, change in note durations, and rests, already a moderate correlation can be obtained, which further improves with our cue "change in timbre". It seems, thus that the model predictions agree with many of the perceptual boundaries.

2.5 General discussion

The purpose of the two current experiments was to gain further knowledge on how subjects perceive structural boundaries in music and, in particular, popular songs. The segmentation experiment explored how subjects segmented six songs into smaller parts, with the result of boundaries with a varying number of boundary indications. In the salience rating experiment a subset of boundaries from the segmentation experiment were presented and subjects were asked to rate the salience of these boundaries.

In a previous experiment Clarke and Krumhansl (1990) asked subjects to rate the salience of the segment boundaries indicated by each individual subject. They, however, did not analyze if there was a correlation between the relative number of boundary indications across all subjects and the given salience rating. The current study shows a significant correlation between the two methods of measuring boundary salience in the two experiments. The number of boundary indications can, therefore, be used reliably as a salience measure.

The correlation between the number of indications of a boundary and the given salience rating suggests two methods to estimate the salience of segment boundaries. One is to ask subjects to indicate segment boundaries (Experiment I), the other is to provide subjects with a set of boundaries and ask them to rate the salience (Experiment II). With these two methods it is not only possible to evaluate the boundary profiles of boundary predictions, but also to evaluate the salience of individual boundaries.

In the salience rating experiment, subjects were not only asked to give a salience rating, but also to describe the cues contributing to the perception of the boundary. The descriptions were classified into "cue classes". In general, the boundary descriptions were similar to the cues used in theoretical models. In fact, out of the nine classes of our "cue classes" five are based on changes. However, in addition to cues based on changes, subjects also mentioned other cues, like harmonic progression, repetition, or global structural descriptions (e.g., "start of chorus"). In particular, the use of global structural descriptions and repetition shows that subjects not only focus on local attributes while listening to the song, but also integrate the local cues into a larger context.

The descriptions subjects gave for the boundary cues and the obtained "cue classes" can be related to the grouping preference rules (GPR) of "A generative theory of tonal music" (GTTM) by Lerdahl and Jackendoff (1983). According to Clarke and Krumhansl (1990) the grouping rules can be classified into three types: Rules based on similarity and changes in the surface structure; a rule based on the deeper-level harmonic structure; and rules based on repetition and symmetry. Several cues from the "cue classes" can be directly related to the rules of GTTM, such as *change in melody* to change in register (GPR 3.a), *change in level* to change in dynamic (GPR 3.b), *break* to rest (GPR 2.a). However, certain cues of the grouping preference rules of GTTM were not mentioned by our subjects, such as articulation (GPR 3.c), length change (GPR 3.d) or attack-point, which is a long note in between two short notes (GPR 2.b). Attack-point and length change are interesting as it seems plausible that subjects perceive a change in length or a long note within short notes and hence describe boundaries with these cues. The results show, however, that subjects do not mention these cues as the main factors for perceiving a boundary, so they might be unaware of these specific cues. Therefore, even though subjects do not use attack-point and length change as boundary cues, these cues have been found to be good predictors for segment boundaries (Frankland & Cohen, 2004). On the other hand, several cues

were mentioned in our experiments as boundary cue descriptions which are not part of GTTM. Our experiment shows, for instance, that subjects very often mentioned cues belonging to the class change-in-timbre. This cue is, thus, one of the most dominant or easily perceived cues for segmentation, which is in line with the findings of Deliège (1987), who found timbre-change to be an important perceptual cue for segmentation.

We also evaluated the correlation of the boundary profiles predicted by musicological models and the perceptual boundaries, which revealed a moderate fit between the predicted boundaries and the perceptual boundaries. Several additional cues could possibly enhance the performance of the models. Repetition, for instance, is often mentioned in music theory (e.g. Lerdahl & Jackendoff, 1983) or in perceptual studies (e.g. Clarke & Krumhansl, 1990) and seems to be an important cue. For model implementations, however, repetition has several difficulties, as emphasized in Cambouropoulos (2006a). The theoretical difficulties lie, for instance, in the definition of what to repeat (i.e., repetition of pitch, rhythm, or both) or on what scale the repetition should take place (i.e., repetition of two notes or repetition of chorus). Another cue, not included in the models but perceptually relevant, is harmony or harmonic progression (Deliège, 1987). In music theory, it is generally assumed that the harmonic cycle in tonal music elicits feelings of tension followed by resolution. Harmonic progressions may influence the perception of boundaries, as shown empirically with probe tone techniques (Krumhansl & Kessler, 1982), and by an experiment where a two-note probe occurred either just before, straddled, or just after a harmonic boundary (Tan, Aiello, & Bever, 1981). Although the tested models do not make use of repetition or harmony, the predictions are still moderately correlated with the perceptual boundaries.

A final issue is, if there is a way to derive a hierarchical structure from the boundary indications. Music theory often assumes a hierarchical organization of the music piece (Schenker, 1935; Lerdahl & Jackendoff, 1983). The question is if and how such a hierarchical organization can be derived from our perceptual profiles. One possible solution is to set thresholds at different salience values and take all boundaries that exceed this threshold as segment boundaries at that level. Such a procedure would also be legal in the framework of GTTM, where overlapping segments are not allowed. How to set the threshold intervals and if the segment boundaries extracted in this manner have any meaning is still to be evaluated.

3 The perception of structural boundaries in polyphonic representations of Western popular music[†]

This chapter presents two experiments on the perception of structural boundaries in Western popular music. In the first experiment, subjects segmented two polyphonic representations of six different songs. In the second experiment, subjects rated the salience of boundaries selected from the first experiment and indicated which musical cues were responsible for the boundaries. The overall frequency of boundary indications was highly correlated with the mean salience rating, suggesting that the number of boundary indications across subjects is a good measure of boundary salience. The strongest cues for boundaries indicated by subjects were harmonic-progression, rhythm-changes, timbre-changes, and tempo-changes. Furthermore, boundary indications and their salience ratings were compared with those from the previous chapter in which monophonic representations of the same songs had been used. Results show that the segment boundaries are perceived at similar time instances and their strengths are highly correlated across the three representations.

[†]This chapter is based on Bruderer, M.J., McKinney M.F., and Kohlrausch, A. "The perception of structural boundaries in polyphonic representations of Western popular music", submitted for publication to Music Perception.

3.1 Introduction

When listening to music, listeners automatically perceive structural properties and are sensitive to structural boundaries in music. While the perception of structural boundaries has been investigated in short excerpts (Deliège, 1987), monophonic pieces of music (Frankland & Cohen, 2004), and monotimbral pieces (Clarke & Krumhansl, 1990; Krumhansl, 1996; Deliège & Ahmadi, 1990), no study has analyzed the perception of segment boundaries in complete songs of polyphonic Western popular music. Moreover, musicological models on segmentation have so far been limited to monophonic melodies (e.g., Tenney & Polansky, 1980; Cambouropoulos, 1997; Temperley, 2001; Frankland & Cohen, 2004).

Previous research has found that structural boundaries in music are conveyed through different cues, including changes in note length (Frankland & Cohen, 2004), changes in timbre (Deliège, 1987), and repetitions (Cambouropoulos, 2006a). The segment boundaries induced by these cues can only be perceived, however, if the listener pays attention to the line containing these cues. Sloboda and Edworthy (1981) proposed a model where the listener processes the music in a figure/ground manner: the attended melody forms the figure while the harmony forms the background. Bigand, McAdams, and Forêt (2000) found such a model to be valid for nonmusicians, but for musicians they proposed a voluntary integration model, where several streams are integrated into a single perceptual structure. Nonmusicians, on the other hand, simply reduced the complexity of polyphonic music by focusing attention on a single voice. These models suggest that, at least for nonmusicians, it should be possible to divide the difficult task of segmentation of polyphonic music into the easier task of segmenting a single voice.

Such a reduction to a single voice can, however, not account for cues based on several simultaneously played lines. Examples are polyrythms, harmonic changes, and cues induced by repetitions occurring in different voices, as in a canon or a fugue. Thus, the different lines of a polyphonic piece are often interrelated and may influence the perception of segment boundaries. The aforementioned examples suggest why the perception of segment boundaries may be influenced by additional properties present only in polyphonic music.

A further issue for experiments investigating the perception of structural boundaries in music is the form in which the stimuli are presented. Music from live performances often contain nuances that could influence the perception of particular attributes.

Typical perceptual experiments, however, often use a more formalized form of music, mainly a rendering of the synthesized score. Here we examine the perception of segment boundaries by using two different polyphonic versions for the stimuli: A polyphonic stimulus synthesized from the MIDI score, and a time-aligned audio recording. The results of these two polyphonic representations are then compared with the results from the monophonic representation shown in the previous chapter, which was also time-aligned to the polyphonic representations.

The differences between these three types of stimuli were the following. The monophonic MIDI stimuli do not contain explicit harmony. Furthermore, they were rendered from the MIDI file with synthesized instrument samples and did not contain vocals and thus lyrics. The polyphonic MIDI stimuli were similar to the monophonic MIDI stimuli in that they did not contain vocals, but with several voices playing at the same time and therefore containing explicit harmony. The audio recordings were polyphonic with explicit harmony and with recordings of real instruments. Five of the six audio recordings used here contained a singing voice and thus also added the further dimension of the semantic information included in the lyrics. Four out of the five songs had lyrics in English and one song had lyrics in Japanese. Both, the stimuli synthesized from MIDI as well as from audio contained expressive information. Although much effort had been made to align the MIDI stimuli to the audio recordings, the expressive information were not exactly the same. Based on these differences of the stimulus types it was expected that subjects would segment most consistently in the audio recording, as this type of stimuli contained most information (explicit harmony as well as human voice) and the monophonic MIDI stimuli would lead to the lowest consistency across subjects because of the missing additional information. It could, however, also be hypothesized that exactly because the monophonic MIDI version is simpler, the segmentation task should be easier and less ambiguous.

The experiments presented here are similar to the ones presented in the previous chapter in that they use the same experimental method but with different stimuli, to investigate the influence of the form of representation on the boundary placement and boundary salience. Furthermore, we identify the perceptual cues used for the segmentation of polyphonic music and compare the results to those obtained earlier using monophonic representations of the same songs. This comparison allows us to identify the influence of polyphony on the placement and the salience of structure boundaries, and on the cues used in segmentation.

3.1.1 Overall experimental design

We conducted two experiments, a segmentation experiment (experiment I) and a salience rating experiment (experiment II). Both experiments were performed with two different stimulus representations, synthesized MIDI and recorded audio. We want to emphasize that the order of discussion of the experiments does not reflect the order of the measurements. We first obtained all data for the synthesized stimuli for experiment I and II, in a complete within-subject design. We then performed all measurements with the audio recording stimuli, which again were done in a within-subject design. The two subject groups, one participating in the experiment with the MIDI stimuli and the other in the experiment with the audio stimuli, had only a small overlap.

3.2 Experiment I: Perceptual segmentation of six popular songs

The aim of the segmentation experiment was to study how subjects segment polyphonic popular songs. In particular we were interested if subjects are consistent across repeated segmentations of the songs and furthermore if different subjects indicate the same boundaries. Here we first present the results and analyses of the individual experiments on polyphonic music and then show the comparison between the current results and our previous work with monophonic stimuli after the two experiments.

3.2.1 Method

Stimuli

For comparison, we used the same set of songs for this study as in in the previous chapter. The songs had been carefully chosen to have a relatively high dispersion of postulated cues for boundary perception (cf. previous chapter). Audio and synthesized MIDI stimuli of each song were obtained and the audio and MIDI versions were manually time-aligned.

The two MIDI versions of “Body and soul” were almost the same, however, one was time-aligned to the audio version from Billy Holiday (vocal) and the other to the audio version from Coleman Hawkins (instrumental). Although much effort was spent on aligning the MIDI and audio versions, the song “Body and soul” (instrumental) was

less well aligned than the others due to differences in the melodies between the two stimulus types.

Procedure

The same procedure was used as in the first experiment of the previous chapter. Subjects were asked to indicate segment boundaries by pressing the space bar on the computer keyboard.

Apparatus

The MIDI stimuli were synthesized offline using Steinberg's Cubase CS internal MIDI synthesizer and the audio stimuli were used as recorded on audio compact disc (CD). The same apparatus as in the previous chapter was used.

Subjects

Twenty-one subjects participated in the segmentation experiment using the MIDI stimuli. Subjects' age ranged from 21 to 40 years (μ=27.1, σ=4.8). The number of years of practical musical training ranged from none to 26 years (μ=6.4, σ=8.5) and from none to ten years of theoretical music training (μ=1.7, σ=2.9). None of the subjects were professional musicians.

Eighteen subjects participated in the segmentation experiment using the audio recordings. Subjects' age ranged from 21 to 37 years ($\mu = 26.5$, $\sigma = 4.2$). Their practical musical training ranged from none to 21 years of practical training ($\mu = 6.8$, $\sigma = 7.0$) and from none to ten years of theoretical musical training (μ=2.1, σ=3.1). None of the subjects were professional musicians.

A few subjects participated in our experiments for more than one stimulus type, including our previous experiment on MIDI melodies (Chapter 2). The two experiments presented here (on polyphonic MIDI and audio) had six subjects in common. The current polyphonic MIDI experiment had three subjects in common with our former experiment on monophonic MIDI melodies (Chapter 2). The current polyphonic audio experiment had five subjects in common with our former experiment on monophonic MIDI melodies.

3.2.2 Results

There was a wide range in the total number of boundary indications for each subject. Across the six songs individual subjects indicated between 2 and 151 boundaries per trial (μ=17.9) for the MIDI stimuli and between 2 and 99 boundaries (μ=23.6) per trial for the audio stimuli. The mean number of boundary indications across subjects and songs was for the MIDI stimuli 19.2 (σ=21.1) for trial 1, 17.4 (σ=16.6) for trial 2, and 17.3 (σ=16.9) for trial 3. For the audio stimuli the mean numbers of boundary indications were 23.1 (σ=16.3), 23.6 (σ=14.7), and 24.2 (σ=16.2) for trials 1, 2, and 3, respectively. We tested the numbers of boundary indications between the three trials with an ANOVA, which showed no significant difference in the number of boundary indications across trials (for MIDI: $F_{2,340} = 1.08$, p = 0.34, and for audio: $F_{2,289} = 0.56$, p = 0.57). No interaction between song and repeated trials was found (for MIDI: $F_{10,340} = 0.51$, p = 0.88, and for audio: $F_{10,289} = 0.56$, p = 0.57). There was, however, a strong influence of the song (for MIDI: $F_{5,340} = 24.6$, $p < 0.001$, and for audio: $F_{5,289} = 35.9$, $p < 0.001$) and an influence of the subject (for MIDI: $F_{20,340} = 23.8$, $p < 0.001$, and for audio: $F_{17,289} = 61.5$, $p < 0.001$) on the number of indicated boundaries, thus individual subjects indicated a significantly different number of boundaries and their number of indicated boundaries was song dependent. These results suggest that subjects segmented the songs similarly in terms of the number of boundary indications across the three trials.

In order to construct a profile of the boundary indications across subjects, we first generated pulse-trains of the boundary indications for each subject and trial and then collapsed them across subjects and trials. The collapsed pulse-train was smoothed by convolving the pulse-train with a Gaussian window to reduce the effect of small time dispersions in the boundary indications. The size of the Gaussian window was chosen to be as large as possible, while minimizing the number of times that more than one boundary indication from the same trial fell into the same window (cf. also Chapter 2). Figure 3.1 shows the estimation of the "optimal" window size based on the analysis of boundary indications for the individual subjects. The optimal window size was defined as the number of windows containing exactly one boundary indication from each trial while minimizing the number of windows containing more than one boundary indication from one trial. The figure shows the analysis for the MIDI complete (left panel) and audio stimuli (right panel). For both representations the optimal window size is between

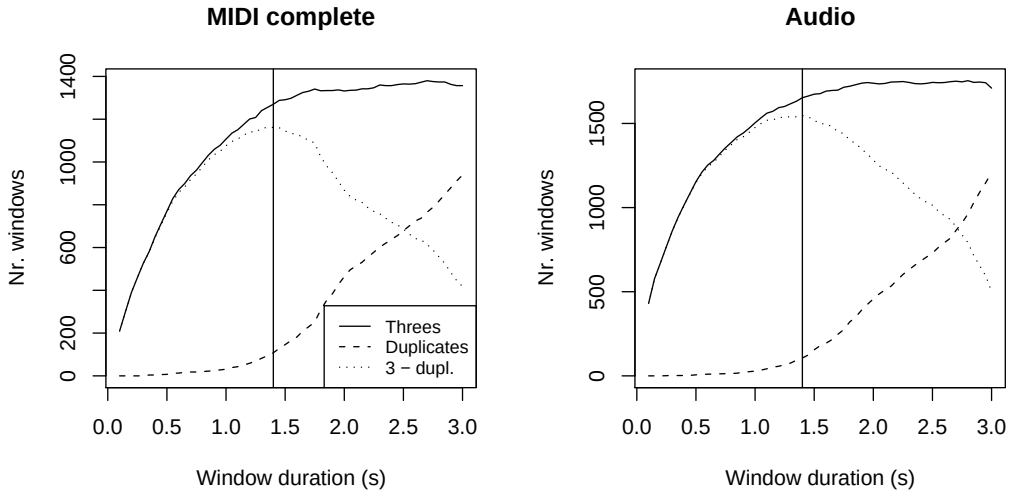


Figure 3.1: The estimation of the optimal window duration. The solid line represents the number of windows containing exactly one indication per trial as a function of the window duration. The dashed line represents the number of windows containing at least two indications from the same trial. The dotted line is the difference between the two with the vertical line representing the maximum of the difference and thus the optimal window duration.

1.0 and 1.5 s. In order to ease comparison of the results from this chapter also with the results of Chapter 2 the "optimal" window size was set to the same value used before. Therefore, the full width at half maximum of the Gaussian was set to 1.25 s. At each peak of the smoothed pulse-train the number of boundary indications within this "optimal" window was counted, while taking no more than one indication per trial, window and subject into account (in order to avoid accidental extra boundary indications). The resulting boundary profiles from the six songs for the MIDI and audio stimuli are shown in Figure 3.2 in columns two and three, respectively. For comparison we added the boundary profiles of the monophonic melodies from the previous chapter in the first column. Boundary profiles are normalized by the maximum possible number of boundary indications, which is given by the product of number of subjects and number of trials. The figure shows that there is a wide range in the number of boundary indications for different boundaries. For each song and stimulus type there are a few boundaries that are indicated by most subjects and trials. It seems also that some of the songs have a more regular structure of salient boundaries, for example both MIDI complete versions of "Body and soul". These two versions of "Body and soul" have similar segmentation patterns, likely due to the fact that these two stimuli were the

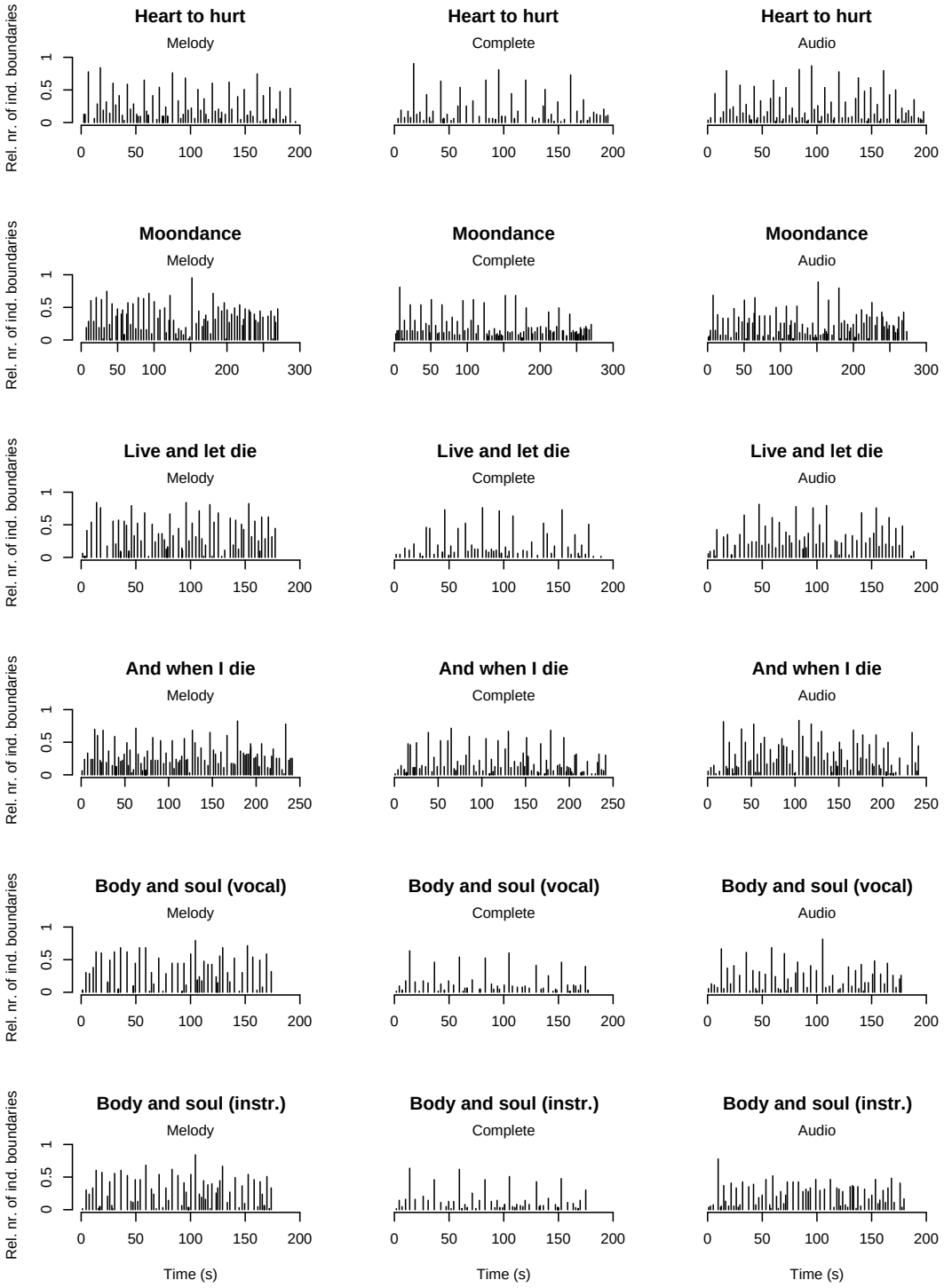


Figure 3.2: Boundary indications over time for all songs and the three stimulus types. The abscissa represents the time in seconds and the ordinate the number of boundary indications within a 1.25 s window, normalized by the theoretical maximum number of indications. For comparison, the first column shows the distribution of the segmentation of the monophonic MIDI stimuli of the six songs (Chapter 2). The second column shows the distribution for the MIDI complete stimuli and the third for the audio stimuli.

same apart from different timings. Overall, it seems, thus, that certain boundaries are perceived by many subjects, and others only by a few subjects.

We performed both within- and across-subject consistency analyzes to investigate if subjects segmented the piece similarly over the course of the three trials (within-subject) and to see whether different subjects segmented the pieces in a similar manner (across-subject). One way of analyzing the within-subject consistency is to calculate the difference in the number of boundary indications across the three trials. The reason for this measure was the following. We wanted to take into account the fact that more emphasis should be given if a subject indicated an additional boundary if it only indicated few boundaries in contrast to an additional boundary of a subject indicating many boundaries. To summarize this, we calculated the mean-difference-per-trial for each subject:

$$\text{mean-difference-per-trial}[i] = \frac{|n_1[i] - n_2[i]|}{N[i]} + \frac{|n_1[i] - n_3[i]|}{N[i]} + \frac{|n_2[i] - n_3[i]|}{N[i]},$$

where $n_1[i]$, $n_2[i]$, and $n_3[i]$ are the number of boundary indications of subject i for trial 1, 2, and 3, and N is the total number of boundary indications of subject i. The mean of the mean-difference-per-trial across songs was 0.18 (σ=0.19) for the MIDI stimuli and 0.18 (σ=0.14) for the audio stimuli, which means that, across trials, subjects indicated on average one additional boundary or one boundary less for every six boundary indications. It seems, thus, that subjects were relatively consistent in the number of boundary indications across the three trials.

An alternative way to estimate within-subject consistency is to calculate the correlation between the smoothed pulse trains for each trial pair per subject. The smoothed boundary profiles for each trial were pair-wise correlated (using the Pearson correlation) with each other, resulting in three within-subject correlations per subject. Averaged correlation values for each song across subjects are shown in Table 3.1. The table shows that subjects are moderately consistent over the three trials and that there was little influence of the song on within-subject correlation. Although not significant, there is a trend that subjects were slightly more consistent between the second and third trial than between the first and second trial, thus becoming more consistent over the course of the three trials. Overall, however, the two within-subject measures show no significant change in the segmentation pattern over the course of the three trials.

Table 3.1: The within- and across-subject correlation between the boundary profiles in the repeated trials for the six songs. The within-subject correlation was calculated by taking the smoothed boundary profiles of the boundary indications of each trial and correlating them pair-wise. In the last column the mean correlation across subjects is shown, calculated by correlating the mean boundary profile of each subject pair-wise across subjects.

Midi complete					
	Within-subject				Across-subject
	Trials 1-2	Trials 1-3	Trials 2-3	Mean (Sd)	Mean (Sd)
"Heart to hurt"	0.56	0.53	0.62	0.57 (0.19)	0.45 (0.19)
"Moondance"	0.52	0.49	0.57	0.53 (0.20)	0.31 (0.27)
"Live and let die"	0.62	0.55	0.60	0.59 (0.20)	0.41 (0.22)
"And when I die"	0.53	0.53	0.57	0.54 (0.18)	0.35 (0.18)
"Body and soul" (v)	0.54	0.51	0.57	0.54 (0.23)	0.35 (0.20)
"Body and soul" (i)	0.56	0.53	0.58	0.56 (0.23)	0.35 (0.21)

Audio					
	Within-subject				Across-subject
	Trials 1-2	Trials 1-3	Trials 2-3	Mean (Sd)	Mean (Sd)
"Heart to hurt"	0.62	0.61	0.68	0.64 (0.20)	0.43 (0.19)
"Moondance"	0.64	0.62	0.66	0.64 (0.20)	0.30 (0.23)
"Live and let die"	0.62	0.57	0.66	0.62 (0.16)	0.36 (0.18)
"And when I die"	0.53	0.57	0.60	0.57 (0.18)	0.38 (0.18)
"Body and soul" (v)	0.60	0.57	0.65	0.60 (0.22)	0.37 (0.17)
"Body and soul" (i)	0.51	0.49	0.55	0.52 (0.21)	0.26 (0.22)

To measure across-subject consistency, the smoothed pulse train of all boundary indications across the three trials was first computed for each individual subject resulting in a per-subject mean boundary profile. These profiles were then pair-wise correlated for each pair of subjects and the mean was taken. Summaries of these correlations are presented in the last column of Table 3.1 for all six songs. The table shows that overall the correlation is rather low, despite the fact that several boundaries were indicated by all subjects. Furthermore, there is no clear consistent pattern across the stimulus types and the across-subject pattern is relatively similar across the six songs for both the MIDI and audio stimuli.

