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

Title: Multi-Scale Modelling of Segmentation : Effect of Music Training and Experimental Task

Year: 2016

Version:

Please cite the original version:

Hartmann, M., Lartillot, O., & Toiviainen, P. (2016). Multi-Scale Modelling of Segmentation : Effect of Music Training and Experimental Task. Music Perception, 34(2), 192-217. https://doi.org/10.1525/MP.2016.34.2.192

Multi-scale Modelling of Segmentation: Effect of Music Training and Experimental Task

Martín Hartmann
University of Jyväskylä, Jyväskylä, Finland

Olivier Lartillot
Aalborg University, Aalborg, Denmark

Petri Toiviainen
University of Jyväskylä, Jyväskylä, Finland

While listening to music, people often unwittingly break down musical pieces into constituent chunks such as verses and choruses. Music segmentation studies have suggested that some consensus regarding boundary perception exists, despite individual differences. However, neither the effects of experimental task (i.e., real-time vs. annotated segmentation), nor of musicianship on boundary perception are clear. Our study assesses musicianship effects and differences between segmentation tasks. We conducted a real-time experiment to collect segmentations by musicians and nonmusicians from nine musical pieces. In a second experiment on non-real-time segmentation, musicians indicated boundaries and their strength for six examples. Kernel density estimation was used to develop multi-scale segmentation models. Contrary to previous research, no relationship was found between boundary strength and boundary indication density, although this might be contingent on stimuli and other factors. In line with other studies, no musicianship effects were found: our results showed high agreement between groups and similar inter-subject correlations. Also consistent with previous work, time scales between one and two seconds were optimal for combining boundary indications. In addition, we found effects of task on number of indications, and a time lag between tasks dependent on beat length. Also, the optimal time scale for combining responses increased when the pulse clarity or event density decreased. Implications for future segmentation studies are raised concerning the selection of time scales for modelling boundary density, and time alignment between models.

Received: December 13, 2014, accepted March 9, 2016.

Key words: music segmentation, music training, segmentation task, segmentation modelling, musical features

Listeners parse the structure of music by focusing attention on musical feature change and repetitions of sequences. They can spontaneously predict and detect relevant changes that demarcate the beginning and end of verses, choruses, and other types of musical structures. Many gaps in our knowledge on temporal processing of perceptual streams such as music, speech, and movement still need to be bridged. Indications of musical change are complex to study, since they stem from our memory-guided perception and cognition of points deemed to be musically salient (Deliège, 2007). The role of musicianship in the listener remains an important question, as it can help explain possible transfer effects of music learning. In addition, the difference between listeners' real-time and non-real-time ("annotation") indications of change is still unclear, although this difference can shed light on the assimilation of musical structure as a temporal process. Moreover, the study of the perceived structure in music can encourage developments in automatic systems to facilitate music editing and playback, such as adding music to family videos.

Perceived contrasts, discontinuities, changes, and repetitions at multiple hierarchical levels commonly serve as heuristics that guide the identification of musical segment boundaries (Addessi & Caterina, 2000). Studies in automatic segmentation often refer to these musical novelty points simply as instants of significant change (Foote, 2000). In this paper we will use segment boundaries and instants of significant change interchangeably, since we will investigate a particular aspect of music segmentation that is more related to musical change than to repetition or similarity.

As a general rule, people share a common sense of the instants at which the music in a piece changes in a significant way (Clarke & Krumhansl, 1990). This assertion is backed by evidence from listening studies on segmentation that shows a consensus despite varying frequency of indications (Bruderer, 2008; Clarke & Krumhansl, 1990; Koniari, Predazzer, & Mélen, 2001). Besides boundary indication time points, analyzed segmentation data in these studies include verbal justifications of segment boundaries, judged time positions, and duration of segments. In particular, boundary indications

Music Perception, volume 34, issue 2, pp. 192–217, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2016.34.2.192

have been defined according to perceived tension (Addessi & Caterina, 2000; Krumhansl, 1996), expectations and closure (Peebles, 2011), descriptors (Bailes & Dean, 2007; Krumhansl, 1996), and grouping rules (Clarke & Krumhansl, 1990; Deliège, 1987; Frankland & Cohen, 2004; Temperley, 2001). Automatic segmentation systems have been implemented in corpus-based studies; these systems were based on musical features (Hargreaves, Klapuri, & Sandler, 2012; Sanden, Befus, & Zhang, 2012; Smith, Chuan, & Chew, 2013), sets of rules (Bruderer, 2008; Cambouropoulos, 2006; Lartillot & Ayari, 2009; Lartillot, Yazıcı, & Mungan, 2013), or probabilistic methods (Ferrand, Nelson, & Wiggins, 2003; Lattner, Grachten, Agres, & Chacón, 2015; Pearce, Müllensiefen, & Wiggins, 2010), and generally compared against ground-truth data (cf. Paulus, Müller, & Klapuri, 2010; Peeters & Deruty, 2009). Bruderer (2008), Wiering, de Nooijer, Volk, and Tabachneck-Schijf (2009), and Pearce et al. (2010) have compared the performance of some segmentation systems. Other work on segmentation includes a neural study on finding working memory triggers (Burunat, Alluri, Toiviainen, Numminen, & Brattico, 2014) and a performance study on improvisational structure (Dean, Bailes, & Drummond, 2014). Outside our scope, work on musical *closure* has explored the role of musicianship and experience on boundary perception of classical music (Peebles, 2011; Sears, Caplin, & McAdams, 2014).

Recently, Bruderer (2008) investigated participants' perceptual segmentation of music in three formats: polyphonic audio, MIDI melodic lines, and polyphonic MIDI. This work tackled the effect of polyphony in music on segmentation, the role of perceived boundary strength on segmentation, and the prediction of perceptual segmentation via different melodic parsing models. The main findings by Bruderer included: 1) a similar pattern of results for all three versions of the stimuli, 2) a positive relationship between the frequency of indications of boundaries and their perceived strength ratings, 3) a positive relationship between the actual segmentation by listeners and three segmentation cues of parsing models: timbral changes, rest onsets, and attack-points (i.e., a long note in between two short notes). Bruderer also investigated the effects of musicianship on segmentation, but the approach was limited mainly by small sample size and a lack of professional musicians in the sample. In addition, the time scale parameter (see below) used for modelling boundary density across participants was adjusted based on multiple segmentation trials. Due to the need of several trials from the same participant, this method could result in rather lengthy data collection tasks if the issue under study does not involve repeated segmentation.

In this study, which can be considered a follow-up to the work by Bruderer (2008), we suggest a novel approach for modelling segmentation boundary density. We apply a comparable methodological approach (i.e., based on *kernel smoothing*) to study effects of music training upon participants' segmentation of polyphonic audio stimuli. We introduce alternatives to find optimal segmentation boundary density parameters (comparison between groups or tasks, and estimation of model-to-data fit; see Results).

Regarding the issue of experimental segmentation tasks, various methods have been used to gather segmentation boundary data, as there is no established approach and data collection method and comparison studies are scarce. Examples of segmentation tasks include listening to the example once followed by three consecutive real-time segmentation trials (Bruderer, 2008), and segmenting into two clusters *online* during listening (Peretz, 1989) or *offline* after listening (Deliège, 1987). Another study asked subjects to listen to the example, segment in real-time, and make changes or deletions to their boundary profiles to obtain a precise, non-real-time *annotation* for use in further experiments (Clarke & Krumhansl, 1990). Previous work on melodic clustering suggests the possibility that the data collection method has an effect on the boundary indications by listeners: Peretz (1989) compared an explicit segmentation task with an offline retrospective recognition memory task and an online prospective probe recognition task. Differences were found between tasks in the role of critical boundaries upon probe identification, suggesting that the mnemonic role of clustering for tune recognition is task dependent, and that similar tasks may, however, capture distinct stages of musical analysis. Several studies investigated the differences between repeated segmentations of the same stimuli, and reported an increase in the number of indications over repeated segmentations of the target stimulus (Bruderer, 2008; Deliège, 1987; Deliège, Mélen, Stammers, & Cross, 1996; Krumhansl, 1996). However, this trend did not reach statistical significance, and it was found for audio but not for MIDI versions of the stimuli (Bruderer, 2008). Frankland and Cohen (2004) asked listeners to parse MIDI melodies in three consecutive trials, and found an increase of within-subject correlation throughout repetitions. Koniari et al. (2001) compared children who listened to stimuli once prior to segmentation with children who had listened to the stimuli three times; no statistically significant effects of familiarization with the target stimuli were found over the segmentation profiles.

Regarding the role of musicianship on boundary perception, studies have reported effects of music training

on subject agreement and on number of indications. Results from studies rooted on the Generative Theory of Tonal Music (GTTM, see Lerdahl & Jackendoff, 1983) suggest the possibility that both musicians and nonmusicians can represent the hierarchical structure of the music from its perceived surface, but these representations would differ due to differences in musical skills (Deliège, 1987; Koniari & Tsougras, 2012; Peretz, 1989). Children (Koniari et al., 2001) and adults (Bruderer, 2008) with music training exhibited higher within-subject agreement: they showed more consistency across repeated segmentations of a target stimulus than untrained listeners. As regards inter-subject consistency, Schaefer, Murre, and Bod (2004) reported higher agreement between musically experienced listeners than between inexperienced ones. In addition, studies focusing on different aspects of segmentation have reported that participants with music training indicate roughly twice the number of boundaries than untrained ones (Bruderer, 2008; Deliège, 1987). Other studies investigated agreement of the segmentation with respect to Gestalt or GTTM rules, with the hypothesis that these rules would better predict musicians' segmentation. Subjects with music training segmented more in accordance with GTTM rules than untrained ones (Deliège, 1987; Koniari & Tsougras, 2012; Peretz, 1989), but the direction was inverse for general Gestalt rules (Schaefer et al., 2004).

An unsolved methodological issue in music segmentation studies is how to combine boundary indication profiles from multiple participants to obtain a representative model. This is not a trivial step since participants can greatly differ from one another with respect to the location of segment boundaries and to the number of indications. Moreover, it can be problematic to systematically match boundaries from different listeners that are close in time, since it requires researchers to determine whether listeners were indicating the same musical change. In order to estimate the temporal proximity between participants' indications that correspond to the same perceived event, the time constant or *time scale* of the segmentation should be optimized; for instance, if listeners' indications of the same musical change are quite distant in time from one another, larger time scales will be required for their aggregation, and vice versa. Since there is no common modelling approach to reliably obtain aggregate distributions of point process data or to measure their similarity (Dauwels, Vialatte, Weber, & Cichocki, 2009), multiple methods have been used in music perception, from sampling responses that are roughly close enough in time (Koniari & Tsougras, 2012; Koniari et al., 2001) to summing indications within each musical beat (Krumhansl, 1996) and note (Deliège, 1987; Deliège et al., 1996; Frankland & Cohen, 2004). These models are best suited for monophonic music, especially for discrete events in the symbolic domain, but not for polyphonic audio music, which involves overlapping events and frequent timbral change.

An alternative approach that has not received enough attention in music segmentation studies is Gaussian kernel smoothing. This method models segmentation data by placing a Gaussian curve at each boundary to estimate an underlying probability density function (Silverman, 1986). The result is a curve of perceptual segment boundary density over time; its local peaks represent regions where multiple boundary indications are close enough in time. The smoothness of this representation can be modified by increasing the width of the Gaussian kernel used. If participants' indications of the same musical change are not close enough, the smoothness parameter of the curve should be increased to reduce its noisiness. However, very high smoothness results in an inaccurate curve that would represent different musical changes with only one peak. To offer an optimal representation of perceived musical change across multiple listeners, an appropriate level of smoothness needs to be found. Segmentations at a high time scale are optimally represented with larger kernel widths, and vice versa.

Smooth density profiles of 1 s (Burunat et al., 2014) and 1.25 s (Bruderer, 2008) have been suggested for modelling the distribution of boundary indications. Burunat et al. (2014) found after repeated optimization trials that a time scale parameter of 1 s could optimally group together motif-level segmentation data of a stimulus. Using six stimuli, Bruderer (2008) found an optimal width of 1.25 s based on differences between individual data for three consecutive segmentation trials. This method yields a length at which most windows include marks for all trials, but least windows include more than one mark within any trial. This approach exhibits some limitations: it requires each participant to segment the same stimulus multiple times, uses an arbitrary number of trials, and assumes similarity of profiles across trials. One of the main findings obtained via this approach was that the estimated boundary density corresponded to boundary strength ratings, since the rated strength of a subset of indicated boundaries correlated strongly with the frequency of indications, as previously predicted by Clarke and Krumhansl (1990) and Frankland and Cohen (2004). Another approach to obtaining a representation of segmentation density would be to use multi-scale models; these have been applied for music visualization and

analysis of structure (Kaiser & Peeters, 2013; Martorell Dominguez, 2013; Mauch, MacCallum, Levy, & Leroi, 2015). Multi-scale models of density offer a more comprehensive representation of hierarchical aspects of segmentation than density profiles.

