Author(s): Hartmann, Martin; Lartillot, Olivier; Toiviainen, Petri

Title: Interaction features for prediction of perceptual segmentation : Effects of musicianship and experimental task

Year: 2017

Version:

Please cite the original version:

Hartmann, M., Lartillot, O., & Toiviainen, P. (2017). Interaction features for prediction of perceptual segmentation : Effects of musicianship and experimental task. Journal of New Music Research, 46(2), 156-174. https://doi.org/10.1080/09298215.2016.1230137

Journal of New Music Research

Interaction Features for Prediction of Perceptual Segmentation: Effects of Musicianship and Experimental Task

Martín Hartmann · Olivier Lartillot · Petri Toiviainen

Abstract As music unfolds in time, structure is recognized and understood by listeners, regardless of their level of musical expertise. A number of studies have found spectral and tonal changes to quite successfully model boundaries between structural sections. However, the effects of musical expertise and experimental task on computational modelling of structure are not yet well understood. These issues need to be addressed to better understand how listeners perceive the structure of music and to improve automatic segmentation algorithms. In this study, computational prediction of segmentation by listeners was investigated for six musical stimuli via a real-time task and an annotation (non real-time) task. The proposed approach involved computation of novelty curve interaction features and a prediction model of perceptual segmentation boundary density. We found that, compared to non-musicians', musicians' segmentation yielded lower prediction rates, and involved more features for prediction, particularly more interaction features; also non-musicians required a larger time shift for optimal segmentation modelling. Prediction of the annotation task exhibited higher rates, and involved more musical features than for the real-time task; in addition, the real-time task required time shifting of the segmentation data for its optimal modelling. We also found that annotation task models that were weighted according to boundary strength ratings exhibited improvements in segmentation prediction rates and involved more interaction features. In sum, musical training and experimental task seem to have an impact on prediction rates and on musical features involved in novelty-based segmentation models. Musical training is associated with higher presence of schematic knowledge, attention to more dimensions of musical change and more levels of the structural hierarchy, and higher speed of musical structure processing. Real-time segmentation is linked with higher response delays, less levels of structural hierarchy attended and higher data noisiness than annotation segmentation. In addition, boundary strength weighting of density was associated with more emphasis given to stark musical changes and to clearer representation of a hierarchy involving high dimensional musical changes.

M. Hartmann
Finnish Centre for Interdisciplinary Music Research
Department of Music
University of Jyväskylä
E-mail: martin.hartmann@jyu.fi

O. Lartillot
Department of Architecture, Design and Media Technology
Aalborg University
E-mail: ol@create.aau.dk

P. Toiviainen
Finnish Centre for Interdisciplinary Music Research
Department of Music
University of Jyväskylä
E-mail: petri.toiviainen@jyu.fi

Keywords segmentation density, novelty detection, musical training, segmentation task, boundary strength

1 Introduction

Humans possess the ability to perceptually parse ongoing streams into discrete events. This perceptual operation, which is called segmentation, makes it possible to understand activities that involve sound and movement, just like it is possible, in a messy room, to recognize each of its objects (Zacks & Swallow, 2007). It has central importance, for instance, in the area of speech perception, as it is needed for language acquisition: infants exploit different speech segmentation cues to identify words in sequences of syllables and to recognize larger groupings such as clauses (Johnson & Jusczyk, 2001; Seidl, 2007). Similar but specialized psychological processes may apply to music listening, since musical events that share related characteristics or high temporal proximity are often grouped into sequences, even in passive listening contexts. This temporal psychological process of integrating musical events into larger units, which has been proposed to be universal (Drake & Bertrand, 2001), can be inversely formulated: listeners segment long musical streams when they perceive changes and repetitions. Musical feature change is a common cue for segmentation: listeners indicate segment boundaries if they easily perceive that there is a contrast, such as a stark change in dynamics or instrumentation. Multiple strategies are exploited by composers (Deliège, 2001), improvisers (Dean, Bailes, & Drummond, 2014) and performers (Poli, Rodà, & Vidolin, 1998) to induce perception of musical changes, and communicate musical structure to the listener. This paper focuses, however, on musical listeners only, and on a particular conception of segmentation. We refer to segmentation in its broad sense, as we understand perceptual segment boundaries as *significant instants of musical change*; implications of this choice are discussed further.

Listeners often indicate long notes and rests as segment boundaries during segmentation of songs (Bruderer, 2008); generally, temporal patterns upon which phrase and metrical units emerge have been deemed a crucial factor in the perception of musical structure (see Dawe, Plait, & Racine, 1994). Also melodic and harmonic changes, including pitch jumps, changes in register, and especially chord changes and modulations have been regarded to influence segmentation decisions. Tonality largely contributes to perceived musical structure, because unimportant events in a tonal hierarchy generate expectations of musical relaxation that are often confirmed when more important events evoke resolution (Bigand, Parncutt, & Lerdahl, 1996). Both metrical structure position and tonal hierarchy are considered to define the relative importance of certain musical events with respect to others within a given time span (Lerdahl & Jackendoff, 1983), and may have an impact on one another: musicians tend to infer metrical structure on the basis of chord changes when note duration and harmony imply different meters (Dawe et al., 1994). In this sense, boundary perception results from an intertwined mix of musical feature changes and it can be challenging to disentangle the contribution of different aspects of segmentation, especially for real-world music.

Music information retrieval (MIR) studies have proposed a variety of automatic segmentation algorithms with a focus on evaluating model performance against ground truth data using accuracy measures such as precision, recall and F-measure (for instance Aljanaki, Wiering, & Veltkamp, 2015); few studies in this area (e.g. Jensen, 2007) have systematically assessed the relevance of different musical features for segmentation. In most cases, automatic segmentation of music in audio format is done via *novelty* detection (Foote, 1997, 1999, 2000) approaches, which roughly consist in the extraction of frame-decomposed musical features and the computation of novelty curves. These curves describe, for each time point, the amount of dissimilarity between a certain number of feature frames before and after that point. For instance, points in the music that are characterized by tonal change would show high novelty for the tonal features.

The potential of combining different acoustic features for segmentation and structural analysis has been mentioned in MIR studies (Turnbull, Lanckriet, Pampalk, & Goto, 2007; Paulus & Klapuri, 2009). Few novelty-based studies (Paulus & Klapuri, 2009; Eronen, 2007; Peeters, 2007) have yielded enhanced automatic structural analyses via the summation of spectral and chroma features; this operation can be considered as a logical disjunction (**OR**), because changes of either or both spectral and chroma features would result in novelty peaks. To our knowledge, no studies in this area have implemented logical conjuction

(**AND**) operations, which would yield novelty peaks only after concurrent change of both features. For example, an interaction feature resulting from a spectral novelty curve and a chroma novelty curve would not register a given spectral change unless it was accompanied by a simultaneous chroma change, and vice versa. From a computational perspective, such a novelty feature interaction approach seems appropriate because it can reduce the effect of spurious novelty peaks derived from high feature sensitivity; it may also be relevant from a perceptual viewpoint, since listeners probably pay most attention to changes that are evoked by more than one musical dimension (see Smith, Schankler, & Chew, 2014).

For evaluation purposes, novelty peaks are compared to the ground truth data, which often involve a set of isolated time points; MIR studies on this area are typically based on a large number of stimuli, so ground truth segmentation data is obtained from at most few annotators (e.g. Smith, Burgoyne, Fujinaga, De Roure, & Downie, 2011). In contrast to MIR ground truth data, studies focusing on listeners' perception of boundaries often collect data from many participants and aggregate their boundary indications (Deliège, 1987; Krumhansl, 1996; Frankland & Cohen, 2004). To maximize estimation accuracy, recent studies (Bruderer, 2008; Burunat, Alluri, Toiviainen, Numminen, & Brattico, 2014; Hartmann, Lartillot, & Toiviainen, in press) have used Kernel Density Estimation (Silverman, 1986), a method that generates a smooth probability density estimate of the data via a Gaussian or other kernel function. This procedure is comparable to drawing a histogram, where each bin would aggregate listeners' responses within a temporal region; roughly, Kernel Density Estimation is like a histogram that is smoothed into a curve. This approach yields more accurate representations of segmentation and allows to perform group comparisons, for instance between musicians and non-musicians.

Musical experience seems to have an impact on listeners' focus of attention during music listening and on their representation of structure. Non-musicians are often considered to pay more attention to aspects related to the musical surface; they often tap with the fastest pulse during finger tapping tasks (Martens, 2011), and tend to place more boundary indications than musicians in segmentation studies (Hartmann et al., in press; Bruderer, 2008; Deliège, 1987), suggesting that non-musicians focus more on changes in timbre, fast rhythmic layers, and pitch jumps. Most research has found that non-musicians focus less on harmonic functions than musicians, for instance in a task that consisted in rearranging musical segments, non-musicians paid more attention to rhythmic and metric aspects than to tonality (Deliège, Mélen, Stammers, & Cross, 1996). Moreover, a rhythm identification study showed that musicians' perception of rhythmic patterns for temporal sequences with harmonic accompaniment was more influenced by location of chord changes than non-musicians', whose answers were less consistent, and biased towards responses that fitted the inferred meter (Dawe, Platt, & Racine, 1995). Based on these findings, it could be posited that non-musicians' segmentation can be more accurately predicted from the audio signal than musicians'; musicians would pay also attention to deeper aspects such as tonal context, which cannot be accurately modelled since they are rooted on implicit knowledge of Western tonal hierarchies. Other studies on processing and perception of musical structure (see Tillmann & Bigand, 2004) however suggest that schematic knowledge (see Justus & Bharucha, 2001) is built through mere exposure to music, as both groups focused on musical surface and deeper aspects of structure during tasks involving harmonic priming and manipulation of global organization of pieces. Hence, it becomes unclear if musically trained listeners are more influenced by schematic expectancies during segmentation than untrained listeners or, conversely, if for both groups few musical events suffice to generate accurate forecasts about mode or upcoming chords in the music (Tillmann & Bharucha, 2002). Thus far, no studies have investigated the prediction of musicians' and non-musicians' segmentation, nor systematically examined whether or not these groups pay attention to same or different acoustic features during segmentation tasks. A deeper understanding on how musical training shapes our perception and understanding of structure and an examination of what musical dimensions listeners are attending to are needed in order to gain further insights on how musical structure is processed.