A question was whether a decreasing number of boundary indications was primarily caused by the effect that each individual listener decreased the number of boundary indications across the three trials, i.e., indicating a boundary only on one or two trials. Alternatively, such a decrease in the number of boundary indications could be caused by the effect that a growing number of subjects did not indicate such a boundary at all, while the remaining subjects still indicated the boundary across all three trials. The first alternative would mean that subjects are very consistent among each other and decrease their perceived boundary level as indicated by the number of boundary indications. The second alternative would mean that different subjects have different opinions whether a specific boundary is present or not. Those for whom the boundary still was present would then ideally give three boundary indications across the three trials, while others would give zero indications.

The best chance to decide which alternative is more appropriate for our data is to look at those boundaries which achieved about 0.5 of the maximal possible number of boundary indications. The analysis shown in Figure 3.3 was thus performed for all boundaries with indication proportions between 0.4 and 0.6. For each of these boundaries, we determined the number of subjects who had indicated this boundary 0,1,2, or 3 times across the three trials. By doing this for all selected boundaries for all six songs, we get a distribution/histogram of frequency of occurrences of 0,1,2, or 3 indications. The first alternative mentioned above would lead to a large number of responses for one or two indications and only few for zero and three indications (inverted "U" shape). The second alternative would result in a pattern with clear maxima for zero and three indications ("U" shape). The observed "U" form, as shown in Figure 3.3, for the MIDI complete stimulus types (left panel) as well as the audio stimulus types (right panel) is clearly more in agreement with alternative two, suggesting that subjects

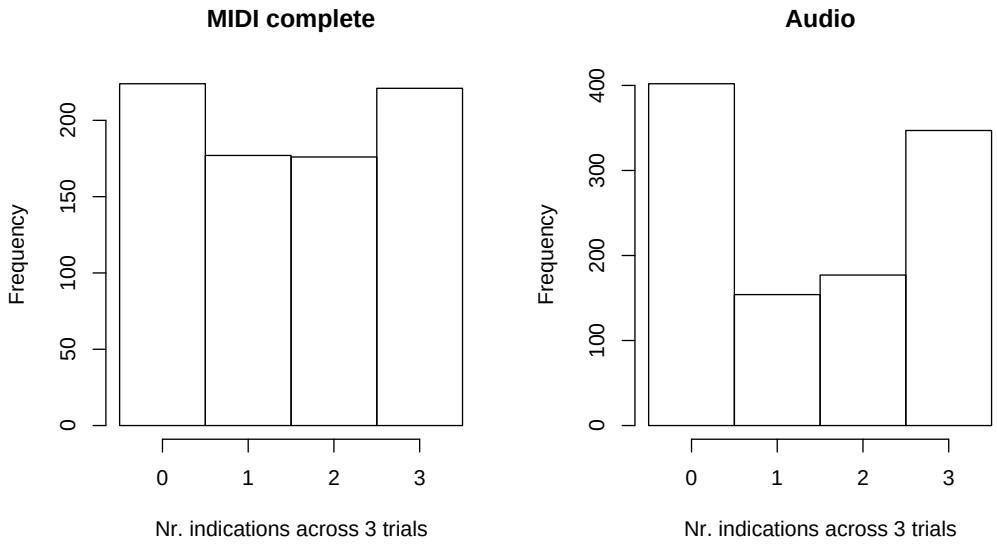


Figure 3.3: Histograms of the number of boundary indications per boundary and per subject across the three trials for the moderate boundaries (relative number of boundary indications between 0.4 and 0.6) across subjects and songs. The abscissa shows on how many trials individual subjects notated these moderate boundaries, i.e., a zero means that the subject did not notate the boundary across the three trials and a three means that the subject notated the boundary across all three trials. The left plot shows the histogram for the MIDI complete stimuli and the right plot shows the histogram for the audio stimuli.

indeed tend to either indicate a boundary across all three trials or do not indicate the boundary at all.

To differentiate these possible scenarios further, we examined the dispersion of the total number of boundary indications at a given time point over subjects and trials. For each boundary we related the number of boundary indications to the number of subjects that indicated this boundary at least once, and to the mean number of boundary indications (across the three trials) of those subjects who indicated this boundary at least once. The number of subjects that indicated a particular boundary and the mean number of trial indications were each correlated with the total number of boundary indications. The correlation between the number of subjects and the total number of boundary indications was 0.98 for the MIDI complete stimuli and 0.97 for the audio stimuli. The correlation between the mean number (as defined above) of boundary indications across the three trials and the total number of boundary indications was 0.75 for the MIDI complete stimuli and 0.83 for the audio stimuli. Thus, the total number

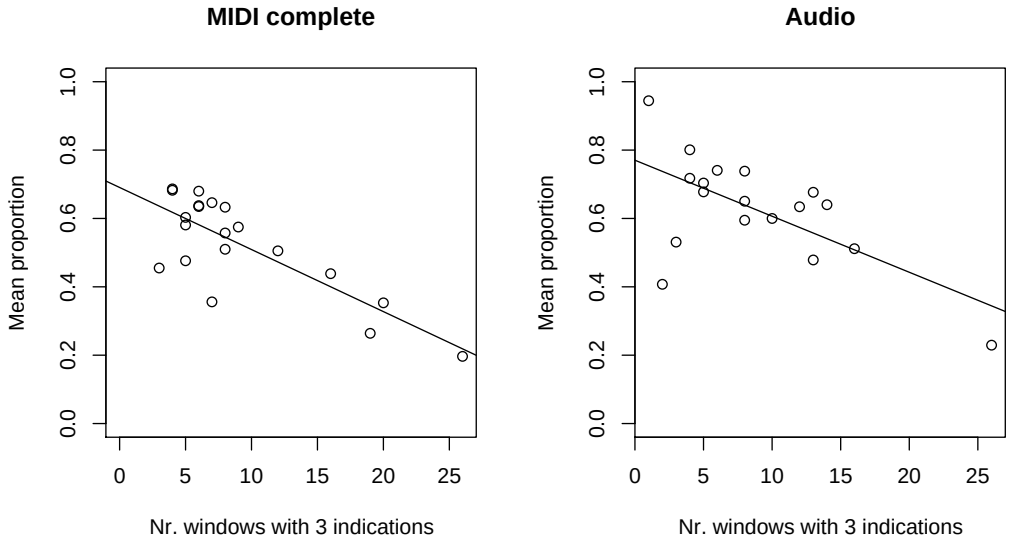


Figure 3.4: Scatterplot for the song "And when I die" of the number of windows containing exactly one boundary indication in each trial and the mean proportion number of these boundaries shown for the MIDI complete (left panel) and the audio representation (right panel). The points indicate the results for the individual subjects.

of boundary indications is correlated much stronger with the number of subjects than with the mean number of indications per trial. This outcome suggests that a decreasing number of boundary indications reflects a decreasing number of subjects indicating a boundary, rather than a decrease in the per-subject mean indication. It seems, thus, that different subjects use different segmentation strategies. The segmentation strategy is then applied consistently across the three trials.

The difference in the number of boundary indications and thus the segmentation strategy may be the result of different individual boundary thresholds. Such a threshold would mean that a subject indicates a boundary only if the boundary strength exceeds the individual threshold. A consequence of the threshold would be that subjects indicating only few boundaries perceive and indicate only the strongest boundaries while subjects using many boundaries for segmentation also indicate boundaries having a lower strength. This hypothesis was tested with the following analysis. For each song and subject the indications for which the subject gave exactly one boundary indication per trial within the optimal 1.25 s window were selected, i.e., the consistent boundaries across the three trials. We then analyzed what the mean proportion number (across

all subjects and trials) of these boundaries was. This mean proportion number should be high for subjects only contributing to strong boundaries, i.e., those with a high threshold. These subjects should overall only indicate few boundaries. There was a significant negative correlation between the number of consistent boundaries and the mean proportion number for four songs of the MIDI complete representation ($-0.74 < r < -0.80$, $p < 0.001$) but not for "Heart to hurt" ($r = -0.35$, $p = 0.12$) and "Body and soul (vocal)" ($r = -0.13$, $p = 0.59$). For the audio representation, the relation was significant for all songs ($-0.64 < r < -0.92$, $p < 0.01$) except "Body and soul (instr)" ($r = -0.55$, $p = 0.12$). Figure 3.4 shows the relation for a typical song, "And when I die", for the MIDI complete (left panel) and the audio representation (right panel). It seems, thus, that there is a negative correlation between the consistent boundaries and the mean proportion number across all subjects. These findings support the idea that subjects indeed have an individual threshold when asked to indicate a boundary.

Another question is whether the lyrics have an influence on the perception of segment boundaries. It was expected that the lyrics support the perception of segment boundaries and that thus the boundaries in the audio representation coinciding with the start or end of lyric phrases are more easily perceived and therefore more often indicated. From the six songs, four had lyrics in English, a language all participating subjects understood. None of our subjects understood the language of the song in Japanese. The four English songs could thus be evaluated for the influence of the lyrics on the perception of segment boundaries. The evaluation was done by creating a boundary profile where a boundary is placed at the start of each phrase of the lyrics as well as another analysis with the lyric boundary placed at the end of each lyric phrase. The lyric boundary profile was then compared to the boundary profiles of the MIDI complete and the audio representation. For each of the lyric boundaries we extracted the highest peak in the MIDI complete and audio profile within a time window placed at the center of each lyric boundary. If either of the two representations did not contain a boundary around the lyric boundary it was counted as a boundary with strength zero. To account for general differences in the number of boundary indications between the MIDI complete and the audio representation, we normalized the number of boundary indications by the number of subjects (21 for MIDI complete and 18 for audio) and the mean number of indicated boundaries across the song ($\mu = 18.0$ for MIDI complete, $\mu = 23.6$ for audio). We then compared the mean number of boundary indications for the boundaries close to the lyric boundaries. To see whether the lyrics have a significant

influence on the strength of segment boundaries, a Welch t-test was performed on the boundary strength of the boundaries close to the lyric boundaries of the MIDI complete and audio representation. All boundaries within a certain time window were considered as being the same. The difference between the lyric boundaries at the start of lyric phrases was significant ($p < 0.05$) for the song "Live and let die" for all window sizes between 1.0 and 2.0 s, for the song "And when I die" it was only significant ($p < 0.05$) for a window size of 1.6 to 1.7 s, and for the song "Body and soul" it was significant with a window size longer than 1.75 s. The difference between the lyric boundaries at the end of lyric phrases was significant ($p < 0.05$) for the two songs "Moondance" for window sizes above 1.35 s and "Body and soul (vocal)" for all window sizes between 1.0 and 2.0 s, except between 1.4 and 1.5 s. In all cases which showed a significant influence, the boundary strengths were higher for boundaries that coincided with the start or end of lyric phrases. These results show that lyrics seem to have a contribution to the perception of segment boundaries, but whether the start or end of the lyric phrase has a higher influence is song dependent.

A further investigation tested whether subjects with practical musical training segmented the pieces differently compared to subjects without musical training. To analyze the influence of musical training we compared the 33% of all subjects with the highest practical musical training (MIDI complete: $\mu = 15.3$ years, $\sigma = 6.05$; audio: $\mu = 15.3$ years, $\sigma = 4.23$) with the 33% lowest training (MIDI complete: $\mu = 0$ years; audio: $\mu = 0.5$ years, $\sigma = 0.84$). A Welch t-test showed a significant influence of musical training on the number of boundary indications (MIDI complete $t = 4.08$, $df = 216.4$, $p < 0.001$; audio $t = 5.95$, $df = 186.6$, $p < 0.001$). Subjects with musical training indicated fewer boundaries (MIDI complete $\mu = 16.4$, audio $\mu = 19.3$) than nonmusicians (MIDI complete $\mu = 26.9$, audio $\mu = 31.8$). Moreover, the within-subject consistency was significantly higher for subjects with high musical training (Welch t-test: MIDI complete $t = -3.31$, $df = 81.1$, $p < 0.001$; audio $t = -5.32$, $df = 81.7$, $p < 0.001$). The across-subject correlation, however, was not significantly different between the two groups of subjects for any of the six songs (using Welch t-tests the p value ranged for MIDI complete representation from $p = 0.21$ to 0.47 and for the audio representation from $p = 0.15$ to 0.64). It seems thus, that subjects with musical training segment less often and are more consistent within themselves. Across subjects, however, subjects with musical training are not more consistent than nonmusicians.

So far we have considered only the boundary indications and the derived boundary

profiles. It is also of interest to see if the boundary profiles and in particular the number of boundary indications within a specific time window obtained in this experiment can be used as a measure of boundary salience and what cues are used to convey the most salient boundaries.

3.3 Experiment II: Salience rating of selected boundaries

The primary goal of this salience rating experiment was to get an explicit measure of the salience for selected boundaries from the previous segmentation experiment. In addition to the salience rating, a second goal of this experiment was to identify explicit musicological cues underlying each of the selected boundaries. A third goal was to see if there exists a correlation between the implicit salience measure derived from data collected in the segmentation experiment and the explicit salience measure collected here. Previous experiments have assumed such a correlation between these two measures (Clarke & Krumhansl, 1990; Frankland & Cohen, 2004) and in the previous chapter we validated this relationship for monophonic melodies. This experiment investigated if the correlation also holds for polyphonic popular music.

3.3.1 Method

From the boundary indications obtained in the segmentation experiment, a selection of boundaries was made for this experiment. We selected the strongest boundaries (boundaries with enough indications to fall within 10 percent of most often indicated boundaries per song), two to three moderately strong boundaries, and two to three weaker boundaries. The selected boundaries provided thus a wide range of boundary strengths. Care was taken that the weaker boundaries were not in the temporal vicinity of stronger boundaries to avoid confusion. In total, between 16 and 22 boundaries were selected for each song for the MIDI complete stimuli and between 13 and 20 boundaries for each song for the audio stimuli.

Procedure

Subjects had the same task as in the previous chapter. They were asked to rate the salience of the selected boundaries and describe the boundary cues.

Subjects

For the MIDI stimuli, 20 out of 21 subjects from the segmentation experiment participated in this rating experiment. Subjects' age ranged from 21 to 40 years (μ=27.3, σ=4.8). The practical musical training ranged from none to 26 years (μ=6.4, σ=8.6) and the theoretical musical training ranged from none to ten years (μ=1.7, σ=2.9).

For the audio stimuli 15 out of 18 subjects participated in this rating experiment. The age of the subjects ranged from 22 to 37 years (μ = 26.9, σ = 4.4). The practical musical training ranged from none to 21 years (μ = 6.7, σ = 7.3) and the theoretical musical training ranged from none to ten years (μ=2.0, σ=3.1).

Apparatus

The same apparatus and setup as in the second experiment of the previous chapter was used.

3.3.2 Results

The across-subject mean boundary salience ratings for the selected boundaries ranged from 0.5 to 5.6 for the MIDI stimuli and from 0.2 to 5.7 for the audio stimuli. Thus, subjects used the whole scale for rating boundary saliences. To obtain an estimate of across-subject consistency, we correlated the boundary salience ratings for individual subjects pair-wise with each other. The correlations between the salience ratings are shown in Table 3.2. Across all songs the mean correlation across the subjects' salience ratings was 0.64 $\pm$ 0.031 for the MIDI stimuli and 0.66 $\pm$ 0.026 for the audio stimuli. The across-subject correlations were not significantly different in both representations (see also Table 3.6). Overall, subjects gave moderately similar salience ratings for both stimulus types.

Subjects also described the perceptual cues of each boundary. To analyze the cues of each boundary further, the provided cue descriptions were classified into "cue classes", which is an extended version of the one that has been developed in the previous study (Chapter 2). The extension took into account that polyphonic music also contains additional features, such as rhythmic instruments and vocals. The "cue classes" consist of the five groups *rhythm*, *timbre*, *vertical/horizontal* and *dynamics* as well as one group describing more complex summaries of music structure, with the five groups

Table 3.2: The mean across-subject correlation of the salience ratings with the standard error of the mean. The salience ratings of each subject were pair-wise correlated with each other subject, with the standard error of the mean estimated by bootstrapping of the salience ratings (N=500).

Stimulus type	Heart to hurt	Moondance	Live and let die	And when I die	Body and soul (vocal)	Body and soul (instr.)
MIDI compl.	0.64±0.044	0.67±0.041	0.72±0.030	0.56±0.045	0.72±0.050	0.61±0.075
Audio	0.64±0.055	0.69±0.042	0.77±0.035	0.60±0.053	0.70±0.048	0.54±0.051

and their classes shown in Table 3.3. Each description of a boundary cue given by subjects was classified into these "cue classes", similarly to the previous chapter. For the MIDI stimuli the classification resulted in a total of 2571 terms mentioned and for the audio stimuli the classification yielded 1768 terms. It should not be overlooked, however, that the two stimulus types had different numbers of selected boundaries and also a different number of participating subjects for this experiment. The MIDI stimuli had a total of 120 boundaries (across the six songs) and 20 subjects, while the audio stimuli had 98 boundaries and 15 subjects. In order to facilitate data comparison across the stimulus types, the total number of terms was normalized by division by 2400 (20 subjects × 120 boundaries) for the MIDI stimuli and by 1470 (15 subjects × 98 boundaries) for the audio stimuli. The resulting values are shown in Table 3.4 as percentages for the MIDI stimulus (in parenthesis) and in Table 3.5 for the audio stimulus (in parenthesis). For the MIDI stimulus, the terms most often given by subjects as boundary cues were *change in timbre(other)* (4.5%, thus in the mean 0.045 times per boundary per subject) and *global structure* (3.4%). For the audio stimuli the most often given terms were *global structure* (4.3%) and *change in timbre(other)* (3.7%). These results indicate that subjects perceive and notate changes in timbre and global structural terms like "beginning of verse" most easily for both types of stimuli.

To analyze whether subjects with musical training were more consistent in their salience rating, we compared the mean salience rating of the third of subjects with the lowest practical musical training with the highest third. For the MIDI complete representation, a Welch t-test showed that subjects with musical training were significantly higher correlated with each other than subjects without musical training ($t = 5.82$, $df = 35.7$, $p < 0.001$), but that there was no significant difference for the

Table 3.3: The "cue classes" used to classify the descriptions given by subjects for the boundary cues.

Group	Cue class
Rhythm	Change in strength
	Tempo change
	Rhythm change
Timbre	Voice
	Drums
	Other timbre changes
Vertical/horizontal	Harmonic progression and tonality change
	Melody change
Dynamics	Level change
	Break
	Global structure
	Repetitions

audio representation ($t = 2.10$, $df = 16.8$, $p = 0.051$). We were also interested whether subjects with musical training, as expected, use different descriptions compared with subjects without musical training. We compared the third of subjects without musical training with the third of subjects having the highest practical musical training. For the MIDI complete representation subjects with musical training described boundaries with significantly more terms ($t = -3.41$, $df = 10.8$, $p < 0.01$) and with more terms from the "cue class" global structure ($t = -2.89$, $df = 8.67$, $p < 0.05$). For the audio representation musically trained subjects described boundaries not with a significantly different total number of terms ($t = -1.23$, $df = 11.1$, $p = 0.25$) but significantly less often using terms from the "cue class" break ($t = 2.67$, $df = 11.4$, $p < 0.05$). It seems, thus, that the influence of musical training is dependent on the representation, at least for our selection of subjects and songs.

In addition to counting the total number of times a term was used, it is also interesting to see which of the cues were used for the most salient boundaries. Therefore, each term was also related to the boundary salience rating given for the described boundary, to compute the "mean term rating" for each cue. The mean term rating for a particular cue is calculated as follows: First, for each cue the boundaries were selected that were described with the cue. Second, the salience ratings from the subjects that used the cue in their boundary descriptions were collected. The mean salience rating across all boundaries to which subjects assigned the cue was taken as the mean term rating. The mean term ratings of each cue are shown in Table 3.4 for the MIDI stimuli

Table 3.4: The mean term rating of the classified descriptions for the MIDI complete stimuli and the normalized number of terms in percentage (in parenthesis). The normalized number of times a certain cue was mentioned per boundary per subject was calculated by dividing the total number of terms mentioned by the number of subjects and by the number of selected boundaries in the song (and by the number of songs for the "Overall" row). For clarity, the normalized number of times a certain cue was mentioned is shown in percentages, i.e., percentage per boundary per subject. In total there were 120 boundaries across the six songs and 20 subjects participated in the salience rating experiment. The labels are Drum: change in timbre(drum); Voice: change in timbre(voice); Other: change in timbre(other); Level: change in the sound level; Progression: harmonic progressions; Melody: change in the melody; Strength: change in rhythm strength; Tempo: tempo change; Rhythm: change in rhythm; Global: global structural descriptions; Break: break or rests; and Repetition: repetition of previous material.

Song	Rhythm strength	Tempo change	Rhythm change	Drum	Voice	Other timbre
Heart to hurt	– (0)	3.65 (0.71)	4.65 (0.96)	– (0)	– (0)	4.84 (3.1)
Moondance	6.00 (0.042)	3.64 (0.92)	4.86 (1.8)	4.67 (0.12)	6.00 (0.042)	4.75 (4.8)
Live and let die	– (0)	5.22 (2.6)	5.47 (2.5)	4.17 (0.25)	– (0)	4.80 (6.9)
And when I die	– (0)	4.79 (3)	4.95 (2.6)	5.38 (0.33)	– (0)	4.87 (6.4)
Body and soul (vocal)	– (0)	2.86 (0.29)	4.44 (0.38)	– (0)	– (0)	4.56 (3)
Body and soul (instr.)	– (0)	2.50 (0.25)	2.50 (0.083)	– (0)	– (0)	4.54 (2.9)
Overall	6.00 (0.0069)	4.55 (1.3)	5.01 (1.4)	4.82 (0.12)	6.00 (0.0069)	4.76 (4.5)

Song	Progression	Melody	Level	Break	Global	Repetition
Heart to hurt	4.89 (0.38)	4.23 (3.5)	4.44 (0.67)	3.77 (1.8)	4.87 (3.6)	4.37 (2)
Moondance	– (0)	5.23 (3.6)	4.22 (0.38)	3.44 (0.75)	4.67 (4.1)	3.95 (4.3)
Live and let die	5.00 (0.12)	5.05 (1.8)	4.78 (1.1)	3.89 (0.38)	4.28 (3.1)	4.45 (2.1)
And when I die	5.17 (0.25)	4.93 (1.9)	5.29 (0.29)	5.10 (2.4)	4.43 (2.9)	4.18 (1.8)
Body and soul (vocal)	3.60 (0.21)	3.42 (1.1)	5.00 (0.12)	2.35 (1.5)	4.17 (3.4)	3.71 (3)
Body and soul (instr.)	5.00 (0.12)	4.29 (1.3)	2.00 (0.083)	2.33 (2.2)	4.07 (3.5)	3.94 (3.3)
Overall	4.73 (0.18)	4.66 (2.2)	4.59 (0.44)	3.52 (1.5)	4.43 (3.4)	4.04 (2.8)

Table 3.5: The mean term rating of the classified descriptions for the audio stimuli (see caption of Table 3.4). The normalized number of times a certain cue was mentioned per subject and per boundary is indicated in parentheses, shown in percentage, i.e., percentage per boundary per subject. In total there were 98 boundaries across the six songs and 15 subjects participated in the rating experiment.

Song	Rhythm strength	Tempo change	Rhythm change	Drum	Voice	Other timbre
Heart to hurt	5.00 (0.14)	3.64 (0.75)	4.83 (0.41)	– (0)	4.14 (2.9)	4.24 (2.3)
Moondance	– (0)	3.14 (0.48)	4.67 (1.8)	4.25 (0.27)	4.39 (1.6)	3.85 (4)
Live and let die	– (0)	4.80 (3.4)	4.91 (1.6)	4.67 (0.2)	4.12 (1.6)	4.46 (5.6)
And when I die	5.00 (0.068)	4.87 (3.5)	5.03 (2.1)	5.06 (1.2)	4.80 (1)	4.80 (5.8)
Body and soul (vocal)	4.67 (0.2)	2.17 (0.41)	5.00 (0.14)	4.67 (0.2)	4.71 (2.4)	4.23 (2)
Body and soul (instr.)	– (0)	2.78 (0.61)	3.14 (0.48)	6.00 (0.2)	– (0)	4.20 (2.4)
Overall	4.83 (0.068)	4.39 (1.5)	4.75 (1.1)	4.97 (0.35)	4.40 (1.6)	4.37 (3.7)

Song	Progression	Melody	Level	Break	Global	Repetition
Heart to hurt	4.69 (1.1)	3.91 (1.6)	3.62 (1.4)	2.57 (2.5)	4.54 (4.8)	3.68 (1.5)
Moondance	5.07 (0.95)	4.08 (1.7)	4.05 (2.5)	2.43 (1.6)	3.93 (8.4)	4.32 (3.8)
Live and let die	5.83 (0.41)	4.80 (0.34)	4.40 (2.7)	2.00 (0.54)	4.08 (2.5)	4.26 (2.1)
And when I die	4.83 (0.41)	4.12 (0.54)	3.95 (2.5)	4.20 (1.4)	5.03 (4.8)	5.45 (1.5)
Body and soul (vocal)	4.50 (1.2)	1.92 (0.82)	2.00 (0.61)	2.26 (1.6)	4.05 (3.8)	5.44 (1.1)
Body and soul (instr.)	4.50 (1.4)	3.50 (0.54)	2.81 (1.1)	1.42 (2.9)	4.12 (1.8)	3.83 (2)
Overall	4.76 (0.91)	3.70 (0.92)	3.82 (1.8)	2.36 (1.7)	4.29 (4.3)	4.39 (2)

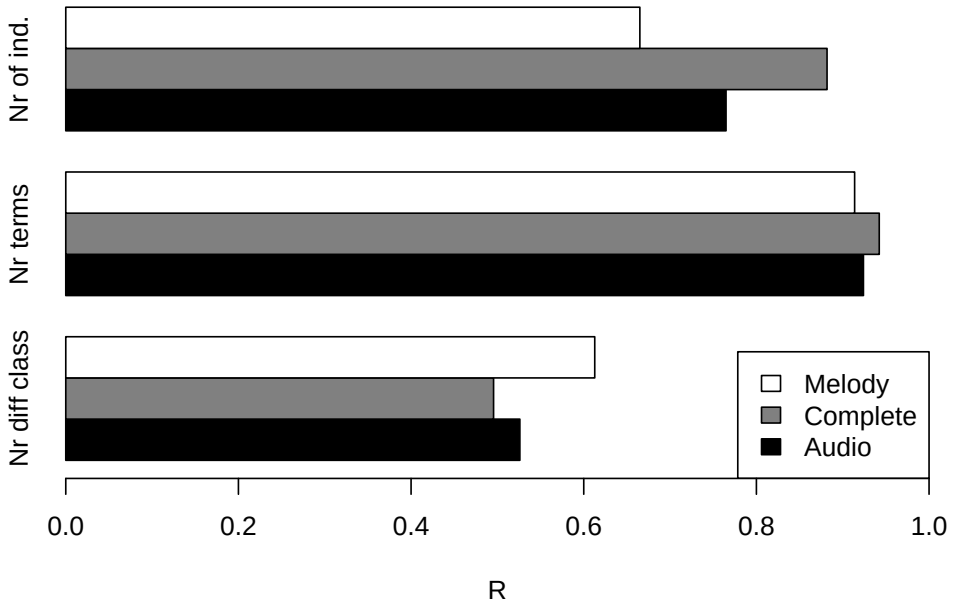


Figure 3.5: The correlation between the mean salience rating and three other measures: the number of boundary indications within a time window (Nr of ind.), the number of terms used for describing the boundary (Nr terms), and the number of different classes used to describe the boundaries (Nr diff class). The first rows (white) represent the results of the previous chapter for the monophonic MIDI stimuli, the second rows (gray) show the results for the MIDI complete stimuli, and the third rows show the results for the audio stimuli.

and in Table 3.5 for the audio stimuli. The tables show a wide range of mean term ratings, thus there are cues mainly used for salient boundaries, such as *change in timbre (drum)*. Other cues are used for less-salient boundaries, such as for instance *breaks*, which is interesting as breaks are generally considered as strong boundary indicators. It seems, thus, that subjects also described less salient boundaries with the cue break. For the MIDI stimuli, *change in timbre (drum)* and *change in timbre (voice)* were only mentioned in a few songs, however, they were associated with high salience ratings when they were mentioned. For both stimulus types there was a wide spread in the mean term ratings across songs for the cue *change in tempo*, which can be attributed to the fact that only the two songs "Live and let die" and "And when I die", had strong changes in tempo.