The literature shows at least three aspects of segmentation that remain to be tackled regarding musicianship, experimental tasks, models, and stimuli. First, the effect of music training remains an open question in phrase-level segmentation. One reason for this is the lack of assessment of differences in music training among participants (Krumhansl, 1996). Another issue is small sample size. Bruderer (2008) included only 7 participants in the sample, none of whom were professional musicians. Other related questions, such as relative delay between participant groups, were not investigated. Understanding the role of music training on participants' segmentation can yield clues on transfer effects of musicianship and guide recruiting of participants for further music listening studies. Second, listeners in segmentation tasks get familiar with target stimuli in initial "listening only" or practice trials. This procedure is based on the assumption that a complete hierarchical mental representation of a stimulus can only be achieved after it is heard in its entirety (Lerdahl & Jackendoff, 1983), hence boundary indication tasks require a familiarization step. According to this principle, real-time segmentation tasks should be preceded with "listening only" trials or repeated multiple times with the same stimulus, and offline segmentation tasks would provide a more complete representation of the perceived structure. It has been shown that repetition of real-time segmentation increases within-subject consistency, suggesting an effect of retrospective aspects upon segmentation. However, to our knowledge few studies have investigated the effect of real-time compared to offline segmentation, particularly when it comes to clustering of relatively large examples into multiple parts. The effect of task should be further explored to, for instance, compare real-time brain activity during music listening against expert annotations of musical structure. Third, few perceptual segmentation models based on indications by participants have been suggested, and versatile strategies are required to find optimal time scales for modelling. The relationship between the optimal segmentation time scale of a stimulus and its musical characteristics also remains a question. Robust models of multiple segmentations oriented towards naturalistic stimuli can provide further insights on perceived structure and be advantageous for automatic structural analyses.

The aims of this study, which investigates the contribution of music training and segmentation task in phrase-level segmentation, and estimates optimal time scales for segmentation modelling, can be condensed into the following questions:

1. What is the effect of music training on the indication of musical segment boundaries by listeners in a real-time type of experimental setup?
2. What are the differences between a first impression of musical structure as it unfolds over time and an offline, more knowledge-driven music segmentation?
3. Which global characteristics of musical stimuli modulate the optimal time scale for modelling perceptual segmentation?

Regarding the first question, we expected to find differences between segmentation profiles due to music training. We hypothesized that nonmusicians' segmentation would be delayed compared to musicians' segmentation, due to lower recognition delay of boundaries found for musicians and attributed to processing of shorter auditory time-spans (Tierney, Bergeson-Dana, & Pisoni, 2008). We also expected that musicians would exhibit higher inter-subject correlation compared to nonmusicians, who would be less likely to segment in accordance with internalized perceptual rules regarding musical form (Koniari & Tsougras, 2012). Also, it was expected that nonmusicians would indicate more boundaries than musicians, as previously suggested by Bruderer (2008) and Deliège (1987). Another specific hypothesis derived from previous studies was that some dissimilarities between musicians' and nonmusicians' multi-scale segmentation models would be exhibited, as previous differences have been shown; for example, in segmentation of short melodies (Deliège, 1987; Peretz, 1989). We also expected to find differences in optimal segmentation time scales due to musicianship: segmentation by nonmusicians would be optimally represented by lower levels of smoothness (short time scales), under the assumption that they would focus predominantly on lower levels of the hierarchical grouping structure, including changes of loudness, timbre, pitch, and duration. In contrast, we expected high smoothness (large time scales) to be more suitable for estimation of boundary distribution from musicians as they would focus not only on dynamics, instrumentation, register, and pace, but also on higher structural levels (chord and key changes, metric modulation, multiple concurrent changes). For instance, a study on perceived closure of classical cadences (Sears et al., 2014) showed that nonmusicians focus mainly on the leading voice, whereas musicians pay attention to multiple voices, suggesting greater salience of harmonic change for musicians. (Tierney et al., 2008).

For the second research question, we expected to find an effect of experimental task on segmentation: the real-time task was expected to prompt more inaccurate and incomplete segmentations than the annotation task. Since certain aspects of segmentation might only be perceived in retrospect, the real-time task should make it difficult for listeners to anticipate development or repetition of ongoing phrases, and hence to decide whether to indicate a boundary or not. Specifically, real-time segmentation contexts should exhibit relatively delayed boundaries due to the time required by participants to recognize musical changes as significant and respond by indicating them: if the musical context does not facilitate boundary anticipation, listeners might need to pay attention to subsequent musical events in order to recognize and indicate a boundary. We also expected that real-time task segmentations would be more dissimilar with each other than non-real-time segmentations due to variation among participants in their ability to anticipate boundaries and in their delay to respond to recent musical changes. Also, the non-real-time task would probably exhibit more boundary indications, such as those prompted by retrospectively perceivable musical changes, whereas in the real-time task only stark musical contrast (e.g., simultaneous change in instrumentation, harmonic function, melodic contour, and rhythmic patterns) would be indicated. We hypothesized, however, that both tasks would share some commonalities. First, the perceived strength ratings of a boundary in the non-real-time task would somewhat reflect the proportion of participants that indicated it in the real-time task. For example, the real-time task would mostly prompt indication of stark and predictable boundaries, which should be among those boundaries perceived as strongest in the annotation task. We also expected that, at a general level, both tasks would exhibit relatively high similarity since the real-time task would still yield a broad representation of the perceived musical structure. These tasks would become comparable when using large time scale parameters, because high levels of smoothness would reduce differences between tasks caused by recognition delay and retrospective aspects of segmentation (which are only compensated in the annotation task). Each task, in this sense, was expected to involve different optimal time scales for its representation: real-time segmentations should describe simultaneous change of multiple musical attributes, which would be optimally estimated with large time scales. In contrast, comprehensive, non-real-time annotation tasks might induce segmentation at multiple hierarchical grouping levels, ranging from beats to larger patterns such as melodic sequences. This type of annotation would be comparable to a GTTM time-span reduction; according to this, a single time scale cannot suffice for density estimation, but small time scales can still offer an appropriate representation of the trend across listeners towards frequent segmentation.

Regarding our third research question, we expected that optimal segmentation time scales for modelling responses across participants would relate to global rhythmic description cues of each stimulus, such as estimated beat length, pulse clarity, duration, and number of note events. The underlying assumption was that the optimal time scale for modelling responses would not be stimulus invariant; it would instead depend on rhythmic properties of each stimulus, such as ability to evoke a sense of beat and meter. For instance, musical pieces with lower rhythmic stability would induce less precise annotations by participants, so larger time scales would be required for modelling segmentation density. Similarly, segmentation of music with a relatively low number of events should hinder listeners' boundary anticipation, resulting in sparser boundary profiles that would require higher levels of smoothness for density estimation. Support for this hypothesis would shed light on the relationship between perceptual boundary data and audio rhythmic features, and lend validity to the proposed modelling approach for estimation of optimal time scales for segmentation.

Method

We conducted two listening experiments on perceptual segmentation at the Music Department of the University of Jyväskylä. Figure 1 illustrates the computer interfaces that were utilized to collect segmentation responses.

Experiment 1: Real-time Task

The first experiment collected significant instants of change that were indicated by participants as they listened to unfamiliar stimuli. Our general aim for this experiment was to capture a fresh, "live" description or first impression of the music as it unfolded over time.

APPARATUS AND STIMULUS MATERIALS

We collected real-time segmentation responses, stimuli familiarity, and background information from subjects via a Max/MSP computer patch. The stimuli used in the experiment were 18 excerpts from 9 multi-instrumental and polyphonic piano musical pieces (see Appendix for abbreviations and information) comprising various styles. The musical examples were mainly excerpts extracted from longer pieces, and their duration ranged

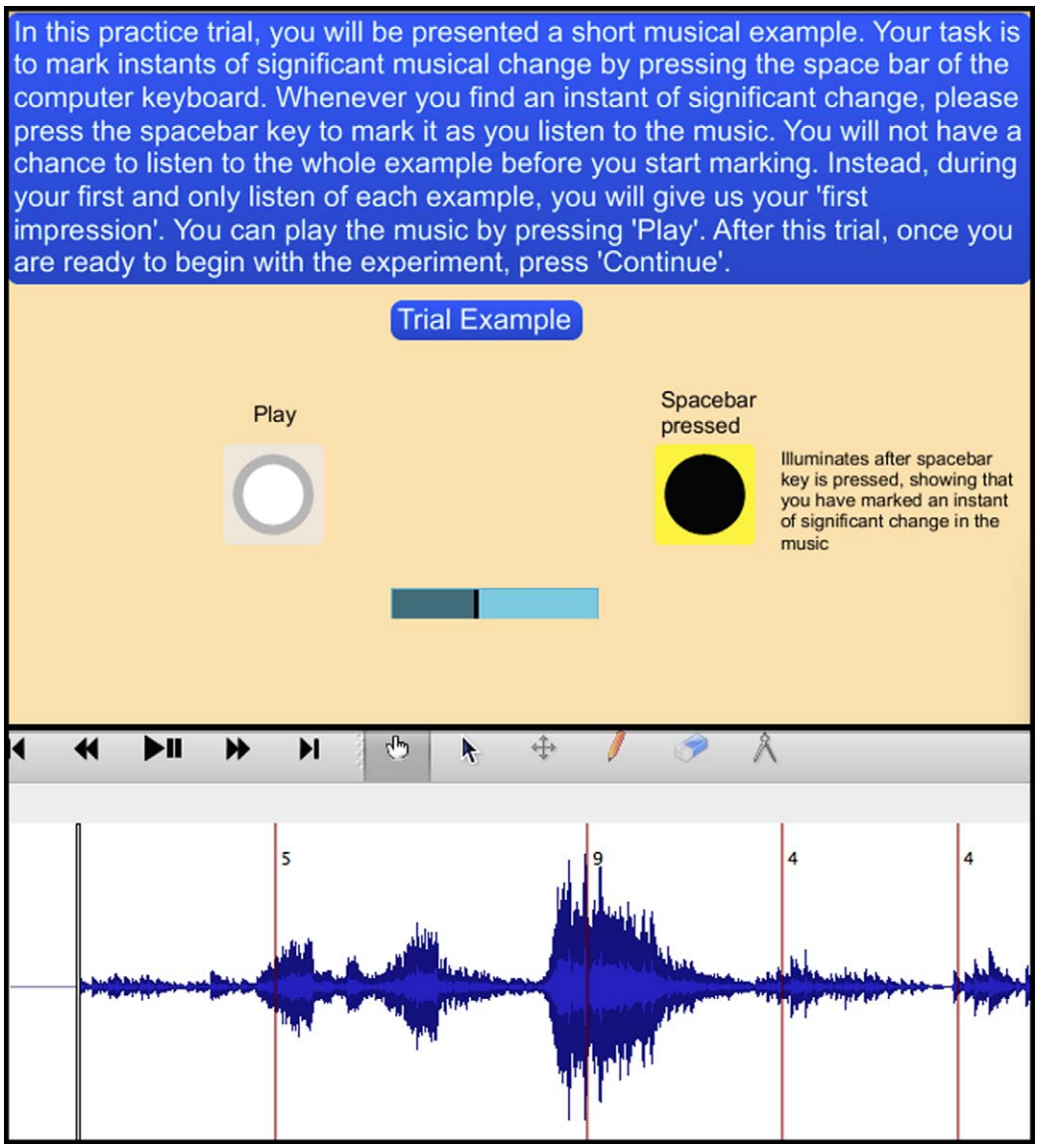

FIGURE 1. Upper image (Experiment 1): Trial instructing listeners to indicate instants of significant change while listening to the music. Lower image (Experiment 2): Part of an annotation segmentation performed by a musician for the stimulus *Ravel*; vertical bars indicate marked boundaries, and numbers situated next to the bars indicate perceived boundary strength ratings.

from 2 to 8 min. We trimmed the 8-min examples into chunks of around 2 min for an even length distribution and to avoid fatigue of participants. In order to contextualize the section ends and beginnings, these were overlapped with each other by 3 s, which corresponds to the duration of the echoic memory store (Toiviainen & Krumhansl, 2003). After the experiment we concatenated the segmentation data from these chunks to obtain sets of boundary data for entire musical examples. The root mean square (RMS) energy level was normalized for the level of the stimulus with the lowest value, and the peak intensity was adjusted for each stimulus. The whole set thus exhibited approximately homogeneous loudness so participants could listen via headphones at comfortable volume levels.

The musical pieces that were selected for the experiments do not only differ in style; the temporal structure varies in quantity and type of dimensions that manifest musical progression: harmony (*Morton*), instrumentation, and harmony (*Dream Theater*), tempo and harmony (*Couperin*), dynamics, instrumentation and harmony (*Genesis*), tempo, instrumentation, and

harmony (*Smetana*), dynamics, tempo, and harmony (*Ravel*), and dynamics, tempo, instrumentation, and harmony (*Dvořák*, *Piazzolla*, *Stravinsky*). Other criteria were also used for selecting the stimuli; we focused on polyphonic material, in the sense of music containing simultaneous note events, to prompt segmentation relying on processes of texture change. We included only music without lyrics since these were found to have an important effect on boundary perception (Bruderer, 2008), and hence would have posed difficulties for estimation of general trends across stimuli. The duration of the stimuli had to be long enough to invoke segmentation (over a minute of music), but short enough to avoid fatigue of participants (our upper limit was of 10 min). Besides the selection of multiple musical idioms, we aimed to obtain more generalizable results by including stimuli with varying structural complexity, and whose boundaries would be induced by different musical elements (timbre, rhythm, harmony) or interactions thereof. We also considered the availability and adequacy of MIDI versions of the stimuli for future work that could take advantage of symbolic musical descriptions (large interonset intervals in *Smetana* and *Dvořák*, long rests in *Morton*, *Ravel,* and *Piazzolla*). We included music that would induce segmentation due to complex processes such as similarity (*Genesis*, *Morton*, *Couperin*, *Dvořák,* and *Piazzolla*), symmetry (*Dvořák*) and texture change (*Genesis*, *Stravinsky,* and *Dream Theater*). Moreover, in most cases the stimuli presented in the real-time task would not be known to the participants in order to reduce artifacts due to familiarity.