Boundary perception is affected by musical expectancies; some boundaries are easier to anticipate as music temporally unfolds in real-time, whereas others are totally unexpected percepts. Listening to the whole stimulus has been posited to provide a better understanding of the musical structure because some boundaries cannot be per-

ceived until they occur, or are perceived retrospectively, i.e. ulterior to the actual musical change (Lerdahl & Jackendoff, 1983). In this respect, different methods to gather segmentation responses from participants have been used in studies on musical structure processing. Hartmann et al. (in press) found differences between real-time and non-real time segmentation in boundary density, number of boundary indications (more boundaries in the annotation task than in the real-time task), optimal segmentation time scales, and also a time lag between tasks; these differences were attributed to the inaccuracy of real-time task data, which contains delayed or "missed" indications, especially for boundaries that are only perceived retrospectively. If annotation tasks are less noisy, they should be more accurately predicted by segmentation systems; probably due to this assumption, annotation task data seems to be regarded as a more reliable ground truth for evaluation of MIR segmentation systems. However, to our knowledge no studies have compared real-time and annotation segmentation tasks with regards to their predictability from the audio signal content. It would be important to shed more light on this possible difference between tasks, because collection of segmentation data from listeners is lengthy, particularly when it comes to annotation tasks; also, both experimental tasks are used (e.g. real-time segmentation is common in brain and music studies) so it would be beneficial to know whether or not they yield similar models to better understand how musical structure is processed.

The third issue, which is related with the previous one, is about perceived boundary strength, its relationship with boundary density and its acoustic basis. Boundary strength ratings seem to be associated to listeners' preference towards certain types of grouping of musical events; for instance, short melodic sequences including contour changes or gaps (e.g. rests) tend not to be heard as groups (Lerdahl & Jackendoff, 1983; Deliège, 1987), but gaps are perceived as stronger boundaries than changes in melodic contour (Deliège, 1987; Clarke & Krumhansl, 1990). It has been also found that listeners generally agree about which musical boundaries are perceived as strongest (Clarke & Krumhansl, 1990). Further, Bruderer (2008) found a positive relationship between the mean strength ratings of a boundary across participants and its relative frequency of indications. This suggests that boundary strength ratings can be estimated from listeners' boundary density; in other words, boundary strength ratings are superfluous data in segmentation tasks involving multiple participants. Hartmann et al. (in press) could not replicate Bruderer's result, suggesting that boundaries perceived as strong are not necessarily more likely to be indicated and vice versa. On top of that, it is currently neither known whether or not weighting boundary density according to boundary strength ratings would have an effect on prediction of segmentation, nor what would be the direction of this effect. Tackling this issue would help clarify what boundary strength ratings inform about perceived musical structure, and what is their relationship with local boundary density and local musical contrast. In particular, it would be interesting to better understand what aspects of musical change are associated to perceived boundary strength in real-world music.

Recently, Hartmann et al. (in press) investigated effects of musicianship, differences between real-time and annotation segmentation tasks, and optimal time scales for comparison between segmentations. This study can be considered a follow-up to Hartmann et al. (in press), because the same boundary data and methodology for aggregation of indications is applied in this study. Our main goal is to investigate prediction of perceptual segmentation, and further study the effect of musicianship and experimental task on segmentation. Due to the complexity of this psychological process, we focused mainly on the study of segment boundaries that are prompted by significant instants of musical change. This paper attempts to shed light on the following research questions:

- To what extent does musicianship affect segmentation, and more specifically, how does computational prediction of segmentation for musicians differ from that of non-musicians?
- What is the effect of experimental task on segmentation, particularly on the modelling of real-time and non real-time segmentation tasks?
- Related to the previous question, what is the contribution of perceived boundary strength ratings on prediction of non real-time segmentation?

As a first hypothesis, we expected to find an effect of musicianship on model prediction, as non-musicians should be more accurately predicted by the segmentation models: they would focus more on perceived local acoustic changes, which could be accurately detected via novelty-based methods. Musicians would instead segment more based on

other aspects, such as learned musical schemata, and find relatively irrelevant surface events to be context and cues for ulterior changes that may be much more significant. Also, more features were expected to be involved in musicians' prediction, particularly more interaction novelty features, because musicians would pay attention to more musical dimensions and to co-occurring feature changes at multiple levels of the musical structure. In addition, we expected smaller response delays for musicians than for non-musicians due to extensive training on sense of timing cues.

Our second hypothesis is that the experimental segmentation task used for data collection has an effect on model prediction rates. We expected the real-time task segmentation to be less accurately predicted because the high cognitive load of the task would lead to imprecise, redundant and missing boundary indications; for instance, real-time tasks should pose difficulties to indicate boundaries as soon as these are perceived, leading to delayed or "missed" boundary indications. In addition, the annotation task prediction models would involve a higher number of musical features, since listeners would have the possibility to focus on more levels of the structural hierarchy, whereas the cognitive load required to complete the real-time task would bias listeners towards a single level. Also, while the annotation task would require little or no time shifting of the boundary data for its optimal modelling, real-time task modelling would benefit from compensation for response delays.

A third hypothesis, connected with the previous one, is that weighting the annotation task according to perceived boundary strength has an effect on model prediction. Boundary strength ratings would yield an increase in segmentation prediction rates because they should describe the amount of perceived musical change more accurately than boundary density. These ratings are likely to correspond with the magnitude of feature discontinuity; for instance, musical boundaries perceived as stark may yield high novelty values because both would stem from discontinuity of musical features. In addition, prediction of models weighted according to boundary strength ratings should involve more novelty interaction features, because strength ratings should describe concurrence of different musical novelty descriptions; in other words, listeners should indicate high strength for boundaries that involve high dimensional musical change, so interaction features should highly contribute to the prediction of strength-weighted segmentation density.

2 Method

The first phase of the experimental design consisted in conducting two listening experiments, a real-time task, and a non real-time task called annotation task. A more thorough description of the experimental procedure, musical stimuli and recruited participants can be found in Hartmann et al. (in press). From the segmentation data collected in these experiments we derived segmentation density curves, which in turn were computationally modelled in a second phase of the design. Figure 1 illustrates the design of this study and highlights the approach used to computationally model the perceptual data.

2.1 Experiment I: Real-time Task

2.1.1 Subjects

18 musicians (11 males, 7 females) and 18 non-musicians (10 females, 8 males) participated in the experiment. The mean age of non-musician participants was 27.28 years (SD = 4.64) and for musicians it was 27.61 years (SD = 4.45). The subjects were local and foreign university students and graduates. The average musical training of musicians was 14.39 years (SD = 7.49); all non-musicians reported being musically untrained.

2.1.2 Stimuli

We used 6 stimuli of around 2 minutes of duration that were relatively unfamiliar to participants and comprised a variety of styles (see A.1); the stimuli considerably differ from each other in terms of musical form, and emphasize aspects of musical change of varying nature and complexity.

2.1.3 Apparatus

The listening experiment interface was designed using Max/MSP; it presented the stimuli through headphones and involved the use of keyboard and mouse to record listeners' responses. The interface included a play bar to show listeners the relative duration of the stimulus and the current time position; each boundary indication triggered a visual feedback.

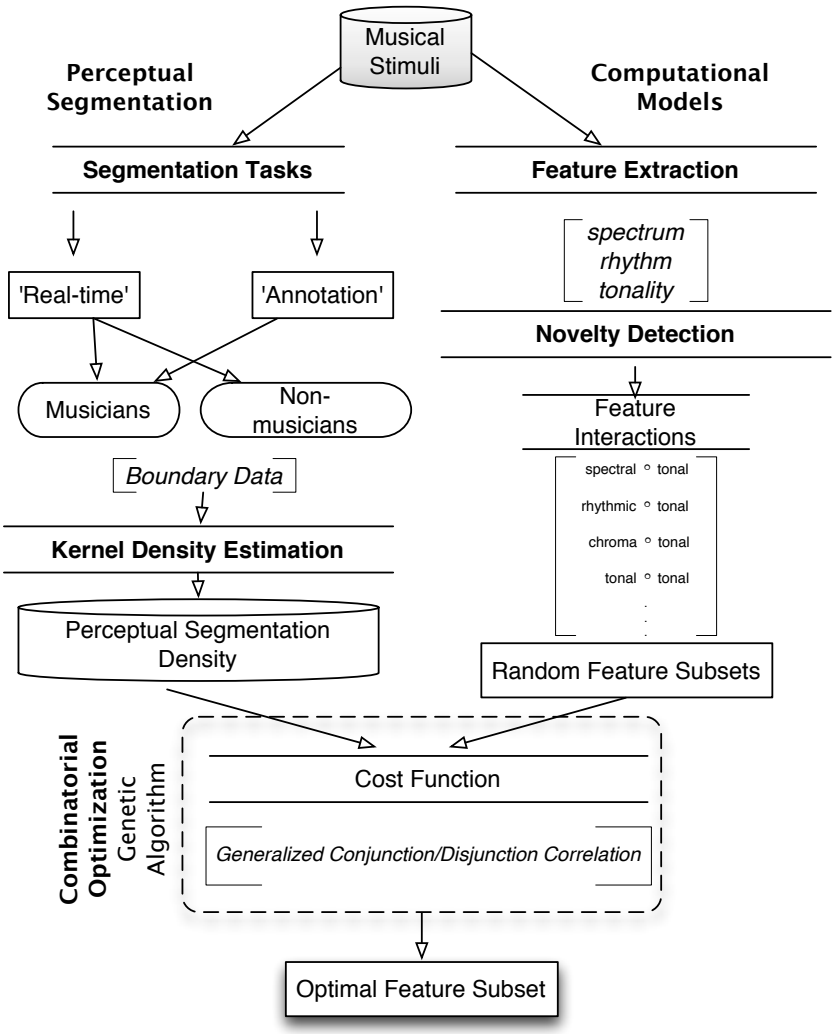

Figure 1: General design of the study

2.1.4 Procedure

Participants were asked to indicate significant instants of change while listening to the music by pressing the space bar key of the computer; the stimuli were presented in random order. For each participant and stimulus the boundary data was recorded in a single pass: they neither had the chance to listen to the stimuli before the segmentation nor were able to modify their boundary indications after the task. The task instructions were as follows: "Your task is to mark instants of significant musical change by pressing the space bar of the computer keyboard. Whenever you find an instant of significant change, please press the space-bar key to mark it as you listen to the music. You will not have a chance to listen to the whole example before you start marking. Instead, during your first and only listen of each example, you will give us your 'first impression"'.

2.2 Experiment II: Annotation Task

2.2.1 Subjects

After Experiment I, we asked all participants if they were familiar with editing software, and while all musicians mentioned having some experience, only four non-musicians expressed familiarity. Since this familiarity was required for the annotation task, we only recruited musicians for Experiment II; all of them had participated in Experiment I.

2.2.2 Stimuli

In this task we utilized the same set of stimuli as in Experiment I.

2.2.3 Apparatus

We used Sonic Visualizer (Cannam, Landone, & Sandler, 2010) to obtain segmentation boundary indications and also ratings of boundary strength. The interface included waveforms of the stimuli to offer visual-spatial cues for indicating boundaries and edit their time locations. The music was played back via headphones, and both keyboard and mouse were used to complete the task.