Two possible measures of boundary salience can be extracted from the "cue classes". One measure is the total number of terms with which a boundary was described, i.e.,

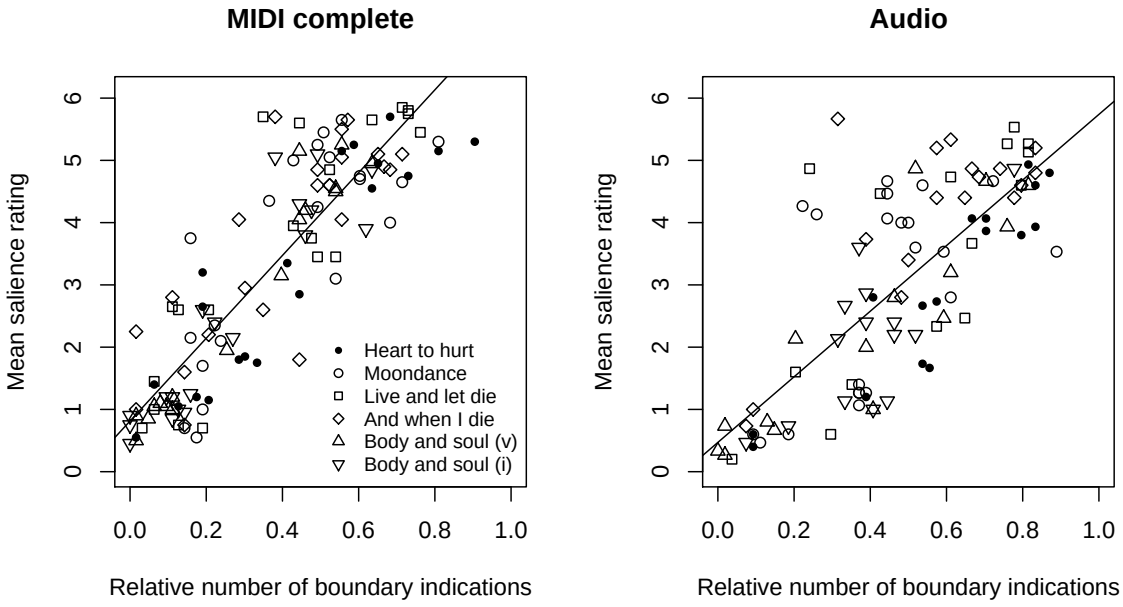


Figure 3.6: Scatterplot of the mean boundary salience ratings versus the relative number of boundary indications for the MIDI complete stimulus type (left) and the audio stimulus type (right) for all boundaries included in experiment II.

the sum of the terms used across all subjects' boundary descriptions. The other measure is the number of *different* classes used to describe a boundary. The idea behind the *different-classes* measure is that there may be a tendency of subjects to describe a salient boundary with cues from different "cue classes". We were interested in whether or not these two measures could be used as additional quantifications of boundary salience. Therefore, the correlations between the salience ratings and the total number of terms and the number of different classes were calculated. The results are shown in Figure 3.5. The correlation between the salience rating and the total number of terms was very high: 0.93 ($p < 0.001$) for the MIDI complete stimuli, and 0.92 ($p < 0.001$) for the audio stimuli. The correlation between the salience rating and the number of different classes was much lower, but still significant: 0.58 ($p < 0.001$) for the MIDI complete stimuli and 0.59 ($p < 0.001$) for the audio stimuli. It seems that the total number of terms used is a much better measure of boundary salience than the number of different classes. It seems that if subjects give a high salience rating to a particular boundary, they also describe the boundary with more terms, but not necessarily with terms from different "cue classes".

A particularly interesting analysis is the relationship between the number of indications of a boundary and the salience rating of the boundary, because it allows to test the assumption that the number of boundary indications (across subjects) is a measure of the boundary salience (Clarke & Krumhansl, 1990; Frankland & Cohen, 2004). Figure 3.6 plots the number of indications versus the salience rating with different symbols for different songs. The figure shows that there exists a strong correlation between the number of boundary indications and the given salience rating, for both stimulus types. The value of this correlation is included in the top of Figure 3.5, which shows that the correlation between the number of boundary indications and the salience rating is indeed very high (0.88 for MIDI and 0.76 for audio). These high correlations suggest that the total number of boundary indications of a particular boundary can indeed be used as a measure of its salience.

In the segmentation experiment we found that, in general, if subjects indicate a boundary they indicate the boundary consistently across all three trials. Subjects, thus, seem to have an either-or perception of segment boundaries, at least in the segmentation process. It is therefore also interesting to analyze whether subjects continue this either-or strategy in rating the boundary salience, i.e., whether they assign a high salience rating for their own indicated boundaries and a low salience rating for boundaries they have not indicated. It should be mentioned that in the salience rating experiment, the individual subjects received no information about their own previous segmentation results. This analysis was done for moderate boundaries within the range of 0.4 to 0.6 of the normalized number of boundary indications, which were 17 boundaries for the MIDI complete stimulus type and 25 boundaries for the audio stimulus type. The number of times a boundary was indicated across the three trials of these moderate boundaries was correlated with the given salience rating of each subject across subjects and songs. The correlation was -0.027 (non significant) for the MIDI stimulus type and 0.017 (non significant) for the audio stimulus type. It seems, thus, that subjects are rather binary in *indicating* the existence or nonexistence of boundaries, while they use a more gradual scale for *rating* the salience of boundaries. The observed high correlation between the total number of boundary indications and the mean salience ratings (cf. Figure 3.5) suggests that the criterion or threshold for indicating a boundary corresponds to different salience scale values for different subjects.

We also found evidence that the start or end of lyric phrases may support the

perception of segment boundaries. Subjects, however, did never explicitly describe the boundaries cues by the lyrics of the song. When they described the vocal content of the piece they used descriptions like "begin/end of singing" or "voice start/stops". It seems, thus, that subjects do not consciously attribute the start and end of lyric phrases as a cue for segment boundaries.

3.4 Summary of experimental findings

The main goal of the salience rating experiment was to compare the segmentation profiles and salience ratings between two different polyphonic representations of the same songs. The polyphonic stimuli consisted of synthesized MIDI files and audio recordings of the same six songs used in the previous study. The observed pattern that a few boundaries are indicated by all or most subjects and others are indicated less often is in line with previous research (Deliège, 1987; Krumhansl, 1996; Frankland & Cohen, 2004). Furthermore, we observed that individual subjects tended to either indicate a certain boundary in all three trials or not at all. Thus, different subjects will, in general, segment a given piece in different ways, with agreement on the very salient boundaries, but disagreement on less salient boundaries. The result is important for algorithms for automatic segmentation, which could, in addition to indicating the place of segment boundaries, also assign a salience score to each boundary.

The aim of the salience rating experiment was to obtain an explicit salience measure and to attach boundary descriptions to the segment boundaries. The high correlation between the salience rating and the number of boundary indications within a time window (obtained in the segmentation experiment) extends our previous results on monophonic music (Chapter 2). Thus, for both, monophonic and polyphonic music, we have two methods to obtain the salience of segment boundaries: the first is to ask subjects to segment the music piece. The number of boundary indications across subjects within a time window is then used as a measure for boundary saliences; the second is to present a selection of boundaries to subjects and ask them to explicitly rate the salience of given boundaries.

Our results show a high correlation between subjects' salience ratings and the total number of terms they used to describe a boundary. There is also a moderate correlation between salience ratings and the number of different classes used to describe the boundaries. These correlations suggest that subjects describe salient boundaries with

more terms, but they do not use a higher number of different classes of terms to describe salient boundaries. Although previous research has only found a minor influence of musical training on segmentation (Deliège, 1987; Deliège & Ahmadi, 1990; Krumhansl, 1996), the boundary cue descriptions could be biased by the musical knowledge of subjects (who were not professional musicians). Subjects may not be able to express precisely what they (subconsciously) perceive as a boundary cue.

The results from the first experiment showed that subjects are more likely to indicate a boundary either across all trials or not indicate the boundary at all. They are, thus, binary in their boundary judgment and given the fact that the number of boundary indications varied highly across subjects, it seems that subjects have rather individual segmentation strategies instead of a common global boundary perception. This finding concurs with previous research, which asked subjects to segment short melodies three times and found a significant correlation between the three trials (Frankland & Cohen, 2004). The moderate to low correlation of the moderately indicated boundaries with the given salience rating, however, indicates that such a binary decision is not further applied in the salience rating task. Furthermore, the global segmentation profile was highly correlated with the mean boundary salience rating. These results suggest that the segmentation profile obtained with a pool of subjects is a good estimate of the boundary saliences.

In summary, the two experiments further confirmed our proposed experimental method for obtaining the salience of segment boundaries of the previous chapter and corroborates that the frequency with which a boundary is indicated and the assigned salience rating are highly correlated, also for polyphonic music. Given that we have the results for three different versions of the songs, two from the present study and another from the previous chapter we will focus in the following on the comparison across the three different stimuli.

3.5 Comparison of the different stimuli

The previous chapter used monophonic representations of the same songs as in the present study to analyze the boundary indications and the salience ratings of selected boundaries with their assigned cue descriptions. In total we have, thus, the results for three different stimulus-types of the same six songs: monophonic and polyphonic stimuli

Table 3.6: Mean within- and across-subject correlations of the smoothed boundary indication profiles and the mean pair-wise correlation of the boundary salience ratings across subjects, shown with the standard deviation in parenthesis.

	MIDI melody	MIDI complete	Audio
Mean within-subject correlation	0.63 (0.15)	0.56 (0.19)	0.60 (0.19)
Mean across-subject correlation	0.37 (0.24)	0.37 (0.22)	0.35 (0.21)
Mean of the pair-wise correlation of the salience ratings	0.56 (0.21)	0.65 (0.19)	0.66 (0.18)

synthesized from the MIDI as well as the recorded audio. Because all three stimuli were time-aligned, the segmentation results can be compared with each other.

We first examined the effect of stimulus type on within- and across-subject consistency to see, for a specific stimulus type, if subjects are more consistent across the three trials or whether there is greater across-subject agreement. We calculated the pair-wise correlations between different smoothed boundary profiles of individual trial pairs (within-subject) and between the mean smoothed boundary profile across the three trials across subject pairs (across-subject). The averages of these values are shown in Table 3.6. A number of conclusions can be taken from the table. First, for the segmentation experiment, the consistency within subjects is considerably higher than across subjects, thus subjects indicated the boundaries relatively similar across the three trials but different subjects tended to indicate different boundaries. Second, the within- and across-subject correlations seem to be fairly similar across the three stimuli and thus the type of stimulus has little influence on these measures. A third observation is that the across subject correlation is higher in the rating experiment than in the segmentation experiment. These results complement the earlier analysis, in which we showed that the idiosyncratic segmentation patterns of different subjects are not reflected in their salience ratings, thus leading to more similar rating behavior across subjects.

The two experiments yielded an implicit and an explicit measure of the boundary saliences. These two measures were the number of boundary indications within a time window obtained in the segmentation experiment (implicit) and the salience rating of selected boundaries (explicit). We wanted to investigate what the effect of the stimulus type was on these two measures. For instance, a qualitative analysis of Figure 3.2 shows that the highly salient boundaries of the song “Heart to hurt” are indicated

at similar points in time across all three stimulus types. We therefore correlated the mean boundary indications profiles, as well as the mean salience ratings, pair-wise across the three different stimulus types. We expected the MIDI melody stimuli to be more correlated with the MIDI complete stimuli than with the audio stimuli, as the melody is a subset of the MIDI complete stimuli and is thus contained entirely in the MIDI complete stimuli. It was also expected that the audio stimuli would be more highly correlated with the MIDI complete stimuli than with the MIDI melody stimuli, because both the audio and the MIDI complete are polyphonic stimuli.

For the correlation of the boundary indications, the smoothed profile was used to compensate for the scattering of the boundary indications. Figure 3.7 shows the pair-wise correlations for each song. The figure also includes the standard error of the mean, which was estimated by bootstrapping the subjects mean boundary profile (across the three trials) across subjects (N=500). Care was taken in the bootstrap repetitions to either include or exclude subjects who participated in more than one experiment. The figure shows that in three of the songs, thus "Live and let die"(3), "And when I die"(4), and "Body and soul (instrumental)"(6), the correlation between the MIDI melody stimuli and the audio stimuli was, as expected, lowest. "Heart to hurt"(1) and "Moondance"(2), on the other hand, have generally a higher correlation across the three pairs of stimulus types, although this difference is not significant. For the instrumental versions of "Body and Soul"(6) the correlation between the MIDI melody stimuli and MIDI complete stimuli was significantly higher than that between either of the MIDI and the audio stimuli. One possible explanation for this observation is that the audio version of this song was less well represented by the MIDI because the melody of the MIDI was not a perfect representation of the melody of the audio recording. Apart from these specific cases there seems to be a moderately high correlation across the different stimulus types. This correlation value may, however, be affected by the between group design of the studies with different stimulus, which had only a few subjects in common.

To estimate the between-group effect of the subjects recruited for the experiments on the correlations between the boundary profiles, we calculated the correlation for one stimulus type between the mean boundary profile of half of the subjects with the mean boundary profile of the other half of the subjects, where the order of the subjects was randomized. This process was repeated 500 times in a bootstrap analysis. The mean correlations across the 500 repetitions were 0.87 (σ=0.03 across songs) for the MIDI

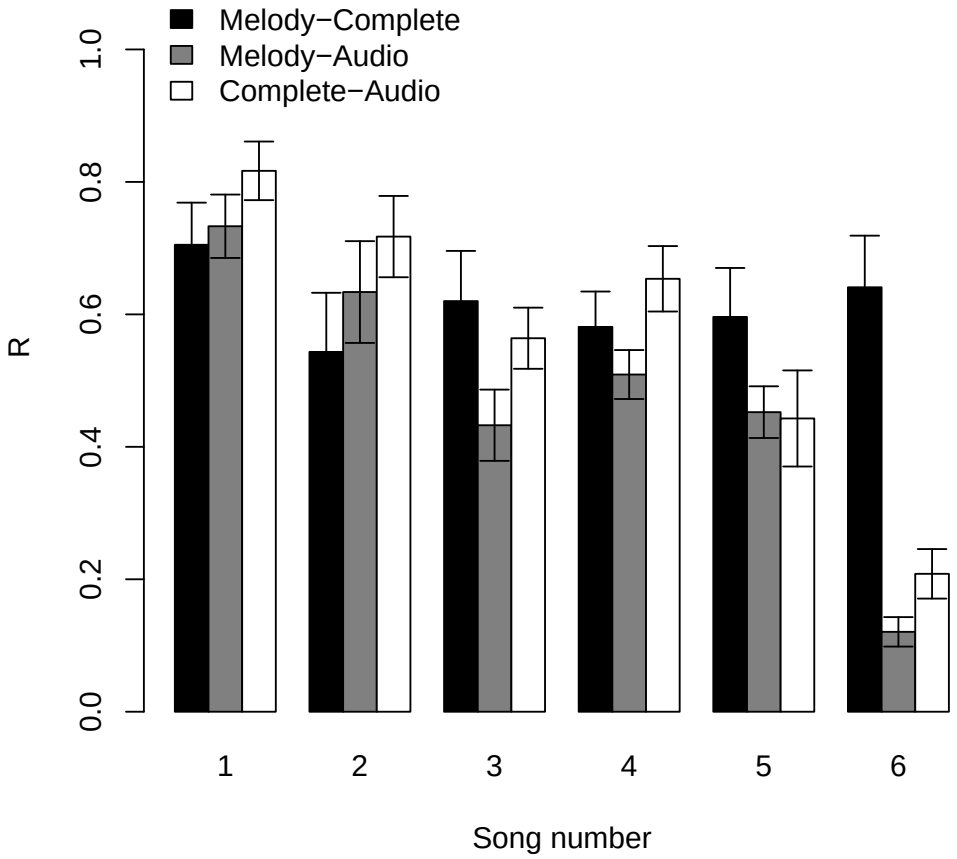


Figure 3.7: Pair-wise correlations between the boundary indication profiles of the different stimulus types across the three stimulus types for the six songs. (1): Heart to hurt, (2): Moondance, (3): Live and let die, (4): And when I die, (5): Body and soul (vocal), (6): Body and soul (instrumental). The mean and the standard error of the mean have been estimated by bootstrapping across individual subject data (500 resamples). The left bar (black) is the result for the MIDI melody stimuli correlated with the MIDI complete stimuli stimuli, the bar in the middle bar (gray) is for the MIDI melody stimuli correlated with the audio stimuli and the right bar (white) is for the MIDI complete stimuli correlated with the audio stimuli.

melody stimuli, 0.85 (σ=0.03) for the MIDI complete stimuli, and 0.83 (σ=0.04) for the audio stimuli. These between group correlations were all higher than the correlations of Figure 3.7 but they are clearly lower than one. Thus, the between group design of the experiment has the consequence that the correlation values depicted in Figure 3.7 are lower than one would expect in a complete within-group design. This result emphasizes the similarity of the boundary profiles across the three stimulus types. This high similarity suggests that the segmentation of polyphonic music could be facilitated by focusing on the melody line of the polyphonic music.

In addition to the boundary indication profiles, we also compared the salience ratings from the explicit salience rating experiment across the three stimulus types. The selected boundaries differed slightly across the three stimuli, but there was substantial overlap. Across the three representation pairs, in total there were 198 boundaries in common, with between 3 and 16 common boundaries per song. The correlation of the mean boundary salience ratings between a stimulus type pair was 0.82 for the MIDI melody and MIDI complete, 0.69 for the MIDI melody and audio, and 0.83 for the MIDI complete and audio. Thus, the MIDI melody and audio were, as expected, less correlated than the other two stimulus type pairs. To investigate how the correlation is distributed across songs, the correlations per song are plotted in Figure 3.8. The figure shows no clear correlation pattern across the six songs. For all songs except "Body and soul (instrumental)"(6), there are at least two comparison pairs that are not significantly different, suggesting that for all three stimulus types there is a certain agreement in the salience ratings. The large errorbars of the songs "Live and let die"(3) and "Body and soul (instrumental)"(6) for the correlation between the two MIDI versions and the audio are most likely related to the fact that they only had few boundaries in common. Overall, the correlation of the salience ratings across the three different stimulus types is high.

We also analyzed whether the same pattern of similarity found between the pair-wise correlation in the number of boundary indications can also be found in the pair-wise correlation of the salience rating. When comparing smoothed boundary profiles across the three representations (Figure 3.7) with the mean salience rating (Figure 3.8), it seems that there are several congruencies. For four songs, "Heart to hurt", "Moondance", "Live and let die", and "Body and soul" (instrumental) the general pattern is similar for both, the boundary profile as well as the salience rating. For the other two songs, "And when I die" and "Body and soul (vocal)" the correlation

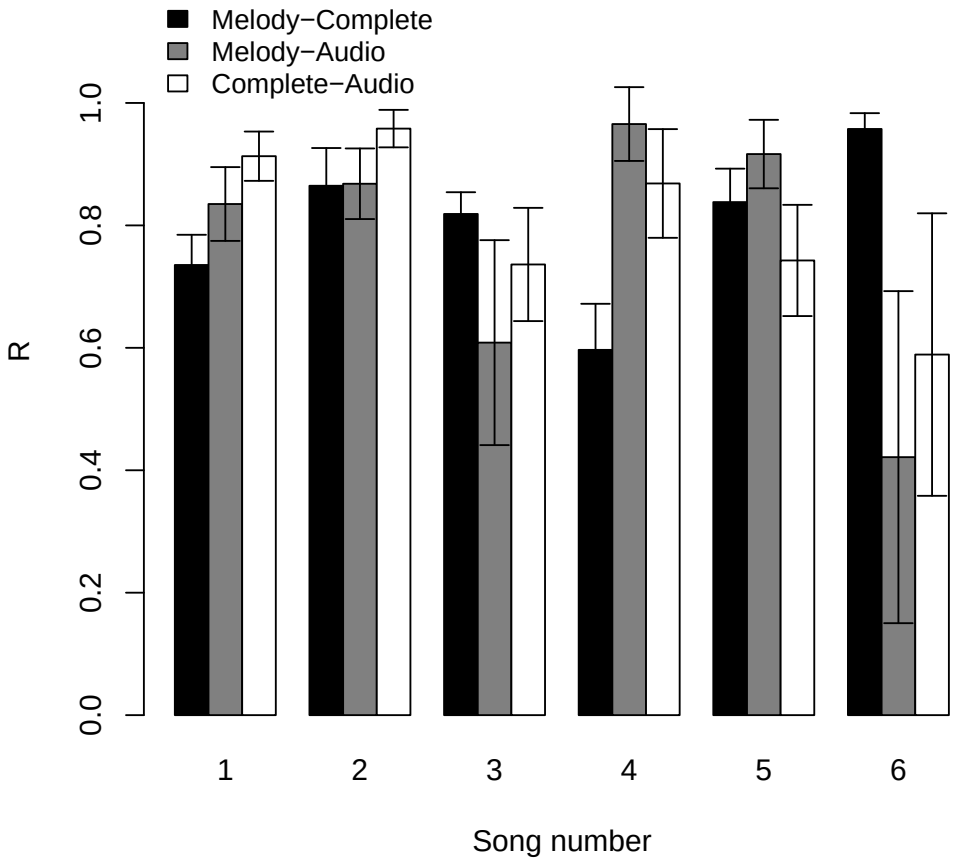


Figure 3.8: Pair-wise correlations of the mean salience ratings of the selected boundaries pair-wise common across stimulus types for the six songs: (1): Heart to hurt, (2): Moondance, (3): Live and let die, (4): And when I die, (5): Body and soul (vocal), (6): Body and soul (instrumental). The standard error of the mean was calculated with bootstrapping of the salience ratings (N = 500). The left bar (black) is the result for the MIDI melody (Mel) correlated with the MIDI complete (Complete) stimuli, the bar in the middle bar (gray) is for the MIDI melody stimuli correlated with the audio (Audio) stimuli and the right bar (white) is for the MIDI complete correlated with the audio stimuli.

of the salience ratings between MIDI melody and audio is much higher compared to the correlation between the boundary profiles of MIDI melody and audio. Overall, however, the pair-wise correlations between the boundary profiles and the salience ratings are fairly similar.

The descriptions of the boundaries were also analyzed for differences between the three stimulus types. We expected subjects to use additional cues for describing the boundaries in the polyphonic stimulus types. The results show that some cues, such as change in timbre (voice), were indeed only found in data from the audio stimulus because those cues only exist in the audio stimulus-type. However, other cues, such as change-in-melody and repetition, were used more often in the two MIDI stimuli, indicating that the audio version emphasized different boundary cues. It is also interesting to notice that the cue harmonic-progression had the strongest mean term rating in the MIDI melody stimulus, thus the cue harmonic progression seems to be more easily perceived for our subjects, who were nonmusicians, in the monophonic version. In the polyphonic versions, subjects made a distinction between different timbre changes. While subjects in the experiment using monophonic stimuli described boundaries with the general term of "changes in timbre", subjects in the experiment using polyphonic stimuli, in addition to the description "changes in timbre", distinguished between drums and vocals. The polyphonic versions, thus, allowed subjects to describe the boundaries with more precise terms or with additional cues only found in polyphonic music. These additional cues, however, did not critically change the segmentation patterns across the different representations.

3.6 General discussion

The aim of the present study was to analyze how subjects perceive segment boundaries in polyphonic songs taken from Western popular music. In the segmentation experiment, subjects were asked to segment six songs each in two different polyphonic forms: (1) a synthesized version based on a score-like (MIDI) representation and (2) an audio recording. In the rating experiment, subjects were asked to rate the salience of selected boundaries taken from the segmentation experiment. We were in particular interested in how the segmentation of the polyphonic stimuli differs from that of monophonic music (Chapter 2). The comparison of stimulus pairs showed that the segmentation pattern and boundary salience ratings are highly correlated across

stimulus types. These results suggest that, if the different stimuli are well time-aligned, there is little influence of polyphony on the perception of segment boundaries, at least for songs taken from Western popular music.

The low influence of the representation on the segmentation pattern also suggests that subjects are able to segment pieces independently of the expressive performance information. Although we tried to align the MIDI stimuli with the audio stimuli, the expressive information of the two are likely to be different. It seems, thus, that the expressive information in the audio recordings underline the general structure, but adds little *new* segment boundaries. Hence, albeit a segmentation algorithm could use the expressive information to segment a piece of music (cf. Chuan & Chew, 2007), the segmentation pattern should be similarly obtained with segmentation algorithms that do not explicitly use expressive information (Cambouropoulos, 2001a; Temperley, 2001; Bod, 2002; Frankland & Cohen, 2004).

The high correlation of the boundary profiles and the salience ratings across stimulus types make it plausible that subjects (who were not professional musicians in the present study) focus indeed on a single line in the music as suggested by Sloboda and Edworthy (1981). Such an attention based segmentation strategy is supported by experiments using functional magnetic resonance imaging (fMRI), which found that the regions in the brain activated by simple and more complex stimuli are very similar (Janata, Tillmann, & Bharucha, 2002) and that phrase boundaries are likely to be associated with memory- and attention-related processes (Knösche et al., 2005). From a practical point of view, however, it is not yet clear how to identify the musical line on which the attention is focused. Extensive research has been done on the segregation of auditory streams (cf., Bregman, 1990) and algorithms have been implemented for automatic music transcription (cf., Klapuri, 2004), as well as voice separation algorithms have been proposed in musicology (Cambouropoulos, 2006b). Nevertheless, although some attempts have been made to segregate audio recordings (e.g., Itoyama, Goto, Komatani, Ogata, & Okuno, 2007), further investigations are necessary to apply these types of algorithms to the segregation of complex audio recordings.