We believe that some of the stimuli might be relatively more challenging to segment, particularly in real-time contexts. *Stravinsky* and *Ravel* are characterized by unexpected but highly contrasting musical changes in loudness, texture, rhythm, and tonality. For *Couperin*, on the other hand, some boundaries are more subtle and can only be anticipated due to underlying tonal context. The rhythmic organization of this piece also induces phrase grouping, but some local temporal discontinuities might be difficult to anticipate in real-time contexts. *Morton* is also characterized by rhythmic discontinuity, here perceived as sudden breaks followed by long pauses, which are likely very hard to anticipate during the first listening. Also, the introduction of *Piazzolla* could sound erratic due to lack of key clarity and abrupt changes and pauses.

SUBJECTS

We obtained segmentation data from 18 nonmusicians (11 males, 7 females) and 18 musicians (10 females, 8 males). One of our aims was to collect data from even-participant samples regarding demographic information (gender and age) and musical styles played by musicians. The mean age was similar across groups: nonmusicians = 27.28 years ($SD = 4.64$), musicians = 27.61 years ($SD = 4.45$). The subjects were local and foreign students and graduates from the University of Jyväskylä and Jyväskylä University of Applied Sciences. The musicians had an average of 14.39 years ($SD = 7.49$) of music training and played classical (12 participants) and non classical musical styles (6 participants) such as rock. The main instruments played by the musicians were piano (5), guitar (4), flute (2), bass guitar, clarinet, saxophone, cello, violin, viola, and voice. All the musician participants considered themselves either semiprofessional (12 participants) or professional (6 participants) musicians with 6 or more years of training. All nonmusicians self-reported as untrained, and none of the participants reported skills in dance or sound engineering.

PROCEDURE

The experiment took place in two sound-attenuated rooms with a computer. The average duration was around 50 min for nonmusicians and 47 min for musicians. The main experiment task was described to participants 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 spacebar 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.'" After reading instructions and completing a trial, they segmented each of the musical stimuli, which were presented in randomized order. Participants did not have an opportunity to listen to the whole example beforehand. The interface had a play bar that offered basic visual-spatial cues regarding the beginning, current time position, and end of the stimuli. After the segmentation of each target stimulus, participants indicated their familiarity with it via a 5-point Likert scale.

After the segmentation of all the target stimuli, participants filled out a questionnaire including demographic and music-related questions. We gathered information regarding music training, weekly frequency of music listening, and favorite musical genres of the participants. Participants who reported music training accessed an additional questionnaire regarding musicianship and including professional status. This questionnaire also asked about main instrument and other instruments played, musical styles played, and number of years of training. This information was further

utilized to match participants from both groups, remove outliers, and include a diverse sample of participants (e.g., different kinds of instrumentalists and styles performed). After this, the experimenter asked subjects for some feedback on the task and rewarded them with a movie ticket.

Experiment 2: Annotation Task

We conducted a second experiment with the purpose of obtaining a more comprehensive and precise set of segmentations from participants. For this experiment, we recruited musicians who had participated in Experiment 1 and who had reported experience in audio editing tasks. We did not include nonmusicians in this experiment because only a small number of them had reported previous audio editing experience. In this experiment, each target stimulus was presented for listening before the segmentation task to prompt more deliberate indications. Subjects were asked to mark instants of significant change while listening to stimuli, similar to what they had done in Experiment 1. The last steps were to correct imprecise time locations or discard unwanted marks, and to rate the perceived strength of each boundary. Participants were asked not to add new marks at that point, under the assumption that they would tend to over-segment while focusing on short excerpts of the stimuli (following Krumhansl, 1996).

APPARATUS AND STIMULUS MATERIALS

We prepared an interface in Sonic Visualiser (Cannam, Landone, & Sandler, 2010) to collect time points and strength ratings of indicated boundaries from 6 musical examples. Participants used headphones to playback the music at a comfortable listening level and a keyboard and mouse for the segmentation task. To keep the total duration of Experiment 2 at around one hour, we used 6 stimuli from Experiment 1 that lasted around 2 min each. We did not include *Piazzolla*, *Dream Theater* or *Stravinsky* in Experiment 2 since these were 6 min longer than the other stimuli.

SUBJECTS

The same 18 musicians of Experiment 1 participated in Experiment 2, and they were all familiar with the use of audio editing software.

PROCEDURE

The experiment was conducted in a room with a computer with the exception of two subjects who participated at the same time in a computer laboratory. Contrary to Experiment 1, which did not require assistance, in this case the experimenter remained in the room during the training to make sure that the task was clear. The experimenter read each step of the instructions together with the participant and occasionally answered questions regarding the task. The participant performed the task via two trial stimuli by following the instructions, and after this the experimenter left the room. The written instructions included a presentation of the interface tools and a task description, which consisted of the following steps:

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 to 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.

The interface showed stimuli waveforms over which subjects would play back the music, add marks, reposition them, and rate their strength. The waveforms could bias participants towards boundary indications based on amplitude changes, so they were asked to focus on the music rather than on visual content. These visual-spatial cues, which are often used for expert annotation of structure in Music Information Retrieval (MIR), were needed due to the detailed audio editing that was needed for the task. After the participants completed the task, which lasted an hour on average, they provided feedback and were rewarded with a movie ticket.

Results

Table 1 includes information about age, training, and listening habits of participants. The mean listening habits (music listening hours per week) of participants were significantly higher for the group of musicians, $t(34) = 2.26$, $p < .05$, although they showed more dispersion in this respect (two musicians explained that they seldom listen actively to music as a primary activity although their whole day is usually consumed with musical activity). Five musicians and one nonmusician were familiar with at least one target stimulus, but nobody reported having performed any of the examples. The mean familiarity rating (1 = *not at all familiar*; 5 = *very*

TABLE 1. *Age, Performance Training, and Listening Habits (Hours Per Week) of Participants*

Group	$\overline{x}$ age (*SD*)	Range	$\overline{x}$ years training (*SD*)	Range	$\overline{x}$ hours/week listening (*SD*)	Range
NM	27.28 (4.64)	20 - 34	0	0	10.7 (8.6)	1 - 30
M	27.61 (4.45)	22 - 36	14.39 (7.49)	4 - 32	19.9 (15.7)	2 - 70

TABLE 2. *Sets of Indicated Boundaries Used For Segmentation Modelling and Their Respective Abbreviations*

	Nonmusicians	Musicians
Real-time Task	*NMrt*	*Mrt*
Annotation Task		*Ma*
Annotation Task$_{\text{boundary strength weights}}$		*Ma*$_w$

Note: NMrt = boundary indications by nonmusicians in the real-time task (Experiment 1). Mrt = boundaries indicated by musicians in the real-time task (Experiment 1). Ma = boundary indications by musicians in the annotation task (Experiment 2). Ma$_w$ = indications by musicians in the annotation task with the addition of perceived boundary strength weights (Experiment 2).

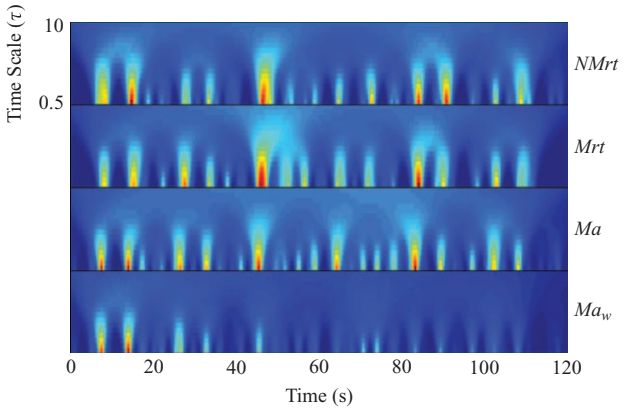

FIGURE 2. Each of the four sets of indicated boundaries was modelled via a multiple time scale approach. The kernel density over time of stimulus *Morton* is represented for 16 time scales.

familiar) per stimulus across participants was 2.4 (mean *SD* = 1.4) for musicians and 2.1 (mean *SD* = 1.1) for nonmusicians. The most familiar pieces for musicians were Stravinsky ($\overline{x}$ = 3.1, *SD* = 1.3), Piazzolla ($\overline{x}$ = 2.9, *SD* = 1.6), and Ravel ($\overline{x}$ = 2.4, *SD* = 1.4). Regarding nonmusicians, they were most familiar with Stravinsky ($\overline{x}$ = 2.7, *SD* = 1.5), Dream Theater ($\overline{x}$ = 2.6, *SD* = 1.1), and Piazzolla ($\overline{x}$= 2.4, *SD* = 1.4).

The responses collected from participants were further processed in order to enable comparisons between the data structures of each task. For the trimmed 8-min examples, we corrected overlapped chunk ends and beginnings by discarding data from the first 3 s of each chunk, except for the initial chunk. For each of these examples, we then concatenated the data across chunks to obtain a set of boundary indications for the full musical example length.

Subsequently, we organized the data as three main sets based on the music training of the participants and the segmentation task that was performed. We allocated 162 segmentations per participant group in the real-time task, since 18 participants per group segmented 9 musical stimuli. For the annotation task set, we allocated 108 segmentations by 18 musicians as each subject segmented 6 musical examples. For brevity's sake we abbreviate the real-time task by nonmusicians to *NMrt* and by musicians to *Mrt*, and for musicians in the annotation task to *Ma* (see Table 2).

To yield global trends across listeners, we utilized a systematic and multi-hierarchical approach. For each group and task we computed segment boundary probability curves using Kernel Density Estimation (KDE, Silverman, 1986). KDEs are comparable to histograms, which are also density estimators, but yield smooth distributions because a kernel function is applied to each data point (in this case each boundary indication) instead of separating data points into bins. For distribution smoothing, we chose a normal kernel function following previous studies (Bruderer, 2008; Burunat et al., 2014). To compare different participant groups and experimental tasks, we obtained perceptual segment boundary density curves at varying smoothing bandwidths; these corresponded to 16 time scales logarithmically ranging from .5 s to 10 s in order to model multiple hierarchical levels. Previous studies (Bruderer, 2008; Burunat et al., 2014) showed that short time scales are optimal for segmentation, so we chose logarithmic scales to efficiently cover these in detail while also providing information regarding larger time scales. We combined single-scale models of different time scales to build matrices in which each row included a perceptual segment boundary density curve at a given time scale, and each column included boundary density for a given time point at different time scales. This multi-scale model of segmentation follows previous work on tonality (Martorell Dominguez, 2013) and musical novelty description (Kaiser & Peeters, 2013; Mauch et al., 2015). We obtained a multi-scale model for each stimulus and segmentation task; Figure 2 shows each of the

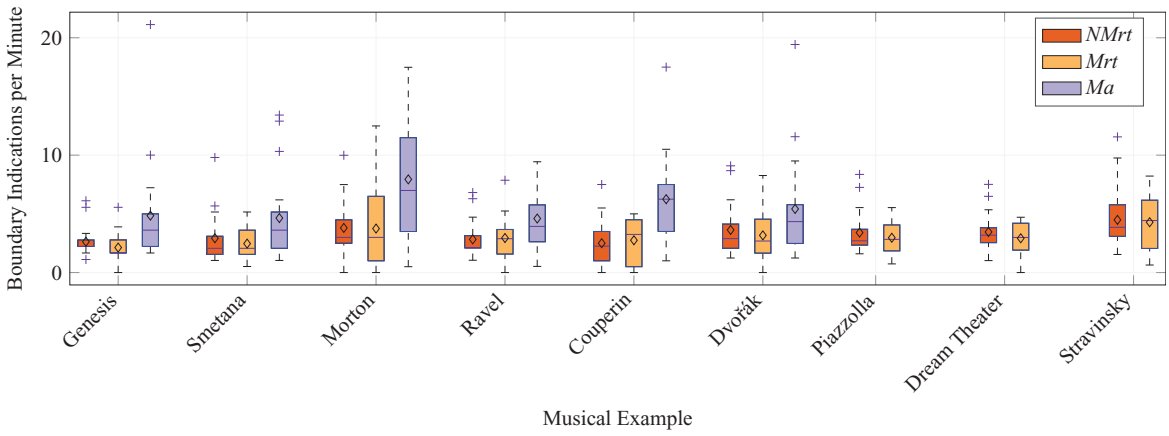

FIGURE 3. Box plot comparing participant groups and segmentation tasks with respect to the number of indicated boundaries per minute for each stimulus.

four multi-scale models obtained for stimulus *Morton*. Within each KDE matrix there are 16 single-scale models, which are ordered along the vertical axis based on their time scale (τ), which ranges from 0.5 s to 10 s.

We included an additional data set with responses by musicians in the annotation task to analyze the role of perceived boundary strength; this set was abbreviated as Ma_w. To generate Ma_w, each of the single-scale models of the annotation task (*Ma*) was weighted based on listeners' boundary strength ratings. This fourth set contained boundary indications at the same time instants as *Ma*, allowing to estimate the boundary strength effect. We mapped for each participant separately minima and maxima strength values to 1 and 10, since only a few subjects used the full range of values.