2.2.4 Procedure

In this task participants were first asked to listen to the whole stimulus. Then, they would listen to the stimulus again and indicate instants of significant change at the same time, just as they had done in the real-time task. Next, they were free to playback from different parts of the stimulus and make their segmentations more precise by adjusting the position of boundaries. In this step, listeners could remove boundaries if these were indicated by mistake. To avoid the tendency to oversegment the stimuli (following Krumhansl, 1996) participants could not add any new boundaries at this stage. Finally, the last step was to rate the perceived strength of each boundary. Since the stimuli waveforms shown in the interface could bias listeners towards segmentation based on amplitude changes, they were verbally asked to focus on the music rather than on visual content. The instructions included a presentation of the segmentation interface and the following task description:

1. Listen to the complete musical example.
2. Listen to the complete example, and at the same time mark instants of significant change by pressing the Enter key.
3. Freely playback the musical example from different time points and correct marked positions to make them more precise, or remove them if these were added by mistake. Do not add any new marks at this stage.
4. Mark the strength of the significant change for each instant with a value ranging from 1 (not strong at all) to 10 (very strong).
5. Move to the next musical example and start over from the first step.

2.3 Perceptual Segment Boundary Density

For each stimulus, the collected boundary indication data from different listeners was aggregated into a perceptual segmentation density curve for each participant group and segmentation task. First, we organized the segmentation data into three groups: musicians in the real-time task, non-musicians in the real-time task, and musicians in the annotation task. Next, we aggregated boundary indications from all participants so as to obtain a single profile of indications per stimulus. Subsequently, we concatenated the boundary data from each stimulus to obtain three boundary profiles spanning a duration of 12 min 5 s each. For each profile we derived a time series of density of segmentation. These segment boundary probability curves were obtained via Kernel Density Estimation (KDE). This approach is illustrated in Figure 4 (upper plot), where segmentation density peaks in the curve imply that multiple participants indicated boundaries at relatively close time points. The amount of closeness required between two boundary indications for them to be represented by the same density peak is defined by the time scale parameter τ, which corresponds to the bandwidth of the KDE Gaussian kernel; in other words, this parameter determines the degree of smoothness of the KDE. We chose a segmentation density time scale of $\tau = 1.5$ s following previous studies that focused on the Gaussian kernel bandwidth for modelling perceptual segmentation (Befus, 2010; Bruderer, 2008); particularly, Hartmann et al. (in press) found a mean optimal time scale for comparison between real-time and annotation task boundary density curves at 1.4 seconds and a mean optimal time scale for comparison between musicians' and non-musicians' boundary density at 1.7 s. The sample rate of the KDE was set to

10 Hz since it was deemed sufficiently accurate for point process data of this nature. Besides the three obtained segmentation density curves, the annotation task data was also modelled taking into account listeners' boundary strength ratings, yielding a weighted boundary density curve. In total, we obtained four curves describing probability density estimates of the boundary data: boundary density for non-musicians in the real-time task (*NMrt*), musicians in the real-time task (*Mrt*), musicians in the annotation task (*Ma*) and musicians in the annotation task with added boundary strength weights (Ma_w).

2.4 Computational Modelling

The structure of the 6 audio stimuli used in the experiments was systematically analyzed via a computational approach based on musical novelty detection that is illustrated in Figure 1 (right side) and Figure 2. Computational models of perceptual segmentation density curves were obtained to estimate the relative predictability of these curves and study which musical features were involved in the prediction.

2.4.1 Feature Extraction

This stage of the experimental design included extraction of musical features from the audio stimuli using MIRtoolbox (Lartillot & Toiviainen, 2007a). We extracted 5 features describing timbre, rhythm, pitch class and tonal context (see A.2). These features were frame-decomposed, in the sense that they were computed on short time frames along the audio stimuli.

2.4.2 Novelty Detection

For each of the features and stimuli, a novelty curve was obtained; to this end, a dissimilarity matrix is first obtained from the audio feature of interest by computing the Euclidean distance between all possible pairs of points in the time series. This matrix is inverted element-wise into a similarity matrix, where important local contrast around the main diagonal represents high dissimilarity between neighboring events (Figure 2). A novelty curve is subsequently obtained via convolution with a Gaussian checkerboard kernel across the main diagonal of the similarity matrix (see Foote, 2000; Lartillot & Toiviainen, 2007a; Paulus, Müller, & Klapuri, 2010, for detailed explanation). The Gaussian checkerboard kernel is illustrated in Figure 3. For each time point t, a novelty value is determined based upon the similarity between the Gaussian checkerboard kernel (centered at t) and the portion of the similarity matrix that is covered by the kernel. The width of this kernel, here understood as the span of the kernel to both directions from the reference point, is a crucial parameter in novelty detection. This is because it determines the smoothness of the novelty curve: larger widths produce smoother representations, and vice versa. To find an optimal novelty kernel parameter we obtained checkerboards for widths ranging between 0.5 s and 13 s in steps of 0.5 s. Next, we concatenated the novelty curves of each stimulus and obtained a time series of 12 min 5 s for each combination of feature and novelty width. In total, we obtained 5 novelty features for each of the 26 novelty widths considered; these are hereafter called basic features (e.g., novelty based on chromagram).

Subsequently, we created 10 interaction features that resulted from the pairwise interaction of basic features; for example, we obtained spectral-tonal, rhythmic-tonal, chroma-tonal and tonal-tonal features. This was done via point-by-point multiplication between each pair of novelty features (Figure 2). Using this method, we obtained for instance a curve via pairwise multiplication between novelty based on fluctuation patterns and novelty based on chroma, which would be called a rhythmic-chroma feature.

To compare novelty features extracted from the audio with boundary density of participants, both basic and interaction novelty features were resampled to 10 Hz to match the length of the boundary density curves; also, the novelty curves were normalized to sum 1. Altogether, we computed a total of 15 novelty features for each of the novelty widths.

2.4.3 Optimal Checkerboard Kernel Width

Next, we examined the relationship between novelty curves at different Gaussian checkerboard kernel widths and segmentation density. The aim was to evaluate segmentation models that would be most comparable to the obtained segmentation density. Boundary density was correlated with each of the novelty curves to find a checkerboard kernel width that would yield segmentation models with optimal prediction rates.

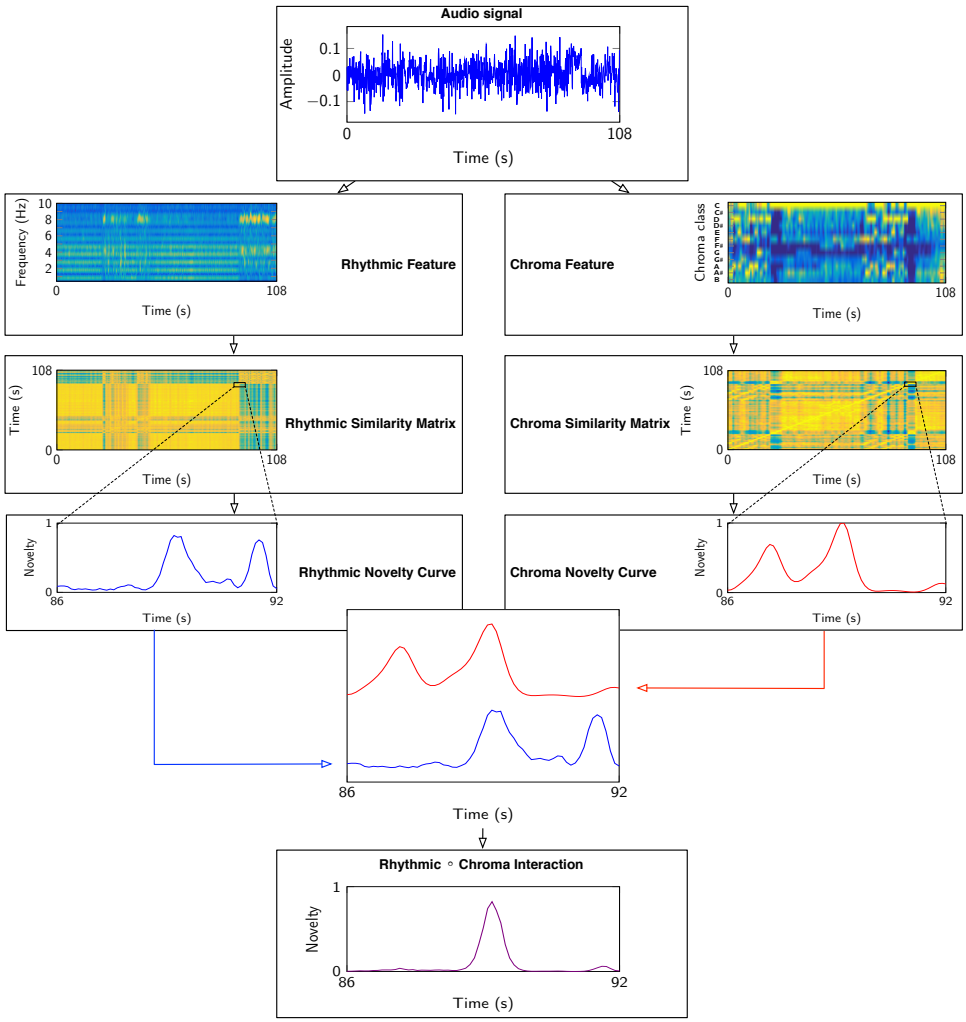

Figure 2: Method used to obtain interaction features via pairwise multiplication between novelty curves.

