JYVÄSKYLÄ STUDIES IN HUMANITIES
303

Martín Ariel Hartmann

Modelling and Prediction of Perceptual Segmentation

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN HUMANITIES 303

Martín Ariel Hartmann

Modelling and Prediction of Perceptual Segmentation

Esitetään Jyväskylän yliopiston humanistisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston vanhassa juhlasalissa S212
tammikuun 17. päivänä 2017 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities of the University of Jyväskylä,
in building Seminarium, auditorium S212, on January 17, 2017 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Modelling and Prediction of Perceptual Segmentation

JYVÄSKYLÄ STUDIES IN HUMANITIES 303

Martín Ariel Hartmann

Modelling and Prediction of Perceptual Segmentation

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Editors
Petri Toiviainen
Department of Music, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

Jyväskylä Studies in Humanities
Editorial Board

Cover picture by Reeta Kosonen.

URN:ISBN:978-951-39-6903-5
ISBN 978-951-39-6903-5 (PDF)
ISSN 1459-4331

ISBN 978-951-39-6902-8 (nid.)
ISSN 1459-4323

Jyväskylä University Printing House, Jyväskylä 2017

To my parents Alicia and Tomás

This part is saddening, right? Because the music is very bendy. . . [1]
— Rafael, 3 years old, watching the film *My neighbor Totoro*

[1] *Esta parte da pena, no? Porque tiene música muy blandita. . .*

ABSTRACT

Hartmann, Martín Ariel
Modelling and Prediction of Perceptual Segmentation
Jyväskylä: University of Jyväskylä, 2017, 94 p.(+included articles)
(Jyväskylä Studies in Humanities
ISSN 1459-4323; 303 (nid.) ISSN 1459-4331; 303 (PDF))
ISBN 978-951-39-6902-8 (nid.)
ISBN 978-951-39-6903-5 (PDF)
Finnish summary
Diss.

While listening to music, we somehow make sense of a multiplicity of auditory events; for example, in popular music we are often able to recognize whether the current section is a verse or a chorus, and to identify the boundaries between these segments. This organization occurs at multiple levels, since we can discern motifs, phrases, sections and other groupings. In this work, we understand segment boundaries as *instants of significant change*.

Several studies on music perception and cognition have strived to understand what types of changes are associated with perceptual structure. However, effects of musical training, possible differences between real-time and non real-time segmentation, and the relative importance of different musical dimensions on perception and prediction of segmentation are still unsolved problems. Investigating these issues can lead to a better understanding of mechanisms used by different types of listeners in different contexts, and to gain knowledge of the relationship between perceptual structure and underlying acoustic changes in the music.

In this work, we collected segmentation responses from musical pieces in two listening experiments, a real-time task and a non real-time task. Boundary data was obtained from 18 non-musicians in the real-time task and from 18 musicians in both tasks. We used kernel density estimation to aggregate boundary responses from multiple participants into a perceptual segment density curve, and novelty detection to obtain computational models based on audio musical features extracted from the musical stimuli.

Overall, our findings provide evidence for an effect of experimental task on perceptual segmentation and its prediction, and clarify the contribution of local and global musical characteristics. However, the findings do not resolve discrepancies in the literature regarding musicianship. Furthermore, this investigation highlights the role of local musical change between homogeneous regions in boundary perception, the impact of boundary indication delays on segmentation, and the problem of segmentation time scales on modelling.

Keywords: musical structure, kernel density estimation, novelty detection, musical features, musical training, perceptual segmentation task

Author

Martín Ariel Hartmann
Department of Music
University of Jyväskylä
Finland

Supervisors

Professor Petri Toiviainen
Department of Music
University of Jyväskylä
Finland

Doctor Olivier Lartillot
Department of Architecture, Design and Media Technology
Aalborg University
Denmark

Reviewers

Associate Professor Emilios Cambouropoulos
Department of Music Studies
Aristotle University of Thessaloniki
Greece

Associate Professor Emilia Gómez
Department of Information and Communication Technologies
Pompeu Fabra University
Spain

Opponent

Associate Professor Emilios Cambouropoulos
Department of Music Studies
Aristotle University of Thessaloniki
Greece

ACKNOWLEDGEMENTS

This research was conducted from October 2011 to September 2016 at the Department of Music at the University of Jyväskylä. The research was supported by the *Finnish Centre of Excellence in Interdisciplinary Music Research* (7118616), *Music Mining Plant: Feature, Structure and Concept Mining for Music Information Retrieval* (218173) Academy of Finland Research Fellowship Project and *Dynamics of Music Cognition* (274037) Academy of Finland Professorship Project, by the *Finland Distinguished Professor (FiDiPro) project Machine Learning for Future Music and Learning Technologies* funded by TEKES, by the *Multidisciplinary network project in higher education: Interaction and eEducation* coordinated by the Faculty of Education of the University of Jyväskylä, by grants and doctoral positions from the Music Department and the Faculty of Humanities of the University of Jyväskylä, and by grants from *Ellen and Artturi Nyyssönen Foundation* and the *Finnish National Programme for Music Research* (Musiikintutkimuksen valtakunnallinen tohtoriohjelma).

I think I have been very lucky with my supervisors, Prof. Petri Toiviainen and Dr. Olivier Lartillot. I am profoundly grateful to Petri since he gave me excellent guidance and support; I want to highlight his genuine enthusiasm, modesty, patience, and good judgement. Petri is a role model for me. I also must mention his unique sense of humor, which lifted my spirits an immeasurable amount of times and made this experience an extremely delightful one! I also want to mention Olivier's reliable guidance and unconditional support; I praise his invaluable dedication and attention to detail, and his generosity and encouraging personality, which is truly inspiring. These lines I am writing are by no means enough to express how thankful I am for having had them as supervisors; this has been an honour for me.