NUMBER OF BOUNDARY INDICATIONS

We looked at the total number of indicated boundaries by each participant with the primary purpose of removing outliers from the sample. We found that all participants were located within 3 standard deviations from the mean, so no sample subjects were removed. Figure 3 compares segmentation tasks and participant groups based on the number of boundary indications per minute for each stimulus. In this and the following box plots, whiskers describe about ± 2.7 *SD* (for normally distributed data), hence covering 99.3% of the total data; mean values are shown with diamond marks. For the first 6 stimuli, the number of boundary indications per example by musicians ranged in the real-time task between 0 and 49, and in the annotation task between 1 and 47. Regarding the real-time task, the number of indicated boundaries for each of the 9 musical stimuli ranged between 0 and 90 for nonmusician participants

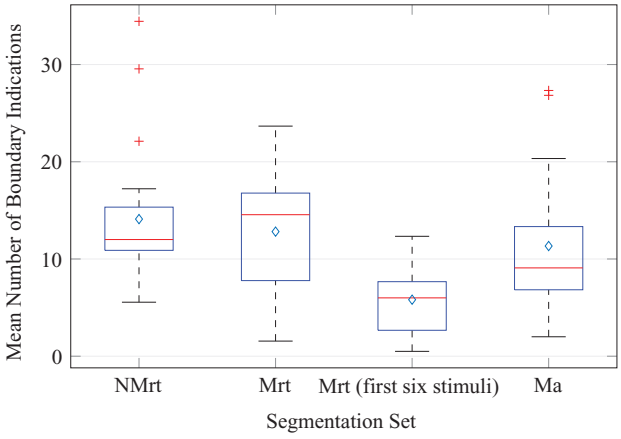

FIGURE 4. Box plot comparing participant groups and segmentation tasks with respect to the mean number of indicated boundaries by each participant.

and between 0 and 64 for musicians. All participants indicated at least a total of 14 boundaries in the real-time task. Some participants mentioned after the task that they indicated few musical changes due to focus on those that were sufficiently significant; four musicians and two nonmusicians segmented once or not at all in some of the segmentation trials, comprising 7% of the 324 collected trials in the real-time task.

We compared the mean number of boundary indications for each segmentation task and participant group (Figure 4). The task comparison showed that participants indicated nearly double the number of boundaries in the annotation task ($\overline{x} = 11.33$, $SD = 8.06$) compared to the real-time task ($\overline{x} = 5.81$, $SD = 4.09$) for the six stimuli that were common to both. We computed paired samples, two-tailed *t*-tests to determine whether the

difference between tasks was statistically significant (H0: mean difference between tasks in the number of boundary indications by participants is equal to zero). We found that musicians indicated significantly more boundaries in the annotation task than in the real-time task for 5 out of 6 stimuli; the difference was significant at $\alpha = .01$ for the stimulus *Couperin*, $t(17) = 3.33$, $p < .01$, and at $\alpha = .05$ for the examples *Genesis*, $t(17) = 2.32$, $p < .05$, *Smetana*, $t(17) = 2.14$, $p < .05$, *Morton*, $t(17) = 2.83$, $p < .05$, and *Ravel*, $t(17) = 2.24$, $p < .05$. However, we did not find a statistically significant difference between tasks for the example *Dvořák*, $t(17) = 1.77$, $p > .05$, since p slightly exceeded .05. The group comparison showed that nonmusicians indicated more boundaries (2285) than musicians (2076), but the difference between groups was not statistically significant for any of the stimuli.

BOUNDARY STRENGTH RATINGS AND LOCAL BOUNDARY DENSITY

Subsequently, we focused on the possible relationship betweeen perceived boundary strength and segment boundary density in order to estimate the external validity of the main finding by Bruderer (2008). We investigated whether musicians' ratings of boundary strength in the annotation task corresponded with the modelled density from the real-time task. For each considered time scale, we correlated the perceived boundary strength values with the real-time task model values at the respective time points (H0: no correlation between boundary strength and segmentation density values of boundary indications); for this analysis we used a version of the real-time task model that was time-aligned with the annotation task model (see below). In addition, we included a time scale of 1.25 s into the KDE matrix for this analysis, since this value was considered optimal by Bruderer (2008) for single-scale modelling of boundary data. We obtained weak mean correlation (around $r = .20$) across stimuli for all the 17 time scales (Figure 5), although the stimulus *Smetana* exhibited moderate correlations —peaking at a time scale of 1.11 s, $r(159) = .54$, $p < .001$ — for time scales below 5.5 s, and weak results above this time scale. We repeated this procedure for the boundary density in the annotation task to find out whether the rated strength of a boundary correlated with its corresponding density value. The overall correlation between perceived strength values and boundary density at the respective time points was in this case very low (around $r = .10$). In sum, the obtained results suggest that boundaries perceived as strong were not more likely to be indicated by participants.

Since these findings contradicted previous research, it was hypothesized that in the annotation task

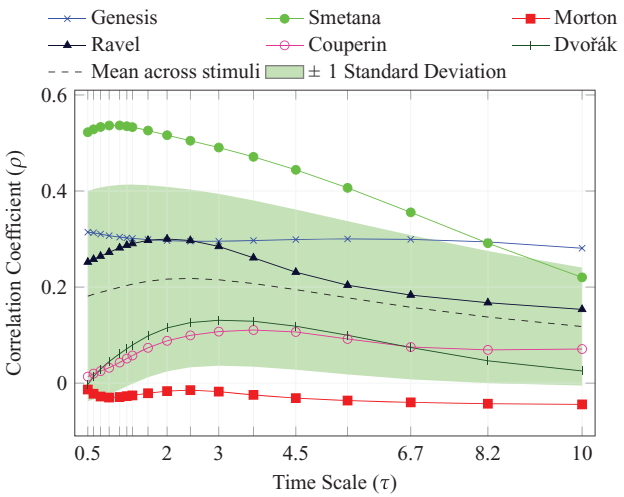

FIGURE 5. Confidence interval plot of correlation between boundary strength ratings in the annotation task and boundary density in the real-time task.

participants did not limit their segmentation to significant instants of change only, but indicated boundaries at multiple hierarchical levels instead. The reason for this would have been that the task induced participants to modify their segmentation strategies, because participants were aware from the instructions that they would have to rate the strength of each boundary after the segmentation. To test this possibility, we calculated the distribution of boundary indications into each strength rating, expecting a large frequency of low strength annotations. The results (1 = 14%, 2 = 11%, 3 = 14%, 4 = 10%, 5 = 15%, 6 = 6%, 7 = 4%, 8 = 8%, 9 = 4%, 10 = 13%) showed indeed a tendency towards low strength boundary indications, since the strength of 49% of the indications was rated between 1 and 4. This suggests that participants tended to indicate all possible boundaries, not only the most significant ones, and thus might explain why boundary strength ratings did not correlate with boundary density. Altogether, we could not find a relationship between boundary strength ratings by participants and boundary density at indicated instants. Because of this, we left the weighted data out of most of the subsequent analyses to focus on the effect of training and task on segmentation.

MEAN INTER-SUBJECT CORRELATION

Next, we examined the degree of cohesion in each segmentation set; to this aim we calculated the mean correlation between subjects within each set and for each example. For each segmentation set and stimulus we computed 18 individual multi-scale models, one model

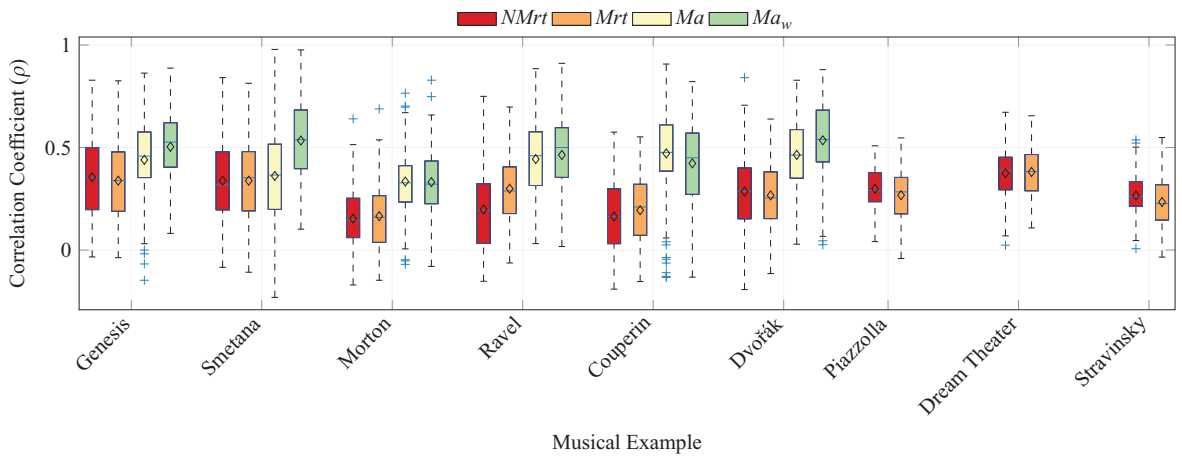

FIGURE 6. Box plot showing inter-subject correlation coefficient per stimulus for each segmentation set; p values (***$p < .001$) obtained via 10,000 Monte Carlo replications and adjusted using Benjamini-Hochberg correction ($q = 0.05$).

per participant, and correlated each pair (H0: the mean inter-subject correlation is equal to the mean of empirical distribution). Figure 6 presents, for each stimulus, the inter-subject correlation coefficient of each segmentation task and group of participants; high mean inter-subject correlation coefficients indicate similar segmentations between most or all participant pairs within a set. Regarding segmentation tasks, the annotation task yielded higher mean inter-subject correlations than the real-time task for all 6 stimuli. Apart from two exceptions, the addition of boundary strength weights to the annotation task led to an increase in cohesion, particularly for 3 stimuli for which the mean inter-subject correlation reached over $r = .50$.

In contrast, the profiles between participant groups were highly alike; nonmusicians, however, exhibited lower mean inter-subject correlations than musicians did for the musical stimulus *Ravel*. All the reported mean inter- subject correlations were significant at $\alpha = .001$ after the adjustment of p values for multiple comparisons via a Benjamini-Hochberg correction procedure ($q = .05$). For each pair of participants and stimulus, p values here indicate the probability of obtaining the actual results if the boundaries corresponding to one of the participants had been randomly placed. To obtain the p values, we performed a Monte Carlo simulation with 10,000 iterations for each of the stimuli and task: 1) we produced 18 random segmentations (the number of boundaries of each segmentation matched the total number of boundaries marked by each participant); 2) we obtained 18 multi-scale models, each one based on a random segmentation; 3) we computed their mean inter-subject correlation. These steps were repeated 10,000 times to generate a random distribution of mean inter-subject correlation. Finally, we calculated how many times this random distribution yielded larger values than the mean inter-subject correlation obtained from participants, and divided this result by the length of the distribution (10,000).

TIME SCALE FOR BEST MODEL FIT TO BOUNDARY INDICATIONS

Subsequently, we focused on which time scales were optimal for obtaining aggregate segment boundary data distributions. To this aim, we estimated the level of smoothing that provided an optimal fit of single-scale model to the boundary data for each segmentation set and musical example. For each subject, we obtained the log-likelihood between each single-scale model and individual data. To find which level of smoothing would offer the best fit to the data, for each time scale we summed the individual estimates together, and subsequently selected the time scale with the maximum sum of log-likelihoods. To avoid overfitting, the estimates were obtained with a leave-one-out procedure, such that for each subject we computed a model that did not include that subject. Figure 7 shows maximum likelihood time scales for each of the 6 stimuli that are common to all segmentation sets.

Comparing groups, musicians exhibited in average higher time scales than nonmusicians. We computed paired samples t-tests to find out whether the maximum likelihood time scales of musicians and nonmusicians were significantly different from each other (H0: mean difference between the maximum likelihood time scales of nonmusicians and musicians is equal to zero). We did not find a significant difference between groups for the first 6 stimuli, $t(5) = 1.02$, $p > .05$, nor for all 9 stimuli, $t(8) = 1.79$, $p > .05$. Comparing segmentation tasks, the

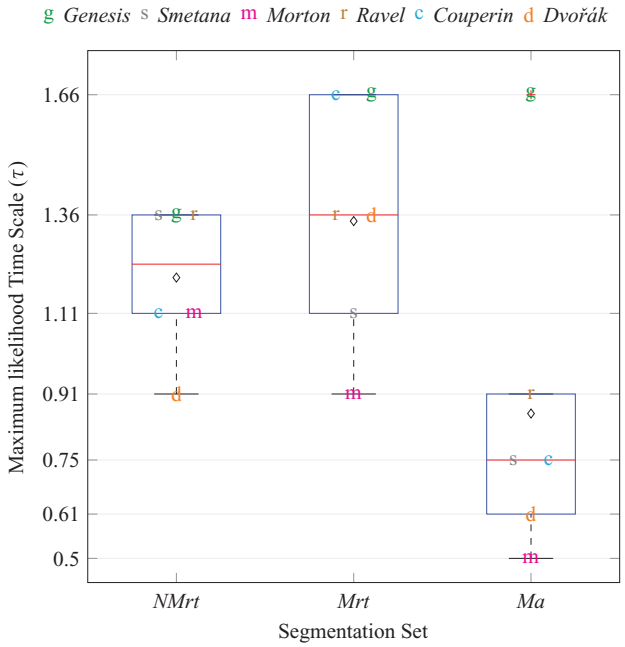

FIGURE 7. Box plot of maximum likelihood time scales for each segmentation set. Each time scale is represented with initials of its corresponding stimulus.

maximum likelihood time scale of each stimulus was larger in the real-time task than in the annotation task. A paired samples *t*-test between tasks was computed to find out whether their maximum likelihood time scales differed (H0: mean difference between the maximum likelihood time scales of real-time and annotation task is equal to zero). We found a significant difference between real-time and annotation tasks, $t(5) = 3.39$, $p < .05$; the optimal time scales were hence significantly larger for the real-time task than for the annotation task.