The small influence of the stimulus type on the segmentation profiles is in direct contrast to the result that subjects use quite different cues to describe boundaries for different stimulus types, for instance cues only found in polyphonic representations. It seems, thus, that the additional cues are not of fundamental influence for perceiving the segment boundaries but may emphasize certain segment boundaries. For the purpose

of segmenting Western popular music it would, thus, be enough to have a monophonic representation of the song, which would favor the algorithms proposed in musicology that are mostly based on monophonic line representation (Lerdahl & Jackendoff, 1983; Cambouropoulos, 2001a; Temperley, 2001). The high agreement between the perceptual segmentation profiles across the three representations, however, may also be related to the choice of stimuli used in the present study. The stimuli were selected from Western popular music, which is in our case often close to homophony, i.e., one melody line with accompaniment. If subjects, thus, focus on the melody line, similarly as extracted and used in the previous chapter, the resulting segmentation pattern would be expected to be similar.

One possible application of the findings of this study is the design of automatic segmentation algorithms. The results of the present study suggest that it should, in principle, be possible to reduce the polyphonic version of Western popular music to a single monophonic structure, which includes changes in timbre. How such a reduction should take place is still to be determined (see, for instance, Cambouropoulos (2006b) for an elicitation of the difficulties of voice separation). The monophonic voice could afterwards be processed with musicological models conceived for melodies to segment this single voice. The segment boundaries could then be placed back to the polyphonic music as segment boundaries. If only the global segmentation of the musical piece is needed, such an algorithm as mentioned above could, at least partly, prevent the need for automatic transcription, which is technically challenging. In summary, for building a segmentation algorithm, we would suggest to first identify the most salient voice, then segment this voice with existing segmentation algorithms, and use the obtained segment boundaries as the boundary estimates for the polyphonic piece.

4 Perceptual evaluation of formal musicological cues for automatic song segmentation[‡]

The present study evaluated how well boundaries predicted by nine musicological cues taken from the models LBDM by Cambouropoulos (2001a) and GTTM by Lerdahl and Jackendoff (1983) quantified by Frankland and Cohen (2004) and the additional cue of timbre-change can predict perceptually obtained boundaries. The predicted boundary profiles were correlated with perceptual boundary profiles of six songs obtained in a previous study shown in Chapter 2. The best individual cue was the LBDM-onset. The optimal combination of three cues comprises LBDM-onset, GTTM-rest, and timbre-change, resulting in a physical correlation of 0.80 to 0.89 between perceptual and model boundary profiles. Analysis of the description of the missed salient boundaries suggests that including the cues tempo-change and harmonic progression could improve the model predictions. The optimal cue combination for segmentation profiles of polyphonic *versions of the same songs as obtained in Chapter 3 was the combination of the cues LBDM-onset, timbre-change, and the start of a rest.*

[‡]This chapter is based on Bruderer, M.J., McKinney M.F., and Kohlrausch, A. "Perceptual evaluation of musicological models for automatic song segmentation", to be submitted for publication.

4.1 Introduction

Listening to music not only comprises the aesthetic experience but also the interpretation of the musical form. Based on the structure of this form, the listener is able to segment a musical piece into smaller parts. The segmentation process has been studied perceptually (Deliège, 1987; Clarke & Krumhansl, 1990; Krumhansl, 1996) as well as musicologically (Lerdahl & Jackendoff, 1983) and different types of cues have been identified that convey the structure of music, including change in note duration, breaks, and change in pitch. Some of these cues have been incorporated into musicological models for music segmentation (e.g., Cambouropoulos, 2001a; Temperley, 2001; Frankland & Cohen, 2004). However, it is not yet clear if these models are also apt to predict boundaries in pieces of longer duration or those taken from musical genres other than classical music.

One difficulty with the evaluation of segmentation algorithms is the scarce availability of evaluation material based on music structure perception. The musicological models have not only been conceived by music theorists but have also mainly been evaluated on datasets annotated by music theorists. For instance Cambouropoulos (2001a) tested his model on a set of 52 melodies, where musicians marked manually preferred punctuation positions on a musical score. The indicated boundaries, thus, rather represent where boundaries *should* be instead of where they are actually perceived. Although this problem has been identified (e.g. by Clarke & Krumhansl, 1990), there are still few empirical data available on perceived segment boundaries.

The goal of this study was to compare the boundaries predicted by cues taken from musicological models with perceptual boundary profiles obtained in the previous two chapters. We evaluated the cues taken from two musicological models, which are all based on local discontinuities in the musical surface and are applied to monophonic score notations. These two particular models have been selected as they can easily be segregated into their different cues, which is, for instance, not straightforward for the model of Temperley (2001). As one of the two models is based on a "Generative Theory of Tonal Music" (GTTM) by Lerdahl and Jackendoff (1983), we first introduce the organization of GTTM.

GTTM is a theory based on four different components: metrical structure, grouping structure, and two types of analytic pitch reductions. The theory defines two types

of rules, well-formedness rules, which define allowed structures, and preference rules, which define the favored structural analysis. For our purpose, the most interesting part of the theory is the grouping structure component, and in particular the grouping preference rules (GPR), comprising a set of seven rules. The GPR rules define plausible segment boundaries of a piece of music.

As GTTM is not a formal theory, GTTM needs an expert to apply the rules to the musical score, using GPR to find segment boundaries. In addition, the theory does not specify how to combine the different rules. Therefore, Frankland and Cohen (2004) quantified four rules from GPR into a form and implemented them in a segmentation algorithm. The quantified rules were rest, attack-point (long note in between two short notes), register change, and length change. They also suggested that in principle it should be possible to combine the outcome of the four quantified rules, but did not give weights to each rule. Furthermore, they also evaluated these quantified rules with the perceptual parsing of six short melodies and found that only two of the rules, attack-point and to a lesser extent rest, were necessary to predict the perceptual boundary profile.

The "Local Boundary Detection Model" (LBDM) by Cambouropoulos (2001a) uses two general rules to define local boundaries, which can be applied to different entities such as change in length, rest, or pitch intervals. The first rule sets a boundary when two successive intervals of an entity are different and the second rule specifies that the boundary is placed on the larger of the two difference-intervals. The model also suggests the assignment of a degree-of-change to each boundary, thus a global boundary profile can be extracted with boundaries having different strengths. The implementation used here applied three cues for evaluating the boundary profile: pitch intervals, inter-onset intervals, and rests (Eerola & Toiviainen, 2004). In the present implementation the model thus uses the temporal and pitch information of the notes in the MIDI file to predict segment boundaries.

4.2 Method

4.2.1 Material and general evaluation method

To evaluate the models, the following cues were extracted from monophonic MIDI representations of the same six songs as used in the second chapter: four cues of GTTM (Lerdahl & Jackendoff, 1983) quantified by Frankland and Cohen (2004) and

three cues of LBDM by Cambouropoulos (2001a). In addition to these cues, we added the cue timbre-change. The exact definitions of these cues are given in Appendix C. The Rest rule of GTTM was quantified in two versions, one where the boundary is set at the beginning of the rest (RestOn) and one where the boundary is set at the end of the rest (RestOff). Figure 4.1 shows, as an example, all cues extracted from a single song along with the perceptual boundary profile obtained from the study described in Chapter 2. The figure shows that the number of extracted boundaries varies considerably across cues. It can also be seen that at several places, boundaries predicted by different cues coincide. In total, nine profiles of predicted boundaries were evaluated: RestOn, RestOff, Attack-point, Register (change), Length (change), LBDM-rest, LBDM-onset, LBDM-pitch, and Timbre (change).

Segment boundary predictions based on each cue were compared with the perceptual segment boundary profiles obtained in Chapter 2 using the same monophonic stimuli. In the previous study subjects were asked to indicate segment boundaries by pressing a key on the computer keyboard while listening to the musical piece. To obtain a global boundary profile, the boundary indications were collapsed across subjects and the frequency of indications within a 1.25 s window was taken as the boundary's salience.

Our basic measure to compare the performance of two different cues was the cross-correlation of their corresponding boundary predictions. Instead of using the precise time position of each boundary, the boundary indications were smoothed to account for possible time shifts between the predicted and perceptual boundary as well as to compensate for the scattering of subjects' boundary indications (cf. Chapter 2). To smoothen the boundary indications, they were collapsed into a pulse train, which was then convolved with a Gaussian window. The full width at half maximum of the Gaussian window was set to 1.25 s which has been shown to be the optimal window size for operations of this type on our data (cf. Chapter 2). The convolved pulse train is referred to as *smoothed boundary profile*. The smoothing process was applied to both the boundaries predicted by the models as well as the perceptual boundaries.

To compare the model predictions and the perceptual boundaries the correlation between the two smoothed profiles was calculated. However, instead of the common Pearson correlation, shown in Equation 4.1, here we used the physical correlation, shown in Equation 4.2. The reason for this choice was that the boundary predictions were all positive. The Pearson correlation removes the mean, which is not sensible

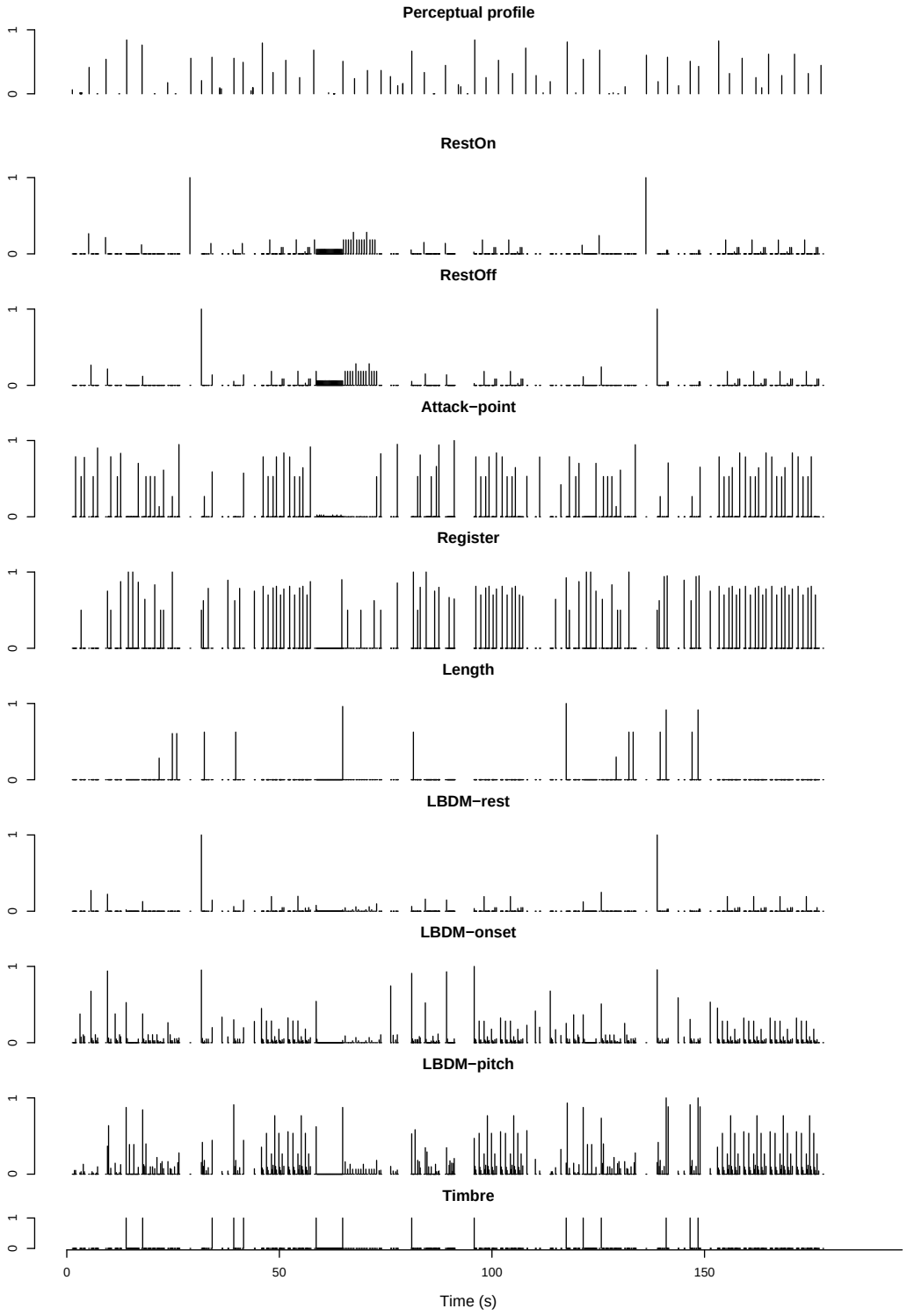


Figure 4.1: An example showing the extracted cues from the song "Live and let die". The first row shows the perceptual boundaries. The other nine rows each show the boundary predictions for one individual cue extracted from the MIDI melody representation of the song. See Appendix C for definition of the cues.

when dealing with nonnegative temporal patterns. Therefore the physical correlation was chosen.

$$\text{Pearson correlation: } r = \frac{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}, \tag{4.1}$$

where r is the Pearson correlation, x_i and y_i are the two smoothed boundary profiles at instance i, and $\bar{x}$ and $\bar{y}$ are the means of x and y, respectively.

$$\text{Physical correlation: } c = \frac{\sum_i x_i^2 y_i^2}{\sqrt{\sum_i x_i^2 \sum_i y_i^2}}, \tag{4.2}$$

where c is the correlation and x_i and y_i are the two smoothed boundary profiles at instance i.

The difference between these two correlation measures can be illustrated with a sinusoidal waveform correlated with a cosine waveform, but both having a positive offset. The Pearson correlation would be zero, as the offset is removed before calculating the correlation, while the physical correlation would still be positive.

4.3 Comparing individual cues with each other

Several of the extracted cues are based on the same underlying properties of the note duration, breaks (offset-to-onset interval), and the pitch intervals, but each implemented in a slightly different way. To analyze whether these similarities between the different cue definitions are also reflected in their boundary predictions, the boundary profile of each cue can be compared with those from other cues. It would, for instance, be expected that the three cues based on rest are highly correlated with each other.

If the boundary predictions of two cues have a low correlation (between them), this means that they predict different boundaries and that they are therefore orthogonal to each other. Orthogonality is a desired property as it allows the separation of the contribution of each individual cue. To study the relationship between the cues, the correlation was calculated between all possible pairs of the predicted boundary profiles for each cue. Figure 4.2 shows the correlations of each cues' boundary profile with each other cue averaged across the six songs. The figure shows that several implemented cues are indeed highly correlated with each other. The boundary profile of LBDM-pitch is highly correlated with the boundary profile predicted by Register, both likely predicting

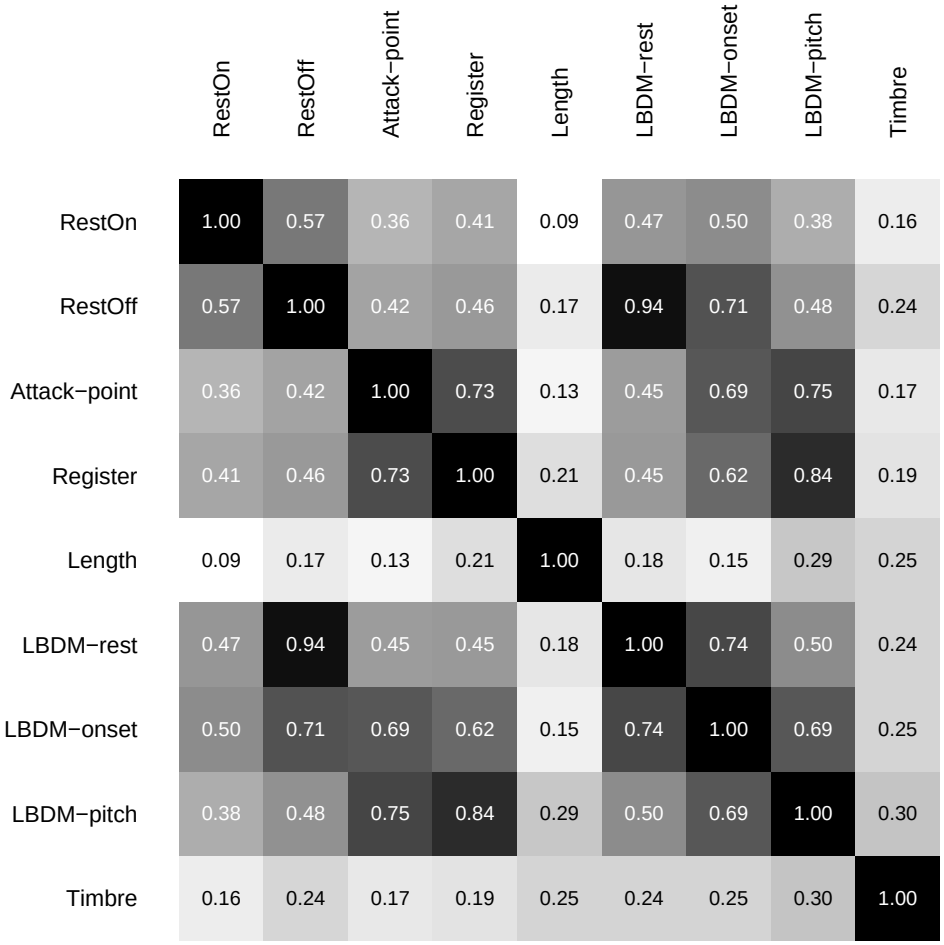


Figure 4.2: Pair-wise correlations of predicted boundary profiles based on single cues. The first five columns are the quantified rules of GTTM by Frankland & Cohen (2004) (RestOn and RestOff, attack-point, length (change), register (change)). The columns LBDM-pitch, LBDM-onset, and LBDM-rest describe cues from the model of Cambouropoulos (2001). The Timbre cue is our implementation of timbre-change. The gray code represents the correlation, where a dark color shows a high correlation between cue pairs and a light color a low correlation.

a boundary if there is a large pitch interval. The predicted boundary profiles of RestOff and LBDM-rest are also highly correlated with each other. It is also interesting to notice that attack-point has a relatively high correlation with LBDM-pitch as the definition of attack-point is based on difference in the durations of consecutive notes while LBDM-pitch is based on pitch intervals. Overall, these results suggest that, although the cues were implemented in different ways, some cues still result in similar boundary predictions.

The most striking observation in Figure 4.2 is that two cues have a very low correlation with the other cues: Length and Timbre. The low correlation of Length is probably caused by its strict definition, which results in relatively few predicted boundaries. The reason for the low correlation of Timbre is also probably related to the fact that this cue leads to only a few predicted boundaries. It seems, thus, that these two cues are orthogonal to the other cues. Whether these two cues, as well as the other cues, predict salient perceptual segment boundaries will be evaluated in the following.

4.4 Performance of individual cues

The main aim of this study was to evaluate how well boundaries from theoretical cues agree with perceptual boundaries. The goal was to figure out which cues and which combinations of cues give the best prediction, and to see how many of the most salient boundaries can be predicted. In the following three analyses are described: 1) First, we evaluated the performance of the individual cues by analyzing the influence of a possible time delay between the predicted and perceptual boundaries. 2) Second, we studied the optimal cue combination. 3) Finally, we tested how well the models can predict the most salient perceptual boundaries.

In order to evaluate the performance, the smoothed boundary profile of each individual cue was first correlated with the smoothed perceptual boundary profile without any time alignment. The performance of each cue is shown by the left-most (black) bars of Figure 4.3. Several cues are more highly correlated with the perceptual profile than others, such as LBDM-onset, which shows a correlation of greater than 0.8. The two cues Length and Timbre, on the other hand, have the lowest correlation with the perceptual boundaries. Overall the boundary profiles of most cues are

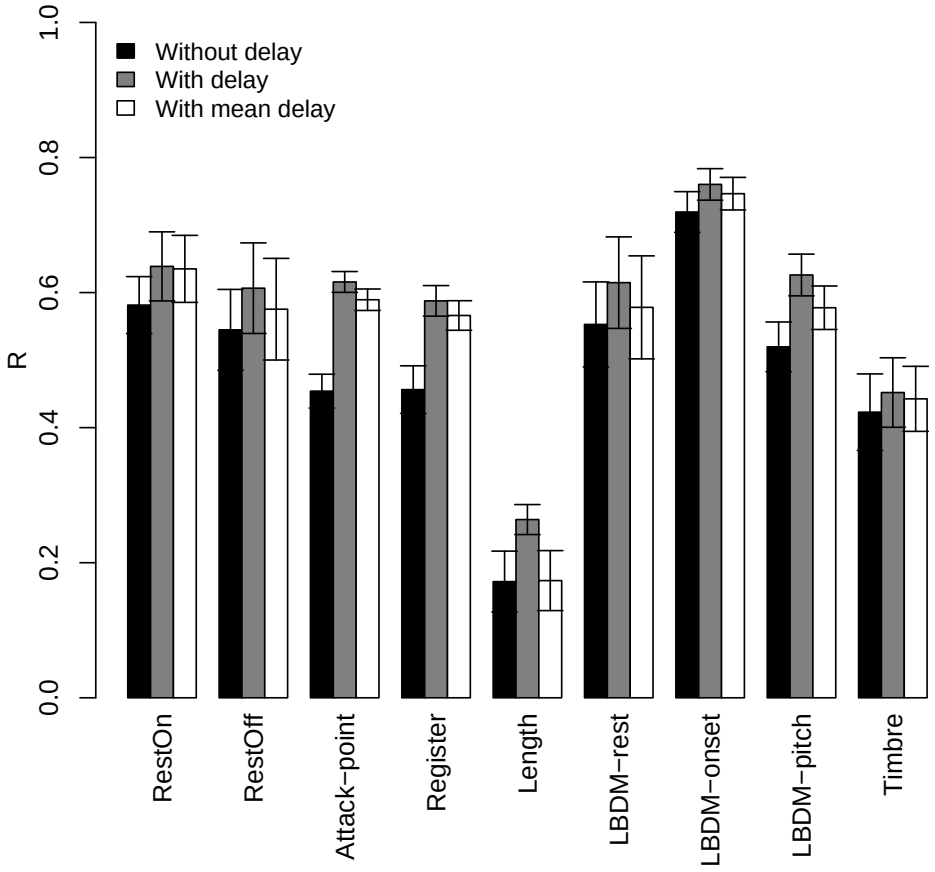


Figure 4.3: The correlation and the standard error of the mean of the individual cues with the perceptual boundaries for three different delays: no delay, the optimal delay, the mean delay across the six songs.

moderately correlated with the perceptual boundary profiles. The best performing cue is LBDM-onset.

4.4.1 Delay compensation

Our performance measure of correlation is sensitive to the precise time alignment of the two boundary profiles in question. If the predicted and perceptual boundary profiles are not well time-aligned the resulting correlation would be lower than when the two are precisely time-aligned. To investigate the existence of possible time-shifts in the profiles, we calculated the cross-correlation function (i.e., the correlation as a function of a delay between the two profiles) between the predicted boundary profile with the perceptual profile of each cue.

There are two possible sources for a time-shift in the boundary profiles. In the perceptual boundary profiles, it is possible that the boundary indications are delayed due to the subject's reaction time, while in the model predictions, there could be a systematic offset due to the algorithmic formulations. In the perceptual experiment, one would predict that, when the boundary is unexpected, subjects indicate it *after* the boundary has occurred. However, when the boundary is anticipated, it could also be indicated *before* it actually appears in the music. In that case it should be closer to the actual position than an unexpected boundary. The other possible source for a delay are imprecise boundary predictions. The models, for instance, define to which note a boundary can be assigned, but they do not specify if the boundary should be placed at the beginning or the end of a note. For both sources, we would expect the delay to be relatively consistent across the six songs for each individual cue. For different cues, however, the delay could vary, either because the model formulations are different or because different cues are processed at different stages in the auditory system. If systematic delays exist, the compensation of the mean delay across songs should lead to a higher correlation between the model predictions and the perceptual boundary profiles.

To evaluate the influence of the delay, the predicted boundaries were cross-correlated with the perceptual boundaries using the physical correlation. The definition of the cross-correlation function is given for continuous functions f and g in Equation 4.3 and for the discrete functions f_i and g_i in Equation 4.4, where f_i and g_i are the normalized boundary profiles. Because the here-used physical correlation function is not changed by adding of zeros at the end of the two vectors, f_i and g_i were zero padded up to the maximal delay.

$$(f * g)(\tau) = \int_t f(t)g(\tau + t)dt \tag{4.3}$$

$$(f * g)_i = \sum_j f_j g_{i+j}. \tag{4.4}$$

Different maximal delays in the cross-correlation function were tested and a maximal delay of ± 5 seconds was found to be sufficient. For each cue, the maximum in the cross-correlation function was identified for each song, and the corresponding delay will be called optimal delay. As an example, the cross-correlation functions of the perceptual boundary profiles for the six songs with the profiles of the cues LBDM-onset and RestOn are shown in Figure 4.4. The figure shows that, in general, there is a clear peak in the cross-correlation function. The figure also shows that a time shift can

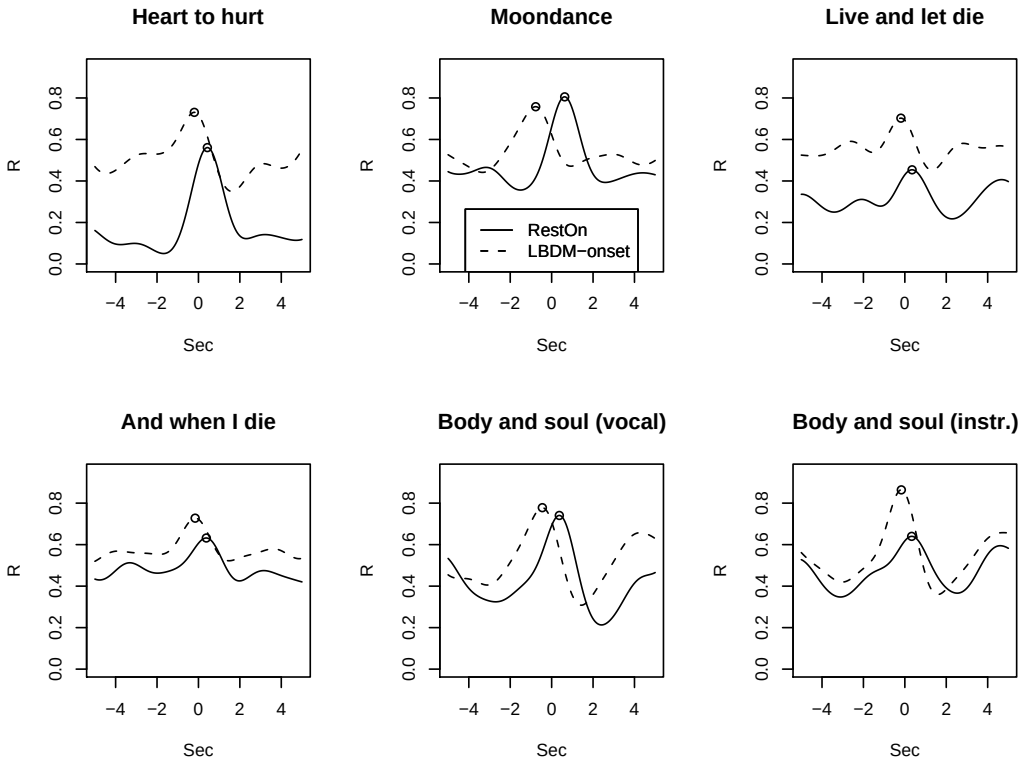


Figure 4.4: Cross-correlation functions of the boundary profiles of two cues, RestOn and LBDM-onset, and the perceptual boundary profiles across the six songs. The abscissa shows the time-shift and the ordinate shows the correlation with the perceptual boundary profile. The cross-correlation was used to estimate a possible delay between model predictions and perceptual boundaries. The circles in the graph denote the optimal delay for the combination of song and cue.

increase or decrease the performance of individual cues considerably.