I wish to express my deepest gratitude to the reviewers of my thesis, Associate Professor Emilios Cambouropoulos and Associate Professor Emilia Gómez. Due to my admiration for their groundbreaking and creative work in this field, I was convinced that they would offer valuable feedback; indeed, their comments and suggestions showed both their serious commitment and their motivation to share their knowledge and expertise. I am humbled that Associate Professor Emilios Cambouropoulos has agreed to be my opponent.

I would also like to thank everyone at the Music Department of the University of Jyväskylä for creating such an amazing work environment. Being with you at Musica is always a joyful experience. Due to my nomadic behavior, I had the opportunity to share office space with various people, all of whom were wonderfully passionate about research. Ibi, thank you for the joint work, for all your energy, initiative, choco-müsli bars, and all types of discussions (epistemological, political, life-related, and so forth). Henna, thank you so much for all your enormous help regarding Finnish language, for being a great emotional buffer, and for all the interesting and/or hilarious topics at coffee breaks. Pasi, almost any random conversation that we have becomes a source for new ideas; thank you for your help, collaboration and support during all these years. Emily, you are awesome, I am so

grateful for all the help that you gave me with proofreading, and especially for all the fun at the office, including, but not limited to, life, politics, APA style, reviewers, and so on. Elsa, thank you for helping me with my English, and for our unstoppable morning chats about life, people, research, music, Finnish language, Donegal, Buenos Aires, and so forth. Olivier, thank you for admirably inspiring positive attitude, resilience and self-confidence, which are paramount at a workplace that involves dealing with trial and error methods, research funding applications, paper submissions, and other possible sources of frustration. Margarida, thank you for your great company, for your wisdom, and for making full use of your conversational skills to spend time on a topic until the probability of the data (given the hypothesis that we would keep sanity if we continued talking) is $p < .001$. Jörg, thank you for the laughs, the long talks, friendly advice, and for your invaluable advice on Tele guitars. Marc, thank you, at least, for the work done together, for your open-minded spirit, perfectly combining scientific discussions, life, and laughter; you know perfectly well that I would thank you for so much more than that but I'm trying not to be too soppy. Vinoo, you are a such an inspiring human being; I'm equally thankful for your kindness, your exemplary diligence, and the delicious food. Jonna, thanks for your initiative, your coordination skills, your clarity and your fantastic sense of humor. Agustín, thank you for your tenacity, curiosity, strive for elegant solutions and great anecdotes. Warm thanks to my dear colleagues Jaakko Erkkilä, Markku Pöyhönen, Geoff Luck, Suvi Saarikallio, Esa Ala-Ruona, Birgitta Burger, Mikko Myllykoski, Will Randall, Katharina Schäfer, Imre Lahdelma, Juan Mendoza, Anna-Kaisa Ylitalo, Nina Loimusalo, Shawn Condon, Marianne Tiihonen, Nerdinga Letulė. My gratitude also goes to amazing people that I unfortunately don't see so often: Tuomas Eerola, Elvira Brattico, Mari Tervaniemi, Minna Huotilainen, Mikko Leimu, Rafael Ferrer Flores, Anemone Van Zijl, Kaisa Johansson, and Tommi Himberg.

I also want to thank the terrific people that gave me warm reception and guidance during my Research Visit to Queen Mary; Marcus Pearce, Elaine Chew, Marc Sandler, Geraint Wiggins, Yading Song, Holger Kirchhoff, Steven Hargreaves, and Katerina Kosta. Very special thanks to five gentlemen that I met during my studies and research in Musica: Shriram Alluri, David Ellison, Alex Reed, Alex Berman, and James Andean. Sincerest gratitude to my Finnish friends Otto and Tuomas, my international friends Li-Tang, Mark, Gaby O. and Pablo T., and my Argentinian friends Pablo B., Boris, Ignacio E., Ignacio S., Damián, Gaby G., Julia, and Luciana.

My deepest thanks to Sofía, Vera, David, Anna, Silvia, Riitta, and Kaisa. Heartfelt thanks to my beloved siblings Alejandro and Irene, to Alejandra, and to my dear nephews and nieces Ciro, Violeta, Lena, Ema, and Rafael. The final and most special thanks goes to my family. Reeta, my dear wife, I love you. I cherish your encouragement and patience. Isla, my beautiful daughter, you are the greatest joy of my life.

Jyväskylä, December 14, 2016

CONTENTS

LIST OF INCLUDED ARTICLES

PI Martín Hartmann, Olivier Lartillot, and Petri Toiviainen. Multi-scale Modelling of Segmentation: Effect of Musical training and Experimental Task. *Music Perception*, 2016.

PII Martín Hartmann, Olivier Lartillot, and Petri Toiviainen. Interaction Features for Prediction of Perceptual Segmentation: Effects of Musicianship and Experimental Task. *Journal of New Music Research*, 2016.

PIII Martín Hartmann, Olivier Lartillot, and Petri Toiviainen. Musical Feature and Novelty Curve Characterizations as Predictors of Segmentation Accuracy. *Manuscript submitted for publication*.

ABBREVIATIONS

NMrt Non-musicians in the real-time task

Mrt Musicians in the real-time task

Ma Musicians in the annotation task

Maw Musicians in the annotation task - weighted boundaries based on listeners' boundary strength ratings

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

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

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

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

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

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

Piazzolla Piazzolla, A. (1959). Adiós Nonino. [Recorded by Astor Piazzolla y su Sexteto]. On *The Lausanne Concert* [CD]. BMG Music. (1993)
Spotify: http://open.spotify.com/track/6X5SzbloyesrQQb3Ht4Ojx.
Excerpt: 0-08:07.968. Used for Experiment I 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 Theater]. On *Train of Thought* [CD]. Elektra Records. (2003)
Spotify: http://open.spotify.com/track/3TG1GHK82boR3aUDEpZA5f.
Excerpt: 0-07:50.979. Used for Experiment I 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: 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 I 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.