ALIGNMENT BETWEEN TASKS AND GROUPS

Our next objective was to examine whether different segmentation models were aligned with each other. We estimated the delay in the real-time task with respect to the boundary placements in the annotation task. To this end, for each musical example we computed a two-dimensional cross-correlation between the real-time and annotation task models. We found that the real-time task was lagged from the annotation task and a mean optimal time lag between tasks across stimuli at 1.05 s ($SD = 0.15$). For subsequent analyses, we shifted backward the real-time task indications by 1.05 s for all stimuli because the optimal time lag variation among stimuli was small (from 0.9 s to 1.3 s).

We also investigated whether musicianship had an effect upon relative lags in the real-time task indications.

TABLE 3. *Optimal Time Lag For Alignment Between Groups*

Stimulus	Optimal Time Lag (τ)	Delayed Group
Genesis	0.6	*M*
Smetana	0	—
Morton	0.2	*NM*
Ravel	0.4	*NM*
Couperin	0.2	*M*
Dvořák	0.2	*NM*
Piazzolla	0.4	*M*
DT	0.2	*NM*
Stravinsky	0.2	*NM*
Mean (*SD*)	0 (0.33)	—

Note: KDE time scale = 1.6 s, M = delay by musicians, NM = delay by nonmusicians.

We therefore compared musicians and nonmusicians in the real-time task via the aforementioned cross-correlation procedure. We found high alignment between segmentations made by musicians and nonmusicians in this task, as shown in Table 3; the mean alignment between groups for each stimulus at a time scale of 1.6 s was 0 s ($SD = 0.33$). The delays found were minimal and did not follow a particular trend, even for other considered time scales, suggesting no time lag between groups.

Continuing, we assessed whether the variability of the optimal time lag among stimuli could be attributed to rhythmic differences between examples. We extracted global rhythmic descriptions from the music (beat length, average note duration, event density, and pulse clarity, using *MIRToolbox* 1.5, see Lartillot & Toiviainen, 2007) and compared these with the optimal time lags of the stimuli between segmentation tasks (H0: no correlation between rhythmic features and optimal time lag). We found a significant correlation, $r(4) = .87$, $p < .05$, between optimal time lag and stimulus global beat length ($\mathrm{BL} = \frac{60}{\mathrm{tempo}}$). This result indicates that real-time and annotation data are more closely aligned to each other for stimuli with shorter beat length, and vice versa. A simple linear regression was done to examine the impact of beat length on the optimal time lag between tasks (H0: beat length does not predict optimal time lag). Beat length significantly predicted optimal time lag, $\beta_1 = .72$, $t(4) = 3.48$, $p < .05$; $\beta_2 = .66$, $t(4) = 5.5$, $p < .01$. Beat length also explained a significant proportion of variance in optimal time lag, adjusted $R^2 = .69$, $F(1, 4) = 12.10$, $p < .05$. The obtained simple linear regression equation ($\tau = .72 \times BL + .66$), and particularly the nonzero intercept suggests that the lag in the real-time task can be explained not only by a delay dependent on beat length, but also by a constant time lag among stimuli. Figure 8 illustrates the prediction of

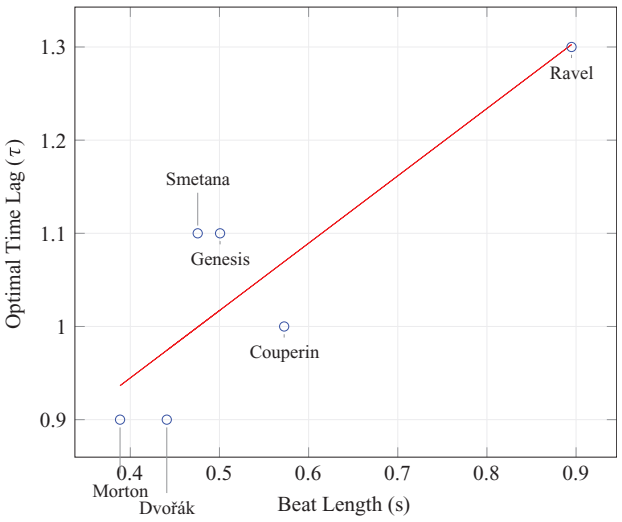

FIGURE 8. Scatter plot of optimal time lag for alignment between tasks as a function of stimuli beat length (BL). Trend line: simple linear regression equation $\tau = .72 \times BL + .66$.

optimal alignment between real-time and annotation segmentations based on beat length. Correlations with other rhythmic features were not significant, although correlation directions were as expected; average note duration: $r(4) = .71$, $p > .05$; pulse clarity: $r(4) = -.31$, $p > .05$; event density: $r(4) = -.25$, $p > .05$.

SIMILARITY BETWEEN TASKS AND GROUPS

Our following analyses focused on the similarity between segmentation sets for different participant groups and tasks; multiple approaches can be implemented to investigate this. One possible way to perform this analysis involves a detailed exploration of the segmentation profiles for particular excerpts based upon GTTM or other rules. For instance, Figure 9 illustrates the location in the score of some of the boundary indications for the example *Morton*. This fox-trot piano piece consists of a 4-bar introduction followed by a 12-bar blues progression. Differences between the profiles of musicians and nonmusicians in the real-time task include a boundary indication from a nonmusician at bar 15 (eleventh bar of the blues progression), which was probably elicited by the V7-I progression of the last two beats. Since the motif of bar 14 is repeated in bar 15, this segmentation is in agreement with GPR 6 (Parallelism), according to which parallel musical segments should be analyzed as parts of groups, and not as forming entire groups. This individual-level difference does not clearly show up from the multi-scale models, because the proposed approach highlights segmentation

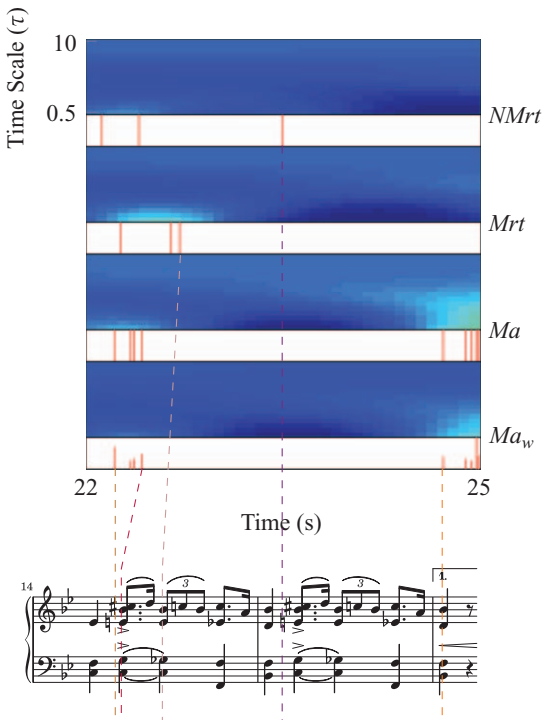

FIGURE 9. Multi-scale analysis for a 3-s extract of the stimulus *Morton*. Solid vertical lines: boundary indications by listeners. Dashed lines: approximate location of boundaries in the score.

responses at a group level. Interestingly, two musicians in the real-time task indicated a boundary at the beginning of the triplet in bar 14, perhaps due to boundary perception evoked by the C9-D9 chord change. In contrast, the annotation task exhibits a rather different multi-scale model and boundary profile, with two distinct boundary regions. The first region lies around the second note of bar 14 whereas the second region, located in bar 16, can be predicted by the parallelism rule; both boundary regions are in agreement with the attack-point proximity rule (GPR 2b). The annotation task profile suggests that boundary indications between these regions in the real-time task correspond to delayed responses, at least in the case of musicians.

CORRELATION BETWEEN MULTI-SCALE MODELS

In this study we opted to focus on a similarity analysis at a global level in search of trends based on whole musical stimuli. This choice is motivated, among other reasons, by the fact that real-world polyphonic music is not optimally suitable for rule-based approaches, or at least not as much as monophonic music in the symbolic domain is. For each musical stimulus, we compared each pair of multi-scale models; Figure 10 presents obtained

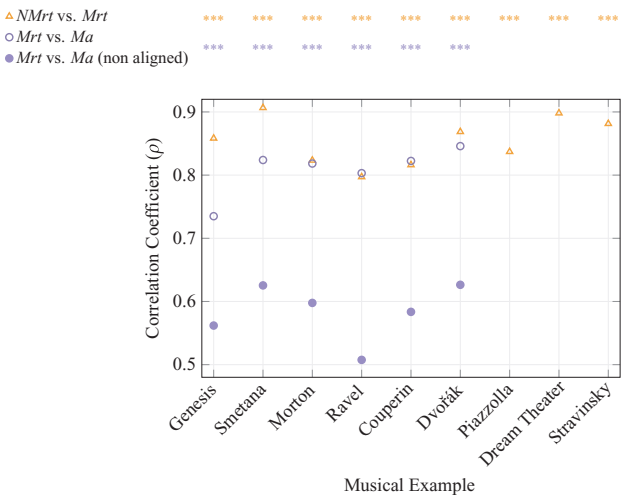

FIGURE 10. Multi-scale model correlation per stimulus comparing participant groups, segmentation tasks and alignment strategies; p values (***$p < .001$) obtained using Monte Carlo simulation and adjusted via Benjamini-Hochberg correction ($q = 0.05$).

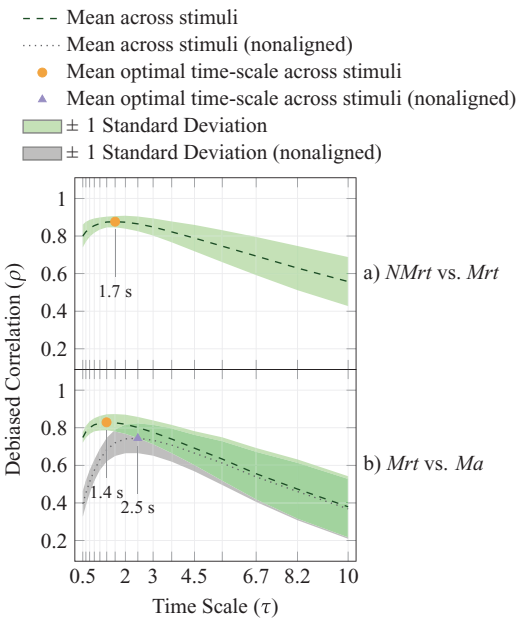

FIGURE 11. Mean correlation across stimuli between sets of indicated boundaries at 16 different time scales. Shaded areas: 1 standard deviation of the mean (estimate of correlation dispersion of different stimuli). Figure *a* compares participant groups in the real-time task. Figure *b* shows mean comparisons between segmentation tasks across stimuli for each alignment strategy.

correlations between groups, between tasks, and between alignment strategies (H0: the correlation between models equals the mean of empirical distribution). For groups, we found strong correlations between multi-scale models corresponding to musicians and nonmusicians. The task comparison also showed mostly strong correlations between real-time and annotation tasks by musicians for aligned segmentation models. For time alignment, the correlations between tasks for nonaligned models were weaker; the mean correlation reached $r = .58$ compared to $r = .81$ for aligned models. The reported p values (***$p < .001$) were drawn from a Monte Carlo simulation and were later adjusted for multiple testing using Benjamini-Hochberg correction ($q = 0.05$).

CORRELATION BETWEEN SINGLE-SCALE MODELS

We also examined the relationship between groups and between tasks at each time scale separately to determine which time scales yielded highest similarity between models. To this end, for each stimulus and time scale we computed correlations between participant groups and between segmentation tasks. A bias in the correlation coefficients caused by the smoothing of boundary indications was removed by using Monte Carlo simulation (10,000 iterations). We computed a correlation baseline for each combination of example and time scale, and then subtracted it from the original correlation. Figure 11 shows the mean and standard deviation of the debiased correlations across musical examples at each time scale; the markers correspond to mean optimal time scales across stimuli for comparison between segmentation models. Figure 11a shows the similarity between musicians and nonmusicians at each of the 16 time scales that were used for segmentation modelling. The mean correlation between musicians and nonmusicians in the real-time task ranged from high to moderate, and peaked at a time scale of 1.7 s. Figure 11b shows the mean debiased correlation between tasks across stimuli for both nonaligned and aligned analysis. The exhibited correlations were higher for aligned models than for nonaligned models, with peaks at time scales of 1.4 s and 2.5 s, respectively. Comparing Figure 11a and Figure 11b, the correlation between groups was higher than the correlation between tasks, which yielded higher dissimilarity for both aligned and nonaligned models.