To estimate a mean delay across songs, the cross-correlation profiles were summed across songs. The maximum in the summed profile was taken as the mean delay and is given in Table 4.1. The most consistent cues were RestOn, and Timbre. The three definitions of rest, Length, LBDM-onset, and Timbre had a relatively low absolute mean delay (below 1 s), suggesting that for these cues there is only a small delay between the predicted boundaries and the perceptual boundaries.

For each cue we obtained two different delays, the optimal delay per song and the mean delay across the six songs. It is interesting to analyze how the use of these delays increases the performance of the individual cues. In order to evaluate the performance, the smoothed boundary profile of each individual cue was correlated with the smoothed perceptual boundary profile, where the cue profiles were shifted with two different

Table 4.1: The mean delay of each cue in s with the standard error of the mean (in parenthesis). The standard error of the mean was estimated by the optimal delays across the six songs. A negative value indicates that the predicted cue profile needs to be advanced in time to obtain a better performance.

	RestOn	RestOff	Attack-point	Register	Length
Mean delay	0.44 (0.045)	-0.37 (0.37)	2.53 (1.1)	-1.73 (1.0)	-0.17 (0.95)

	LBDM-rest	LBDM-onset	LBDM-pitch	Timbre
Mean delay	-0.32 (0.38)	-0.31 (0.1)	-2.26 (1.1)	0.26 (0.071)

delays: With the optimal delay compensation per song and per cue as well as with the mean delay compensation across songs for each cue. The performance of each cue with these two delay compensations is also shown in Figure 4.3 (gray and white bars). For Attack-point and Register, both, the optimal and mean delay compensation significantly increased the performance. For the cues Length and LBDM-pitch the performance with the optimal delay was significantly higher than without the delay. For the other cues there is little influence of the delay on the performance. The compensation with the optimal delay per song for each individual cue reached by definition the highest correlation. However, the application of the mean delay also improved the performance of each cue compared with no delay compensation. In the following, therefore, the mean delay was applied to the boundary profile of each cue before further analyses are performed.

4.5 Combination of cues

A further goal of this study was to find which *combination* of cues can best predict the perceptual boundaries. Previous studies, in fact, often proposed the combination of different cues. Frankland and Cohen (2004), for instance, suggest a linear combination of four cues from GTTM: Rest, Attack-point, Length, and Register. Cambouropoulos' (2001) LBDM suggests a cue combination where LBDM-onset receives twice the weight of the other two cues, LBDM-rest and LBDM-pitch. It is, thus, likely that a combination of different cues can improve the boundary predictions.

There are several possible ways to test which cues and which combinations of cues give the most accurate boundary prediction. The simplest way is to test every possible combination of cues against the perceptual boundary profile. This approach, however, is of complexity N^2 and is thus computationally expensive with an increasing number

of cues. A more efficient way is to start with a combination of all cues and to remove individual cues depending on their influence on the performance. Alternatively, it would also be possible to start with a single cue, test the model with each additional cue, and add the cue which improves the model performance most. This process is then repeated until all cues are included in the combination. However, the subtraction method has the advantage that cue combinations that give a good performance are held in the process of eliminating the cue with the least influence.

The subtraction algorithm starts by taking all cues and then successively eliminates the cue that has the least influence on the mean performance across all songs. The performance results during the successive elimination of cues are shown in Figure 4.5. The figure shows that the performance stays relatively constant during the removal of the first seven cues and only starts to decrease when removing one of the last three cues. The final remaining cue is LBDM-onset. Thus, the best combination of three cues across songs comprises RestOn, LBDM-onset, and Timbre. The combination of these three cues yields a correlation with the perceptual boundary profile of between 0.80 to 0.89 for the six songs ($\mu = 0.85$, $\sigma = 0.0051$). An example of the perceptual profile along with the predicted profile obtained using these three cues is shown in Figure 4.6. The figure shows that many of the high boundaries were correctly identified, however, often with different salience (heights) than in the perceptual profile.

We estimated the contribution of each of these cues to the global prediction of the perceptual boundary profile. To do this, we calculated the optimal linear combination of the three cues, RestOn, LBDM-onset, and Timbre, and concatenated the boundary profiles from each song into one long profile. The same concatenation was applied to the predicted boundary profiles. Then, the optimal linear combination of the three cues, RestOn, LBDM-onset, and Timbre was calculated with linear regression. The optimal combination was 0.29 × Timbre + 0.25 × LBDM-onset + 0.29 × RestOn, thus the influence of each of the three cues was nearly identical.

It is also interesting to analyze whether the choice of the window size used to smoothen the boundary profiles has a strong influence on the best combination of cues. We, therefore, tested the combination of the best three cues with the elimination method described above for different window sizes. Furthermore, we calculated the performance of the concatenated boundary profiles with the three best cues for each window size. The best three cues and the correlation between the predicted boundary profiles with the perceptual boundary profiles are shown in Table 4.2. The table shows that with all different window sizes the three optimal cues are the same, Timbre,

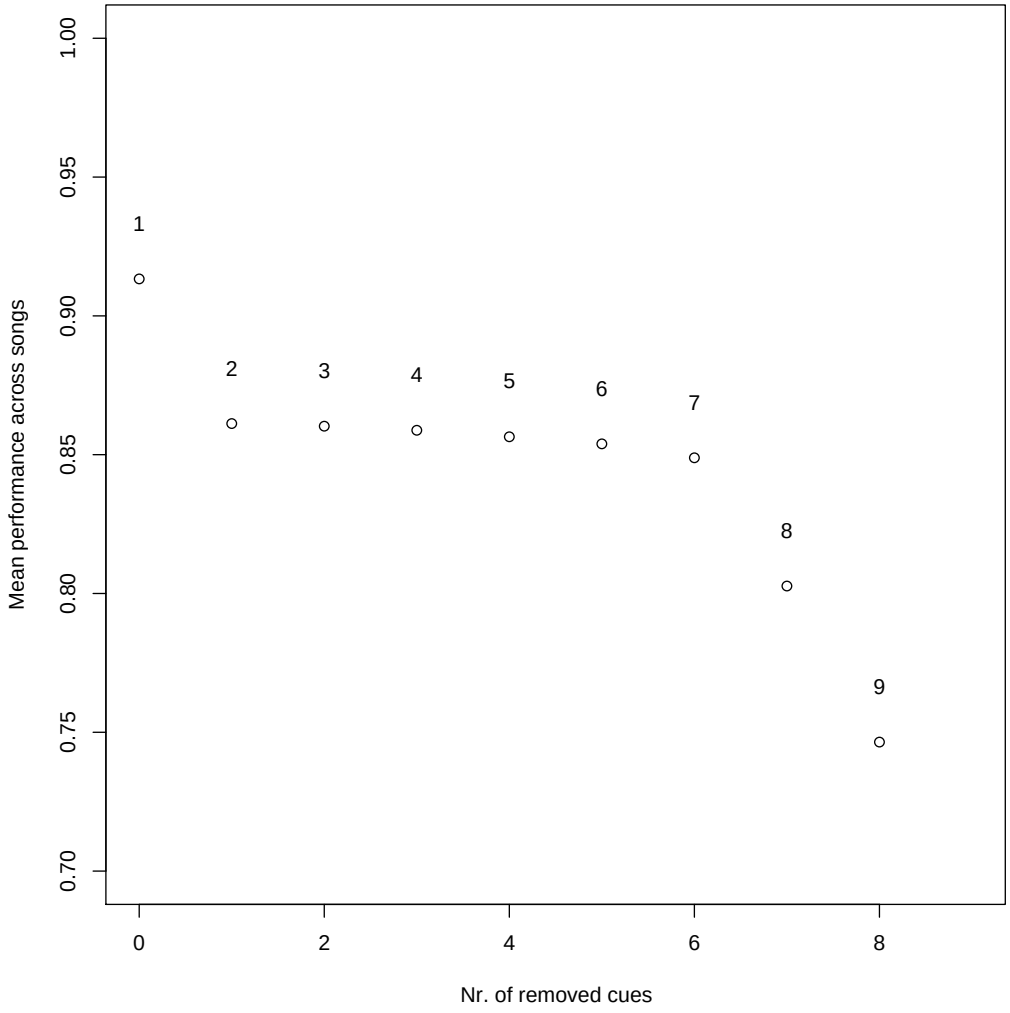


Figure 4.5: The performance of a combination of cues with the successive removal of the least important cue. The figure shows the mean performance across songs, with the removed cue written above the performance with the remaining cues. The number indicate which cue has been left out. 1: none, 2: attack-point, 3: Length, 4: LBDM-pitch, 5: LBDM-rest, 6: Register, 7: RestOff, 8: RestOn, 9: Timbre. The final cue is LBDM-onset. The best three cues are, thus, LBDM-onset, Timbre, and RestOn reaching a correlation of 0.85.

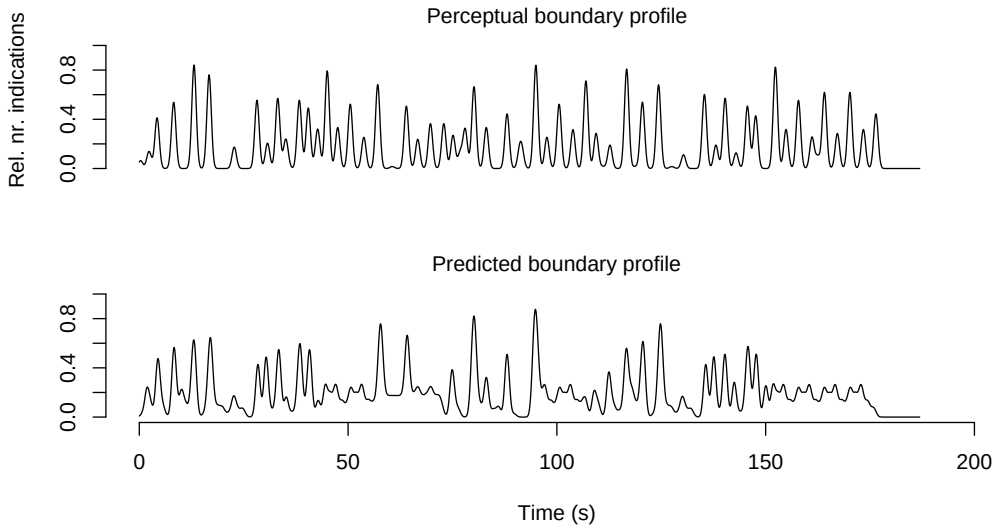


Figure 4.6: An example of the boundary profile predicted by the best combination of the three cues RestOn, LBDM-onset, and Timbre for the song "Live and let die". The figure at the top shows the normalized smoothed boundary profile of the perceptual boundaries. The bottom figure shows the smoothed boundary profile of the linear combination of the three cues fitted to the normalized perceptual boundary profile.

Table 4.2: The optimal combination of cues estimated with the elimination process (first three rows) and the performance, i.e., correlation of the predicted boundary profile of these three cues with the perceptual boundary profiles.

Window size	0.5	1.0	1.5	2.0
2nd last elimination	Timbre	Timbre	RestOn	RestOn
Last elimination	LBDM-onset	RestOn	Timbre	Timbre
Remaining cue	RestOn	LBDM-onset	LBDM-onset	LBDM-onset
Performance	0.75	0.84	0.87	0.86

LBDM-onset, and RestOn, but in a slightly different order depending on the window size. The last row of the table shows that the correlation between the predicted and perceptual boundary profiles is lower with a shorter window size. These results suggest that, although the correlation is slightly different, the window size does not critically influence the performance of the three optimal cues. It also seems that the general optimal combination of cues, i.e., timbre-changes, LBDM-onset, and start of a rest, is independent of the window size.

The correlation performance measure we have used so far may have been overloaded with the many boundary indications and predictions at low saliences. For practical applications, it may be more useful to examine only the most salient boundaries, and thus, in the following we analyze the prediction of boundaries with a high salience.

4.6 Thresholding the perceptual and predicted boundaries

The results of the perceptual experiment showed that certain boundaries were indicated by more subjects than other boundaries, thus that different boundaries have different salience values (cf. Chapter 2 and 3). We were interested whether the most salient boundaries of the predicted boundary profiles correspond to the most salient boundaries of the perceptual profile.

Two methods were used to analyze how well the models can predict salient boundaries, one using a correlation measure, the other using a modified measure for precision and recall. The first measure is a correlation of the boundary profile of the most salient perceptual boundaries with the profile of the most salient boundaries of each cue. Because the salience measure for the boundaries is normalized to one, we could choose the most salient boundaries by setting a threshold criterion. For both, the model and the perceptual boundaries, we calculated the correlation with different

thresholds, each ranging from 0 (no threshold) to 0.9 (all boundaries higher than 0.9). The results show that for all cues, except Timbre, the correlation between the model prediction profile and the perceptual boundary profile is highest when the threshold is set to 0, i.e., when the analysis includes all boundaries. This result may be related to the fact that it is more difficult to reach a high correlation with profiles comprising only a few boundaries. For the cue Timbre the thresholding of the model predictions had no influence because this cues is binary in its boundary prediction – it does not attribute a salience value to each boundary. For Timbre, the highest correlation is reached when all boundaries above 0.8 of the perceptual boundaries are considered, suggesting that changes in timbre are associated with salient segment boundaries. We also applied the thresholding to the optimal combination of cues, RestOn, Timbre, and LBDM-onset. Figure 4.7 shows the mean correlation between the smoothed boundary profiles of this optimal cue combination and perceptual boundaries across the six songs. Similar to the results obtained for most individual cues, the correlation between the smoothed boundary profile of the predicted boundaries and the perceptual boundaries is highest when no thresholding is applied.

A second method, more orientated on precision and recall measures, was used to evaluate the prediction of the most salient boundaries. From the boundary profiles we extracted the X highest perceptual boundaries and the same number of most salient boundaries from the model predictions. We then compared the selected boundaries for coincidences to calculate the precision measures.

$$\text{Relative number of correct boundaries} = \frac{\text{correct boundaries}}{\text{selected boundaries}}, \tag{4.5}$$

where the selected boundaries are the X highest perceptual boundaries. Because we selected the same number of boundaries from the perceptual and the predicted boundary profiles, the precision is equal to the recall. If, for example, from a selection of ten boundaries eight are correctly predicted within a 1.25 s window and two are missed, the relative portion of correctly predicted boundaries of the model would be 0.8. The relative correct boundary coefficients for the six songs using the boundary predictions of the combination of the three best cues Timbre, LBDM-onset, and RestOn is shown in Table 4.3. The table shows that the number of correctly predicted boundaries varies strongly across different songs. For instance, the model is good at predicting the most salient perceptual boundaries in the song “Heart to hurt”, but fails in the

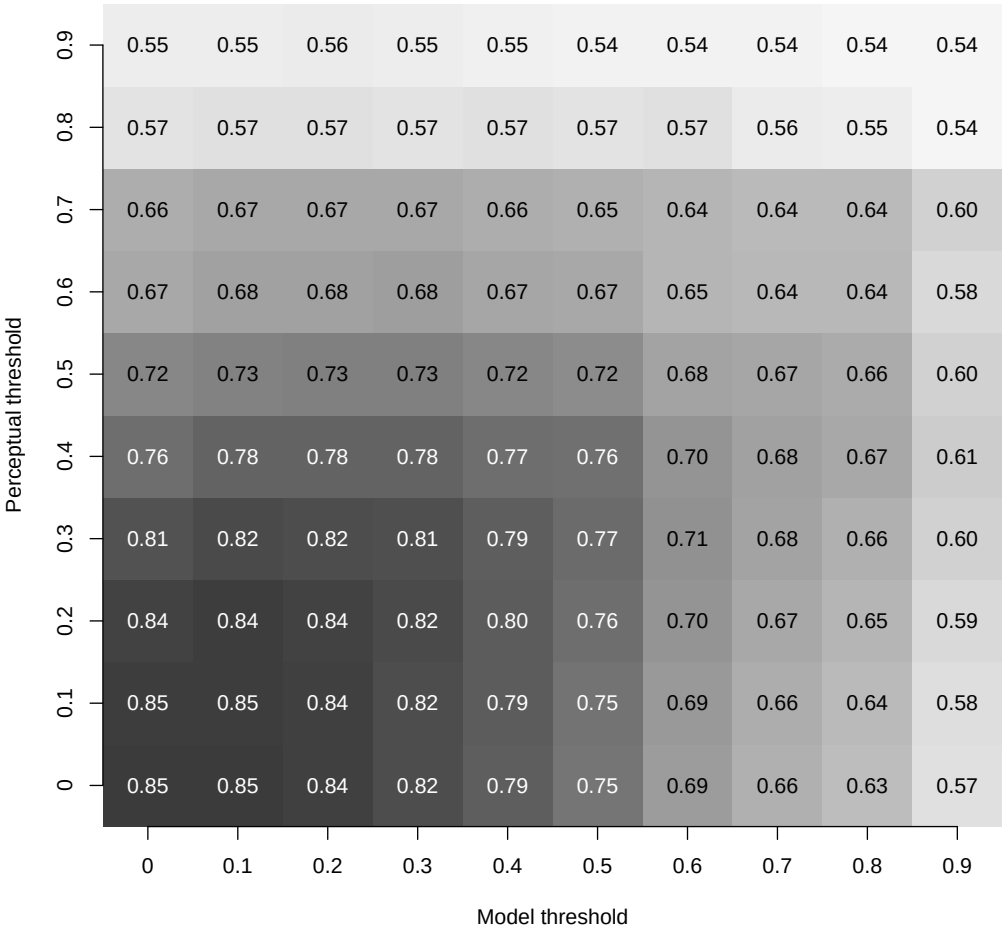


Figure 4.7: The correlation between the thresholded model prediction profile and the thresholded perceptual boundary profiles for the best linear combination of the cues RestOn, Timbre, and LBDM-onset.

Table 4.3: The relative number of correctly predicted boundaries, where only the boundaries with a salience in the highest 20th to 35th percentile are taken (with the number of boundaries in parenthesis).

Percentile	Heart to hurt	Moondance	Live and let die	And when I die	Body and soul (v)	Body and soul (i)
0.8	0.62 (13)	0.20 (15)	0.45 (11)	0.61 (18)	0.70 (10)	0.45 (11)
0.7	0.80 (20)	0.32 (22)	0.53 (17)	0.67 (27)	0.67 (15)	0.82 (17)
0.65	0.87 (23)	0.38 (26)	0.53 (19)	0.71 (31)	0.70 (17)	0.75 (20)

song "Moondance". For the other songs it can predict between half to two-thirds of the most salient perceptual boundaries with a temporal accuracy of 1.25 s.

The cues extracted by the musicological models evaluated in the present study can be related to the "cue classes" developed in a previous study (Chapter 2 and 3). In this earlier study, subjects, who were not professional musicians, were asked to describe the boundary cues of the most salient boundaries (as well as for some less often indicated boundaries). The descriptions given by the subjects were then classified into "cue classes", which consisted of the classes harmonic progressions, melody changes, tempo changes, rhythm changes, timbre changes, level changes, breaks, and two more complex descriptions, which we called here global structure and repetition. In the following, we will try to relate the cues used in the models to these "cue classes" to see which of the boundary descriptions are covered by the extracted musicological cues. Rest could be translated to the "cue classes" level-change or break. Timbre would correspond to the "cue classes" timbre-change, but also to global structure because a change in timbre coincides in the stimuli with the start and end of the melody. The cue LBDM-onset has no directly corresponding cue, the "cue class" that would probably best represent it is melody-change. Several other "cue classes" were used for describing boundary cues that have no corresponding cue in the models evaluated here. For instance, the models do not extract repetition or changes in rhythmic cues (tempo- or rhythm-change). Overall, the evaluated models do, thus, only partly cover the cues used by subjects to describe the boundaries.

One reason for the difficulties in predicting salient boundaries could be that the "cue classes" not covered by the model cues are important for perceiving the segment boundaries. Since we obtained the description of the cues of the most salient boundaries in Chapter 2, we can study the cue descriptions of the boundaries that the model could

not predict. We, therefore, counted for each cue in the "cue classes" the number of times the cue was used in the descriptions of the predicted boundaries and compared it with the number of times the cue was used for a *not* predicted boundary. To compare the two, we normalized the number of times the cue was mentioned by the number of found/not-found boundaries to have a frequency of each term per boundary found/not-found. The most interesting cues are those that have a high difference in the relative frequency of cue descriptions. We expected, for instance, that the cues that are extracted by the tested models are more often used as boundary descriptions of the boundary hits than of the missed boundaries. We, therefore, analyzed the boundaries that were above the 65th percentile, as for these boundaries the mean correct prediction was highest. The cue classes that changed more than twice in their frequency were tempo change and change in timbre (other). The values for the cue timbre are 1.42 and 8.36 for the not predicted and the predicted boundaries, respectively, and the values for tempo change are 1.27 for the boundaries not predicted and 0.36 for the predicted boundaries. Change in tempo was not extracted by our evaluated cues. In contrast, the cue Timbre is much more associated with predicted boundaries, and it is included in the model cues. It seems, thus, that especially the cue timbre-change is an easily perceived cue that is also well predicted by the model. Furthermore, the model performance could most likely be improved by including algorithms that extract tempo changes.

4.7 Comparison of perceptual profiles from monophonic and polyphonic music

So far we have only examined the models' abilities to predict perceptual segment boundaries in monophonic melodic stimuli. It would be interesting to see how well model predictions of segment boundaries in melodic lines translate to perceptual segment boundaries in full polyphonic representations of the same songs. In the third chapter, we compared the perceptual segmentation patterns of three different stimulus types with each other: two stimuli synthesized from MIDI (MIDI melody, which is a monophonic representation of the song, and the complete polyphonic MIDI), and the polyphonic audio recording. The results showed that the stimulus type has relatively little influence on the perceptual segment boundaries when the different stimulus types are well time-aligned. Thus, a model operating on a melody line should be able to predict the perceptual segment boundaries in a full polyphonic song. To measure how

Table 4.4: The mean delay (in s) of each cue for the two polyphonic representations MIDI complete (MIDI) and audio. A negative value indicates that the predicted cue profile needs to be advanced in time to obtain a better performance. The upper table shows the cues from the GTTM model (RestOn, RestOff, Attack-point (AP), Register, and Length). The lower table shows the three rules from LBDM (LBDM-rest, LBDM-onset, LBDM-pitch), and Timbre.

	RestOn	RestOff	AP	Register	Length
Mean delay (MIDI)	0.77	0.21	-1.75	-1.56	0.25
Mean delay (Audio)	0.25	-0.59	-2.43	4.21	0.76

	LBDM-rest	LBDM-onset	LBDM-pitch	Timbre
Mean delay (MIDI)	0.24	-0.01	4.56	0.50
Mean delay (Audio)	-0.54	-0.43	3.75	0.13

well the models predict the boundaries in these other stimulus types, we correlated each individual cue with the perceptual boundaries obtained for the three different stimulus types. For each representation the corresponding mean delay was applied, which is listed for MIDI melody in Table 4.1 and for the two polyphonic representations in Table 4.4. The tables show that, except for the two cues Register and LBDM-pitch, the mean delays are relatively similar for all cues. The correlation between the perceptual boundary profiles and the predicted boundary profiles across representations are shown for each cue in Figure 4.8. The figure shows that the predicted boundaries have the highest correlation with the boundaries obtained using the MIDI melody stimuli, which was expected, as the features are based on this representation. The correlation between the predicted boundary profiles and the two polyphonic representations is generally lower, but, except for the three cues Attack-point, Register, and LBDM-pitch not significantly different for the MIDI complete and audio representation.

To further analyze the model's ability to predict perceptual boundaries in polyphonic stimuli, we also compared the prediction of combinations of cues with the three perceptual boundary profiles. The correlation for a combination of the three cues Timbre, LBDM-onset, and RestOn averaged across the six songs was 0.85 $\pm$ 0.0051 for MIDI melody, 0.70 $\pm$ 0.011 for MIDI complete, and 0.69 $\pm$ 0.012 for the audio representation. We also tested which combination of cues resulted in the best performance for each stimulus type using the successive subtraction method. The best combination were again for all three representations LBDM-onset, Timbre, and RestOn. To calculate the optimal weights for these cue combinations, we concatenated the

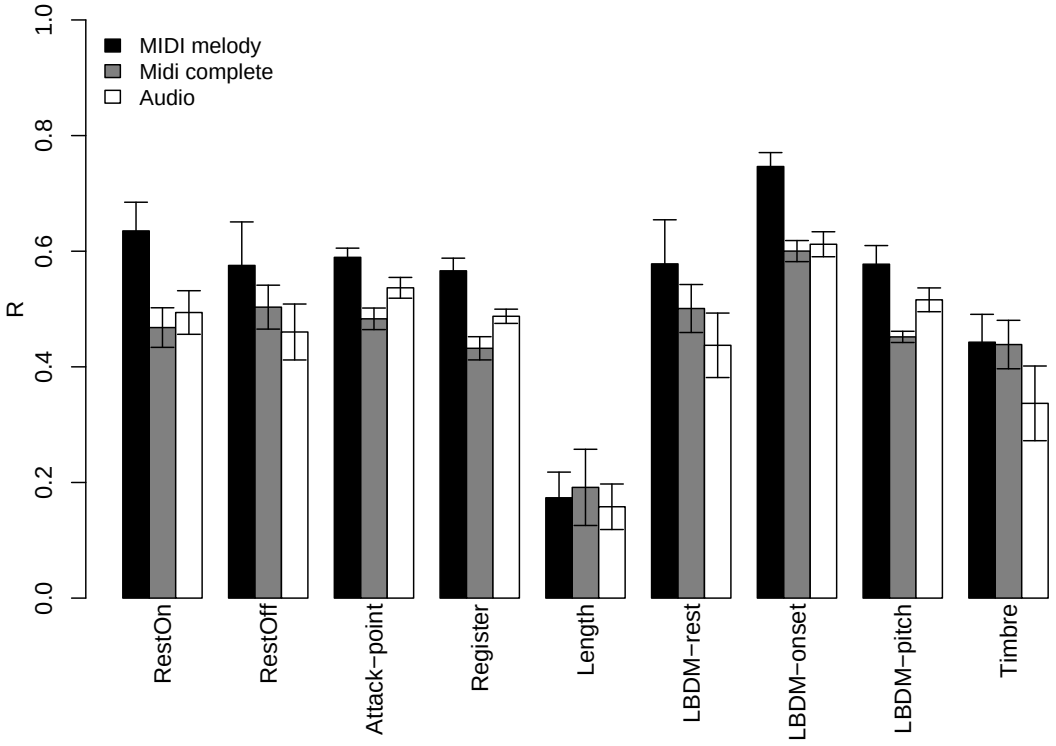


Figure 4.8: The correlations of the individual cues with the perceptual boundary profiles obtained in segmenting three different types of stimuli: MIDI melody, MIDI complete, and audio stimuli.

six boundary profiles of the predicted boundaries and calculated a linear regression to the concatenated perceptual profiles. The weights were 0.15 × RestOn + 0.19 × LBDM-onset + 0.27 × Timbre for the Midi complete stimuli and 0.21 × RestOn + 0.27 × LBDM-onset + 0.26 × Timbre for the audio stimuli. These results suggest that for all three stimulus types the combination of change in note-length (LBDM-onset), timbre-change, and the start of a rest results in the best performance of the predicted boundary profiles.