GPR Grouping Preference Rule

GTTM Generative Theory of Tonal Music

KDE Kernel Density Estimation

MDSP Mean Distance between Subsequent Peaks

MIR Music Information Retrieval

1 INTRODUCTION

Humans possess the ability to perceptually parse ongoing streams into discrete, meaningful events. This perceptual operation, which is called segmentation, makes it possible to understand continuous information or activities that involve sound and movement, just like it is possible, in a messy room, to recognize each of its objects (Zacks & Swallow, 2007). Besides being a general necessity regarding human perception and cognition that applies to different modalities, segmentation has central importance, for instance, in the area of speech perception, as it is needed for language acquisition (Johnson & Jusczyk, 2001; Seidl, 2007). In everyday music listening experiences, musical events that share related characteristics or high temporal proximity are often grouped into sequences. It is noteworthy that different music listeners can often distinguish the same sections in popular songs.

Temporal psychological processes of integration of musical events into larger units might be involved and could even be universal in music listening (Drake & Bertrand, 2001; Mungan, Yazıcı, & Kaya, in press). An inverse formulation of these processes would be that listeners segment long musical streams when they perceive unexpected changes. Besides other cues related to, e.g., similarity and repetition, musical feature change commonly prompts segmentation: listeners indicate segment boundaries if they easily perceive that there is a contrast, such as a stark change in dynamics or instrumentation. Multiple strategies are exploited by composers, improvisers and performers to induce perception of musical changes and communicate musical structure to the listener (Deliège, 2001; Dean, Bailes, & Drummond, 2014; Poli, Rodà, & Vidolin, 1998).

Segment boundaries can be understood as a representation of the perceptual structure, which involves hierarchies assigned to different events in music. In this sense, boundaries indicated by listeners may differ from each other regarding their salience due to acoustic characteristics of perceived changes, such as contrasts in instrumentation. Besides the temporal location and salience of the change, other relevant aspects regarding indicated boundaries include associated acoustic phenomena, time scale of changes, and reliability of segmentation (i.e., level of agreement between subjects). In sum, a number of variables associated to perceptual boundaries can be studied with regards to music listening: temporal

location (when), perceptual salience (how strong), temporal scale (how often), description cues (what type), and perceptual agreement (how reliable). Our investigation addresses these aspects using a particular conception of perceptual segmentation. We refer to segmentation in its broader sense, understanding perceptual segment boundaries as *instants of significant musical change*.

A number of scholars have attempted to understand the possible role of musical expertise on boundary perception. Lerdahl and Jackendoff (1983) claim that it is unlikely that a particular piece is heard by listeners in exactly the same way, although they should all agree on the most natural ways of hearing a piece. Results from experimental studies rooted in this theory suggest that both musicians and non-musicians are able to represent the hierarchical structure of the music from its perceived surface, but that differences in musical skills have an effect on these representations (Koniari & Tsougras, 2012; Peretz, 1989; Deliège, 1987). Indeed, musically trained children (Koniari, Predazzer, & Mélen, 2001) and adults (Bruderer, 2008) exhibited higher within-subject agreement in segmentation, showing more consistency across repeated segmentations of a target stimulus than untrained listeners. In addition, studies focusing on different aspects of segmentation have reported that musically trained participants indicate more boundaries than untrained ones (Bruderer, 2008; Deliège, 1987). A possible reason for these differences might be that musicians rely more on musical schemata during segmentation than non-musicians. Thanks to explicit schematic knowledge, musicians might be able to anticipate subsequent changes in the music, and group together a melodic line, thus indicating fewer boundaries instead of stumbling on local surface discontinuities elicited by embellishments.

However, there has been skepticism regarding the effect of musicianship on segmentation, since the literature presents conflicting findings and results that do not reach statistical significance (Bruderer, 2008). For example, although musically trained subjects have been found to segment more in accordance with musicological rules than untrained ones (Koniari & Tsougras, 2012; Peretz, 1989; Deliège, 1987), the opposite was found for more general rules (Schaefer, Murre, & Bod, 2004). Also, studies on processing and perception of structure (see Tillmann & Bigand, 2004) showed that both groups focus on "musical surface" (melodic contour and rhythm) and deeper aspects of structure during tasks involving harmonic priming and manipulation of global organization of pieces. Another issue is small sample size; Bruderer (2008), for example, included only 7 participants in the sample, none of whom were professional musicians.

Various methods have been used to gather boundary data in segmentation studies. 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). Other studies asked subjects to listen to the example, then segment in real-time, and finally make changes or deletions to their boundary profiles to obtain a precise, non real-time annotation for use in further experiments (Wiering, de Nooijer, Volk, & Tabachneck-Schijf, 2009; Ayari & McAdams, 2003; Clarke & Krumhansl, 1990). Non real-time (or offline) segmentation has been

posited to provide a different understanding of the musical structure because some boundaries cannot be perceived until they occur, or are perceived retrospectively, i.e., ulterior to the actual musical change (Lerdahl & Jackendoff, 1983). In this respect, boundary perception should be affected by musical expectancies; some boundaries are easier to anticipate as music temporally unfolds in real-time whereas others can be totally unexpected. That is, heard musical events not only resignify previous events and boundary indication decisions, but prompt predictions about possible future events (Hansen & Pearce, 2014). In this regard, only one study has investigated segmentation in real-time and non real time contexts (Peretz, 1989); this work on clustering of short melodies compared an online indication and an offline probe recognition task, aiming to understand possible differences between online organization of sequences and the representation of structure in our memory.

The literature shows that the role of both musicianship and experimental task (i.e., online vs offline) in perceptual segmentation remain a question. Furthermore, other possibly associated issues, such as relative delay between participant groups or tasks, have not yet investigated. Understanding the role of musical training in participants' segmentation can yield clues about transfer effects of musicianship and guide recruiting of participants for further music listening studies. Also, the effect of task on segmentation should be further explored to, for instance, compare real-time brain activity during music listening against expert annotations of musical structure.