LINK BETWEEN OPTIMAL TIME SCALE FOR SET COMPARISON AND RHYTHMIC FEATURES

Following this analysis, we investigated the possible relationship between optimal time scales for segmentation and global rhythmic descriptions of each stimulus. We calculated the similarity between optimal time scales found for comparing tasks and four acoustic features.

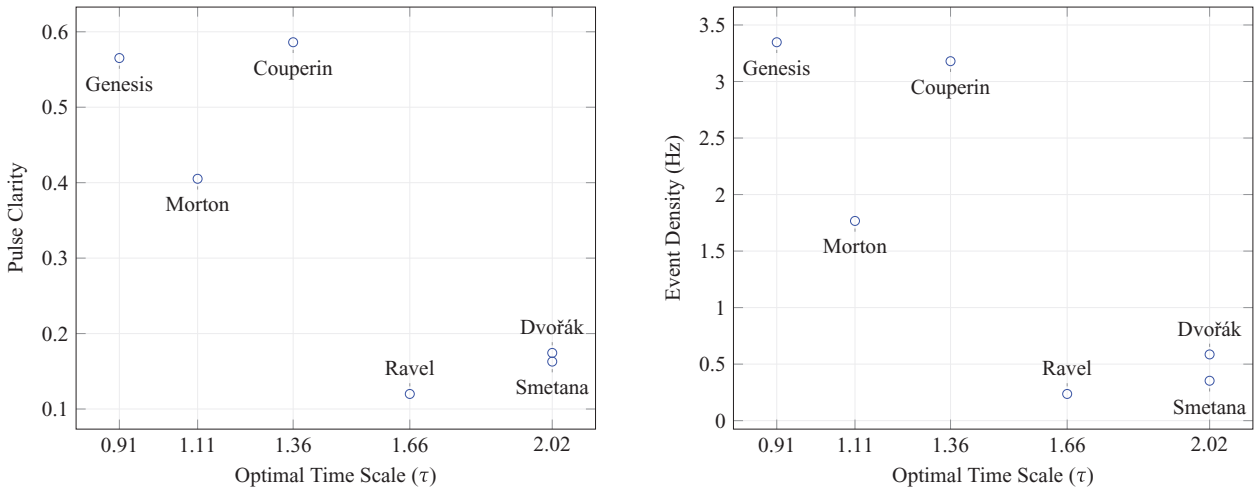

FIGURE 12. Relationship between two rhythmic descriptors and optimal time scales for correlation between tasks. a) Negative link, $r(4) = -.83$, $p < .05$, between pulse clarity and optimal time scale to compare real-time task and annotation task. b) Negative relationship, $r(4) = -.82$, $p < .05$, between frequency of events and optimal time scale for task comparison.

For each musical stimulus we estimated pulse clarity (underlying rhythmic pulsation), event density (average frequency of events), average note duration (inverse of event density), and global tempo using *MIRToolbox 1.5*; subsequently we correlated the optimal time scales for task comparison with each feature (H0: no correlation between optimal time scales for task comparison and rhythmic features). We obtained strong negative correlations between the optimal time scales for task comparison and both pulse clarity, $r(4) = -.83$, $p < .05$, and event density, $r(4) = -.82$, $p < .05$; the left and right plots in Figure 12 show the inverse link between optimal time scales and pulse clarity and event density, respectively. We obtained lower correlations with the other rhythmic features, namely average note duration, $r(4) = .66$, $p > .05$, and tempo, $r(4) = -.11$, $p > .05$, and these results did not reach significance.

Discussion

From a methodological viewpoint, this study contributes to state of the art research in boundary perception on a number of accounts. We introduced a real-time data collection task in order to analyze spontaneous boundary indications. Compared to previous work, here the target stimuli were heard for the first time in the segmentation step, rather than during previous listening only conditions or practice trials. Another novel aspect, which aimed to illuminate the difference between intuitive and more conscious boundary indications, was to thoroughly compare how the examples were segmented by the same listeners in this task and in an annotation task that resembles previous data collection methodologies (Clarke & Krumhansl, 1990; Wiering et al., 2009). In addition, we expanded previous studies on musicianship by collecting spontaneous indications from musicians and nonmusicians using diverse stimuli. Unlike previous work that included only a small number of nonprofessional musicians (Bruderer, 2008), we aimed to reach optimal validity for group comparisons by using stringent criteria for musicianship.

Another contribution of our study for music segmentation was the implementation of a multi-scale analysis approach to represent the boundary indications of the participants as Kernel Density Estimation matrices. In comparison to the approach used by Bruderer (2008), we did not need to obtain repeated segmentations of the same stimulus from each participant to find an optimal time scale of the segmentation, because multi-scale modelling allows the estimation of which time scales offer optimal fit based on a single segmentation trial. In contrast to previous studies, via this approach we investigated how optimal time scales for segmentation and inter-task delays are linked to rhythmic characteristics of the audio stimuli.

NUMBER OF INDICATIONS

Our analysis of mean number of boundary indications for each group and task revealed no significant differences between participant groups (Figure 3). We did not find a significant effect of music training on the number of indications per minute for any of the examples. It must

be noted, however, that nonmusicians indicated in total 9.1% more segments than musicians (which could be partly attributed to the outliers of Figure 3). Although the median of participants in Figure 4 showed an opposite trend for the mean number of indications across stimuli, this result should not be disregarded. For instance, Bruderer (2008) reported, for a smaller participant pool, that musically trained participants indicated significantly fewer boundaries. Also, following the event segmentation theory (see Peebles, 2011), it could be that some nonmusicians have difficulties predicting goals and intentions in the music, and hence segment into shorter units. For example, the exposition of the *Piazzolla* theme (1:08 - 1:26) is highly ornamented, which camouflages its symmetry and underlying melodic parallelism (equal duration and durational values but different note pitches). Nonmusicians probably failed to integrate non-neighboring patterns, since they tended to cluster the ornaments, and divided the theme into more fragments than musicians. In contrast, musicians' schematic knowledge might have enabled them to anticipate future changes and group the melodic line together, instead of stumbling on local surface discontinuities elicited by embellishments. However, musicians segmented more than nonmusicians in *Couperin*, a stimulus that exhibits few musical changes other than those prompted by underlying tonal context. Regarding this, it is possible that nonmusicians tend to segment into larger units if they have difficulties discovering changes in the music. Overall, we did not find effects of musicianship at a global level of analysis, but some effects may be evidenced via exploration of specific musical passages.

Regarding the task comparison, we found that musicians indicated more boundaries in the annotation task than in the real-time task for all 6 stimuli, a trend that reached significance for most examples; this suggests an effect of the data collection task upon the number of boundary indications. We put forward three possible explanations for these (and other) differences between segmentation tasks. The first one is that during the execution of the second task, participants discovered other plausible boundaries for indication, perhaps due to familiarity with the underlying musical structure; in this vein, the number of listeners' judgments of section ends has been found to increase throughout progressively shorter presentations of the same piece (Krumhansl, 1996).

Another possibility is that the annotation task instructions biased listeners towards frequently indicating boundaries to be later able to give different ratings of boundary strength. The salience rating instructions may have influenced listeners in the annotation task to annotate as many boundaries as possible and at multiple time scales, whereas in the real-time task listeners may have indicated boundaries at a single time scale, perhaps at a rather large one due to focus on significant changes. If this was the case, then listeners may not have utilized the same concept of segmentation across tasks, lowering the validity of the annotation segmentation data; this might explain the extreme outliers in the annotation task (Figure 3). To address this, future segmentation task instructions should ask participants to indicate boundary strength ratings only after they have segmented all the stimuli, and they should not be informed about the salience rating step beforehand.

A third explanation is that the real-time task not only involves more sustained attention and concentration than the annotation task, but also hinders segmentation based on repetition and other retrospective aspects of segmentation. Some musical events are recognized *a posteriori* as instants of significant change due to the effect of ulterior events; for example, two motives can be identical except for a local difference (e.g., an alteration) in the middle of the second motif that, when perceived, prompts boundary perception between motives. Also, the use of ornamentation during cadences such as the trills in *Morton* (0'25") might disguise imminent musical changes, which become more evident retrospectively; this might partly explain the notable difference between tasks shown in Figure 3. Future work could analyze which particular time points of the stimuli exhibit high contrast in boundary density between tasks by subtracting segmentation models from one another; also, initial and final positions of boundaries in the annotation task can be recorded to explore boundary replacements.

BOUNDARY STRENGTH AND DENSITY

We next examined the relationship between boundary strength ratings and boundary density. We investigated whether model density, which is a local estimate of frequency of boundary indications, correlated with boundary strength ratings. The mean correlation across stimuli was low for all time scales (Figure 5), although *Smetana* exhibited moderate correlations. This suggests in principle no relationship between rated strength and frequency of indications, although this could be contingent on the stimuli.

We calculated the distribution of boundary indications into boundary strength ratings under the hypothesis that the annotation task instructions indirectly induced participants to indicate all possible boundaries so that these could be assigned different strength ratings. We found that about half of the indicated boundaries were given

relatively low strength ratings, which suggests that the correlation between strength ratings and density is low because participants indicated both highly significant boundaries and less salient ones. This result may also explain why participants indicated twice as many boundaries in the annotation task than in the real-time task.

Since the annotation task seemed to include additional "weak" boundaries compared to the real-time task, we investigated whether subjects agreed more about the location of boundaries rated as strong than about those rated as weak. We then correlated strength ratings in the annotation task with boundary density in the same task at the respective time points. We obtained a lower correlation than for the comparison between strength in the annotation task and density in the real-time task. This suggests that boundary strength informs more about the frequency of boundary indications for strong boundaries than for weak boundaries, and that participants agreed more about the location of boundaries rated as strong than about boundaries rated as weak. We remark that a visual comparison of the segmentation models suggested that all the boundaries with high density in the annotation task with added weights also showed high density in the real-time task; future studies should restrict the annotation task model to boundaries with the highest strength ratings to find out whether this creates an increase in the correlation between tasks.

According to our findings, the relative frequency of boundary indications does not predict the boundary strength ratings. Bruderer (2008) compared frequency of indications for a subset of boundaries within a window of 1.25 s with mean ratings of boundary salience. In contrast to our findings, Bruderer did find moderately high to high correlations across diverse musical stimuli, but our analysis is not identical. Bruderer restricted the analysis to a subset of boundary peaks with different indication frequencies, whereas we analyzed complete boundary data. In addition, we did not choose boundary indications via analysis windows but picked density values corresponding to each boundary. Also, most of the stimuli utilized by Bruderer (2008) were popular music with lyrics; this could have induced a relatively high agreement regarding boundary strength ratings, and could partly explain why we found difficulties in replicating his finding with instrumental and more varied musical stimuli. Overall, it is possible that participants were biased by the task instructions or that they had difficulties in assigning relative weights to boundaries. Alternatively, it could be that the frequency of indications does not inform about boundary strength: a stark drum pattern change should be indicated by multiple participants, but in order to be rated as strong it may need to be accompanied by silences, modal change, changes of instrumentation, musical novelty, or other aspects evoking boundary perception. Also, boundaries prompted by diminished triads or musical quotation may be indicated by few participants but still be rated as strong.

INTER-SUBJECT CORRELATION

We next compared the relationship between subjects for different groups and tasks via the correlation of pairs of individual segmentation models. Regarding tasks, we found the annotation task to exhibit higher mean inter-subject correlation than the real-time task (Figure 6). This suggests an effect of task on inter-subject correlation: in the real-time task, probably participants could not anticipate some boundaries and missed indicating them if the stimuli were relatively unpredictable (unlike *Smetana*, which exhibited similar inter-subject correlation across tasks). We also observed an improvement of the inter-subject correlation for the annotation task with added weights for *Smetana*, *Dvořák,* and *Genesis*. This suggests that for these stimuli listeners assigned similar strength ratings (or gave weak strength ratings to boundaries that others did not indicate). In contrast, a more 'ambiguous' stimulus (*Couperin*) exhibited an opposite trend: inter-subject correlation dropped for the model with added strength.

Comparing groups, both musicians and nonmusicians exhibited very similar mean inter-subject correlation. This result, in line with previous findings (Bruderer, 2008), suggests that musicianship may not have an effect on inter-subject correlation. It should be noted, however, that the relatively complex stimulus *Ravel* exhibited relatively higher inter-subject correlation for musicians. An exploration of the multi-scale models for a subtle tonal change that is induced by rapid arpeggios (1' 34", *Un peu marqué*) shows differences between groups. The boundary density corresponding to this change is relatively higher for musicians than for nonmusicians, suggesting higher consensus between musicians. Hence, no effects of musicianship were found but further qualitative analyses are required to observe possible effects when focusing on particular musical motives.

Overall, for both groups and tasks, we found low mean inter-subject correlations and great variability of the obtained coefficients. In principle, this suggests that participants were attending to different features (such as change in timbre or tonality) or to different hierarchical levels of segmentation; however, the approach used is sensitive to small timing variations between profiles.

This is because, unlike multi-scale models across participants, individual multi-scale models involve high density peaks of relatively short time spans. Because of this, small differences in the perceptual delay of participants can have a considerable effect on their inter-subject correlation; hence, participants who exhibited very high inter-subject correlation did not only segment the same musical changes, but were also highly synchronized with each other. In any case, future studies should further examine variance in inter-subject correlation and in number of boundary indications, considering that different participants may pay attention to different hierarchical levels (suggested by Bruderer, McKinney, & Kohlrausch, 2006), musical features, interactions of features, or top-down structural aspects (similarity, symmetry, and so forth). For instance, participants could be clustered into subgroups to explore the validity of grouping them based on i.e., musicianship or instrument.