2.4.4 Non-linear Modelling

We investigated the prediction of perceptual data from combinations of novelty curves via a non-linear modelling approach. The approach consisted in finding a subset of novelty curves whose 50th percentile (median ordinal position) would optimally correlate with the segmentation density curve. This procedure involves a non-linear aggregation of novelty features that assigns weights to features for each time point based on ranked values. From the perspective of soft computing,

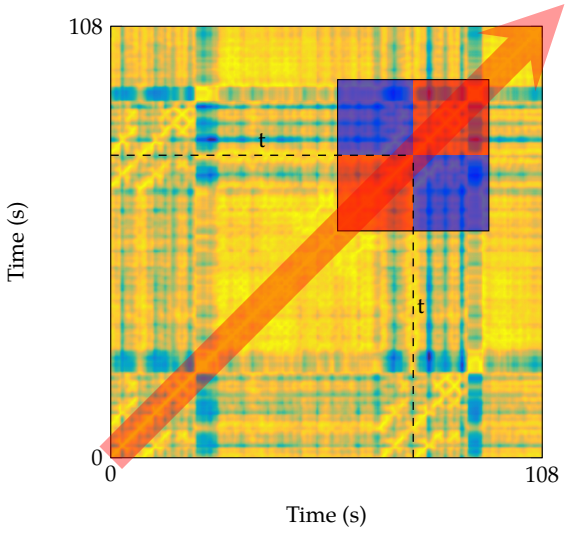

Figure 3: Convolution of a Gaussian checkerboard kernel along the main diagonal of a chromagram-based similarity matrix.

the percentile aggregation involves a monotonically increasing mapping that follows a continuous logic function called conjunction/disjunction function (Dujmović & Larsen, 2007). Roughly, the 0th percentile (equivalent to the min function) can be understood as a pure logical **AND** conjunction ("all criteria are satisfied") because if the minimum among features is high, then all features should have high values; conversely, the 100th percentile (max function represents pure **OR** disjunction ("at least one criterion is satisfied") because a high maximum value among features implies that at least one of the features has a high value. Following this logic, 1th–99th percentiles lie between **AND** and **OR**, exhibiting varying levels of *orness* (closeness to maximum). Hence, taking the 50th percentile across features would be comparable to a 'majority judgement', because it would only result in high values if at least half of the features exhibited high values. Several statistics, including arithmetic mean, median, min, max and percentile belong to the family of ordered weighted averaging (OWA) operators (Yager, 1988, 2006), but have different characteristics; for example in arithmetic mean aggregation all data elements get equal weights, whereas percentiles use only one argument to determine the aggregated value (for an odd number of arguments).

2.4.5 Combinatorial Optimization

In order to find an optimal subset of features for computational modelling we performed discrete combinatorial optimization. Via this approach we searched for a combination of novelty features whose percentile-based model would yield highest prediction rates, i.e. maximum correlation with the perceptual segmentation density. A generalized conjunction/disjunction correlation was used as a cost function criterion within a combinatorial optimization routine. The cost function finds the optimal value of the correlation coefficient y by minimizing the negative of the correlation between actual and predicted density, $y_{opt} = \underset{y}{\text{argmin}} - \text{corr}(x, p_\alpha)$, where x is the segmentation density and p_α is the α-percentile along features of a given subset. The reason for using combinatorial optimization was the high number of possible feature combinations per perceptual segmentation density curve (2^{15}). We used a Genetic Algorithm search heuristic to find an optimal feature subset for each perceptual segmentation density curve. The optimization cost function was initialized with a random subset of features. Since the algorithm employs a stochastic selection at each iteration, it tends to avoid local optimal solutions, i.e. subsets that are only best within the context of neighboring combinations. As a result, we obtained for each segmentation density curve an optimal percentile model, the correlation between these two curves and an optimal subset of features for computing the model. Correlation p-values (H0: *no correlation between observed and predicted segmentation density*) were obtained via Fisher's z-transformation of r, with standard scores adjusted

for effective degrees of freedom (i.e., corrected for temporal autocorrelation, see Pyper & Peterman, 1998; Alluri et al., 2012).

3 Results

We conducted three main analyses via the proposed experimental design: a comparison between perceptual segmentation sets based on model prediction rates, an examination of the novelty features involved in the prediction models, and an assessment of the model prediction rates for time lagged perceptual segmentation density. Figure 4 illustrates the main outcomes of the approach: for non-musicians' segmentation of 2 min 20 s of music (stimulus *Dvořák*) in the real-time task, it compares boundary indication data, perceptual segmentation density, selected novelty features and computational model prediction.

3.1 Novelty Kernel Width

To find accurate novelty curves for computational modelling, we initially examined the effect of modifying their kernel widths. To this end, we computed correlations between segmentation density curves and novelty curves for each of the 26 novelty widths obtained. Figure 5 shows the correlation profiles of the novelty features for each segmentation density curve. The global maxima of each curve, highlighted with markers, tend to be situated at large novelty widths in all cases. To find an optimal kernel width for further prediction of segmentation density, we computed a mean optimal novelty width across curves for each of the 4 segmentation densities, and finally a mean novelty width across segmentation densities. Via this method we found an optimal width of 11 s across novelty features and segmentation density curves (please refer to A.3 for correlation values at this width). We also obtained z-values for these correlation profiles to estimate significance of correlation, although a figure is not included for succinctness; z-values around 4, indicating significant results at the $p < .001$ level.

We further tested if a novelty width of 11 s would be appropriate for prediction of density. The mean temporal distance between peaks of each density curve was estimated; given the results of the aforementioned correlations, we expected that this distance would be around 11 seconds. For each density curve, we picked each time point that had a larger density value than its two neighboring time points and than 20% of the maximum density value in the curve. We found that the temporal distance between peaks in the density curves tended to be about as large (*NMrt*: 13.07 s ±8.16 SD; *Mrt*: 12.82 s ±8.72; *Ma*: 10.13 s ±7.33; *Maw*: 11.27 s ±8.90) as the optimal novelty kernel width. The requirement of a minimum peak height was used to disregard peaks with very low density values, since these would correspond to indications from few listeners. Without this restriction, the temporal distance between peaks was still relatively large (*NMrt*: 8.19 s ±2.97 SD; *Mrt*: 9.46 s ±4.64; *Ma*: 7.54 s ±3.04; *Maw*: 7.86 s ±3.10).

Comparing density curves, Figure 5 shows that the annotation task density curve with added weights tended to yield the highest correlations for most features. Adding weights to the annotation task lead to an increase in correlation (with respect to *Ma*) for all but three features when using a novelty width of 11 s (A.3). A possible reason for this correlation increase could have been the larger variance of the boundary density in the annotation task with added weights, which might have increased similarity with novelty curves due to their high variance. If the increase in correlation was the result of simply adding variance to the boundary density via addition of weights, then a random set of weights would be likely to yield a density curve that would result in increased correlation with respect to the weighted annotation task density. To test this possibility, we performed a Monte Carlo permutation (20000 iterations). At each iteration, 1) a random vector of boundary weights (between 1 and 10) of length equal to the number of boundary indications in the annotation task was generated, and a kernel density curve of the annotation task that included the random vector of weights was correlated with each of the 15 novelty curves. This resulted in a correlation distribution per novelty feature; for each distribution, the sum of the values that were equal or higher than the correlation reported in the study (A.3) for *Maw* was divided by the length of the distribution. Features that showed an improvement after adding weights to the annotation task tended to yield correlations for *Maw* that were unlikely to be reached by using a random set of strength weights ($p < .001$ for 8 features; $p < .01$ for 1 feature; $p > .05$ for Key Strength, Chromagram ∘ Key Strength, and Key Strength ∘ Tonal Centroid). This suggests that higher variance of the bound-

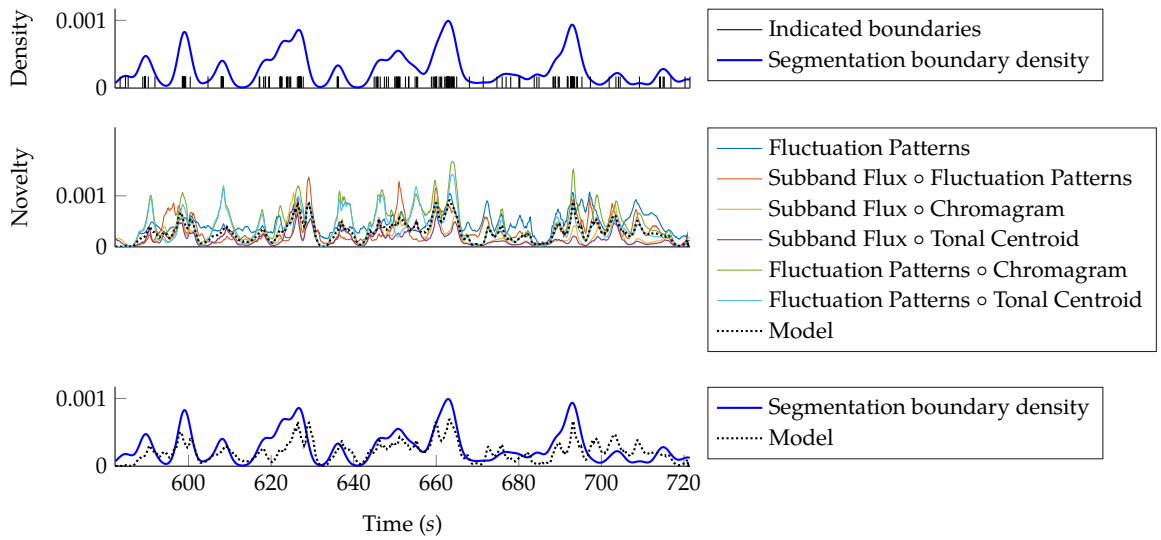

Figure 4: Perceptual segment boundary density and computational segmentation model for non-musicians in the Real-time task (stimulus Dvořák). Upper graph: Boundary indication data and segmentation boundary density. Middle graph: Model predictors and computational model prediction. Lower graph: Perceptual segmentation boundary density and computational model prediction. The model was computed using a time lag of 1.7 s.

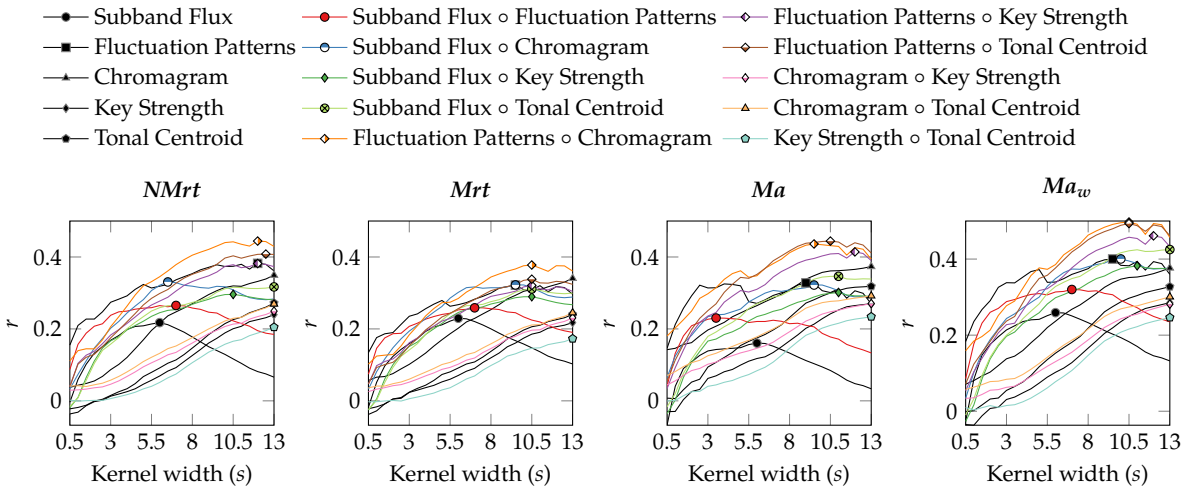

Figure 5: Correlation between perceptual segment boundary density and novelty curves at novelty widths ranging from 0.5 s to 13 s. Maximum points for each curve are highlighted with markers.

ary density was probably not an important factor in the correlation increase obtained from listeners' boundary strength ratings; in other words, boundary strength ratings from listeners include relevant information that lead to an increase of segmentation prediction accuracy.