Audio-based automatic segmentation algorithms are currently widely investigated, as they have many applications in music information retrieval, including music summarization (or *thumbnailing*), chord detection, music transcription, and music classification. Music Information Retrieval (MIR) studies have proposed a variety of automatic segmentation algorithms with a focus on evaluating model performance against ground truth data using accuracy measures (Aljanaki, Wiering, & Veltkamp, 2015). For evaluation purposes, predicted segmentation is compared to ground truth data, which often involve a set of isolated time points; studies on this area are typically based on a large number of stimuli, so ground truth segmentation data is obtained from at most few annotators. Only relatively recent work has focused on the possibility of designing new types of ground truths that involve multiple annotators, musical dimensions, and temporal scales (Nieto, 2015; Smith, Burgoyne, Fujinaga, De Roure, & Downie, 2011; Peeters & Deruty, 2009). In contrast to MIR ground truth data, studies focusing on listeners' perception of boundaries often collect data from many participants and aggregate their boundary indications (Deliège, 1987; Krumhansl, 1996; Ayari & McAdams, 2003; Frankland & Cohen, 2004); the most common aggregation method is based on the proportion of listeners responding within a beat, note event, or fixed time window.

Despite the increasing interest in audio-based prediction for a large number of listeners (e.g., Nieto, 2015; Müller, Chew, & Bello, 2016), to our knowledge this has not yet been done. Furthermore, the role of musical training and experimental task in prediction of segmentation remains to be clarified. For example, it is plausible

that automatic prediction models may yield higher segmentation accuracy for non-musicians than for musicians. This would be the case assuming that non-musicians would tend to indicate more boundaries due to lack of explicit schematic knowledge, and that bottom-up prediction models focusing on local discontinuities may be rather sensitive to acoustic changes in the music. Regarding segmentation tasks, one could expect that the annotation task would yield higher prediction accuracy than the real-time task based on the assumption that annotation segmentations would involve higher temporal precision with respect to changes in the musical signal. In this regard, utilizing ground truth data based on many listeners can help to increase its reliability, and investigating the role of musicianship and conducted experimental task in prediction can lead to a better understanding of what types of acoustic musical features are processed by different listeners and in different listening contexts during segmentation.

Related to this, studies have investigated the role of different cues evoking boundary perception, and the use of musical features for prediction of segmentation. Listeners group sequences of musical events when these are delimited by temporal gaps, changes in register, or dynamics (Bruderer, 2008). Studies related to segmentation prediction have proposed rules to find boundaries based upon different musical features (e.g. Cambouropoulos, 2001; Bohak & Marolt, 2016). In the audio domain, the performance of the prediction has been observed to depend more on musical stimuli than on the algorithm used or the choice of parameters (Peiszer, Lidy, & Rauber, 2008). This means, in principle, that the characteristics of different dimensions of individual pieces are key for optimal prediction of segmentation accuracy. However, the extent to which musical stimuli characteristics related to, e.g., spectral, rhythmic or pitch related change, have an effect on prediction accuracy of segmentation models still needs to be investigated. Addressing this issue would allow for the possibility to automatically select a feature or set of relevant musical features for optimal segmentation prediction based on distinctive characteristics of each stimulus.

The following research questions are addressed in our investigation:

1. What is the role of musical training and experimental task on perceptual segmentation?
2. What factors related to musicianship and conducted experimental task may have an important role in prediction of perceptual segmentation?
3. What is the contribution of intrinsic aspects of musical stimuli (e.g., spectral, rhythmic or pitch related change) to segmentation model prediction accuracy?

Our first main hypothesis was that musical training and experimental task has an effect on listeners' segmentation; for instance, more indication delays are expected for non-musicians compared to musicians, and for real-time contexts. Our second hypothesis was that bottom-up audio-based models would more accurately predict non-musicians' responses due to differences in schematic knowledge, and that non real-time responses would yield higher accuracy than real-time responses due to more precise indication data. Our third hypothesis was that musical pieces characterized by high local continuity of a given musical feature such as spectrum,

rhythm or tonality would yield higher segmentation prediction rates for that feature: the perceived strength of changes in a given musical dimension should be higher if these are more infrequent.

At this point, it is worth mentioning some aspects of segmentation that are not covered in this approach, as it will help to circumscribe the scope of our endeavor. One of them is repetition in music, which has a clear effect on listeners' representations of structure, but it is difficult to study because it involves a complex operationalization of perceptual similarity. In addition, hierarchical aspects of structure cannot be fully investigated with this approach, because we focused on segment boundaries, which can be considered a flat projection of a more complex representation.

The remainder of this thesis is organized as follows. Chapter 2 presents the theoretical foundation of our work, including music-theoretical approaches to segmentation, behavioral and computational studies related to the analysis of structure and some of the general challenges that are faced in this area of research. Chapter 3 recapitulates the aspects of segmentation that are covered in this thesis by describing its main goals and how they are addressed in the studies. Chapter 4 focuses on methodological procedures that were employed in our studies, namely data collection, perceptual and computational modelling, analyses regarding comparisons between groups and tasks, and the issue of temporal scales in segmentation modelling. Chapter 5 offers a concise account of the three articles included in this dissertation. Chapter 6 elaborates on theoretical and practical implications of our main findings, gives a remark on limitations regarding our approach, and proposes a number of suggestions for future research in music segmentation. Finally, we complete this dissertation with general conclusions (Chapter 7).

Author's contribution: The author was the primary contributor for the experimental design, collection of empirical data, implementation, data analysis, and writing of the three articles. The coauthors of the papers, Olivier Lartillot and Petri Toiviainen, advised on experimental design, suggested how the data could be analyzed, helped with the interpretation of the results, and contributed to writing.

2 A REVIEW OF THE LITERATURE ON MUSIC SEGMENTATION

The complex acoustic phenomena that determine boundary perception have been a center of attention in the realm of music segmentation. For example, behavioral studies have analyzed verbal description cues associated with musical changes indicated by listeners (Deliège, 1987; Clarke & Krumhansl, 1990; Bruderer, 2008). Sets of musicological rules to parse scores of musical pieces based on their content have also been proposed (Lerdahl & Jackendoff, 1983; Tenney & Polansky, 1980). These and other formalizations of musical structure have been included in symbolic computational models for automatic detection of segment boundaries or to estimate boundary probabilities of events (Wiering et al., 2009). Beyond MIDI-based models, which have often been oriented towards monodic pieces, the study of structural analysis for polyphonic audio has gained momentum over the last decade (Paulus, Müller, & Klapuri, 2010). This chapter aims to present the state of the art of music segmentation, focusing on various study areas and emphasizing topics that are pertinent to our investigation, such as musicianship and segmentation tasks.