TIME SCALE FOR BEST MODEL FIT TO INDICATIONS

We next estimated an optimal time scale for each example and set for modelling boundary data across participants; these optimal time scales correspond to the segmentation models that obtained the best fit to participants' boundary indications. The group comparisons for the real-time task (Figure 7) showed higher mean optimal time scales for musicians, although we did not find significant differences. The result is difficult to interpret since two stimuli exhibited opposite trends, but it could be that for most stimuli musicians focused on higher levels of the hierarchical grouping structure (such as changes of key, and of rhythmic and metrical patterns), or that they were less isochronous in their indications. The results were clearer for the task comparison, since we found larger optimal segmentation time scales for the real-time task and this difference was significant. This suggests that participants segmented at multiple levels of grouping or at relatively lower levels in the annotation task compared to the real- time task. We highlight, however, two outliers that exhibited a similar pattern across segmentation sets: the optimal time scale for *Genesis* was the largest for all segmentation sets, whereas *Morton* exhibited relatively low time scales for all sets. The music of *Genesis* combines multiple experimental sounds and effects within relatively long, homogeneous melodic-harmonic sections, and in this respect a segmentation at large time scales would be expected. Conversely, shorter time scales for *Morton* could be explained by ambiguity in harmonic progression, which has been found to decrease feeling of completion (Cuddy, Cohen, & Mewhort, 1981) and hence might induce boundary perception due to expectancy violation.

ALIGNMENT BETWEEN TASKS AND GROUPS

We investigated the degree of alignment between real-time and annotation task segmentation; this possible lag in the real-time task compared to the annotation task is evidenced in Figure 2, which shows that the models are not perfectly aligned. We found that the indications obtained from the task were delayed, and a mean optimal time lag across stimuli at 1.05 s for alignment of real-time and annotation tasks. In other words, it took participants an average of 1.05 s in the real-time task to recognize perceived boundaries and respond to these by pressing a key on the computer, suggesting that in this task they were usually unable or did not intend to anticipate upcoming musical changes.

Another goal was to find out whether the optimal time lag between tasks was dependent on temporal characteristics of the stimuli. Our results showed that global beat length (and, equivalently, global tempo) of the stimuli can predict the dispersion of the optimal time lag (Figure 8). This means that faster stimuli with shorter beat length would yield higher alignment between real-time and annotation task segmentation, and vice versa. In addition to the regression coefficient from which we derived this interpretation, the regression equation included a nonzero constant term, in other words a stimulus invariant time lag. This suggests that boundary indications in the real-time task are delayed at least by a number of beats (stimulus dependent) plus a constant time lag (stimulus independent). A plausible interpretation of the regression equation ($\tau = .72 \times BL + .66$) is that the real-time segmentation lag might stem from a *recognition delay* of around ¾ of a beat and a *response delay* of about ⅔ of a second: listeners possibly required less than a beat (between 0.4 s and 0.9 s depending on stimulus) to pass in order to recognize a perceived change as significant, and over half a second to respond to the change by indicating a boundary. Future work should compare the time lag between tasks for different portions of the stimuli to find out if the lag is reduced as engagement with the stimulus increases during real-time segmentation.

We also analyzed the level of alignment between musicians' and nonmusicians' segmentation models. We expected nonmusicians to be delayed compared to musicians, due to the effects of music training in auditory working memory. For example, musicians seem to be faster in capturing the statistical structure of perceived streams (François, Jaillet, Takerkart, & Schön, 2014) and exhibit larger auditory memory spans

(Tierney et al., 2008) than nonmusicians. However, we found the overall lag between musicians and nonmusicians to be practically zero, and focusing on the lag for each stimulus did not show a trend towards any particular group. These results suggest that music training has no effect upon indication time lag, as the negligible lags reported in Table 3 could be attributed to noise. This should, however, be explored in future studies including more varied stimuli such as highly predictable pop ballads and contemporary classical music with unexpected changes, and also assessing whether the delay increases in the initial stimuli sections but progressively decreases.

SIMILARITY BETWEEN TASKS AND GROUPS

Correlation between multi-scale models. We examined the relationship between groups, tasks, and alignment strategies by computing correlations based on the multi-scale models of the collected boundary data. It was found (Figure 10) that boundary data from musicians and nonmusicians yielded very similar multi-scale models for all stimuli. This result suggests that music training did not have an effect on the real-time multi-scale segmentation models, and that musicians and nonmusicians indicated similar structural descriptions, at least at a general level.

Regarding tasks, we observed for the aligned multi-scale models that musicians segmented very similarly in both real-time and annotation tasks. This suggests that if both tasks are time-aligned, the effect of segmentation task is not that evident. We also observed that the correlations were overall lower for the aligned task comparison than for the group comparison. Possible dissimilarity factors in the annotation task include the chance to indicate perceivable boundaries retrospectively, reduce perceptual delays via reposition of boundaries, and also the task instruction requirement to rate perceived strength, which could have led to the aforementioned bias. Another finding regarding the effect of alignment strategy was that the similarity between tasks was notably lower for nonaligned models; this suggests that alignment is needed for comparisons between real-time and annotation tasks in order to compensate for the latency of participants in the real-time task.

Correlation between single-scale models. We further investigated mean similarity between segmentation models at each time scale. As shown in Figure 11, the optimal time scale for comparison between tasks was larger for the nonaligned models (2.5 s) than for the group comparison (1.7 s). In other words, relatively large time scales are optimal for comparison between tasks, whereas smaller segmentation time scales yield dissimilarity between tasks, which is probably due to recognition delay and retrospectively perceivable boundaries. Also, we found that the peak correlations were higher for the group comparison than for both aligned and nonaligned task comparisons. This means that the similarity between participant groups was higher than the similarity between tasks, which suggests effects of segmentation task but no effects of group. In addition, we obtained mean optimal time scales across stimuli for group comparison and aligned task comparison at 1.4 s and 1.7 s respectively. These rather low optimal time scales suggest that both participant groups focused on chord, dynamics, pulse, and other relatively frequent changes rather than key or melodic boundaries. Also, these mean time scales are possibly indicative of the relative variance between subjects regarding the indication of single boundaries; for instance, indications within a 1.7 s span may relate to the same boundary, whereas those that are further apart might correspond to different perceived boundaries.

Optimal time scale for set comparison and rhythmic features. We additionally investigated whether there was a relationship between optimal time scales for task comparison and musical rhythm descriptors. We found moderate to strong links between the optimal time scales of the stimuli and three descriptors: pulse clarity, event density, and average note duration. Our results suggest that the time scale for comparison between segmentation tasks can be measured in terms of rhythmic clarity, event density, or average note duration rather than in seconds: short time scales are optimal to compare segmentations for music, characterized by a clear pulse and a relatively large number of short note events. It can be further argued that music with high global pulse clarity and event density facilitates forecast of boundaries because large interonset intervals and long rests (common cues for melodic segmentation) may appear more contrasting. Future work could test this possibility by estimating whether pulse clarity and event density predict segmentation model entropy, although other structural features such as loudness, instrumentation, cadences, and tonal closure might play a more prominent role.

Finally, we investigated a possible link between musical tempo and optimal time scale for comparison between tasks. It was expected that music with fast tempo would exhibit short optimal time scales for task comparison, and vice versa. We found only a weak negative correlation between optimal time scales and global tempo, although the direction of the relationship was according to our expectation and in line with findings

suggesting general increase of asynchrony with lower metronome tempo in finger tapping tasks (Repp & Su, 2013).

General Discussion

Regarding the first hypothesis of the study, our findings did not provide support for an effect of music training on musical segmentation. Musicians exhibited very high model alignment with nonmusicians, which is inconsistent with our prediction that nonmusicians would be delayed compared to musicians and also with findings suggesting differences in auditory memory spans between groups (Tierney et al., 2008). Another unexpected result was the similar inter-subject correlation for both groups; musicians did not exhibit higher consensus than nonmusicians, hence musicians' schematic knowledge may not increase group homogeneity regarding segmentation. Furthermore, multi-scale model similarity analyses showed very strong resemblance between musicians and nonmusicians in the real-time task, suggesting a relatively similar pattern of segmentation responses between groups. We also found that musicians and nonmusicians' time scales for optimal model-to-data fit were similar, which, unlike our expectation, suggests that both groups segmented at similar time scales. Moreover and also contrary to our expectations, we did not find a significant difference between groups in the number of boundary indications, although nonmusicians indicated more boundaries than musicians in the real-time task. This suggests no effect of musicianship on number of boundaries, although future studies should investigate in what musical contexts nonmusicians segment more often than musicians, and the contribution of expectation violation to this phenomenon.

In sum, we could not find sufficient evidence to reject the null hypothesis (no difference between segmentation from musicians and nonmusicians); hence, only limited implications can be derived from these findings. Perhaps musical boundary data analysis in the context of real-time segmentation does not reveal effects of music training despite having different representations of musical structure; it could also be that effects of musicianship can be only identified for shorter musical passages; alternatively, music training may not modulate musical structure representations. The first possibility assumes that differences due to musicianship only become apparent in implicit segmentation scenarios (Bigand & Poulin-Charronnat, 2006). The second alternative is supported by findings indicating group effects in melodic segmentation of short tunes (Peretz, 1989). On the other hand, the third possibility implies that structure boundary perception is independent of instrument skills and of cognitive loads associated with intensive training. We remark, however, that nonmusicians indicated more boundaries than musicians; future studies should gain further understanding on the additional boundaries indicated by nonmusicians via analysis of boundary taxonomies. Related to this, musicians tended to exhibit overall larger optimal time scales for data modelling; this finding requires future investigation since it suggests that musicians could pay attention to higher-level musical features.

In line with our second hypothesis, we did find effects of experimental task on perceptual segmentation. As expected, the real-time task set was delayed with respect to the annotation task. This finding suggests that during real-time segmentation listeners did not segment impulsively but ensured themselves that their predictions were correct before indicating a boundary. Related to this finding, rhythmic characteristics (global beat length) of the stimuli had an effect on the magnitude of the real-time task lag. This suggests that the latency of participants' responses in the real-time task consists of a recognition delay dependent on stimulus beat length, plus a constant response delay. We also found that listeners' segmentations were more similar to each other in the annotation task than in the real-time task. This suggests that they indicated boundaries less isochronously in the real-time task, because some boundaries could only be retrospectively perceived, or because of individual differences in perceptual delay. Moreover, listeners indicated significantly more boundaries in the annotation task than in the real-time task. This is a highly expected result since the annotation task offers more time to determine boundaries in finer detail, but also suggests that listeners focused not only on a single and large time scale, but also on other time scales of the segmentation, providing support to the aforementioned GTTM postulate (Lerdahl & Jackendoff, 1983). We also found that the alignment of real-time and annotation task models notably increased the correlation between them. This suggests that the real-time task lag made a major contribution to the task effect; further studies should consider other segmentation alignment strategies as well. In addition, single-scale model analyses showed that relatively long time scales (2.5 s) were optimal for comparison between segmentation tasks. This result suggests that time scales below 2.5 s are not smooth enough for task comparison, probably due to response delays in the real-time task and retrospective aspects of segmentation. Furthermore, we found that the time scale for optimal fit of the single-scale models

to the data was shorter in the annotation task than in the real-time task. This suggests that boundaries tend to be indicated in the real-time task within a larger time span, following simultaneous change of multiple musical features. In contrast, the annotation task may prompt clustering patterns at different hierarchical levels. A related issue, for which we could not replicate previous findings (Bruderer, 2008), was the relationship between perceived strength in the annotation task and the density of the real-time task segmentation model at the respective time points. Since we failed to find a link between these two, the frequency of indications of a boundary may not necessarily inform about its mean salience rating, but about acoustic or contextual aspects of segmentation. However, our results elicited questions about a possible bias due to the annotation task instructions, which might at least partly explain differences between tasks and require further investigation. Overall, among the main contributors to the effect of task we could find the real-time task lag and the differences in number of boundary indications. The lag depends to some extent on rhythmic characteristics of the stimuli, whereas the differences in number of boundaries are due to the impossibility to indicate boundaries retrospectively in the real- time task, and possibly to the strength rating task, which encouraged over-segmentation.

Regarding optimal segmentation time scales for task comparison, we found, in accordance with our third hypothesis, a dependence on global rhythmic pulsation, on amount of events, and on duration of events. This suggests that the time scale for modelling perceptual segmentation could be measured in terms of these rhythmic characteristics rather than in seconds; for instance, segmentation of music with unclear pulse and few note events of usually long duration requires to be modelled at large time scales. Noteworthy, rhythmic features extracted from the audio stimuli can be used to systematically predict aspects of segmentation from participants, as evidenced by analyses on optimal time scale and task alignment. Further work on alternatives to fixed time scales such as variable density estimation methods could gain new insights regarding this issue, because rhythmic features are not static but dynamic.