4.8 General discussion

The aim of the present study was to compare the predictions of several musicological cues with the perceptual profiles obtained in the previous two chapters. The perceptual profiles were obtained by asking subjects to segment six pieces taken from Western popular music. The main results of the study were that a high correlation between annotated and predicted segment boundaries can be obtained by a linear combination of timbre-change, the onset rule of LBDM (Cambouropoulos, 2001a), and the quantized rest rule of GTTM (Frankland & Cohen, 2004). Using these three cues, a physical correlation of about 0.85 between predicted and perceived boundary profiles can be reached. This combination of cues is in line with the model implementation of the LBDM (Cambouropoulos, 2001a), which gave the highest importance to the onset rule. It also concurs with the findings of Frankland and Cohen (2004), who found attack-point – a long note in between two short notes – as being the most predictive cue for perceptual segmentation, and the rest rule as being the second most predictive rule. The high influence of timbre-change that we found in our results may be the result of our stimulus construction, which consisted of the melody or the most salient accompaniment in the absence of melody. In our stimuli, a change in timbre, thus, coincided with the beginning or end of the melodic line. Nevertheless, timbre-changes have also been found in a different study to be an important perceptual boundary cue for segmenting short excerpts (Deliège, 1987). It is also notable that none of these studies found a strong influence of pitch intervals for predicting segment boundaries. In summary, these findings suggest the use of three cues to model perceptual segment boundaries in musical melodies: change in timbre, change in note length, and begin of rests.

A further analysis of the predictive power of these three cues revealed that they are useful in predicting boundaries over the full range of perceptual salience. If only the more salient boundaries are taken into account, the correlation between model and data decreases. A possible source for the lower correlation between the most salient boundaries and model predictions is that certain boundary cues are not implemented in the evaluated models. Examples of such cues are tempo-changes and rhythm-changes. Other cues not implemented in the models need a larger context than a few notes to be identified, such as repetition or harmonic progressions. The importance of repetition as a boundary indication has been identified (cf. Cambouropoulos, 2006a) and a model has been proposed that extracts boundaries based on previously encountered segments from a corpus of musical pieces with annotated phrase boundaries (Bod, 2002). The annotations, however, were done by musicologists and may therefore be different if the pieces were perceptually annotated. Another cue, harmonic progressions, is also likely to need a larger context than a few notes to be properly extracted. Principles of music theory suggest that the global harmonic structure influences the location of local boundaries (Schenker, 1935; Lerdahl & Jackendoff, 1983). Recent research, however, found that the local context often overrules the global structure (Tillmann & Bigand, 1998; Tillmann et al., 1998; Tillmann & Bigand, 2001). Correspondingly, music theorists and psychologists have introduced the term *perceptual present*, which is the idea that listeners do not analyze the piece as a whole but rather evaluate the content of a sliding window. The maximum size of the sliding window is considered to be between 5 to 30 seconds (cf. Lalitte, Bigand, & Poulin-Charronnat, 2004). The idea of the perceptual present suggests that listeners evaluate a time window for segment boundaries with the duration of the temporal window likely to be longer than a few notes. The model implementations used here only evaluate three to four consecutive notes to predict segment boundaries.

Despite these qualifications, the present study shows that with only three locally-derived cues, a high correlation can be found between the predicted boundary profile and perceptual boundary profiles obtained in segmentation experiments.

5 Summary and conclusions

The aim of the research presented in this thesis was to gain a better understanding on how humans perceive musical structure and in particular how humans perceive structural boundaries in music. A particular interest was to identify the main cues contributing to the perception of segment boundaries in Western popular music. A further aim of the present thesis was to evaluate which theoretical cues have the highest contribution to predict segment boundaries.

To investigate the perception of structural boundaries, an experimental setup was developed, consisting of two experiments. In the first experiment subjects were asked to segment six songs by pressing a key on a computer keyboard. In the second experiment a subset of marked boundaries was taken and subjects were asked to rate the salience of the given boundary on a scale from 0 to 6 and to describe the perceived musical cues of each boundary. This experimental setup was used to study the perception of structural boundaries in monophonic (chapter 2) as well as in polyphonic music (chapter 3) taken from Western popular music.

In chapter 4 of the present thesis the segmentation profiles obtained with the above method were used to evaluate different cues proposed in musicological models. The cues were based on changes in successive notes having the difference in one (or more) of the following four cues: pitch interval, note duration, breaks and rests, and timbre changes. These cues were then compared individually as well as in combination with the perceptual profiles of the six songs.

5.1 Summary of findings

The experiments showed several important findings. Previous studies assumed that the number of subjects indicating a segment boundary within a time window could be used as the salience of the boundary, however, without giving any formal correlation (Frankland & Cohen, 2004; Krumhansl, 1996; Schaefer et al., 2004; Spiro, 2007). Our method explicitly asked subjects to indicate the salience of a selection of boundaries previously indicated by our subjects. Our results showed a high correlation between the implicit (number of boundary markings within a time window) and the explicit salience measure. The high correlation thus offers two possible ways to obtain the salience of segment boundaries. One is to ask subjects to segment pieces of music. The other is to

take a set of boundaries and ask subjects to rate the salience of the given boundaries. Based on our results, both methods should lead to similar results and, thus, both methods are apt to provide a valid ground truth for evaluating automatic segmentation algorithms. Our empirical data also suggest that the perception of segment boundaries is not binary, i.e., that there is no "correct" segmentation pattern but rather an assembly of boundaries associated with a salience value.

A methodological contribution of the present thesis is the way that boundary indications were integrated to obtain a boundary profile across subjects and trials. Previous research, for example, asked subjects to indicate perceptual boundaries on a score notation (Clarke & Krumhansl, 1990) with the effect of limiting subjects to musicologically trained persons. Other studies integrated the indicated boundaries within a certain time-frame based on the score notation, for instance a one-beat window (Spiro, 2007) or two-beat window (Krumhansl, 1996). The difficulty with the windowing method is how to treat boundaries that are anticipated or delayed and thus do not fall exactly within the boundaries of the rectangular window. Furthermore, it is not clear how to apply the assignment to a temporal window based on the score if the score is not easily available, as it is often the case for music taken from Western popular music. The method presented here, and developed independently by Ferrand (2004), where the boundary indications are convolved with a Gaussian window, avoids these problems. The convolution of the boundary indications with a Gaussian window provides a smoothed boundary profile where boundaries further away are given less weight. We also developed a method to estimate the optimal window size used for the Gaussian smoothing. The optimal window size was defined as the longest window that includes exactly one boundary indication of each trial while minimizing the number of double indications within one window. The obtained optimal window size of 1.25 s concurs with the window size used by other studies to integrate the intrinsic scattering of the boundary indications (Clarke & Krumhansl, 1990; Frankland & Cohen, 2004; Schaefer et al., 2004). The smoothed boundary profile was also used to find the optimal placement of a temporal window instead of relating it to the metrical structure, as the above studies did.

From a methodological point of view it is also interesting to analyze the relation of the boundary indication across the three trials. Previous studies found the segmentation patterns over the course of, for instance, three trials to be "reasonably consistent" to be averaged across the second and third trial (Frankland & Cohen, 2004). Another study

asked subjects to press a key for the duration of a phrase, by pressing the key down at the start of the phrase and lifting it at the end. She reported (Spiro, 2007, p. 85): "[The figures] show that there is little change in the proportion of responses in the high response areas from the first to the final listenings for the PSs [phrase starts]. None have the pattern of gradual 'improvement' over the three listenings indicating that over the three hearings, there is no 'learning'. The PE [phrase end] show a similar pattern to that of the PS responses." The best way to evaluate whether subjects are consistent across the three trials is to evaluate the moderate boundaries, i.e., boundaries that were not indicated by all subjects and trials. Principally these moderate boundaries can be caused by two possible scenarios. 1) Half of the subjects perceive the boundary and indicate it across all three trials and the other half of the subjects do not perceive the boundary and do not indicate it therefore. Or, 2) subjects are not sure about these moderately strong boundaries and indicate them only once or twice across the three trials. Our data show that there is a strong tendency that there are two groups of subjects, one perceiving these moderately often indicated boundaries and indicating these boundaries across all three trials, while the other group does not perceive the boundary at all, i.e., subjects are binary in their boundary indications across the three trials. Different subjects, however, indicate a different number of boundaries, which concurs with previous studies (Deliège & Ahmadi, 1990; Ferrand, 2004; Frankland & Cohen, 2004; Krumhansl, 1996). This finding suggests that subjects have individual thresholds which are kept over the course of the three trials.

A further interest is whether subjects give a high salience rating for their own indicated boundaries. Our data, however, showed no correlation between the individually indicated boundaries and the given salience ratings. These findings suggest that the two experiments are independent and that subjects can rate the salience of boundaries even for boundaries not indicated by themselves in the segmentation experiment.

Another finding was that subjects used some cues more often than others to describe segment boundaries, such as "change in timbre", "global structural" descriptions like "start of chorus/verse", and "changes in rhythm". The frequency of terms with which a boundary was described also had a high correlation with the boundary salience, thus the higher the boundary salience, the more cues are used to describe the boundary. As an additional measure of the most important cues for segment boundaries, the mean term rating was developed (cf. page 35), which relates the boundary salience rating with

the cue descriptions. The highest mean term rating was found for the cues "harmonic progressions" and "change in timbre" and both cues were relatively independent of the song. In addition to the cues identified in previous studies, the cues "change in timbre", "change in tempo", and "change in rhythm" were also found to be important boundary indicators, both in the mean term rating as well as in the frequency of terms.

The comparison between the monophonic and polyphonic representations of the songs showed that when the different representations are well time-aligned, the segmentation patterns are correlated with each other. This findings suggests that subjects perceive segment boundaries of monophonic and polyphonic representations independent of the polyphonic cues for Western popular music, thus that they mainly focus on the melody. It has to be added, though, that our pool of subjects did not contain professional musicians. These subjects may not be as sensitive to more complex facets of music as professional musicians. Nevertheless, these findings suggest the use of attention-based models to estimate the most salient monophonic line in polyphonic music and using this melody line to estimate the segment boundaries.

Finally, we evaluated how well cues and cue combinations extracted by formal musicological models can predict the perceptual boundary profiles. These cues were extracted from two models, the LBDM by Cambouropoulos (2001a) and four rules of GTTM by Lerdahl and Jackendoff (1983) quantified by Frankland and Cohen (2004). These cues have previously only been evaluated on pieces annotated by musicologists or perceptually on short melodies. Our study extended the evaluation to complete pieces with a duration of several minutes and in monophonic as well as polyphonic representations. The results suggest that two to three cues are enough to have a high correlation between the perceptual boundary profiles and the predicted boundary profiles. The combination of cues that best predicted the perceptual boundary profiles were the "onset rule" of the model by Cambouropoulos (2001a), a "change in timbre", and the "rest rule" of GTTM. Using these three cues, a mean (physical) correlation between perceptual and predicted boundary profiles of 0.85 could be reached. Furthermore, for predicting boundaries in polyphonic music, a similar cue combination of "change in note length", "timbre change", and "rests" has shown to perform best. The analysis of the description of salient boundaries that the model could not predict suggests that two additional cues, "tempo changes" and "harmonic progressions", could possibly improve the performance of the predictions.

5.2 Concept of an algorithm

The results of the presented experiments have led to a possible algorithm for segmenting polyphonic music. The proposed approach is to first identify a monophonic representation of the polyphonic mix, which should ideally be the melody if present. Such an algorithm could use voice extraction methods (Cambouropoulos, 2006b) and then take the most salient voice as the melody. The monophonic representation would ideally be the most salient melodic line. The melody would then be segmented with a segmentation algorithms, similar as proposed by Cambouropoulos (2001a). The algorithm uses the cues longer notes, breaks, and timbre-change (with maybe the three additional cues of tempo-change, rhythm-change, and harmonic progressions) to find possible segment boundaries with an assigned salience.

The extracted boundaries could then be placed back as the segment boundaries on the polyphonic piece. If a binary segmentation profile is needed, i.e., where the piece can be segmented, a threshold can be set and all boundaries with the salience exceeding the threshold are taken as segment boundaries.

5.3 Future work

In addition to the cues evaluated in Chapter 4, several additional cues could possibly enhance the prediction performance. The reason these cues have not been incorporated into existing models is partly because it is not clear how to implement them. One of the cues likely to improve the performance of models predicted segment boundaries is repetition. Repetition has been mainly used in audio segmentation algorithms, where the self-similarity matrix of a feature has been used to segment the audio into smaller segments (Foote, 1999). In music theory, the cue repetition has often been identified as an important segment cue, however, so far no complete algorithm has been proposed (cf. Cambouropoulos, 2006a). Furthermore, the *perception* of repetition has not received much attention. The perceptual research on repetition has been limited to studies on Auditory Scene Analysis using very short stimuli (Bregman, 1990). These stimuli cannot represent the complexity of repetition found in complete music pieces. Thus, it is not yet clear how repetition is perceived and what kind of repetition has the main impact in recognizing segment boundaries.

Another cue to be further researched is "harmonic progression". The perception of harmony seems to be governed by two principles. First, the sense of harmony is not defined by the order of the chords. The harmony stays intact, even for scrambled sequences (Tillmann & Bigand, 2001), thus the harmonic context is spread over time. Second, the local harmony overrules the global harmonic structure (Tillmann et al., 1998). These studies suggest the existence of a time window, during which the harmony is integrated. Another study, for example, showed that the preceding chord influenced the rating of "how well does the second chord follow the first" (Bharucha & Krumhansl, 1983). One possible approach to implement harmonic changes could be to estimate the cords used in a local context to then predict local cadenzas, which define points of relaxation in the music and are thus often used by composers as predictors of phrase ends.

Another line of future research could focus on the influence of hierarchy on the perception of musical structure. Such an influence has been assumed by music theory, which often suggest a tree-like reduction (Schenker, 1935; Lerdahl & Jackendoff, 1983). There exist some indications that for tonal music a reduction based on hierarchical properties indeed is preferred over arbitrary reductions (Dibben, 1994). However, other studies concluded that the local context is more relevant for the perception of structural boundaries (Tillmann et al., 1998; Bigand, Madurell, Tillmann, & Pineau, 1999; Tillmann & Bigand, 2001). In addition, it may be useful to implement the influence of the local context into the models, as it has been done partly in the model of Temperley (2001).

Finally, the metrical position of the boundary could influence the relative salience of the boundary. Two possible ways can be imagined for such an influence. One is that the meter modulates the boundary salience and that, thus, boundaries are perceived more salient if they are on a strong beat and less salient if they fall on a weak beat. Another influence could be that the strong beat attracts the boundary indications in the sense that even though a boundary would be perceived (or predicted by musicological rules) to be on a weak beat in the measure, subjects are drawn to indicate the boundary on the strong beat. A possible experiment to test the influence of meter on the perception the boundary's salience would be to create two different melodies, one where a salient boundary is being placed on a strong metrical position and another stimuli where the salient boundary is placed on a weak metrical position. The salient boundaries could be, for instance, a long note or a timbre change. By comparing the number of

indicated boundaries for these salient boundaries for the two stimuli, the influence of the metrical position could be established. Furthermore, the experiment would allow to verify whether a strong metrical position attracts boundary indications in analyzing whether subjects tend to indicate the introduced segment boundary close to the strong beat.

Despite these directions for future research, a general result of this thesis is that perceptual segment boundaries in Western popular music can be predicted quite well with relatively few cues based on local features.

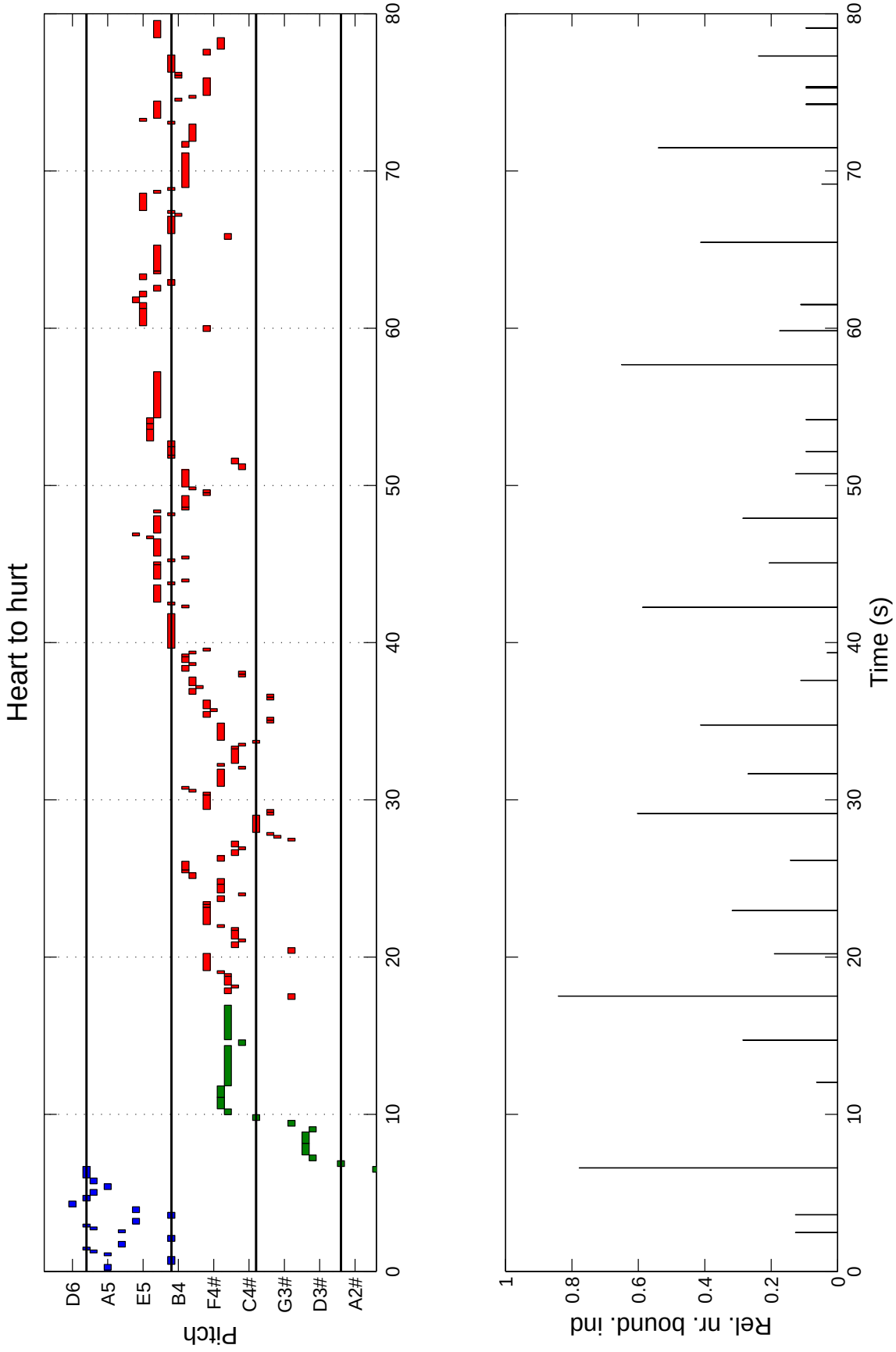
A Excerpts of the MIDI stimuli used

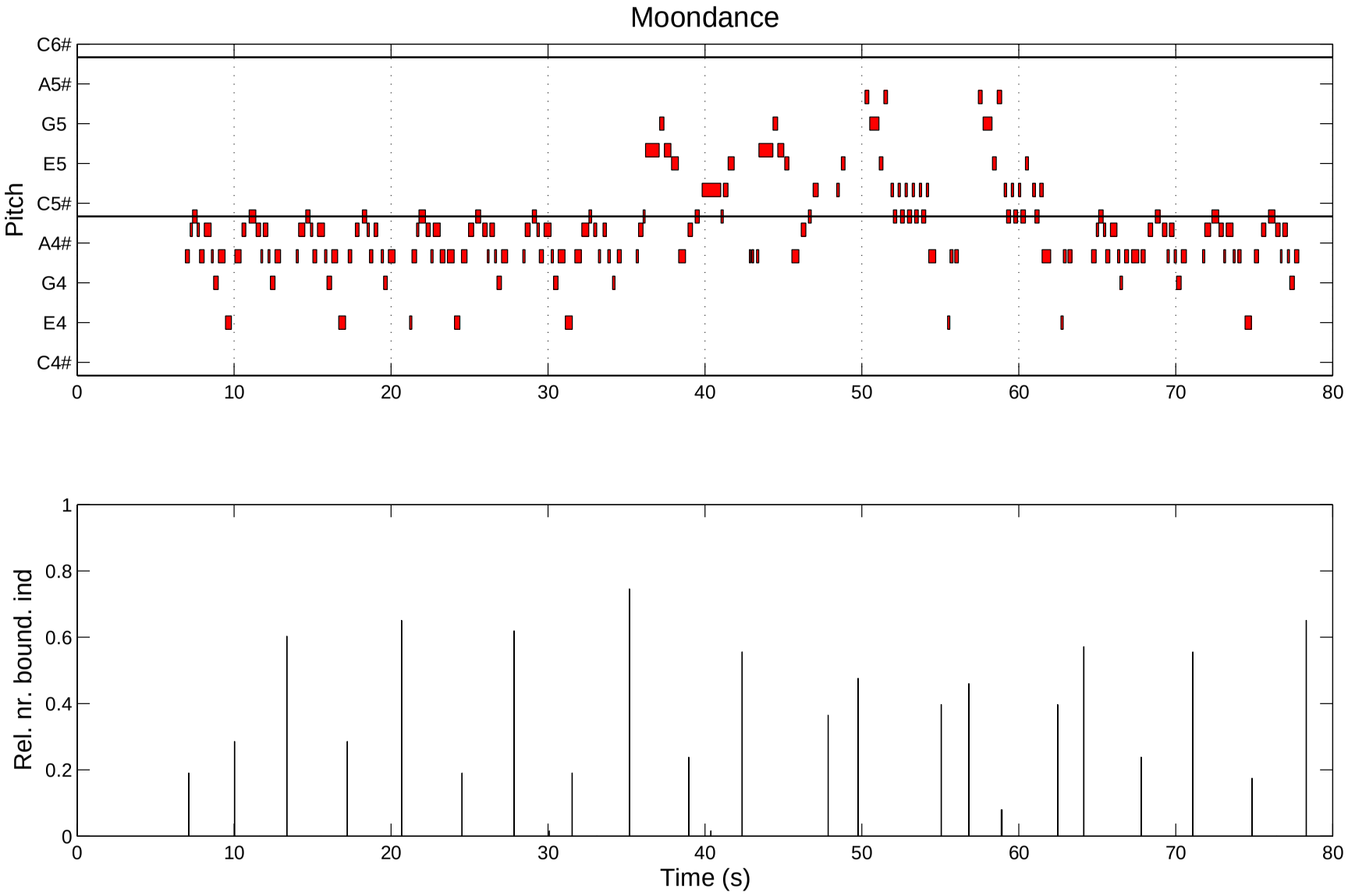
The following five pages show excerpts of the first 80 s of each song, plotted in the piano-roll representation for the monophonic representation (first row) with the perceptual boundary profile of the monophonic representation (second row). The changes in the color of the squares, thus the notes, show a change in the played instrument. The order of the song is "Heart to hurt", "Moondance", "Live and let die", "And when I die", and the version of "Body and soul" aligned to the vocal audio recording. As the instrumental version of "Body and soul" is very similar to the vocal version, here only the vocal version is shown.