2.1 Music-theoretical analysis of segmentation

Music listeners perform some sort of organization during music listening; sequences of auditory events are usually heard as organized units rather than as isolated sounds (Krumhansl, 2001). The mechanisms involved in the organization of sensory information are difficult to study regardless of the perceptual field under investigation, but it is generally agreed (e.g., Deutsch, 1999; Deliège, 1987) that auditory processes of segregation, grouping, and segmentation (Bregman, 1994) should apply to music. We can identify different voices (or streams) and the instruments that play them, we can notice that temporally long entities such as motifs and ostinati are being repeated, even in altered form as patterns or variations, and we also can agree that certain series of musical events, even very long ones, are different enough from other series and may be called phrases, sections, etc. The

grouping phenomena involved have been associated (Handel, 2006) with general perceptual organization rules (e.g., *proximity*, *similarity*, and *good continuation*) that have classically been explored in the visual field (Koffka, 1935). Musical events share similar characteristics in this respect, although the properties that contribute to perceptual grouping often cannot be directly related with those in visual and auditory perceptual organizations. For instance, sequences of musical events are grouped if delimited by temporal gaps (following the *proximity* principle), or by changes in register or dynamics (following the *(dis)similarity* principle).

Segmentation and grouping are linked together: the perception of distinct groups in music, such as melodies and musical phrases, often involves the identification of instants of musical change, which serve as delimiters. These instants can be referred to as segmentation points or segment boundaries. To give an example, a long silence in between two groups of notes may be perceived as a segment boundary; this change follows the Gestalt rule of proximity, which states that temporally (or spatially) close events are grouped together. Another example, which follows the similarity rule (*elements that share similar attributes are grouped together*) is that dissimilarity between sequences due to different instrumentation or harmony would induce boundary perception.

These principles can be condensed into one of the most basic Gestalt principles, the law of *Prägnanz*, which states that, under given conditions, perceptual organization will be as "good" as possible; that is, percepts tend to have the simplest, most stable and most regular organization that fits the sensory pattern (Koffka, 1935). *Maximum-minimum* simplicity is characteristic of "good" perceptual organizations, as in a soap bubble, which has the highest possible volume for its surface and the lowest possible surface for its volume (Koffka, 1935). One could imagine, for instance, a musical piece characterized by long and uniform segments with respect to instrumentation that would be delimited by few, though stark timbral changes. These highly contrasting timbral changes would probably be indicated as boundaries according to the *Prägnanz* principle.

The cue abstraction model (Deliège, 1989; Deliège, Mélen, Stammers, & Cross, 1996; Deliège, 2001, 2007) elaborates on the organization processes involved in real-time music listening and their influence on, e.g., grouping. Briefly, cues (or indices) are comparable to input tags, signposts, or salient elements that emerge at the musical surface. The most relevant cues are extracted to act as boundaries of the grouping that is being formed and aid in the organization of information (e.g., localization), and to obtain abbreviations with the aim of reducing the amount of stored information. This model has different phases, such as perception of same/difference (similarity) relations within a musical piece as a group formation process, formation of memory imprints or prototypical patterns, and progressive development of a schema of the piece; this framework has been investigated in empirical studies (Deliège, 1987, 1989; Deliège et al., 1996; Deliège, 2007).

Another approach to the understanding of perceptual organization of music, in this case aimed towards non real-time segmentation, is the Generative Theory of Tonal Music (GTTM, Lerdahl & Jackendoff, 1983). Their investigation of musical structure concerns the search for a *musical grammar* and its rules; here, musical

grammar is understood as a hierarchical model of the relationship between the musical surface of a given piece and the perceptual structure that it generates. Lerdahl and Jackendoff (1983) proposed a set of Grouping Preference Rules (GPR) as local and global considerations that determine listeners' perception of structure; each grouping would be delimited by event transitions that should be heard as boundaries. These rules have been categorized into three types (Clarke & Krumhansl, 1990): acoustic and temporal rules of proximity (GPR 2: temporal gaps) and similarity (GPR 3: change in register, dynamics, etc), "deeper" rules (GPR 7: tonal structure), and abstract rules (GPR 6: parallelism). Some of these rules, such as GPR 2, are defined in more detail, as they are supposed to be more associated with the musical surface (see Cambouropoulos, 2010, for a critical discussion on this assumption), whereas other rules, including GPR 5, which deals with issues associated with musical parallelism, are less formalized. This approach aims to offer a systematic alternative to structural analysis that could be applied to any Western musical work, without the need of amendments according to the particular piece under study (Lerdahl & Jackendoff, 1983). In practice, this can be very difficult due to conflict between rules. Deliège (1987) discusses these situations: for instance, there are cases in which one rule may suggest a certain segmentation point due to preference to group events that, for example, are joined by legato (GPR 2), whereas another rule could suggest another boundary instead due to a change in register (GPR 3). It may not be possible to superimpose these two segmentations to form three groups, because one group would only include a single event, which is not allowed by another rule (GPR 1).