CONSIDERATIONS FOR FUTURE RESEARCH

An assessment of the validity of our findings should note that these are restricted to segmentation based on musical contrast, and to the assumption that significant musical changes prompt perception of structural boundaries. Future work could compare our operational definition of musical boundaries (*significant instants of change in the music*) with more complex definitions including metaphors ("landmark points while taking a walk in an unfamiliar forest," Deliège et al., 1996; "listen to the music as if it was a story and mark its punctuation," Koniari et al., 2001; "tell how strong the punctuation was," Deliège, 2007) and musicological terms ("press space-bar when you hear a segment boundary [phrase, section, passage]," Bruderer et al., 2006); the effect of a given definition on the resulting boundary profiles should be analyzed. In addition, work on implicit tasks related with segmentation (see Peretz, 1989) could provide insights on retrospective, memory, repetition-based, and other top-down processes that underlie explicit segmentation. For example, we should further examine whether perception of short musical material should suffice to prompt higher-level groupings of longer material (see cue-abstraction theory, Deliège et al., 1996).

Further segmentation studies should overcome methodological issues concerning the validity of the participant sample by including an established questionnaire, such as the Goldsmith's Musical Sophistication Index (Gold-MSI, see Müllensiefen, Gingras, Musil, & Stewart, 2014), which has been recently used for assessments in musicianship studies (Carey et al., 2015; Schaal, Banissy, & Lange, 2015). This can be helpful not only for comparing research findings but also for improving recruitment and classification: Gold-MSI takes into account that training may not determine musical abilities such as perception of form (Bigand & Poulin-Charronnat, 2006; Lalitte & Bigand, 2006), and also that some musical skills do not result from formal music training (Müllensiefen et al., 2014). It is also recommended for future studies to employ full factorial designs to investigate effects of musicianship and experimental task. Collecting segmentation data by nonmusicians in the annotation task would enhance our understanding of commonalities and differences between groups and tasks. Although our sample of nonmusicians reported having no experience in audio editing software, it is very likely that they could have completed the task without problems. Many youth and adults possess the editing skills required for structural annotation tasks as it is common to record and edit videos for web sharing and social networking. Also regarding the effect of musicianship on segmentation, many confounding variables, including level of attention, current state of participants, and aspects of musical structure could have contributed to our negative results. In our view, replication with other participant samples and more musical stimuli is required to understand whether these findings are generalizable to other scenarios. It is possible that local group differences did not show up in the reported global results; for instance, it could be that specific musical

passages may show interesting group differences with respect to accordance with grouping preference rules or indication delays. These and other results remain to be approached at a finer scale to allow for more musically interesting insights; also, new experiments should be devised to understand the role of specific local Gestalt rules and other factors upon segmentation of a rich real-world dataset. Another issue that deserves further study is why the correlations between groups in both single-scale and multi-scale similarity analyses are not higher; as illustrated above, small dissimilarities between models from musicians and nonmusicians might derive from systematic differences between groups with respect to particular aspects of segmentation such as parallelism, instead of from data noise.

Future studies are needed to also clarify the role of experimental task on segmentation, since methodological issues could have hampered our results. Contrary to the annotation task, in the real-time task listeners did not hear the music before responding, they could not amend their responses after segmentation, and they were not asked to rate boundary strength. Future studies should compare different versions of the real-time task that vary only in one way to understand the contribution of different factors. For instance, four real-time segmentation versions could be compared: 1) real-time segmentation, 2) familiarization with stimulus followed by real-time segmentation, 3) real-time segmentation and subsequent boundary reposition, and 4) real-time segmentation followed by boundary strength indication.

New perceptual segmentation modelling approaches should be developed to clarify the interpretation of our results regarding optimal segmentation time scales and contribution of musical features. Our multi-scale modelling method yields ambiguous results, because small optimal time scales for segmentation may indicate any or both of these propositions: 1) participants pay attention to low hierarchical levels of the musical structure, 2) participants are isochronous in their indications. It is difficult to know whether participants pay attention to low grouping levels (e.g., segmenting each note), or exhibit little timing dispersion in their indications; also both cases could also be correct. Further research should also focus on which specific rhythmic, metrical, and grouping structure rules are emphasized via our modelling approach. Finally, systematic time series comparisons between different musical features and perceptual segmentation models could provide thoughtful insights upon description cues involved in segmentation for different tasks and groups.

Author Note

The authors would like to thank Jordan B. L. Smith and two anonymous reviewers for their insightful comments on earlier versions of this manuscript. Thanks also to Emily Carlson for proofreading the paper. This work was financially supported by the Academy of Finland (project numbers 272250 and 274037).

Correspondence concerning this article should be addressed to M. Hartmann, Finnish Centre for Interdisciplinary Music Research, Department of Music, University of Jyväskylä, P.O. Box 35, FI-40014 University of Jyväskylä. E-mail: martin.hartmann@jyu.fi

References

Addessi, A. R., & Caterina, R. (2000). Perceptual musical analysis: Segmentation and perception of tension. *Musicæ Scientiæ*, *4*(1), 31-54.

Bailes, F., & Dean, R. T. (2007). Facilitation and coherence between the dynamic and retrospective perception of segmentation in computer-generated music. *Empirical Musicology Review*, *2*(3), 74-80.

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

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

Bruderer, M., McKinney, M., & Kohlrausch, A. (2006). Perception of structural boundaries in popular music. In M. Baroni, A. R. Addessi, R. Caterina, & M. Costa (Eds.), *Proceedings of the 9th International Conference on Music Perception and Cognition* (pp. 157-162). Bologna: ICMPC.

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*, 249-268.

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In A. Del Bimbo & S. Chang (Eds.), *Proceedings of the ACM Multimedia International Conference* (pp. 1467-1468). Firenze, Italy: ACM Multimedia International Conference.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81-105.

Clarke, E., & Krumhansl, C. (1990). Perceiving musical time. *Music Perception*, 7, 213-251.

Cuddy, L. L., Cohen, A. J., & Mewhort, D. (1981). Perception of structure in short melodic sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(4), 869-83.

Dauwels, J., Vialatte, F., Weber, T., & Cichocki, A. (2009). On similarity measures for spike trains. In M. Köppen, N. Kasabov, & G. Coghill (Eds.), *Advances in Neuro-Information Processing* (pp. 177–185). Auckland: Springer.

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*, 4, 325-359.

Deliège, I. (2007). Similarity relations in listening to music: How do they come into play? *Musicae Scientiae*, *11*(9), 9-37.

Deliège, I., Mélen, M., Stammers, D., & Cross, I. (1996). Musical schemata in real-time listening to a piece of music. *Music Perception*, *14*, 117-159.

Ferrand, M., Nelson, P., & Wiggins, G. (2003). Unsupervised learning of melodic segmentation: A memory-based approach. In R. Kopiez, A. Lehmann, I. Wolther, & C. Wolf (Eds.), *Proceedings of the 5th Triennial ESCOM Conference* (pp. 141-144). Hanover: ESCOM.

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

François, C., Jaillet, F., Takerkart, S., & Schön, D. (2014). Faster sound stream segmentation in musicians than in non-musicians. *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*, 499-543.

Hargreaves, S., Klapuri, A., & Sandler, M. (2012, December). Structural segmentation of multitrack audio. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(10), 2637-2647.

Kaiser, F., & Peeters, G. (2013). Multiple hypotheses at multiple scales for audio novelty computation within music. In R. Kreidieh Ward (Ed.), *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. Vancouver: ICASSP.

Koniari, D., Predazzer, S., & Mélen, M. (2001). Categorization and schematization processes used in music perception by 10-to 11-year-old children. *Music Perception*, *18*, 297-324.

Koniari, D., & Tsougras, C. (2012). The cognition of grouping structure in real-time listening of music. A GTTM-based empirical research on 6 and 8-year-old children. In E. Cambouropoulos, C. Tsougras, P. Mavromatis, & K. Pastiadis (Eds.), *12th International Conference on Music Perception and Cognition*. Thessaloniki, Greece: ICMPC.

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

Lalitte, P., & Bigand, E. (2006). Music in the moment? Revisiting the effect of large scale structures. *Perceptual and Motor Skills*, *103*(3), 811-828.

Lartillot, O., & Ayari, M. (2009). Segmentation of Tunisian modal improvisation: Comparing listeners' responses with computational predictions. *Journal of New Music Research*, *38*(2), 117-127.

Lartillot, O., & Toiviainen, P. (2007). A Matlab toolbox for musical feature extraction from audio. In S. Marchand (Ed.), *Proceedings of the Tenth International Conference on Digital Audio Effects* (pp. 237-244). Bordeaux: ICDAE.

Lartillot, O., Yazıcı, F., & Mungan, E. (2013). A more informative segmentation model, empirically compared with state of the art on traditional Turkish music. In P. van Kranenburg, C. Anagnostopoulou, & A. Volk (Eds.), *Proceedings of the Third International Workshop on Folk Music Analysis* (p. 63). Utrecht: Meertens Institute, Department of Information and Computing Sciences, Utrecht University.

Lattner, S., Grachten, M., Agres, K., & Chacón, C. E. C. (2015). Probabilistic segmentation of musical sequences using restricted Boltzmann machines. In O. Bandtlow & E. Chew (Eds.), *Mathematics and Computation in Music*. London: MCM.

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

Martorell Dominguez, A. (2013). *Modelling tonal context dynamics by temporal multi-scale analysis* (Unpubished doctoral dissertation). Universitat Pompeu Fabra, Barcelona.

Mauch, M., MacCallum, R. M., Levy, M., & Leroi, A. M. (2015). The evolution of popular music: USA 1960–2010. *Royal Society Open Science*, *2*(5).

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS One*, *9*(2), e89642.

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

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, *39*(10), 1367-1391.

Peebles, C. (2011). *The role of segmentation and expectation in the perception of closure* (Unpublished doctoral dissertation). Florida State University.

Peeters, G., & Deruty, E. (2009). Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In S. Baumann, J. J. Burred, A. Nürnberger, & S. Stober (Eds.), *Proceedings of the 3rd Workshop on Learning the Semantics of Audio Signals* (pp. 75-90). Graz: LSAS

Peretz, I. (1989). Clustering in music: An appraisal of task factors. *International Journal of Psychology*, *24*(1-5), 157-178.

Repp, B. H., & Su, Y.-H. (2013). Sensorimotor synchronization: A review of recent research (2006-2012). *Psychonomic Bulletin and Review*, *20*(3), 403-452.

Sanden, C., Befus, C. R., & Zhang, J. Z. (2012). A perceptual study on music segmentation and genre classification. *Journal of New Music Research*, *41*(3), 277-293.

Schaal, N. K., Banissy, M. J., & Lange, K. (2015). The rhythm span task: Comparing memory capacity for musical rhythms in musicians and non-musicians. *Journal of New Music Research*, *44*(1), 3-10.

Schaefer, R. S., Murre, J. M., & Bod, R. (2004). Limits to universality in segmentation of simple melodies. In S. Lipscomb, R. Ashley, R. Gjerdingen, & P. Webster (Eds.), *Proceedings of the 8th Conference on Music Perception and Cognition*. Adelaide: Causal Productions.

Sears, D., Caplin, W. E., & McAdams, S. (2014). Perceiving the classical cadence. *Music Perception*, *31*, 397-417.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Boca Raton, FL: CRC Press.

Smith, J. B. L., Chuan, C., & Chew, E. (2013). Audio properties of perceived boundaries in music. *IEEE Transactions on Multimedia*. *16*, 1219-1228.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Tierney, A. T., Bergeson-Dana, T. R., & Pisoni, D. B. (2008). Effects of early musical experience on auditory sequence memory. *Empirical Musicology Review*, *3*(4), 178-186.

Toiviainen, P., & Krumhansl, C. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*(6), 741-766.

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

Appendix

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.

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.

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.

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

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

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

Piazzolla Piazzolla, A. (1959). Adiós Nonino. [Recorded by Astor Piazzolla y su Sexteto]. On *The Lausanne Concert* [CD]. BMG Music. (1993)
Spotify link: http://open.spotify.com/track/6X5Szbloyesr QQb3Ht4Ojx
Excerpt: 0-08:07.968
Used for Experiment 1 only. Presented to participants as four musical examples: 0-02:00, 01:57-03.57, 03:54-05:54, 05:51-08:07.968

Dream Theater Petrucci, J., Myung, J., Rudess, J. & Portnoy, M. (2003). Stream of Consciousness (instrumental). [Recorded by Dream The- ater]. On *Train of Thought* [CD]. Elektra Records. (2003)
Spotify link: http://open.spotify.com/track/3TG1GHK82boR3aUDEpZA5f
Excerpt: 0-07:50.979
Used for Experiment 1 only. Presented to participants as four musical examples: 0-02:00, 01:57-03.57, 03:54-05:54, 05:51-07:50.979

Stravinsky Stravinsky, I. (1947). The Rite of Spring (revised version for Orchestra) Part I: The Adoration of The Earth (Introduction, The Augurs of Spring: Dances of the Young Girls, Ritual of Abduction). [Recorded by Orchestra of the Kirov Opera, St. Petersburg - Valery Gergiev]. On *Stravinsky: The Rite of Spring / Scriabin: The Poem of Ecstasy* [CD]. Philips. (2001)
Spotify link: http://open.spotify.com/album/22LYJ9orjaJOPi8xl4ZQSq (first three tracks) Excerpts: 00:05-03:23, 0-03:12, 0-01:16 - total duration: 07:47.243.
Used for Experiment 1 only. Presented to participants as four musical examples: 00:05-02:05, 02:02-04:02, 03:59-05:59, 05:56-07:52.243