3.2 Model Prediction Rates

We next examined the prediction obtained from novelty-based computational models for different participant groups and segmentation tasks. To achieve this, we performed combinatorial optimization using generalized conjunction/disjunction correlation as a cost function. We further investigated at this stage the novelty kernel width parameter by obtaining 26 computational models at varying novelty widths. Figure 6 shows that prediction rates tend to increase as a function of novelty but gradually reach a plateau; novelty curves based on a kernel width of 11 s yielded the highest overall prediction rates. Table

	NMrt	Mrt	Ma	Maw
Subset	Fluct. Pat. Chromagram Tonal Centroid Subband Flux ∘ Fluct. Pat. Fluct. Pat. ∘ Chromagram	Fluct. Pat. Key Strength Subband Flux ∘ Fluct. Pat. Subband Flux ∘ Tonal Centroid Fluct. Pat. ∘ Chromagram	Subband Flux Fluct. Pat. Tonal Centroid Subband Flux ∘ Tonal Centroid Fluct. Pat. ∘ Chromagram Fluct. Pat. ∘ Tonal Centroid	Fluct. Pat. Tonal Centroid Subband Flux ∘ Fluct. Pat. Subband Flux ∘ Tonal Centroid Fluct. Pat. ∘ Chromagram Fluct. Pat. ∘ Tonal Centroid
Category	Rhythmic Chroma Tonal Spectral ∘ Rhythmic Rhythmic ∘ Chroma	Rhythmic Tonal Spectral ∘ Rhythmic Spectral ∘ Tonal Rhythmic ∘ Chroma	Spectral Rhythmic Tonal Spectral ∘ Tonal Rhythmic ∘ Chroma Rhythmic ∘ Tonal	Rhythmic Tonal Spectral ∘ Rhythmic Spectral ∘ Tonal Rhythmic ∘ Chroma Rhythmic ∘ Tonal
r	.47***	.43***	.48***	.56***

***$p < .001$

Table 1: Correlations between perceptual segmentation density and computational models' predictions obtained via percentile optimization. *NMrt*: non-musicians in the real-time task; *Mrt*: musicians in the real-time task; *Ma*: musicians in the annotation task; *Maw*: musicians in the annotation task (weights added based on boundary strength ratings). *P*-values adjusted for effective degrees of freedom.

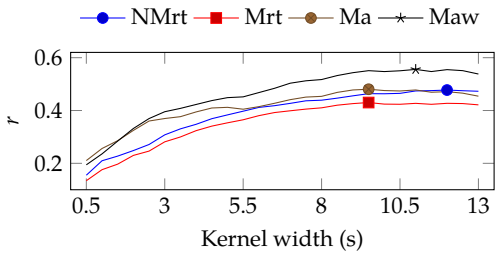

Figure 6: Correlation between perceptual segmentation density and computational model prediction obtained via percentile optimization for novelty widths ranging from 0.5 s to 13 s. Maximum points for each curve are highlighted with markers.

1 shows the correlation between optimal models and segmentation density for each participant group and segmentation task at a novelty width of 11 s; interaction features include the symbol ∘, which indicates pairwise multiplication between two basic features. Notably, prediction rates were lower for musicians than for non-musicians in the real-time task. This result suggests that musicians' segmentation relies more on schematic knowledge than in the case of non-musicians. Comparing experimental tasks, we found higher prediction rates for the annotation task. This suggests that some boundaries are difficult to anticipate in real-time segmentation, and are hence either indicated after longer delays or not indicated at all, leading to more noisy segmentation data. Related to this finding, the effect of experimental task was clearer for the annotation task density curve with added weights, which yielded the highest prediction rates. This indicates that the strength attributed by listeners to boundaries aids to the computational prediction and suggests a positive relationship between musical novelty and perceived boundary strength.

3.3 Selected Feature Subsets

We examined the categories of musical novelty features that were involved in the computational models' predictions. Table 1 presents the musical feature subsets that were selected via combinatorial optimization. Compared to non-musicians' model, the predicted segmentation density for musicians involved all the extracted musical features (i.e., key strength was not included in non-musicians' model). This suggests that, compared to non-musicians, musicians paid attention to more levels of the structural hierarchy during segmentation, and that local key context changes had a larger influence in musicians' segmentation. In addition, the model for musicians involved more interaction features than the model for non-musicians. This suggests that musicians paid more attention to high dimensional features, in other words, to simultaneous change of multiple features. It is also noteworthy that the annotation task model involved more features than the real-time task model; rhythmic and tonal features in particular had more representation in the subsets. This result suggests that in the annotation task listeners followed a more complex pattern of segmentation and focused on multiple hierarchical levels of metrical and tonal structure. In addition, we found that the model for annotation task density with added weights involved the largest amount of feature interactions. This finding sug-

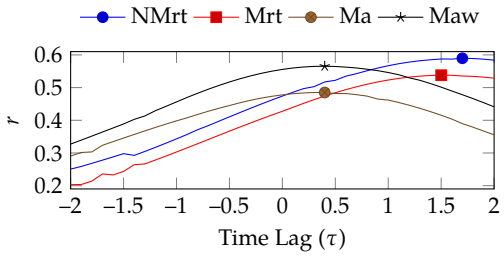

Figure 7: Correlation between perceptual segment boundary density and models' predictions obtained after time lags ranging from −2 s to 2 s, incremented by steps of 100 ms. Positive time lags indicate delay of novelty curves with respect to perceptual segment boundary density, and vice versa. Maximum points for each curve are highlighted with markers.

gests the possibility of a positive relationship between dimensionality of musical feature change and perceived boundary strength.

3.4 Time Lag Between Actual and Predicted Density

Our next step aimed to examine whether or not boundary indication delays had an effect on the model prediction. To approach this goal, we computed prediction models for different time lags of the segmentation density curves. We used 41 lag values ranging between −2 s and 2 s and incremented by steps of 100 ms. Figure 7 shows the correlation between segmentation density and computational model prediction for different segmentation density time lags. Each global peak corresponds to the optimal time lag for a given segmentation set. We found a larger optimal time lag for non-musicians (1.7 s, r = .59) than for musicians (1.5 s, r = .54). A larger (200 ms) optimal lag for comparison between actual and predicted segmentation density suggests a larger response delay during segmentation for non-musicians. Comparing tasks, the annotation task exhibited an optimal lag of 400 ms (r = .48), which is over a second shorter than the real-time task (1.5 s); this finding was replicated for a curve of log-likelihood as a function of time lag, which was not included here for brevity. This suggests that listeners' response delay in real-time segmentation can often reach 1.5 s in the real-time task, whereas in the annotation task the delay is unsurprisingly shorter (due to task characteristics including boundary reposition and increased familiarity with stimuli) but it is still observable and can be addressed. Noteworthy, the prediction rate of most models increased after applying the optimal time lag, illustrating the importance of accounting for listeners' response delays for optimal segmentation modelling.

4 Discussion

This section will discuss three hypotheses that have been formulated with regard to prediction of music segmentation. It is important to highlight at this point that the approach presented in this article is tailored to an understanding of segment boundaries as instants of significant change in the music. The main advantage of this circumscription of the notion of music segmentation is that it allows for a systematic analytic approach ultimately based on correlation between two time series. However, an important shortcoming should be mentioned at this point: musical segments are viewed as built upon boundary indications, whereas as a matter of fact, segments are concomitants of hierarchical representations of musical structure (Marsden, 2005). Moreover, our approach is not conceptually driven, as it disregards higher-level notions of musical motives, phrases, melodies, and themes, which embrace the complexity inherent in musical structures and point to the necessity of taking musical repetition and variation (i.e., parallelism) into account (see Cambouropoulos, 2006; Lartillot & Toiviainen, 2007b). In other words, this paper only partially addresses segmentation as a multi-level problem, because the hierarchical architecture of musical structure gets reduced to a single dimension. The second issue is that aspects related to recurrence in musical structure and perception of motivic patterns are omitted. Although a broader model is clearly required, such reductionism may be justifiable for analytic purposes, and could help to elucidate the applicability of some music-theoretic predictions to actual segment boundary perception. Furthermore, our approach includes current methodology in music information retrieval, but for a different aim: our main focus is on listener's perception of local musical changes rather than system evaluation or comparison between human and algorithmic performance.

4.1 Musicianship

We obtained three main results supporting our first hypothesis, which asserts that musicianship has an effect on segmentation model prediction. First, the segmentation models for non-musicians yielded higher prediction rates than for musicians, so overall prediction based on novelty curves is presumably more reliable for non-musicians (Figure 5, Figure 7 and Table 1). This suggests that segmentation by non-musicians is more guided by "bottom-up" acoustic local change (as detected via novelty curves) than for the case of musicians, who probably relied more on schematic knowledge; in other words, non-musicians yielded higher prediction rates because novelty curves do not model schematic knowledge. Second, prediction of musicians' segmentation involved more musical features (key strength was selected in musicians' model but not in non-musicians') and more novelty interaction features than for non-musicians. This suggests that musicians focus on high dimensional musical change and more levels of the structural hierarchy; for example they may focus on more obvious changes such as instrumentation and rhythm, but also on subtle changes in tonality even if these are implied changes. Several studies support the notion that musicians pay attention to more aspects determining musical change; for instance, musicians' ratings of tension of chords within sequences were mostly influenced by both tonal functions and specific roughness, whereas non-musicians' ratings tended to be mostly prompted by horizontal motion, i.e. melodic arrangement between successive chords due to voicing and use of inversions (Bigand et al., 1996). In addition, a study on perceived cadences (Sears, Caplin, & McAdams, 2014) showed that, compared to non-musicians, musicians do not only pay attention to the most salient melodic line, but also to complex texture changes involving multiple voices. Our third result regarding musicianship was a larger optimal time lag for segmentation prediction in the case of non-musicians than of musicians, which points to a negative relationship between musical training and response delay in segmentation. This effect of musicianship on speed of detecting and indicating segment boundaries is partly not surprising because musicians are explicitly trained to follow musical cues that trigger their entrance during performances; however this result still suggests that non-musicians process perceived musical structure at a slower rate than musicians. In this line, effects of musical training on auditory working memory have been previously shown, since faster ability to capture the statistical structure of perceived streams (François, Jaillet, Takerkart, & Schön, 2014) and larger auditory memory spans (Tierney, Bergeson-Dana, & Pisoni, 2008) have been found for musicians when compared to non-musicians. A direct comparison between boundary density curves via cross-correlation (Hartmann et al., in press) showed that non-musicians were delayed with respect to musicians for most of the stimuli, although it did not result in differences between groups based on the mean lag across stimuli.