Tightly associated with the issue of perceptual organization, work on musical expectation investigates the effect of likelihood of events upon segmentation; for instance, in the audiovisual domain, boundary perception occurs when changes of a salient sensory (e.g., color) or conceptual (e.g., cause-effect relationship) feature are unpredictable (Zacks, Speer, Swallow, Braver, & Reynolds, 2007). In the context of music, and within a real-time perspective, the implication-realization model (Narmour, 1992) focuses on listeners' expectations regarding future events, which may or may not be confirmed by the music. According to this model, these expectations emerge based upon musical implication; for instance, one of the theoretical principles is that a large interval implies a change of direction (Schellenberg, 1997). This implication may be realized in the music, in which case listeners' expectations will be confirmed. If this implication does not occur (i.e, if change of direction does not occur after a large interval), listeners will perceive a sense of closure due to violation of expectation. In the context of segmentation, violation of expectation would be associated with boundary indication. This theoretical approach proposed two general constants, $A + A \rightarrow A$, meaning that successive events that are identical or similar with each other generate expectation of identical or similar events, and $A + B \rightarrow C$, which denotes that dissimilarity between successive events leads to the expectation of another event that is dissimilar to the preceding ones. One could therefore associate expectation violation with two scenarios: $A + A \rightarrow B$ and $A + B \rightarrow A$. Cambouropoulos (2006) has questioned the generality of these principles: in many cases, two subsequent events do not

suggest anything about the nature of a third event, or may even suggest an outcome that would contradict Narmour's constants. Although these constants illustrate the common tendency of repetition to imply repetition and of difference to imply difference, it is problematic to treat them as general cognitive principles.

One of the commonalities of these three theoretical approaches is that they consider the hierarchical aspects involved in the representation of structure. For instance, the similarity between abstracted cues determines the formation of groupings of groups (Deliège, 1989). Further, metrical structure position and tonal hierarchy are considered to define the relative importance of certain musical events with respect to others within a given time span (Lerdahl & Jackendoff, 1983). Tonality also largely contributes to perceived musical structure due to expectancy: unimportant events in a tonal hierarchy generate expectations of musical relaxation that are often confirmed when more important events evoke resolution (Bigand, Parncutt, & Lerdahl, 1996; Margulis, 2005).

The hierarchical representation of structure can be associated with the concept of schemata. This notion stems from the psychological construct of *memory schema*, which can be defined as the formation of "general or associative semantic representations" (Agres & Wiggins, 2015) of new information based on previous experience. In this regard, it has been proposed that knowledge about goals and intentions underlying sequences has an effect on segmentation, particularly upon grouping at larger scale (Kurby & Zacks, 2008). Some examples of musical schemata include gap-fill and changing note archetypes (Rosner & Meyer, 1982), deemed to generate schematic expectations (as opposed to veridical, which are expectations that are intrinsic to specific musical pieces only, see Justus & Bharucha, 2001); another example is cadential closure (Sears, Caplin, & McAdams, 2014; Peebles, 2011). Since some patterns or groupings can be better associated with schematic knowledge than others, this implies that not all of them are equally important, which relates to the idea of hierarchies in segmentation (Hard, Tversky, & Lang, 2006). It could be stated in this regard that the concept of schemata intertwines functional and temporal hierarchies.

One of the common questions regarding musical schemata in the area of music concerns the role of enculturation and training; for instance, due to schematic knowledge, musicians would tend to segment at hierarchically superior levels, whereas a finer segmentation would be expected for non-musicians. From a music-theoretical point of view, it is important to define who is the listener; for instance, Deliège (1989) investigated how the cue abstraction model applies to different groups of listeners, and also explored the relationship between the composers' intentions and the perceptual structure (see Deliège, 2001, on imprint formation). Musical training has been regarded, in this respect, to facilitate the formation of prototypical patterns due to a higher ability to memorize musical events (Deliège, 1989). In contrast, GTTM focuses on a listener that may be experienced, but depending on the *artistic issue* she may be less sophisticated, closer to perfection, and so on (Lerdahl & Jackendoff, 1983); in other words, some musical works would yield higher agreement between listeners regarding grouping preferences. In practice, however, it may be difficult to establish how much

agreement between listeners could be considered to be enough agreement, and what types of commonalities or dissimilarities between different representations of structure should be taken into consideration.

2.2 Perceptual segmentation of music

Experimental work on the psychological problem of how temporal sequences can be represented and understood by listeners has often been addressed via real-time segmentation paradigms (e.g. Newtson & Engquist, 1976). As pointed by Zacks and Swallow (2007), real-time segmentation is characterized by the following: I) it occurs for both spatial and temporal information processing and shapes memory and learning; II) it is an operation that generally does not require special knowledge; III) it occurs spontaneously, and can be considered to be an automatic process in some cases; IV) it is a hierarchically structured representation of events, which means that multiple events with high level of detail can be grouped together into larger units (Kurby & Zacks, 2008; Hard et al., 2006), and not all event boundaries have the same level of importance; V) it may involve specialized neural mechanisms. In the study of music and other information streams, the distinction between real-time and non real-time segmentation is particularly interesting: it can help to explain how structural representations of a piece are gradually constructed. In everyday life, we may take advantage of both types of music segmentation: we segment in real-time when dancing to music that we never have heard before in order to change our dance moves in synchrony with the beginning of musical sections, and we segment in non real-time while trying to find the time position of our favorite part of a musical piece by navigating through a recording.

In theory, temporary and final states of listeners' understanding are separable. It would be possible to make segmentation predictions that would correspond to that of an 'ideal listener' whose intuitive representation of musical structure would not be subject to concomitants of real-time processing. Real-time perception can be associated with a temporal logic: the confidence that a given instant of change should (or should not) be indicated as a boundary varies over time. In practice, these aspects are not so easy to study because temporary states of listeners' understanding may not be well represented by real-time indications: real-time segmentation involves the *anticipation* of future boundaries, and also *critical retrospection* regarding past boundaries or events. In this sense, real-time perception does not follow the linear sequence of the acoustic phenomena but rather a non-linear path. This is somehow associated with other questions regarding the order of occurrence of events on segmentation. Studies on the effect of event order on perceptual musical structure (Tillmann & Bigand, 1998, 2004; Lalitte & Bigand, 2006) show that coherence between local structures is much more important for the experience of a listener than coherence between global structures. This would imply, in the context of boundary perception, that anticipation and critical retrospection of boundaries involve relatively short temporal contexts, such

that listeners' representation of structure may more or less resemble real-time indications. That is, listeners probably build more hypotheses regarding future boundaries that are proximal in time than with respect to temporally distant ones. Although experimental segmentation studies on music segmentation have not investigated differences between indications in real-time and in non real-time, they have often focused on differences between repeated segmentation trials for the same stimulus, finding an increase in the number of indications over repeated segmentations of the target stimulus (Deliège, 1987; Bruderer, 2008; Deliège et al., 1996; Krumhansl, 1996). This trend did not reach statistical significance, however, and Bruderer (2008) found differences for audio but not for MIDI versions of the stimuli.