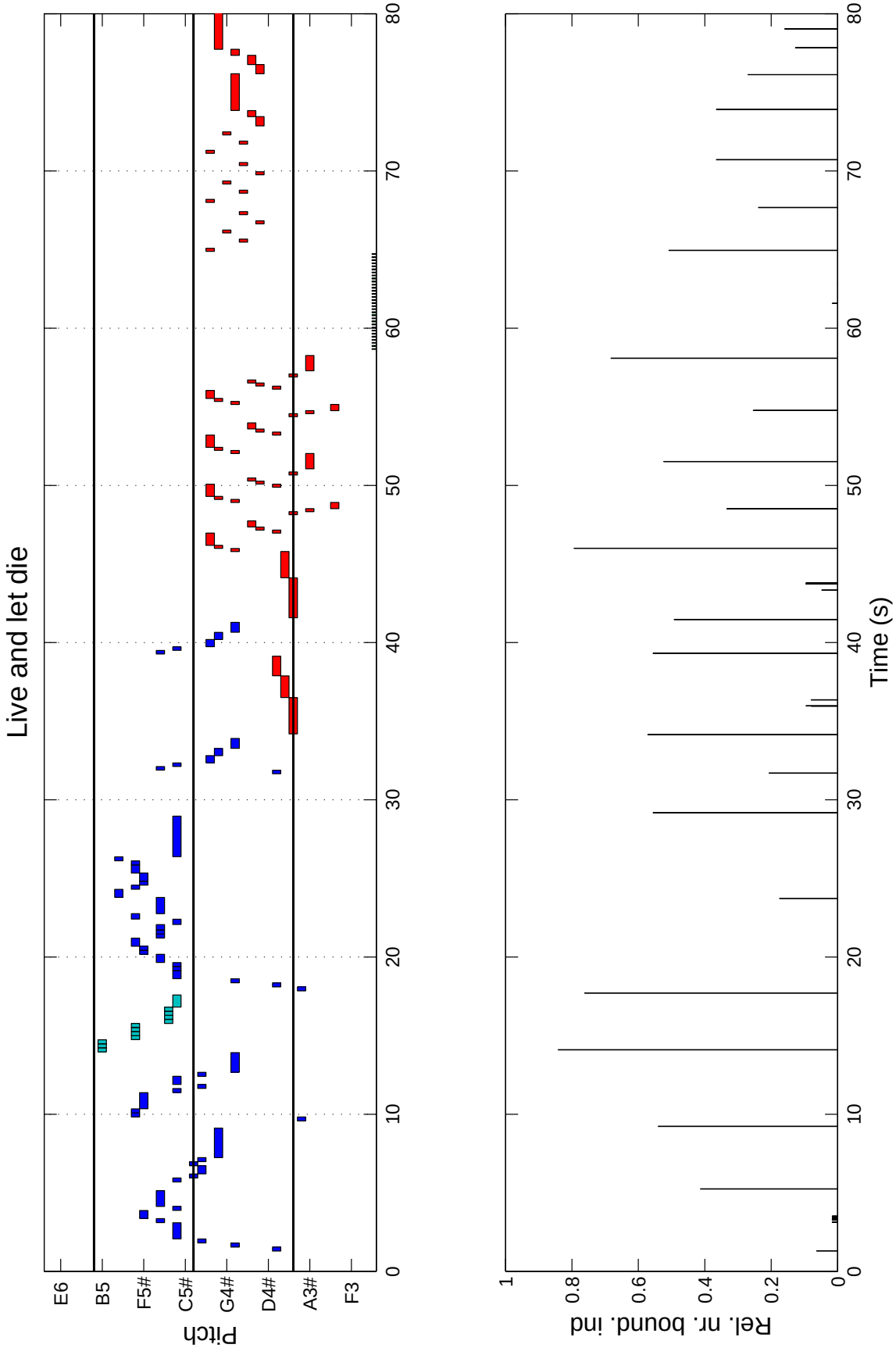
The original source of the MIDI files were the following. The MIDI file for the song "Heart to hurt" was included in the RWC database (Goto et al., 2002). The other songs were taken from the following publicly available internet sources:

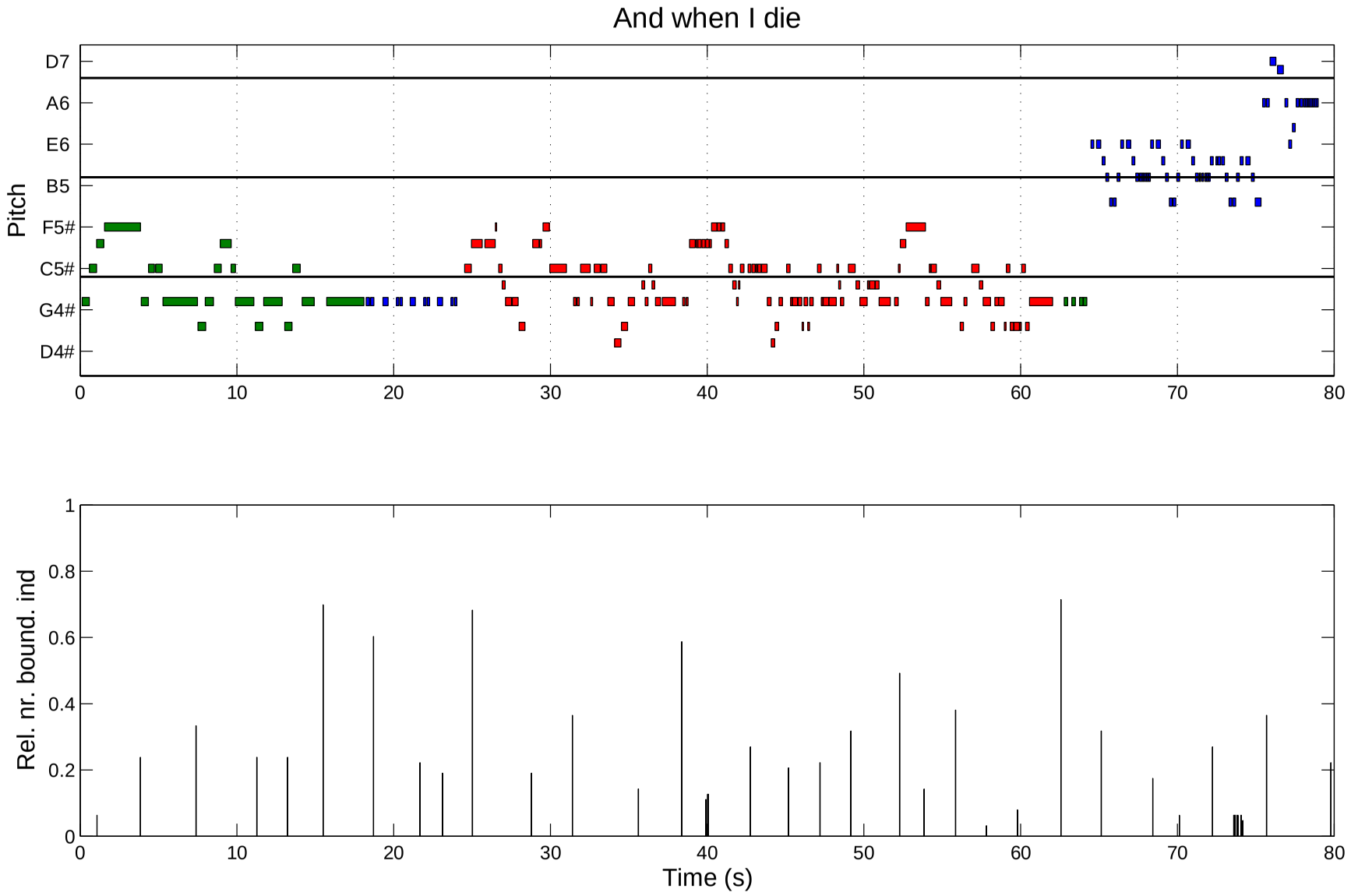
`http://www.mgeurts.tmfweb.nl/Moondance_1_Van_Morrison.mid`
`http://209.197.86.65/19841988/pop/paulmccartney/Live_and_Let_Die.mid`
`http://www.garyrog.50megs.com/midi/andwhenidie.mid`
`http://www.geocities.com/BourbonStreet/Delta/5853/BodyandSoul.mid`

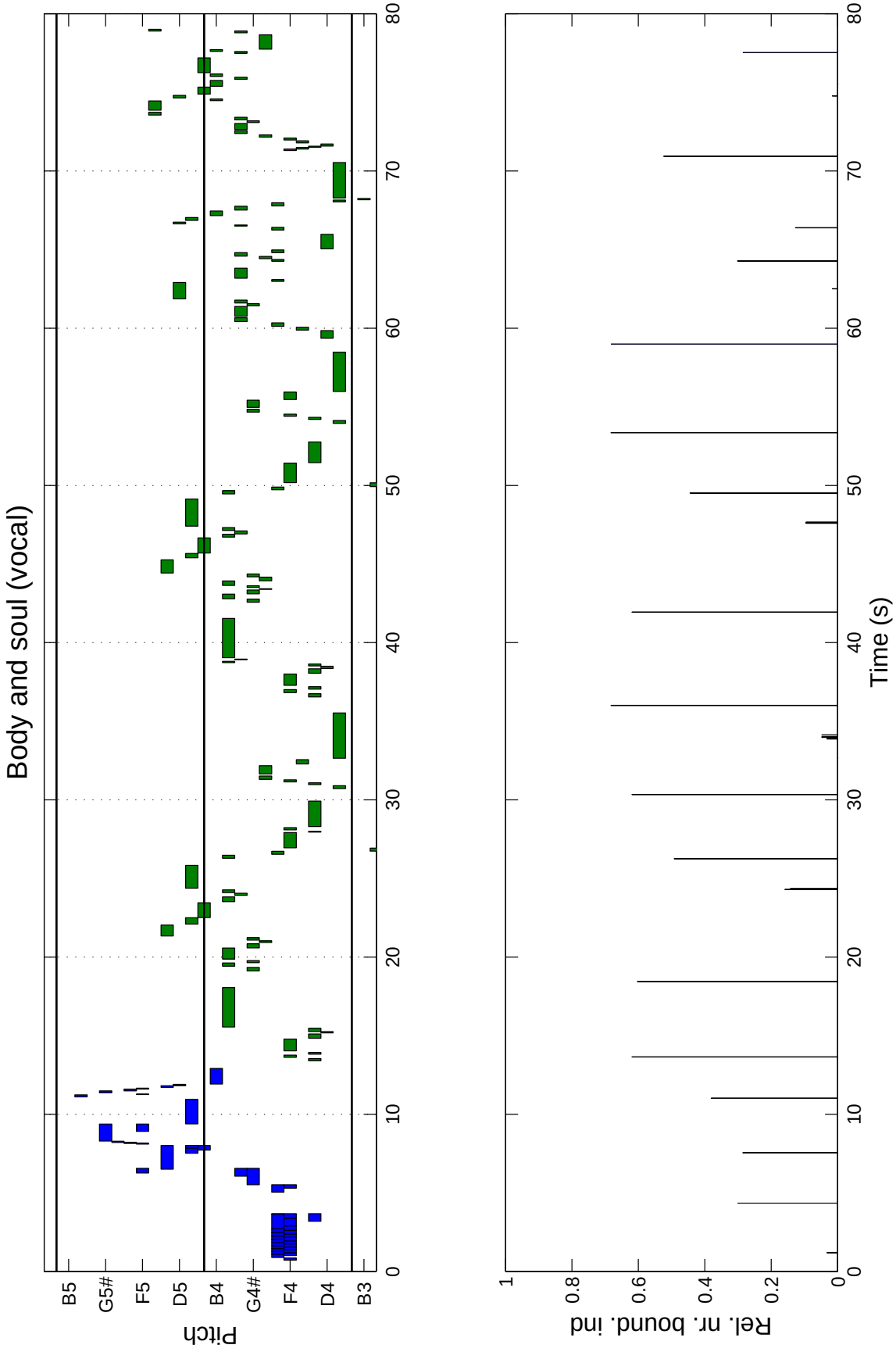
These MIDI files were manually cleaned and time aligned to the audio recording to resemble the audio recording as close as possible. The MIDI files were not deadpan, because we tried to match the audio recording as much as possible and the audio recording included temporal changes. Furthermore, we changed the timbre and the level of each voice of the MIDI voices to better represent the audio recordings.


Heart to hurt
Pitch
D6
A5
E5
B4
F4#
C4#
G3#
D3#
A2#
0
10
20
30
40
50
60
70
80
Rel. nr. bound. ind
0
0.2
0.4
0.6
0.8
1
Time (s)


Moondance
Pitch
C6#
A5#
G5
E5
C5#
A4#
G4
E4
C4#
0
10
20
30
40
50
60
70
80
Rel. nr. bound. ind
1
0.8
0.6
0.4
0.2
0
Time (s)


Live and let die
Pitch
E6
B5
F5#
C5#
G4#
D4#
A3#
F3
0
10
20
30
40
50
60
70
80
Rel. nr. bound. ind
0
0.2
0.4
0.6
0.8
1
Time (s)


And when I die
Pitch
D7
A6
E6
B5
F5#
C5#
G4#
D4#
0
10
20
30
40
50
60
70
80
Rel. nr. bound. ind
1
0.8
0.6
0.4
0.2
0
Time (s)


Body and soul (vocal)
Pitch
B5
G5#
F5
D5
B4
G4#
F4
D4
B3
0
10
20
30
40
50
60
70
80
Rel. nr. bound. ind
0
0.2
0.4
0.6
0.8
1
Time (s)

B Classification of the boundary cue descriptions

B.1 Introduction

In the following we show how we classified the boundary cue descriptions for the MIDI melody given by subjects in the salience rating experiments. The descriptions shown in the following should give an idea which terms were classified into which "cue class". The number in the parenthesis represents the number of times subjects used this wording. Within each class, the terms are ordered according to how often they were mentioned.

The "cue classes" were selected to have as few classes as possible while still be able to classify the different descriptions. The classes "harmonic progression", "melody changes", "tempo changes", "rhythm changes", "timbre changes", "level changes", and "breaks" are basic musicological descriptions of music. The remaining two classes, "global structure" and "repetitions" are musicological descriptions of the musical structure. These two classes are, in contrast to the other terms, not basic cue descriptions, but rather complex summaries of how subjects perceive the musical structure. These two classes, therefore, are not mutually exclusive from the other two classes.

The goal of the following tables is not only to show how we classified them, but also to demonstrate the different terms subjects used for describing boundary cues. Some descriptions may seem arbitrarily classified into our "cue classes". The descriptions given by all subjects for a specific boundary were classified together and thus if several subjects described the boundary cue as a "change in melody" without mentioning any level changes, the cue description of another subject only containing "low -> high" was then interpreted also as a "change in melody" and not as "change in level".

Harmonic progression and tonality change (41)

The class "progression" includes all descriptions related to the "vertical" structure of the music, i.e., harmonic or tonality changes. It also includes the terms cadenza.

- change in tonality (16)
- change in harmony, harmonic change, other harmony (15)
- perfect cadenza (3)

- very/quite conclusive (3)
- other tonality (2)
- harmonic preparation for some change (1)
- semi cadenza (1)

Melody change (325)

The class "melody" includes all the descriptions related to the "horizontal" structure in the sense of changes in the pitch or durations of successively played notes.

- (major/minor/big/small) change in melody (156)
- long tone/note (65)
- change in tone, different tone (22)
- new (piece of) melody (21)
- higher, it continues much higher (8)
- pitch change (7)
- end of melody/note (4)
- change/jump in register (3)
- instrument hangs (2)
- strange tone acting as marker (2)
- disrupting tone (2)
- heightens/higher pitch (2)
- annoying tone (1)
- other melody (1)
- some variations in notes (1)
- the kind of depressing ending of the previous section takes an uplifting turn (1)
- modified melody toward summit (1)
- the previous section is ended by two isolated notes (1)
- more dense (notes) (1)
- a distinctive melody starts which is completely different from the staccatolike chords (1)
- two isolated rather low notes, followed by sudden high notes (1)

- a melody is ended and is followed by a repetition of notes working themselves up the ladder (1)
- the previous theme is suddenly taken with higher notes (1)
- very high tone (1)
- low - high (1)
- new phrase with higher pitch starts (1)
- from the rather monotonous previous section the following is higher (1)
- start of high part in the melody (1)
- a series of slow low notes is now followed by a series of fast high notes (1)
- new phrase at new height (1)
- tone ladder going down (1)
- higher notes (1)
- a long note is followed by a short high one (1)
- a long note is followed by a series of higher and shorter notes (1)
- a new instrument starts. The previous was high-pitched, the next low (1)
- second part of melody; jump down in register (1)
- the melody has reached its climax (high note) and is slowly working down for the ending of the section (1)
- last notes go down (1)
- new phrase with lower pitch (1)
- tone ladder going down
- the music goes from very high to low in a few notes (1)
- the sound gets lower (1)
- lower (1)

Tempo change (54)

This class comprises all terms related to tempo and temporal changes in the stimuli.

- change in tempo/speed/pace (38)
- faster (tempo), song/melody speeds up, fast section starts again, acceleration (11)
- slow down (of melody), slower, the fast guitar session now becomes a series of slow lonely notes (5)

Rhythm change (113)

The class "change in rhythm" includes all terms comprising descriptions of rhythmic changes.

- (major) change in rhythm, different rhythm, change in rhythmic pattern/structure (90)
- new rhythm (17)
- change rhythmic feeling (2)
- rhythm (2)
- the rhythm jumps up to (1)
- start of song main rhythm (1)

Timbre change (530)

The class "Timbre change" includes all descriptions that described a change in the instrumentation.

- instrument change, change (in/of) instrument (321)
- new instrument (112)
- different/(an)other instrument (24)
- new timbre, change in timbre, different timbre (14)
- extra instrument kicks in (2)
- a new high pitched instrument plays (2)
- short piece of earlier instrument again (2)
- the old instrument is back again, which the same old theme (2)
- new sound of instruments come in (1)
- change back to a less annoying instrument (1)
- instrument voice change (1)
- changing the tone of the instrument (1)
- strange tone (1)
- complete change (instrument + rhythm) (1)
- instrument goes crazy (1)
- the first instrument plays again and answers the previous instrument (1)
- one more instrument gets in (1)

- the slow series of notes are now followed again by the distinctive guitar thing (1)
- instrument (1)

Level change (261)

The class "level change" includes all the descriptions that were directly related to a change in the dynamics of the music.

- change in level (239)
- end of sound/song (8)
- music stops (the end) (2)
- music to silence (2)
- silence to music (2)
- clearly the end, a very long note followed by... nothing. (1)
- end (2)
- play stops (1)
- stop of the song (1)
- change of volume (1)
- fading (1)
- music start again (1)

Break (450)

The class "break" includes all terms describing a pause, break, and silence in the music.

- pause (196)
- silence (77)
- short/small pause (27)
- long/big pause (25)
- (just a) short/small break (23)
- short/small gap (19)
- short silence (15)
- gap (14)
- music stops, instrument stop (14)
- big/huge gap (12)

- medium break (9)
- long silence (7)
- big/long break (6)
- break (3)
- no more sound but still intro (1)
- break not very significant (1)
- weird gap (1)

Global structure (450)

The class "global" includes all terms describing the general musicological structure of a song, such as begin of phrases, sections, and chorus-verse. This class, therefore, does not really represent the cues causing the segment boundary, but rather a complex summary of how subjects interpret and analyze the song.

- (start of) new phrase (90)
- (begin of) new part (of song) (52)
- end of (musical) phrase (45)
- start of chorus, refrain starts, new chorus, to refrain, going to chorus, lead in to chorus, start of ritornello, refrain begins (41)
- new section (starting), this section starts again how the previous section, next section (25)
- return to A section, end of A section, part B2 as before, part A, repetition of A, restart of A section, passage to bridge B, the B part of the song, start of A part, repetition of section A, return to B, back to B part, B2 part, A part again (22)
- verse begins, (to) verse, end of verse, new verse, going back to verse, repetition of verse part (20)
- back to the theme, same old theme, theme starts again, again the theme, return to main theme, theme comes in (18)
- repetition of (1st) phrase (13)
- chorus, refrain (13)
- end of (a small) (previous) section (13)

- end of intro(duction), sounds to me like end of intro, sololike introduction, end of introduction to the theme, clearly the playful introduction with repetitious notes ended here (11)
- solo, end of solo, end of a part starting solos (9)
- end of first (major) part, end of a part with "hook line", end of part (5)
- second part, second part of melody, third part of melody (5)
- repetition of (main) melody, last repetition of current melody (5)
- (start of) new melody (5)
- begin/start of main melody (5)
- return to known part, part already heard, repeated part (4)
- phrase (4)
- begin/end of bridge, building up bridge with different rhythmic structure (3)
- end of chorus (3)
- repetition of chorus (3)
- (start/begin of) outro (3)
- next part, another part (3)
- change of section (3)
- (main) theme begins, start of main theme (3)
- new segment (2)
- end of improvisation (2)
- going from verse to chorus (2)
- strong phrase (2)
- melody ends and new one starts, end of melody begin of new (2)
- start of hook line (2)
- pause indication next section (2)
- next section and strange tone acting as marker (1)
- previous section was ended with the already mentioned familiar ending piece (1)
- change in part (1)
- back to main theme (1)
- repetition of intro (1)
- middle of the section but end of a conclusive phrase (1)

- the previous theme is suddenly taken with higher notes (1)
- here the theme of the beginning starts (1)
- there is a long note denoting the end of a section (1)
- end of the piano solo + beginning of the sax solo (end of a solo section, beginning of a new one) (1)
- begin of short solo (1)
- another beginning of the music (1)
- the following section is a variation of the melody in the previous section (1)
- end very conclusive main theme (1)
- a theme starts which does not sound like a mere introduction (1)
- end of short inserted melody and back to first part (1)
- solo with the familiar tune (1)
- sounds like a boundary between two parts of a verse (1)
- melody start after intro (1)
- to interlude (1)
- a familiar theme starts again afterward (1)
- start of intermezzo (1)
- end of theme (1)

Repetitions (301)

The class "repetition" includes all terms comprising descriptions of repeating elements and patterns.

- (start/end of) repetition, repeat (129)
- repetition of (previous) phrase (32)
- repetition of (previous/main) melody, repetition of part of melody, melody restarts (29)
- repetition of boundary X (19)
- known/repeated/familiar theme (16)
- . . . again (16)
- repetition of section/part A/B, return to A (10)
- variation of/in melody, repetition with variation (10)

- known/repeated part (8)
- back/return to main-part/refrain/chorus (7)
- a series of repetitious notes/piece (3)
- repetition of verse (3)
- repetitious notes, repetition of notes (3)
- . . . back (2)
- repeat high part, repetition of high notes (2)
- repetition of intro (1)
- two bars of repeated patterns (1)
- piece that is repeated twice (1)
- same as build-up part already heard before (1)
- repetition of previous musical sequence (1)
- a new instrument starts with the theme from the beginning (1)
- repetition of earlier stuff (1)
- familiar ending piece (1)
- very similar to section (1)
- there is fading and a gap but the song resumes like playing from the beginning (1)
- a piece starts which is recognized because it has returned now for the 3rd time (1)
- the repetitious tune is clearly ended by a series of rhythmic notes (1)

B.2 Classification of descriptions as reported in Clarke & Krumhansl (1990)

In the following the classification of the cues as reported in Clarke and Krumhansl (1990) are described.

Harmonic progressions and tonality change (20)

- change of key (minor to major) (8)
- new material (end of cadenza) (5)
- change of harmony (cadenze) (4)
- arrival of coda (3)

Melody change (34)

- change of register (expansion, more restricted) (22)
- new material (chords changing to melody) (5)
- new material (change of pitch content) (4)
- change of pitch content (2)
- change of melody (1)

Tempo change (12)

- change of tempo (12)

Rhythm change (9)

- change of rhythm (9)

Timbre change (19)

- change of texture (thicker) (15)
- introduction of trill (2)
- change of articulation (1)
- change of tone (due to pedal) (1)

Level change (23)

- change of dynamic (20)
- change of dynamic contour (3)

Break (8)

- pause (silence) (8)

Global structure (33)

- end of previous material (6)
- return of first material (chordal) (5)
- new material (lyrical) (5)
- new material (dramatic) (5)

- new material (end of cadenza) (5)
- new material (isolated block chords) (4)
- start of development (3)

Repetitions (28)

- return of previous material (6)
- return of material (chordal) (6)
- return of first material (chordal) (5)
- return of material (chromatic run) (5)
- return of material (chordal with new pitches) (4)
- fragment of earlier material (2)

C The definitions of the individual model cues

The cues from three theoretical models plus one additional cue have been extracted. In the following we will give the precise definitions of each cue, as described in the original literature. Figure C.1 shows the boundaries predicted by the various models, or model cues, for a short section of the song "Heart to hurt", which had also been represented in Figure 2.3.

C.1 The quantified rules of GTTM

One of the most influential contemporary works on the theory of music structure is the book "A Generative Theory of Tonal Music" (GTTM) written by Lerdahl and Jackendoff (1983). The book is a "formal description of the musical intuitions of a listener who is experienced in a musical idiom" (Lerdahl & Jackendoff, 1983, p. 1) and gives a set of components on how music is analyzed by the listener.

An important distinction between grouping and meter is made: Grouping is the process of segmenting music into smaller parts. Meter is the regular alternation of strong and weak elements. GTTM treats these two as independent of each other, however, the best grouping has its boundaries on the strong beats of the meter. This means the boundaries are best placed if both, grouping and meter, coincide.

Grouping according to the theory of GTTM is mainly a study of sectioning. The "Grouping Structure" is applied on monophonic music scores and therefore on the surface of music. This "Grouping Structure" is divided into two groups, the "Grouping Well-Formedness" rules, which define legal grouping structures and the "Grouping Preference Rules" which define preferred legal structures. The "Grouping Preference Rules", shown in Table C.1, define the notation of a group and what the conditions are that a group should satisfy. The "Grouping Preference Rules" are hypothesized to be the rules that humans apply in order to hear boundaries in the musical surface.

As the "Grouping Preference Rules" are not explicitely stated in a way that can be implemented in a computer program, Frankland and Cohen (2004) quantified four rules. The four quantified rules were rest, attack-point, length change, and register change. Their definition of rest was: "A whole-note rest was coded as a boundary potential of 1.0, with other rest values being proportionally, so that boundary strength ranges from

Table C.1: Grouping preference rules as defined in GTTM (pp. 43-52).

Rule 1		Strongly avoid groups containing a single event.
Rule 2	Proximity	Consider a sequence of four notes $n_1n_2n_3n_4$. All else being equal, the transition $n_2 - n_3$ may be heard as a group boundary if
	a. Slur/Rest	the interval of time from the end of n_2 to the beginning of n_3 is greater than the end of n_3 to the beginning of n_4, or if
	b. Attack/Point	the interval of time between the attack points of n_2 and n_3 is greater than that between the attack points of n_1 and n_2 and that between n_3 and n_4.
Rule 3	Change	Consider a sequence of four notes $n_1n_2n_3n_4$. All else being equal, the transition $n_2 - n_3$ may be heard as a group boundary if
	a. Register	the transition $n_2 - n_3$ involves a greater intervallic distance than both $n_1 - n_2$ and $n_3 - n_4$, or if
	b. Dynamics	the transition $n_2 - n_3$ involves a change in dynamics and $n_1 - n_2$ and $n_3 - n_4$ do not, or if
	c. Articulation	the transition $n_2 - n_3$ involves a change in articulation and $n_1 - n_2$ and $n_3 - n_4$ do not, or if
	d. Length	n_2 and n_3 are of different lengths and both pairs n_1, n_2 and n_3, n_4 do not differ in length.
Rule 4	Intensification	Where the effects picked out by Rules 2 and 3 are relatively more pronounced, a larger-level group boundary may be placed.
Rule 5	Symmetry	Prefer grouping analysis that most closely approach the ideal subdivision of groups into two parts of equal length.
Rule 6	Parallelism	Where two or more segments of the music can be construed as parallel, they preferably form parallel parts of groups.
Rule 7	Time-Span and Prolongational Stability	Prefer a grouping structure that results in more stable time-span and/or prolongational reductions.

1/64 to 1.0. A 64th note is the smallest temporal value specifiable in standard music notation. Rests longer than a whole note should be assigned a value of 1.0. [...] The location of the rest defines the location of the boundary"(Frankland & Cohen, 2004, p. 504-505). We, therefore, used the following quantification:

$$\text{boundary strength} = \begin{cases} 0 & \text{if } r < 1/64 \\ \frac{r-1/64}{1-1/64} & \text{if } 1/64 \leq r \leq 1 \\ 1 & \text{if } r > 1 \end{cases}, \tag{C.1}$$

where r is the duration of the rest in whole notes.