A general implication of these findings is that both participant groups pay attention to local discontinuities in the music, so specific knowledge of structure may not be required for perception of segment boundaries that emerge due to novelty; in this respect Tillmann and Bigand (2004) suggested that, regardless of musical training, the succession of local structures prevails over the succession of global structures in music processing. However, our results suggest that musicians may pay less attention to local discontinuities than non-musicians; so global structures could have a greater role for musicians, who might build more veridical expectancies (see Justus & Bharucha, 2001) for events that are likely to occur in a given piece of music.

4.2 Experimental Task

Three results were found to support our second hypothesis, which states that the conducted experimental task has an effect on model prediction. First, prediction rates for the annotation task were higher than for the real-time task, but controlling for delays inverted this result. This suggests that listeners' delayed indications are responsible for the relatively lower prediction rates in the real-time task, and that once these are compensated, this task yields higher similarity with "bottom-up" novelty-based predictions since listeners neither know with certainty about the unfolding patterns and developments of a piece of music, nor can clearly estimate the relative significance of a given musical change. The second result concerning segmentation tasks is that prediction of annotation task involved more novelty features, particularly rhythmic and tonal features. This result suggests that listeners pay attention in this task to more levels of the structural hierarchy. The third

result with respect to this hypothesis is that the annotation task exhibited a shorter optimal time lag for segmentation prediction than the real-time task. This result is highly expected mainly because the annotation task allowed participants to modify the position of boundaries, but it is noteworthy that the alignment between segment boundaries and instants of musical novelty leads to an increase in prediction rates for offline segmentation tasks as well.

4.3 Boundary Strength Weights

Finally, two main results were found supporting the third hypothesis, which posits that weighting the annotation task segmentation density has an effect on model prediction. First, we found that adding weights to the annotation task increases the model prediction rates. This suggests that the novelty detection approach predicted perceived boundary strength ratings, which is a plausible interpretation because the most stark musical changes should often coincide with high discontinuity of musical features. Moreover, the improvement of prediction rates shows that the strength of a boundary is not equivalent to its density, which suggests that boundary strength weights aid to the prediction of listeners' segment boundaries. This result might seem surprising, considering that Bruderer (2008) found a relationship between frequency of indications of a boundary and mean ratings of boundary strength. However, Bruderer's task instructions referred specifically to the indication of phrases, sections, and passages, whereas our task instructed listeners to indicate significant instants of change, which would have prompted more frequent indications. Possibly, the addition of strength weights in the annotation task highlighted points of relatively high acoustic local change, which could have increased prediction accuracy for musical features that could have been sensitive to these changes. The second result found regarding this hypothesis was that adding weights to the annotation task also increases the number of feature interactions involved in models. This suggests that listeners' boundary strength ratings relate to different interactions, resulting in a hierarchy of high dimensional features; for instance rhythmic-tonal musical novelty could be perceived as more perceptually salient than spectral-rhythmic novelty.

4.4 General Discussion

We may now recapitulate the main conclusions reached here. Regarding musicianship, our results suggest that musicians' schematic knowledge is a potential factor in lower prediction rates compared to non-musicians'; in addition, musicians may pay attention to more dimensions of musical change spanning multiple hierarchical levels of structure, and seem to respond faster to perceived musical change than non-musicians. Comparing experimental tasks, listeners' response delays in the real-time task seem to be a major factor in lower model performance with respect to the annotation task; they may also pay attention to more hierarchical levels of structure in the annotation task, particularly regarding rhythmic and tonal descriptions of change, which possibly make a major contribution in perceptual segmentation. Also, boundary strength ratings in the annotation task may be more associated with perceived concurrence of multiple descriptions of musical change.

The models presented in Table 1 can be sorted based on their prediction rates to find the most satisfactory scenarios for novelty-based prediction of segmentation. For instance, annotation task models yielded higher prediction rates than real-time task models, a result that makes sense because novelty detection does neither account for listeners' response delays nor for difficulties to indicate retrospectively perceivable boundaries. In particular, adding weights to the annotation task boundary density led to a clear increase of prediction rates, showing that novelty curves can model listeners' assignment of hierarchies to boundaries, which might depend on the number of perceived dimensions of musical change. Although the frequency of indications of a boundary (which is equivalent to its density) should to some extent also describe this hierarchy of events, boundary strength weights contribute to the description of boundaries' relative structural importance. In contrast to the annotation task results, real-time segmentation (not adjusted for response delays) resulted in lower prediction rates, especially for musicians (Table 1 and Figure 5), even though their segmentations were less delayed than those from non-musicians. This further supports the interpretation that schematic knowledge had a larger influence on musicians' segmentation decisions, or at least that they paid more attention to aspects such as repetition and musical parallelism instead of solely focusing on local discontinuity.

The compensation for response delays had an effect on the real-time task model performance because novelty detection provides immediate feedback for a given context, whereas listeners' responses to perceived musical change are not instantaneous; the annotation task did not greatly benefit from this compensation because listeners repositioned their boundary indications. A different interpretation of the results is required for optimal models that account for response delays (Figure 7), because real-time task models exhibited a clear increase in prediction rates, and the difference between tasks in this respect became smaller. Overall, larger prediction rates show the need for controlling for response delays in novelty-based segmentation modelling, especially when it comes to real-time segmentation and to non-musicians. Two other contributors to differences between optimal models have been schematic knowledge, which cannot be modelled by the novelty curves and could explain lower prediction rates for musicians' segmentation, and boundary strength ratings, which yielded density curves that emphasized obvious, probably high dimensional musical changes.

A general result to highlight concerning the features involved in the prediction models is the contribution of feature interactions, which suggests that listeners pay attention to high dimensional musical change; for instance, simultaneous change in rhythm and tonality or in timbre and tonality seemed to often evoke listeners' perception of segment boundaries. In particular, the feature interaction *Fluctuation Patterns* ∘ *Chromagram* was selected in all models, suggesting that listeners pay attention to simultaneous changes in pitch class and rhythm during segmentation.

Regarding the proposed non-linear combination approach, it resulted in improved prediction rates with respect to any of the novelty curves extracted (A.3). This means that the combined novelty detected by a majority of the features at each time point yielded better performance than any novelty feature alone, which results from the fact that the contribution of different features to perception of musical change varies over time and over stimuli. For instance, some boundary indications may be represented more by rhythmic than by tonal change, whereas others may exhibit the opposite trend.

4.5 Considerations for Future Research

Our findings suggest that an ideal scenario for accurate boundary density prediction via novelty detection would be based on indications not only of high time precision (i.e. compensated for response delays) and describing only local discontinuities, but also weighted based on perceived strength. To better understand the relative importance of these factors, non-musicians should also be recruited to segment in an annotation task; this addition to the experimental design is feasible because the skills required in an annotation task can be quickly learned. Possibly, an offline annotation would further increase non-musicians' prediction rate with respect to the delay-compensated real-time task.

Future studies on annotation segmentation tasks should systematically study the effect of different task instructions upon segmentation. For instance, allowing addition of new boundaries during the reposition stage of the task might lead to more detailed representations of structural change. In addition, a focus on the final state of an annotation should not ignore other relevant information that can be collected in this task: steps such as boundary reposition and removal should be recorded in order to better understand, for instance, the extent to which a shorter optimal time lag in the annotation task compared to the real-time task could be attributed to boundary reposition or to other factors such as familiarity with the stimuli and task.

In regards prediction rates, the proposed approach, which consisted in computing interaction novelty features and non-linear modelling, yielded up to moderately high correlations with boundary density. These results outperform those reported in a preliminary version of this article (Hartmann, Lartillot, & Toiviainen, 2015), in which a smaller novelty kernel width was used and the effect of response delay was disregarded. Our evaluation of prediction performance was, however, not an end but rather a means by which we could compare different listener groups and segmentation tasks. Benchmark studies on segmentation could further explore compensation for response delays, which led to highest prediction rates.

Focusing further on listeners' response delays, our findings showed that segmentation data can often exhibit up to 1.7 s delays with respect to musical changes; this compensation for response de-

lays increased prediction rates in all models except for the annotation task without added weights. In this regard, retrieval evaluation of boundary detection systems is commonly based on both 0.5 s and 3 s thresholds (Ehmann, Bay, Downie, Fujinaga, & De Roure, 2011), however according to our findings, 3 s would yield overly optimistic results, especially considering that the segmentation ground truth data for these evaluations is collected via annotation tasks; future research on music information retrieval should consider hit rate evaluation only at a time threshold of 0.5 s.

In regards the effect of segmentation boundary strength weights, we believe that further exploration is needed to understand its impact for novelty-based prediction; for instance boundary strength could be correlated with musical novelty at the respective time points in order to better understand their similarity and explore what musical dimensions prompt perception of stark boundaries. This is an important issue to tackle, not only because boundary strength seems to offer descriptions that do not necessarily relate to boundary density, but also because it clearly contributed to the computational prediction and might offer new insights about the structural hierarchy of perceived musical boundaries.

As a methodological consideration, we remark that the novelty kernel width used in this study was rather large. An optimal kernel width spanning large time regions was needed due to noisiness of novelty curves, and to the ample distance between the main peaks in density curves. Although the use of short window lengths and high overlapping between frames are necessary for highly accurate feature extraction, this leads to very detailed similarity matrices, which in turn produce noisy novelty curves. Future studies should consider the use of smoothing filters (e.g. Serrà, Muller, Grosche, & Arcos, 2014) to improve computational efficiency of the models. A related issue pertains to the aggregation of multiple novelty features based upon a single novelty width; for instance, spectral and rhythmic features tended to yield lower optimal kernel widths than chroma and tonal features, so it is difficult to choose a novelty width that gives justice to various features operating on different temporal contexts. To address this issue, it is possible to compute an optimization model for each density curve that could involve a subset of novelty curves with different kernel widths; this promising approach would require finding, for each feature, a novelty kernel width that yields optimal correlation with the density curves. Another matter of concern regarding novelty widths is their relationship with the Gaussian bandwidth of the segmentation density, which was a fixed parameter in this study and requires further assessment using different musical features to better understand the relationship between these two parameters. It should also be remarked that the need to choose a novelty kernel width can be circumvented; for instance, a recently proposed multi-granular method (Lartillot, Cereghetti, Eliard, & Grandjean, 2013) detects novelty by considering both the amount of contrast between neighboring homogeneous passages and the temporal scale of the preceding passage.