Following efforts by Lerdahl and Jackendoff (1983) to formalize the underlying rules behind the cognition of musical structure, a number of perceptual segmentation studies focused on their perceptual validation (Deliège, 1987; Peretz, 1989; Clarke & Krumhansl, 1990; Krumhansl, 1996; Frankland & Cohen, 2004; Bruderer, 2008; Koniari & Tsougras, 2012). Different methodologies were used, for instance, regarding the duration of the stimuli to segment or the number of allowed boundary indications, but an underlying question was common for all studies: how do grouping preference rules relate to segmentation boundary indications? Often, the approach consisted of finding whether the locations of listeners' indications coincide with the GTTM predictions (e.g., Koniari & Tsougras, 2012); another method was to collect justifications behind participants' indications (verbal description cues) which are then compared with the types of rules that would predict the boundaries (e.g., Deliège, 1987). Since the GTTM predictions are attributable to an *ideal listener*, musicianship (Deliège, 1987; Frankland & Cohen, 2004; Bruderer, 2008) and age (Koniari & Tsougras, 2012; Koniari et al., 2001) have been investigated as possible factors. All in all, significantly more segmentation boundaries in accordance with GTTM rules were indicated by 8-year-old children when compared to 6-year-olds, but regarding the role of musicianship, which has been investigated more, findings have been rather inconclusive. For instance, both musicians and non-musicians seem to segment based on aspects related to musical 'surface' (pauses, changes in register, dynamics, see Bruderer, 2008), although musicians' segmentation may better correspond with musicological rules (Deliège, 1987) and with offline expert annotations (Mungan et al., in press).

Another pertinent issue is that the cultural background of listeners could determine segmentation strategies; musicians from different cultural backgrounds seem to differ in their segmentations (see Ayari & McAdams, 2003; Lartillot & Ayari, 2011), but this does not seem to be the case for non-musicians (Mungan et al., in press). It is also plausible that musicians and non-musicians differ in the speed at which they recognize boundaries in order to indicate them. In this respect, psychological studies reported faster capture of statistical structure of perceived streams (François, Jaillet, Takerkart, & Schön, 2014) and larger auditory memory spans (Tierney, Bergeson-Dana, & Pisoni, 2008) in musicians.

Focusing on the stimulus to be segmented, one could suspect that different styles of music would lead to similar segmentation strategies from listeners; for

instance, although tonality, timbre, and repetition of sequences may be very important aspects in music segmentation, listeners may focus on other types of change (e.g., solely on rhythm) for certain pieces. This relates to the distinction between veridical versus schematic expectations (Justus & Bharucha, 2001); such a differentiation between event models and event schemata is plausible but not well supported by experimental tasks (Zacks et al., 2007). It is also important to consider the extent to which the use of naturalistic stimuli is crucial for studying segmentation; according to a study on Western music (Bruderer, 2008), different versions of the same musical piece (a synthesized melodic line, a synthesized polyphonic line, and a real polyphonic example) lead to essentially the same segmentation, meaning that listeners focus mainly on melodic lines. It is possible, in this respect, that perceptual processing into musical *stream segments* (Cambouropoulos, 2010; Rafailidis, Nanopoulos, Manolopoulos, & Cambouropoulos, 2008), which might be involved in the representation of structure, may often lead to similar segmentations for polyphonic and monophonic versions of the same piece.

2.3 Computational segmentation of music

The possibility of modelling listeners' segmentation and structural annotations with fair accuracy has been of great interest among music researchers coming from different areas. Since musical structure is an essential attribute associated with the perceptual integration of different musical elements, the prediction of perceptual structure in music can help explain the extent to which, given a context, specific features of music can systematically prompt a representation of structure. By estimating the prediction accuracy of different segmentation models, we can gain knowledge of the role of perceptual organization rules, and memory-based and attentional processes on the perception of structure. In addition, automatic segmentation of music has a plethora of applications: it can used, for example, for music navigation and visualizations, or to facilitate other automatic tasks such as music transcription or genre classification.

The first attempts to approach segmentation prediction were carried out over monophonic music in the symbolic domain. The two approaches that encompass most of the work on segmentation prediction in the symbolic domain are rule-based and data driven (Ellis, 1996; Parncutt, 1998). The greatest difference between them is that rule-based approaches are short-term and whereas data driven are long-term. Rule-based approaches offer a description of musical structure based on a set of principles of perceptual organization, which may reinforce each other or enter into conflicts depending on the characteristics of the percept under analysis. Out of the many ways to structure the perceptual field, organizations characterized by their simplicity are given privilege, following psychological principles from Gestalt theory. To give an example, the Temporal Gestalt Units (TGU) segmentation model (Tenney & Polansky, 1980), which is based on pitch, duration, dynamics and timbre, follows the assumption that elements in a sequence that differ from

previous elements by an interval that is greater than preceding and following intervals are perceived as change with respect to a given musical dimension. The minimal context required to define a segmentation is four notes (three intervals), although the approach can be extended for the analysis at larger time scales, forming a hierarchical representation. Here, the simplest organization would group together subsequent events that do not differ much between each other with regards to a given dimension.