The following three rules use four consecutive notes n_1, n_2, n_3, and n_4 to predict segment boundaries. Attack-point was quantified as (cf. Frankland & Cohen, 2004, p. 505):

$$\text{boundary strength} = \begin{cases} 1.0 - \frac{n_1+n_3}{2\times n_2} & \text{if } n_2 > n_1 \text{ and } n_2 > n_3 \\ 0 & \text{otherwise} \end{cases}, \tag{C.2}$$

where the n's are the lengths of the notes. The rule also requires n_1 to n_4 to be notes. Because in the MIDI file there is no easy distinction between notes and rests, this condition was omitted, thus following notes were used independent of whether there was a long offset-to-onset interval.

Length change was quantified as:

$$\text{boundary strength} = \begin{cases} 1.0 - \frac{n_1}{n_3} & \text{if } n_3 > n_1, n_1 = n_2, \text{ and } n_3 = n_4 \\ 1.0 - \frac{n_3}{n_1} & \text{if } n_1 > n_3, n_1 = n_2, \text{ and } n_3 = n_4 \\ 0 & \text{otherwise} \end{cases}. \tag{C.3}$$

Because $n_1 = n_2$ and $n_3 = n_4$, we quantized the duration of the notes to 1/32 notes before applying the length change rule.

Register change was defined as:

$$\text{boundary strength} = \begin{cases} 1.0 - \frac{|n_1-n_2|+|n_3-n_4|}{2\times|n_2-n_3|}, & \text{if } |n_2-n_3| > |n_1-n_2| \\ & \text{and } |n_2-n_3| > |n_3-n_4| \\ 0 & \text{otherwise} \end{cases}, \tag{C.4}$$

where the n's are the pitch heights in MIDI notation.

These four rules were then used to predict segment boundaries. Figure C.1, row two to five show the predicted segment boundaries for these quantified rules of GTTM.

C.2 The rules of LBDM

The LBDM by Cambouropoulos (2001a) is based on two general rules to predict segment boundaries:

> **Change rule**: Boundary strengths proportional to the degree of change between two consecutive intervals are introduced on either of the two intervals (if both intervals are identical, no boundary is suggested).
>
> **Proximity rule**: If two consecutive intervals are different, the boundary introduced on the larger interval is proportionally stronger.

A melodic sequence is converted into a sequence of independent parametric interval profiles P_k for the parameters pitch intervals, inter-onset intervals (ioi) and rests. In his paper, Cambouropoulos (2001a) also suggests to set an upper threshold on the intervals. We therefore set the maximum inter-onset interval and the maximum offset-to-onset interval (rest) to four beats and the maximum pitch interval to one octave.

$$P_k = [x_1, x_2, \cdots, x_n] \quad \text{where: } k \in \{pitch, ioi, rest\},\ x_i \geq 0 \text{ and } i \in \{1, 2, \cdots n\}. \tag{C.5}$$

The degree of change r between two successive interval values x_i and x_{i+1} is given by:

$$\begin{aligned} r_{i,i+1} &= \tfrac{|x_i - x_{i+1}|}{x_i + x_{i+1}} && \text{if } x_i + x_{i+1} \neq 0 \text{ and } x_i,\ x_{i+1} \geq 0 \\ r_{i,i+1} &= 0 && \text{if } x_i = x_{i+1} = 0. \end{aligned} \tag{C.6}$$

The strength of the boundary s_i for interval x_i is affected by both the degree of change to the preceding and following intervals, and is given by the function:

$$s_i = x_i(r_{i-1,i} + r_{i,i+1}). \tag{C.7}$$

For each parameter k, the sequence $s_k = [s_1, s_2, \cdots s_n]$ is calculated, and normalized in the range $[0, 1]$.

The overall local boundary strength profile for a given melody is a weighted average of the individual strength sequence S_k. The weights used in Cambouropoulos

(2001a) and implemented in the MIDI-toolbox (Eerola & Toiviainen, 2004) are $w_{pitch} = 0.25$, $w_{ioi} = 0.5$, and $w_{rest} = 0.25$. The cues used in the present study were called LBDM-onset, LBDM-pitch, and LBDM-rest, with an example of the predicted boundaries shown in row six to eight of Figure C.1.

C.3 Timbre change

The definition of change in timbre is directly related to the stimuli used in the perceptual experiment. The stimuli used in the perceptual experiment (Chapter 2) consisted of the melody and if the melody was not present of the most salient accompaniment. Although the stimuli were monophonic, each instrument in the MIDI file was played on a different track but with the tracks being non-overlapping. Timbre change was then defined as a change in the MIDI track between two successive notes.

C.4 The Melisma model

Melisma is a collection of models developed by Temperley (2001) and described in his book the "The Cognition of Basic Musical Structure". The aim of the model is to describe in a formalized way how listeners extract basic types of musical information, such as meter, phrase structure, counterpoint, pitch spelling, harmony, and key. The models use an approach similar to GTTM (Lerdahl & Jackendoff, 1983), where two types of rules are defined. One are "well-formedness rules", defining legal structure, and two, the "preference rules" that define which legal structures are preferred over all possible legal structures. Using these two types of rules the models then try to best satisfy these rules.

The part used here is the "melodic phrase structure" model. The melodic phrase structure model consists of three preference rules. The first preference rule, the gap rule, defines the global boundary profile in adding the inter-onset interval (the duration of the begin of a note to the beginning of the next note), with the offset-to-onset interval (the duration of the end of a note to the beginning of the following note), resulting in a global boundary profile. The second preference rule, the phrase length rule, applies a penalty if the length of a phrase is not eight notes. The third preference rule, the metrical parallelism rule, applies a penalty if a boundary is not at a parallel point in the metrical structure.

To apply this third rule, the "metrical structure" model has to be run on the melody. The metrical structure model is, similarly to the melodic phrase structure model, based on a set of well-formedness preference rules. Although he states nine preference rules, out of the nine only five were implemented in the computer program. The five rules were 1) to prefer a structure that aligns with event-onsets, 2) to prefer strong beats to longer events, 3) to evenly space the beats at each level, 4) to prefers strong beats at the beginning of groups, and 5) to prefer double over triple relationships.

The implementation of the models for both components, the melodic phrase as well as the metrical structure model, uses dynamic programming. This type of algorithm works in running through the piece twice. In the first pass, the analysis is made from the beginning to the end of the piece, where *possible* boundaries (or metrical levels) are estimated. In the second pass, the possible boundaries are tested from the back to the beginnig, where the optimal boundaries are selected. These optimal boundaries then define the segment boundaries of the piece. An example of the predicted boundaries is shown in the last row of Figure C.1.

Temperley (2001) tested the metrical model on 46 excerpts taken from a theory workbook. Two versions were tested, one being the quantized MIDI representations, the other being unquantized, performed MIDI representations. The author reports a correct metrical structure from between 71.5 to 94.4% correct measures across the five different metrical levels.

The melodic phrase structure model was tested on the Essen folksong database (Schaffrath, 1995), which is an annotated corpus of folksongs. The annotation was done by musicologists and not perceptually. The author reports a rate of 75.5% correctly identified phrase boundaries.

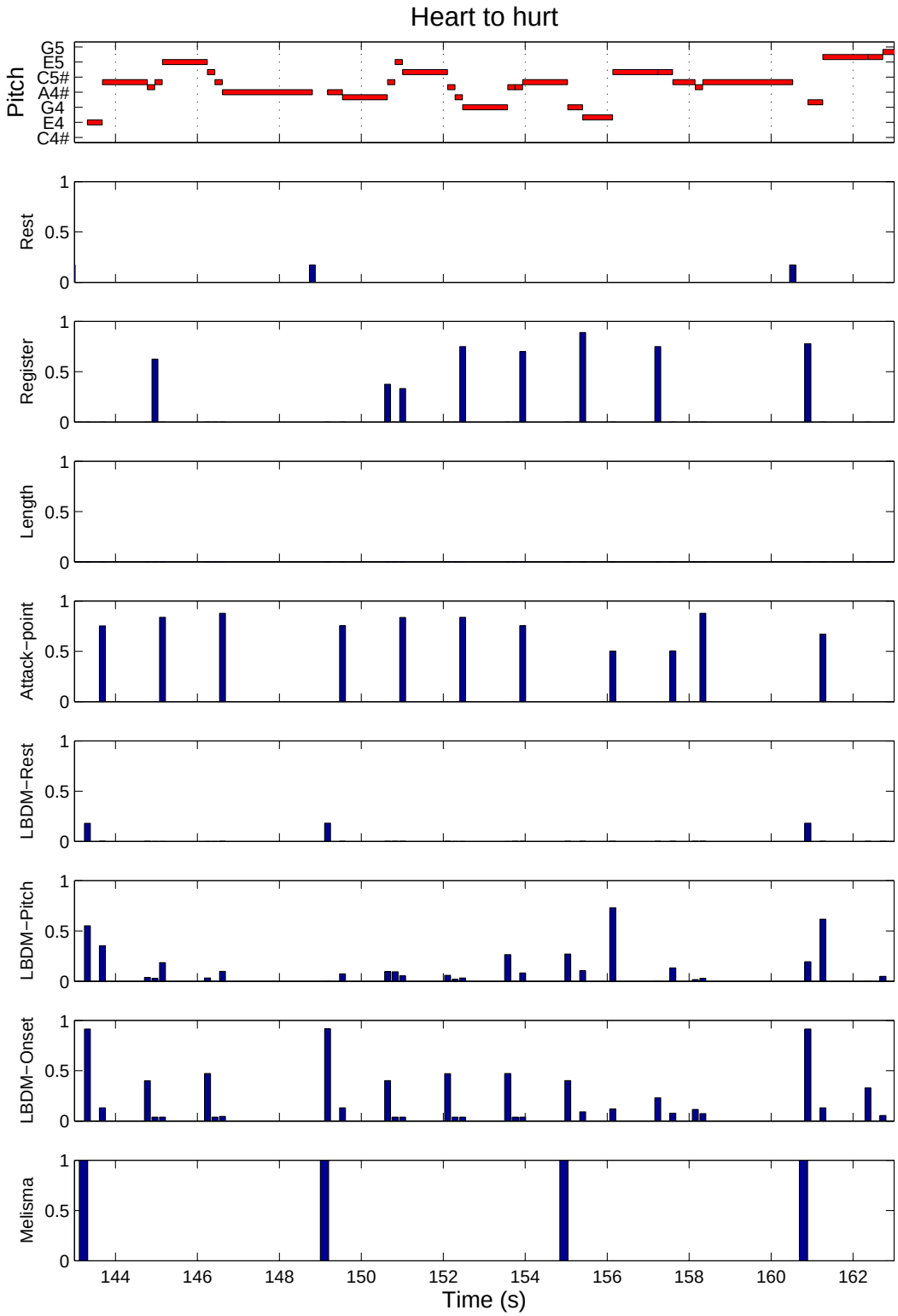


Figure C.1: An example of an excerpt of the song "Heart to hurt" with the corresponding predicted segment boundaries by the different cue implementations. The first row shows the pianoroll representation of the excerpt. The next four rows show the predicted boundaries of the quantified rules of GTTM. The following three rows show the three rules of LBDM. The last row shows the predicted boundaries of the Melisma model.

References

Bartsch, M. A., & Wakefield, G. H. (2001). To catch a chorus: Using chroma-based representations for audiothumbnailing. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (p. 15-18). New Platz, NY, USA.

Bharucha, J., & Krumhansl, C. L. (1983). The representation of harmonic structure in music: Hierarchies of stability as a function of context. *Cognition*, *13*, 63-102.

Bigand, E., Madurell, F., Tillmann, B., & Pineau, M. (1999). Effect of global structure and temporal organization on chord processing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 184-197.

Bigand, E., McAdams, S., & Forêt, S. (2000). Divided attention in music. *International Journal of Psychology*, *35*(6), 270-278.

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100-130.

Bod, R. (2002). Memory-based models of melodic analysis: Challenging the gestalt principles. *Journal of New Music Research*, *31*(1), 27-36.

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2006a). Perception of structural boundaries in popular music. In *Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9)* (p. 157-162). Bologna, Italy.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2006b). Structural boundary perception in popular music. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006)* (p. 198-201). Victoria, Canada.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2008a). The perception of structural boundaries in polyphonic representations of Western popular music. *Submitted for publication to Music Perception*.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2008b). Perceptual evaluation of formal musicological cues for automatic song segmentation. *To be submitted for publication*.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (in press). The perception of structural boundaries in melody lines of Western popular music. *Accepted for*

publication in Musicae Scientiae.

Cambouropoulos, E. (1997). Musical Rhythm: A Formal Model for Determining Local Boundaries, Accents and Metre in a Melodic Surface. In M. Leman (Ed.), *Music, Gestalt, and Computing* (p. 277-293). Berlin: Springer-Verlag.

Cambouropoulos, E. (1998). *Towards a General Computational Theory of Musical Structure*. Unpublished doctoral dissertation, University of Edinburgh, UK.

Cambouropoulos, E. (2001a). The Local Boundary Detection Model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC2001).* Havana, Cuba.

Cambouropoulos, E. (2001b). Melodic cue abstraction, similarity, and category formation: A formal model. *Music Perception*, *18*(3), 347-370.

Cambouropoulos, E. (2006a). Musical parallelism and melodic segmentation: A computational approach. *Music Perception*, *23*(3), 249-267.

Cambouropoulos, E. (2006b). 'Voice' Separation: Theoretical, perceptual and computational perspectives. In *Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9)* (p. 987-997). Bologna, Italy.

Chai, W. (2005). *Automated Analysis of Musical Structure*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Chuan, C.-H., & Chew, E. (2007). A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007).* Vienna, Austria.

Clarke, E. F., & Krumhansl, C. L. (1990). Perceiving musical time. *Music Perception*, *7*(3), 213-252.

Deliège, I. (1987). Grouping conditions in listening to music. *Music Perception*, *4*(4), 325-360.

Deliège, I., & Ahmadi, A. E. (1990). Mechanisms of cue extraction in musical grouping: A study of perception on Sequenza VI for Viola Solo by Luciano Berio. *Psychology of Music*, *18*, 18-44.

Deliège, I., Mélen, M., Stammers, D., & Cross, I. (1996). Musical schemata in real-time listening to a piece of music. *Music Perception*, *14*(2), 117-160.

Deutsch, D. (1999). Grouping mechanisms in music. In D. Deutsch (Ed.), *The Psychology of Music* (2nd ed., p. 299-348). New York: Academic Press.

Deutsch, D., & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review*, *88*(6), 503-522.

Dibben, N. (1994). The cognitive reality of hierarchic structure in tonal and atonal music. *Music Perception*, *12*(1), 1-25.

Eerola, T., & Toiviainen, P. (2004). *Midi toolbox: Matlab tools for music research.* University of Jyväskylä: Kopijyvä, Jyväskylä, Finland. (Available at http://www.jyu.fi/musica/miditoolbox/)

Ferrand, M. (2004). *Data-driven, memory-based computational models of human segmentation of musical melody.* Unpublished doctoral dissertation, School of Arts, Culture and Environment, University of Edinburgh, UK.

Foote, J. (1999). Visualizing music and audio using self-similarity. In *Proceedings of the seventh ACM International Conference on Multimedia (part 1) (MULTIMEDIA99)* (p. 77-80). New York, USA: ACM Press.

Foote, J., & Cooper, M. (2003). Media segmentation using self-similarity decomposition. In M. M. Yeung, R. W. Lienhart, & C.-S. Li (Eds.), *SPIE Storage and Retrieval for Multimedia Databases* (Vol. 5021, p. 167-175). San Jose, California, USA.

Frankland, B. W., & Cohen, A. J. (2004). Parsing of melody: Quantification and testing of the local grouping rules of "Lerdahl and Jackendoff's A Generative Theory of Tonal Music". *Music Perception*, *21*(4), 499-543.

Goto, M. (2006). A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech & Language Processing*, *14*(5), 1783-1794.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC music database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002)* (p. 287-288). Paris, France.

Hamanaka, M., Hirata, K., & Tojo, S. (2006). Implementing "A Generative Theory of Tonal Music". *Journal of New Music Research*, *35*(4), 249-277.

Itoyama, K., Goto, M., Komatani, K., Ogata, T., & Okuno, H. G. (2007). Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals. In *Proceedings of the 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)* (p. I-57-60). Honolulu, USA.

Janata, P., Tillmann, B., & Bharucha, J. J. (2002). Listening to polyphonic music recruits domain-general attention and working memory circuits. *Cognitive, Affective, and Behavioural Neuroscience*, *2*(2), 121-140.

Klapuri, A. P. (2004). Automatic music transcription as we know it today. *Journal of New Music Research*, *33*(3), 269-282.

Knösche, T. R., Neuhaus, C., Haueisen, J., Alter, K., Maess, B., Witte, O. W., et al. (2005). Perception of phrase structure in music. *Human Brain Mapping*, *24*, 259-273.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford: Oxford University Press.

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas. *Music Perception*, *13*(3), 401-432.

Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, *89*, 334-368.

Lalitte, P., Bigand, E., & Poulin-Charronnat, B. (2004). The perceptual structure of thematic materials in The Angel of Death. *Music Perception*, *22*(2), 265-296.

Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Logan, B., & Chu, S. (2000). Music summarization using key phrases. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)* (Vol. 2, p. 749-752). Istanbul, Turkey.

Lu, L., Want, M., & Zhang, H.-J. (2004). Repeating pattern discovery and structure analysis from acoustic music data. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR 2004)* (p. 275-282). New York, USA: ACM Press.

Maddage, N. C. (2006). Automatic structure detection for popular music. *IEEE MultiMedia*, *13*(1), 65–77.

Meyer, L. B. (1956). *Emotion and Meaning in Music*. Chicago: University of Chicago Press.

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: the influence of context and learning. *Music Perception*, *23*(5), 377-405.

Peeters, G., Burthe, A. L., & Rodet, X. (2002). Toward automatic music audio summary generation from signal analysis. In *Proceedings of the 3rd International*

Conference on Music Information Retrieval (ISMIR 2002). Paris, France.

Rauber, A., Pampalk, E., & Merkl, D. (2002). Using psycho-acoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarities. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR 2002).* Paris, France.

Repp, B. H. (1992). Probing the cognitive representation of musical time: Structural constraints on the perception of timing perturbations. *Cognition*, *44*, 241-281.

Schaefer, R. S., Murre, J. M., & Bod, R. (2004). Limits to universality in segmentation of simple melodies. In *Proceedings of the 8th International Conference on Music Perception & Cognition (ICMPC8)* (p. 247-250). Evanston, USA.

Schaffrath, H. (1995). The essen folksong collection in the humdrum kern format. In D. Huron (Ed.), *The humdrum kern format.* Menlo Park, CA: Center for Computer Assisted Research in the Humanities.

Schenker, H. (1935). *Neue Musikalische Theorien und Phantasien, Vol. III: Der Freie Satz.* (English translation 1979. "Free composition", New York, Longman)

Sloboda, J., & Edworthy, J. (1981). Attending to two melodies at once: The effect of key relatedness. *Psychology of Music*, *9*, 39-43.

Smith, J. D. (1997). The place of musical novices in music science. *Music Perception*, *14*(3), 227-262.

Spiro, N. (2007). *What contributes to the perception of musical phrases in western classical music?* Universiteit van Amsterdam: Doctoral Dissertation.

Spiro, N., & Klebanov, B. B. (2006). Application of a new method for consistency assessment and grouping of listeners real-time identification of musical phrase parts. In M. Baroni, A. R. Addessi, R. Caterina, & M. Costa (Eds.), *Proceedings of the 9th International Conference on Music Perception & Cognition (ICMPC9).* Bologna, Italy.

Tan, N., Aiello, R., & Bever, T. G. (1981). Harmonic structure as a determinant of melodic organization. *Memory and Cognition*, *9*(5), 533-539.

Temperley, D. (2001). *The Cognition of Basic Musical Structure.* Cambridge, MA: MIT Press.

Tenney, J., & Polansky, L. (1980). Temporal gestalt perception in music. *Journal of Music Theory*, *24*, 205-241.

Tillmann, B., & Bigand, E. (1998). Influence of global structure on musical target detection and recognition. *International Journal of Psychology*, *33*(2), 107-122.

Tillmann, B., & Bigand, E. (2001). Global context effect in normal and scrambled musical sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 1185-1196.

Tillmann, B., Bigand, E., & Madurell, F. (1998). Local versus global processing of harmonic cadences in the solution of musical puzzles. *Psychological Research*, *61*, 157-174.

Todd, N. P. M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*, *91*(6), 3540-3550.

Todd, N. P. M. (1995). The kinematics of musical expression. *Journal of the Acoustical Society of America*, *97*(3), 1940-1949.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung*, *4*, 301-350.

Perception and modeling of segment boundaries in popular music

---

Summary

While attending to music, the listener automatically perceives the structure in the piece, which is conveyed through cues embedded in the music. Such cues are, among others, changes in the musical surface, such as a pitch interval or change in the note duration, and repetition of previously played material. These cues have been proposed in music theoretical frameworks and have been partially assessed perceptually. The aim of the present thesis was to gain a better understanding of the processes involved in music structure analysis, in particular the perception of segment boundaries and to model the perceptual boundaries with cues taken from formal musicological models.

To investigate the perception of segment boundaries in music, a series of two-part experiments were conducted. In the first part of each experiment (the segmentation part), subjects were asked to segment six representations of pieces taken from Western popular music. In the second part of each experiment (the salience rating part) they were asked to rate the salience of selected boundaries from the first experiment and to describe the perceptual or musicological cues that they associated with each boundary. Both parts were performed three times using three different stimulus types, 1) the melodic line only, synthesized from the MIDI representation, 2) full polyphonic, synthesized from MIDI, and 3) the audio recording. All three representations were time-aligned. The results of each experiment were analyzed and comparisons were made between the two parts and across the three stimulus types, in order to evaluate the influence of polyphony on segmentation. Furthermore, the perceptual boundaries were compared with boundaries predicted by music-theoretical models.

The first part of the thesis deals with the results from the experiments that used the monophonic stimulus types. A method was developed to accumulate individual boundary indications to a global boundary profile. This is based on smoothing the indications with a Gaussian window of a duration between one and two seconds. The results of the segmentation experiment show that there is a wide range in the frequency with which boundaries were indicated. Previous research assumed a correlation between

the frequency with which a specific boundary was indicated and the explicit salience rating. Our results show that this correlation is indeed high, thus, both methods can be used to obtain an estimate of boundary salience. The descriptions of the boundaries were grouped into a number of cue classes to derive the perceptual cues responsible for specific segment boundaries. In addition to the number of times a certain cue was mentioned we developed an additional measure, called mean term rating, which relates the mean boundary salience rating to the assigned boundary cue descriptions. The cues "change in timbre" and "harmonic progressions" had the highest mean term rating and are thus likely to be important cues for segment boundaries. Furthermore, we took three musicological segmentation algorithms to segment the six pieces and compared the predictions with the perceptual boundaries. The segmentation algorithms were compared individually and with the additional cue of "change in timbre", which is not part of any of these models. The comparison showed that the smoothed boundary profile predicted by these models is moderately correlated with the smoothed perceptual boundary profile.

The second part of the thesis extends the first part by examining responses to polyphonic stimuli. The aim of the study was to evaluate the influence of polyphony on segmentation. The same experimental method was used as in the first part. Two different types of stimuli were selected, the synthesized score and the audio recording of the same songs used in the first part. Results show again a high correlation between the implicit salience measure, the number of boundary indications within a time window, and the explicit salience measure. Further findings are that, in addition to the cues found in the monophonic stimuli, subjects perceive additional cues, like "change in tempo" and "change in rhythm" as important boundary indicators. However, although subjects perceive the additional cues for polyphonic representation, the segmentation profiles and the salience ratings were similar across the three stimulus types. It seems, thus, that the boundary notations are little influenced by the stimulus type for music taken from Western popular music if the different representations are accurately time-aligned.

The last part of the thesis deals with comparing the boundaries predicted by segmentation models with our obtained perceptual boundaries. Instead of taking the complete models, the models were dissected into their individual cues and the contribution of each cue was studied separately. One analysis investigated if there was a time delay between the boundaries predicted by the model-cues and the perceptual

boundaries with the result that for the cues yielding the highest performance there was a constant delay across the six songs. A further study examined which combination of cues can best predict the perceptual boundaries. A set of three cues seemed sufficient to have a high correlation between the predicted boundary profiles and the perceptual boundary profiles. These three cues were "changes in timbre", the beginning of a rest, and a long note in between short notes. A final evaluation tested how well the predicted boundaries corresponded to the polyphonic boundary profiles. The results show that, although with a lower correlation than observed for the perceptual boundary profile for the monophonic MIDI stimuli, the predicted boundaries are also highly correlated with the perceptual boundary profiles of the polyphonic renderings of the pieces.

Acknowledgments

The subject of this Ph.D. project has been a change in what I have been studying previously. Coming from Computer Science and engaging in music perception is not the most obvious thing to do. Without the support of many persons this would not have been possible.

First of all, I'm grateful to Armin Kohlrausch for giving me the opportunity to do this research project, for his help and support. Next, I would like to thank Martin McKinney for his weekly discussions and useful comments on all aspects of this thesis. I have learned a lot from you both.

I am thankful to Emilios Cambouropoulos, Don Bouwhuis, Henkjan Honing, and Mark Sandler for giving feedback on the draft of my thesis and participating in my Ph.D. committee. Moreover, I thank all the people participating in my listening experiments. And I am grateful to Philips research for funding this project.

Furthermore I would like to thank Adrienne Heinrich and Nicolas Le Goff for being my "paranimfen" (seconds) as well as all my colleagues in the office, Tom Goossens, Alberto Novello, and Tobias May, for the good atmosphere, the higher noise level (Bob Marley), and all the nonsense. A special thanks to Othmar Schimmel, also a colleague from the office, who first helped me to settle down in Eindhoven and with all the administration of the university and later improved the quality of my articles through endless re-reading of preliminary versions of my articles. I enjoyed your jokes guys and your company.

Finally, I would like to thank my family as well as Karen and Ana for their constant encouragement and motivation. Moreover, I am thankful to all my friends. These were the people who kept me going when things were not as desired.

Curriculum Vitae

Michael Bruderer was born in 1978 and has Herisau/AR in Switzerland as his place of origin ("Heimatort"). After following school education in Switzerland and the United States he studied Computer Science at the Federal Technical Institute of Technology in Lausanne (EPFL). During his studies he participated in the ERASMUS student exchange program and studied for one year at the Escuela Téchnica Superior de Ingenerías Informatica y de Telecomunicación (ETSII) in Granada, Spain. He received his Master in Computer Science in March 2003, with his Master project being done at the Fraunhofer Institute for Digital Technology (IDMT) in Ilmenau, Germany with the title "Automatic Music Instrument Recognition". From May 2003 until December 2003 he worked as a freelancer for the same institution. In March 2004 he started his Ph.D. project at the Technical University of Eindhoven (TU/e). The project, financially supported by Philips Research Laboratories Eindhoven, had the goal of better understanding the perception and modeling of structural boundaries in popular music.