Regarding the non-linear optimization approach used in this study, other strategies including alternative cost functions could be implemented; we have utilized mean-based optimization and cross-entropy minimization as alternatives to percentile-based correlation optimization, but these yielded lower prediction rates. In addition, further work on percentile-based optimization could focus on the improvement of prediction rates using various percentiles (though we observed that 50th percentile offered higher rates than 25th and 75th percentiles) or other summarizing statistics, including computation of aggregations that specify different weights to features depending on their rank (Yager, 2006). Other combinatorial optimization algorithms are also possible; we also experimented with simulated annealing and forward-backward feature selection; but these approaches yielded models with lower prediction rates than the genetic algorithm method. We assumed that this method did not stumble on local minima, however other methods might get closer to the global minimum of the solution space.

A question that may arise is whether or not a linear modelling approach could have resulted in comparable results. Stepwise regression models offer the possibility to rank selected features based on standardized beta coefficients, however these models assume a constant contribution from each feature across time and musical stimuli. We computed the same analysis via this approach, which yielded a similar pattern of results, but these were left out from our analyses due to the presence of negative coefficients in the models. A reason for this is that some interaction novelty features highly correlate with each other, for instance *Chromagram* ∘ *Key Strength* is highly similar to *Chromagram* ∘ *Tonal Centroid* (r = .98); future work could

perform feature selection based on collinearity as a prior step to stepwise regression.

It should also be mentioned that model prediction rates might be optimistic due to relatively low amount of musical stimuli and correlating novelty features, which puts the optimization at risk of yielding an "optimal" subset that may be equally optimal to other subsets, and of generating optimal subsets and models that are highly affected by trivial modifications of the segmentation density curves. Besides the elimination of redundant features, cross-validation with other stimuli or with other groups of listeners should be used in future studies to overcome model over-fitting and increase robustness.

We also remark that, depending on the musical stimulus and especially on musical style, listeners should probably use different segmentation strategies. Hence, it is possible that a methodological approach focused on individual stimuli would have led to different results; e.g., individual stimuli may require different feature subsets for optimal prediction, and variation in prediction accuracy could occur; some of these issues, which are crucial for the development of segmentation systems that automatically adjust their parameters depending on various characteristics of the target stimulus, are currently under investigation (Hartmann, Lartillot, & Toiviainen, 2016).

Finally, we should highlight the differences reported in this study between musicians and non-musicians; a clear trend was found in this respect and the results seem plausible. First, higher prediction rates for non-musicians imply that they focus more on local acoustic change than on other aspects such as schematic expectations. Second, more features in prediction models for musicians, particularly more interaction features, suggest that they pay attention to more musical dimensions and levels of the musical structure. Third, differences in response times between groups could reflect a faster processing of perceived structure in musicians. Although explicit segmentation tasks are not enough to investigate how underlying musical structures are processed, it is possible that learning processes involved in intensive musical training and development of motor skills for musical performance have an effect on the perception of musical structure. A plausible explanation is that musical training leads to different expectations between groups; musicians' anticipation of future events may be facilitated e.g. by schemata that cannot be learned from mere exposure to music, resulting in increased attention to specific types of musical change, such as those prompted by interaction of different acoustic features. Further work should further explore this possibility by comparing experienced musical listeners and musicians in their processing of musical structure.

Acknowledgements The authors would like to thank Alan Marsden and an anonymous reviewer for giving us helpful comments on an earlier version of this paper. Thanks also to Emily Carlson for proofreading the paper. This work was financially supported by the Academy of Finland (project numbers 272250 and 274037).

References

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2015). Emotion based segmentation of musical audio. In *Proceedings of the 15th conference of the International Society for Music Information Retrieval (ISMIR 2014)*.

Alluri, V. & Toiviainen, P. (2010). Exploring perceptual and acoustical correlates of polyphonic timbre. *Music Perception*, *27*(3), 223–241.

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*(4).

Befus, C. (2010). *Design and evaluation of dynamic feature-based segmentation on music* (Doctoral dissertation, Dept. of Mathematics and Computer Science, University of Lethbridge).

Bigand, E., Parncutt, R., & Lerdahl, F. (1996). Perception of musical tension in short chord sequences: the influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, *58*(1), 125–141.

Bruderer, M. (2008). *Perception and modeling of segment boundaries in popular music* (Doctoral dissertation, JF Schouten School for User-System Interaction Research, Technische Universiteit Eindhoven).

Burunat, I., Alluri, V., Toiviainen, P., Numminen, J., & Brattico, E. (2014). Dynamics of brain activity underlying working memory for music in a naturalistic condition. *Cortex*, *57*, 254–269.

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

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: an open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia International Conference* (pp. 1467–1468). Firenze, Italy.

Chew, E. (2002). The spiral array: an algorithm for determining key boundaries. In *Music and artificial intelligence* (pp. 18–31). Springer.

Clarke, E. & Krumhansl, C. L. (1990). Perceiving musical time. *Music Perception*, 213–251.

Dawe, L. A., Plait, J. R., & Racine, R. J. (1994). Inference of metrical structure from perception of iterative pulses within time spans defined by chord changes. *Music Perception: An Interdisciplinary Journal*, *12*(1), 57–76.

Dawe, L. A., Platt, J. R., & Racine, R. J. (1995). Rhythm perception and differences in accent weights for musicians and nonmusicians. *Perception & psychophysics*, *57*(6), 905–914.

Dean, R. T., Bailes, F., & Drummond, J. (2014). Generative structures in improvisation: computational segmentation of keyboard performances. *Journal of New Music Research*, *43*(2), 1–13.

Deliège, I. (1987). Grouping conditions in listening to music: an approach to Lerdahl & Jackendoff's grouping preference rules. *Music Perception*, 325–359.

Deliège, I. (2001). Similarity perception categorization cue abstraction. *Music Perception*, *18*(3), 233–243.

Deliège, I., Mélen, M., Stammers, D., & Cross, I. (1996). Musical schemata in real-time listening to a piece of music. *Music Perception*, 117–159.

Drake, C. & Bertrand, D. (2001). The quest for universals in temporal processing in music. *Annals of the New York Academy of Sciences*, *930*(1), 17–27.

Dujmović, J. J. & Larsen, H. L. (2007). Generalized conjunction/disjunction. *International Journal of Approximate Reasoning*, *46*(3), 423–446.

Ehmann, A. F., Bay, M., Downie, J. S., Fujinaga, I., & De Roure, D. (2011). Music structure segmentation algorithm evaluation: expanding on MIREX 2010 analyses and datasets. In *Proceedings of the 12th conference of the International Society for Music Information Retrieval (ISMIR 2011)* (pp. 561–566).

Eronen, A. (2007). Chorus detection with combined use of MFCC and chroma features and image processing filters. In *Proceedings of the 10th International Conference on Digital Audio Effects* (pp. 229–236). Citeseer.

Foote, J. T. (1997). Content-based retrieval of music and audio. In C.-C. J. K. et al. (Ed.), *Proceedings of SPIE Multimedia Storage and Archiving systems II* (Vol. 3229, pp. 138–147).

Foote, J. T. (1999). Visualizing music and audio using self-similarity. In *Proceedings of 7th ACM International Conference on Multimedia (Part 1)* (pp. 77–80).

Foote, J. T. (2000). Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo* (Vol. 1, pp. 452–455). IEEE.

François, C., Jaillet, F., Takerkart, S., & Schön, D. (2014). Faster sound stream segmentation in musicians than in nonmusicians. *PloS one*, *9*(7), e101340.

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.

Fujishima, T. (1999). Realtime chord recognition of musical sound: a system using common lisp music. In *Proceedings of the International Computer Music Conference* (Vol. 1999, pp. 464–467).

Harte, C., Sandler, M., & Gasser, M. (2006). Detecting harmonic change in musical audio. In *Proceedings of the 1st ACM workshop on audio and music computing multimedia* (pp. 21–26).

Hartmann, M., Lartillot, O., & Toiviainen, P. (2015). Effects of musicianship and experimental task on perceptual segmentation. In J. Ginsborg, A. Lamont, M. Philips, & S. Bramley (Eds.), *Proceedings of the Ninth Triennal Conference of the European Society for the Cognitive Sciences of Music*. Manchester.

Hartmann, M., Lartillot, O., & Toiviainen, P. (2016). Estimating novelty-based predictability of perceptual segmentation with musical features. *Manuscript in preparation.*

Hartmann, M., Lartillot, O., & Toiviainen, P. (in press). Multi-scale modelling of segmentation: effect of musical training and experimental task. *Music Perception*.

Jensen, K. (2007). Multiple scale music segmentation using rhythm, timbre, and harmony.

EURASIP Journal on Applied Signal Processing, *2007*(1), 159–159.

Johnson, E. K. & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567.

Justus, T. C. & Bharucha, J. J. (2001). Modularity in musical processing: the automaticity of harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(4), 1000.

Krumhansl, C. L. (1990). Cognitive foundations of musical pitch. (Chap. 4, Vol. 17). Oxford University Press New York.

Krumhansl, C. L. (1996). A perceptual analysis of Mozart's Piano Sonata k. 282: segmentation, tension, and musical ideas. *Music Perception*, 401–432.

Lartillot, O., Cereghetti, D., Eliard, K., & Grandjean, D. (2013, June). A simple, high-yield method for assessing structural novelty. In G. Luck & O. Brabant (Eds.), *Proceedings of the 3rd international conference on music & emotion (icme3)*. Jyväskylä, Finland.

Lartillot, O. & Toiviainen, P. (2007a). A Matlab toolbox for musical feature extraction from audio. In *International Conference on Digital Audio Effects* (pp. 237–244). Bordeaux.

Lartillot, O. & Toiviainen, P. (2007b). Motivic matching strategies for automated pattern extraction. *Musicae Scientiae*, *11*(1 suppl), 281–314.

Lerdahl, F. & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, M.A.: The MIT Press.

Marsden, A. (2005). Generative structural representation of tonal music. *Journal of New Music Research*, *34*(4), 409–428.

Martens, P. A. (2011). The ambiguous tactus: tempo, subdivision benefit, and three listener strategies. *Music Perception*, *28*(5).

Pampalk, E., Rauber, A., & Merkl, D. (2002). Content-based organization and visualization of music archives. In *Proceedings of the tenth ACM international conference on Multimedia* (pp. 570–579).

Paulus, J. & Klapuri, A. (2009). Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1159–1170.

Paulus, J., Müller, M., & Klapuri, A. (2010). State of the art report: audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 625–636).

Peeters, G. (2007). Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the 5th International Conference on Music Information Retrieval* (pp. 35–40).

Poli, G. D., Rodà, A., & Vidolin, A. (1998). Note-by--note analysis of the influence of expressive intentions and musical structure in violin performance. *Journal of New Music Research*, *27*(3), 293–321.

Pyper, B. J. & Peterman, R. M. (1998). Comparison of methods to account for autocorrelation in correlation analyses of fish data. *Canadian Journal of Fisheries and Aquatic Sciences*, *55*(9), 2127–2140.

Sears, D., Caplin, W. E., & McAdams, S. (2014). Perceiving the classical cadence. *Music Perception: An Interdisciplinary Journal*, *31*(5), 397–417.

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language*, *57*(1), 24–48.

Serrà, J., Muller, M., Grosche, P., & Arcos, J. (2014). Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, *16*(5), 1229–1240.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. CRC press.

Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., De Roure, D., & Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 555–560).

Smith, J. B. L., Schankler, I., & Chew, E. (2014). Listening as a creative act: meaningful differences in structural annotations of improvised performances. *Music Theory Online*, *20*(3).

Terhardt, E. (1979). Calculating virtual pitch. *Hearing research*, *1*(2), 155–182.

Tierney, A. T., Bergeson-Dana, T. R., & Pisoni, D. B. (2008). Effects of early musical experience on auditory sequence memory. *Empirical musicology review: EMR*, *3*(4), 178.

Tillmann, B. & Bharucha, J. J. (2002). Effect of harmonic relatedness on the detection of tempo-

ral asynchronies. *Perception & Psychophysics*, *64*(4), 640–649.

Tillmann, B. & Bigand, E. (2004). The relative importance of local and global structures in music perception. *The Journal of Aesthetics and Art Criticism*, *62*(2), 211–222.

Turnbull, D., Lanckriet, G. R., Pampalk, E., & Goto, M. (2007). A supervised approach for detecting boundaries in music using difference features and boosting. In *Proceedings of the 5th International Conference on Music Information Retrieval* (pp. 51–54).

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics*, *18*(1), 183–190.

Yager, R. R. (2006). A human directed approach for data summarization. In *IEEE International Conference on Fuzzy Systems* (pp. 707–712). IEEE.

Zacks, J. M. & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, *16*(2), 80–84.

A Appendices

A.1 Musical Stimuli - List of Abbreviations

Genesis Banks, T., Collins, P. & Rutherford, M. (1986). The Brazilian. [Recorded by Genesis]. On *Invisible Touch* [CD]. Virgin Records. (1986)
Spotify link: http://open.spotify.com/track/7s4hAEJupZLpJEaOel5SwV
Excerpt: 01:10.200-02:58.143. Duration: 01:47.943

Smetana Smetana, B. (1875). Aus Böhmens Hain und Flur. [Recorded by Gewandhausorchester Leipzig - Václav Neumann]. On *Smetana: Mein Vaterland* [CD]. BC - Eterna Collection. (2002)
Spotify link: http://open.spotify.com/track/2115JFwiNvHxB6mJPkVtbp
Excerpt: 04:06.137-06:02.419. Duration: 01:56.282

Morton Morton, F. (1915). Original Jelly Roll Blues. On *The Piano Rolls* [CD]. Nonesuch Records. (1997)
Spotify link: http://open.spotify.com/track/6XtCierLPd6qg9QLcbmj61
Excerpt: 0-02:00.104. Duration: 02:00.104

Ravel Ravel, M. (1901). Jeux d'Eau. [Recorded by Martha Argerich]. On *Martha Argerich, The Collection, Vol. 1: The Solo Recordings* [CD]. Deutsche Grammophon. (2008)
Spotify link: http://open.spotify.com/track/27oSfz8DKHs66IM12zejKf
Excerpt: 03:27.449-05:21.884. Duration: 01:54.435

Couperin Couperin, F. (1717). Douzième Ordre / VIII. L'Atalante. [Recorded by Claudio Colombo]. On *François Couperin : Les 27 Ordres pour piano, vol. 3 (Ordres 10-17)* [CD]. Claudio Colombo. (2011)
Spotify link: http://open.spotify.com/track/6wJyTK8SJAmtqhcRnaIpKr
Excerpt: 0-02:00.863 Duration: 02:00.863

Dvořák Dvořák, A. (1878). Slavonic Dances, Op. 46 / Slavonic Dance No. 4 in F Major. [Recorded by Philharmonia Orchestra - Sir Andrew Davis]. On *Andrew Davis Conducts Dvořák* [CD]. Sony Music. (2012)
Spotify link: http://open.spotify.com/track/5xna3brB1AqGW7zEuoYks4
Excerpt: 00:57.964-03:23.145. Duration: 02:25.181

A.2 Extracted Musical Features

Basic novelty curves were obtained from similarity matrices of musical features (cf. 2.4). To this end, the following five musical features describing spectral, rhythmic, chroma and tonal attributes were extracted from the musical signal:

Spectral

- Subband flux (Alluri & Toiviainen, 2010) : 10-dimensional feature describing spectral fluctuations at octave-scaled subbands of the audio signal. First, ten second-order elliptic filters are used to divide the signal into subbands. For each frequency channel, a spectrogram is computed using a window length of 25 ms and 50% overlapping. Finally, dissimilarity between successive spectral frames is computed via pairwise normalized Euclidean distance (spectral flux). Unlike other common spectral features such as Mel-frequency cepstral coefficients (MFCCs), subband flux features have been found to predict perceptual aspects of musical polyphonic timbre such as activity, brightness and fullness.

Rhythmic

- Fluctuation patterns (Pampalk, Rauber, & Merkl, 2002): Psychoacoustics-based representation of rhythmic periodicities in the audio signal via estimation of spectral energy modulation over time at different frequency bands. First, a spectrogram in dB scale with frequencies bundled into 20 Bark bands is computed using a window length of 23 ms and a hop rate of 80 Hz. Following an outer ear model (Terhardt, 1979), frequencies between 2000 Hz and 5000 Hz are emphasized, whereas energy at frequency range extremes is attenuated. Further, the spectrogram is weighted based on a perceptual model of spectral masking that, given a high-energy frequency band, attenuates energy at a region of frequencies below that band. Subsequently, for each separate Bark band, a second spectrogram is computed (window length 1 s, hop rate 10 Hz) where the highest frequency taken into consideration is 10 Hz (600 beats per minute). This yields, for each Bark band and each frame, a description of loudness modulation. Each modulation coefficient is weighted based on a psychoacoustic model of fluctuation strength sensation to emphasize modulation frequencies that are optimal for the perception of a strong fluctuation such as a steady beat. Finally, for each frame, the modulation coefficients are summed together. The result is a description of the dynamic evolution of periodicity for each modulation frequency.

Chroma

- Chromagram (pitch class profile, see Fujishima, 1999): 12-dimensional feature describing the energy distribution of each pitch class per spectrogram frame. First, a spectrogram for the highest energy over a range of 20 dB and for frequencies ranging between 100 Hz and 6400 Hz is computed. Frequency bins are then combined into chroma, corresponding to the different absolute pitches. To each chroma is associated a central frequency cl, which is calculated as $cl = 12 \times log_2(\frac{f}{cf})$, where cf is the central frequency related to C4 (set to 261.6256 Hz). The audio waveform is normalized before the spectrogram computation, and each frame of the resulting chromagram is also normalized by the maximum local value. The chromagram is then wrapped into one octave, by summing together chroma values of same pitch classes, leading to a 12-dimensional feature. The spectrogram was computed using a 3 s window length and 100 ms overlapping to obtain a sufficiently high time resolution. The following two features use chromagram as input.

Tonal

- Key strength (Krumhansl, 1990): 24-dimensional feature that represents how well the chromagram fits the different tonal profiles for major and minor keys. The key profiles are based on the probe-tone experimental method and represent the contribution of each of the 12 chromatic tones to a given key. The key strength values of each frame are estimated via correlation between the pitch class profile and each of the 24 key profiles.
- Tonal centroid (Harte, Sandler, & Gasser, 2006): 6-dimensional feature that describes a projection of the pitch class profile onto interior spaces of the circle of fifths, the circle of minor thirds and the circle of major thirds, which derive from a toroidal representation of the harmonic network (*Tonnetz*). The spaces are derived from the Spiral Array model (Chew, 2002) for key boundary detection. For each frame, the chromagram is multiplied with the basis of a 6-dimensional pitch space in order to obtain three co-ordinate pairs, one per circularity inherent in the harmonic network.

A.3 Correlations Between Perceptual Segment Boundary Density and Novelty Features

Feature Type	Basic Feature	NMrt	Mrt	Ma	Maw
Spectral	Subband Flux	.10	.14*	.07	.17**
Rhythmic	Fluctuation Patterns	**.38*****	**.32*****	.31***	**.39*****
Chroma	Chromagram	.32***	.31***	**.36*****	.35***
Tonal	Key Strength	.21***	.19**	.25***	.26***
	Tonal Centroid	.23***	.21***	.31***	.30***

*$p < .05$; **$p < .01$; ***$p < .001$

Table 2: Correlations between perceptual segmentation density and basic features. Maximum coefficients of each set are indicated in boldface. Coefficients from features selected via optimization are highlighted. *P*-values adjusted for effective degrees of freedom, and for multiple comparisons via Benjamini-Hochberg correction ($q = 0.05$).

Type	Feature Interaction	NMrt	Mrt	Ma	Maw
Spectral ∘ Rhythmic	Subband Flux ∘ Fluctuation Patterns	.21***	.22***	.17**	.27***
Spectral ∘ Chroma	Subband Flux ∘ Chromagram	.30***	.30***	.31***	.39***
Spectral ∘ Tonal	Subband Flux ∘ Key Strength	.29***	.28***	.30***	.38***
Spectral ∘ Tonal	Subband Flux ∘ Tonal Centroid	.31***	.31***	.35***	.42***
Rhythmic ∘ Chroma	Fluctuation Patterns ∘ Chromagram	**.44*****	**.37*****	.43***	**.49*****
Rhythmic ∘ Tonal	Fluctuation Patterns ∘ Key Strength	.37***	.31***	.41***	.45***
Rhythmic ∘ Tonal	Fluctuation Patterns ∘ Tonal Centroid	.40***	.33***	**.44*****	.49***
Chroma ∘ Tonal	Chromagram ∘ Key Strength	.22***	.20***	.25***	.26***
Chroma ∘ Tonal	Chromagram ∘ Tonal Centroid	.23***	.21***	.28***	.28***
Tonal ∘ Tonal	Key Strength ∘ Tonal Centroid	.17**	.15**	.22***	.22***

**$p < .01$; ***$p < .001$

Table 3: Correlations between perceptual segmentation density and feature interactions. Maximum coefficients of each set are indicated in boldface. Coefficients from features selected via optimization are highlighted. *P*-values adjusted for effective degrees of freedom, and for multiple comparisons via Benjamini-Hochberg correction ($q = 0.05$).