Another rule-based example is the Local Boundary Detection Model (Cambouropoulos, 1998, 2001). Roughly, it is obtained as a weighted sum between three measures of boundary strength, which are absolute pitch interval, inter-onset interval, and offset-to-onset interval. The main difference with the previous model is that here, the focus is on any change between two subsequent intervals. Compared to Tenney and Polansky (1980), this model does not require an interval to be preceded and followed by shorter intervals to increase boundary likelihood. Instead, the likelihood of indicating a boundary at a given target event increases if it introduces a change with respect to both past and future events, unless the events are characterized by ascending or descending equidistant steps (such as a regular increase or decrease in pitch or length of events). The temporal context for determining boundary likelihood is also shorter, as the Local Boundary Detection Model uses three events (two intervals) at a time in the estimation. In addition, we mention a rule-based approach proposed by Temperley (2001) called Grouper, which consists of a gap rule (preference to locate boundaries at large inter-onset intervals and offset-to-onset-intervals), a phrase length rule (preference for phrases having 8 notes), and a metrical parallelism rule (preference to begin successive groups at parallel points in the metrical structure). Also Frankland and Cohen (2004) quantified some of the grouping preference rules proposed in GTTM with the purpose of validation via perceptual experiments.

Data driven approaches (Bod, 2002c, 2002b, 2002a; Pearce, 2005; Pearce & Wiggins, 2006; Pearce, Müllensiefen, & Wiggins, 2010) are based on the idea that statistical regularities determine perceptual grouping. Under this approach, segmentation may occur due to a change associated with a feature, but hearing a rare sequence or relationship between notes should also contribute to boundary perception. Listeners focus on simplicity but are biased by likelihood; for instance, Gestalt rules predict that large pitch jumps are supposed to be separated by boundaries, but this may not apply for musical pieces characterized by frequent pitch jumps. Since here, the likelihood of occurrence of intervals and sequences determine grouping, these approaches are considered to be *memory-based*: statistical regularities are 'learned' from other (e.g., previous) parts of a piece or from a large musical corpus. These approaches follow the idea that the simplest organizations are preferred but the most likely organizations prevail (Bod, 2002b). Hence, while rule-based approaches follow a priori assumptions to determine whether a sequence should be grouped together or clustered, data driven approaches use the a posteriori assumption that if a series of events occurs frequently, it is more likely to be grouped by listeners. In both cases, however, the quantification of relationships between events (e.g., inter-onset intervals) is required. Since data

driven models can be trained using previous knowledge, they can be used to study schematic and veridical expectations in music: for instance, a major second is in principle unlikely to be a significant change because it is rather ubiquitous in music, and a motif is likely to be grouped together as it reappears in a musical piece. Another advantage of this approach is that it potentially offers a better account of short-term memory processes in real-time segmentation, as it can consider recent occurrences of material. Also, since different music corpora can be used for training, data driven approaches may be applied for modelling segmentation from different types of listeners (e.g., rock and jazz enthusiasts), provided that the model is based on actual relationships between events that listeners focus on. A possible problem with data driven approaches is that they are often based on the repetition of notes or of relationships between notes; exact repetition is often not enough for accurate prediction of musical structure, because the statistical regularity of an input is also determined by the similarity between events or their relationships. In this sense, data driven approaches may offer only a partial solution to the problem of musical parallelism: for example, motifs of equal length, similar durational values and different pitch may yield inaccurate patterns of melodic expectation.

Studies on motivic pattern extraction (Lartillot & Toiviainen, 2007b; Cambouropoulos, 2006; Conklin & Anagnostopoulou, 2001) are related to data driven approaches as both can be based on pattern matching techniques (Bod, 2002c, uses stochastic grammar instead), but they differ in their target (to find motifs vs. to estimate expectancy of events); despite their differences, both approaches can be used to find segment boundaries. Motivic pattern analysis techniques also incorporate perceptual organization rules, for instance, to avoid obtaining redundant long pattern candidates when sequences are repeated (see the problem of cyclic patterns in e.g. Lartillot & Toiviainen, 2007b), or two match sequences that contain variations, for instance in their endings. Again, issues of similarity and parallelism pose challenges to pattern analysis: for example, two patterns might be identical except for differences in ornamentation (Cambouropoulos, 2006). Another issue in pattern analysis is polyphony, which can increase the complexity of heuristics; geometric pattern discovery has been applied over pitch-time representations to address this issue (Meredith, Lemström, & Wiggins, 2002; Meredith, 2015). The extension of these methodologies to the audio domain is difficult, but has great potential with regards to other problems including music transcription.

Prediction of musical structure in the audio domain is currently a much investigated issue, especially with regards to the development of automatic segmentation and structural analysis algorithms, but also within the contexts of music summarization (or *thumbnailing*), chord extraction (Mauch, Noland, & Dixon, 2009), music transcription (Maddage, Xu, Kankanhalli, & Shao, 2004), and music classification (Barbedo & Lopes, 2007). The main advantage of such an approach is the use of real-world polyphonic signals: the automatic analysis of real performances has a very clear musicological, industrial, and societal impact. Currently, the focus of research in this area is mainly on segmentation into structural sections that can be labelled, such as intro, verses and choruses (see Paulus et al., 2010). Several approaches to automatic structural analysis of musical audio have

been proposed. These have been categorized into three types (Paulus et al., 2010; Serrà, Muller, Grosche, & Arcos, 2014): homogeneity-based or "state" approaches, repetition-based or "sequence" approaches, and novelty-based approaches.

The first of these methods is based on Hidden Markov Models (HMMs), the main assumption of which is that sections such as verses and choruses should be homogeneous with respect to a given musical property, so the challenge involves finding what makes a section homogeneous in order to detect transitions between sections. For instance, given a specified number of possible states (Rabiner, 1989), each of them corresponding to a possible combination of instruments in a piece of music, state transition probabilities can be used to determine the points at which the combination changes over time (Aucouturier & Sandler, 2001); most often, the interest is in finding points of change between structural sections (Levy & Sandler, 2008). By assigning a different state to each time point, HMMs can be used to directly find boundaries without the need of peak picking. In these approaches, the number of states must be defined in order to model probabilities of being in a state. Hence, HMMs are more suitable for categorical data (such as key and chord labels, see Pauwels, Kaiser, & Peeters, 2013); however, varied implementations based on HMMs have been used in structural analysis (Paulus et al., 2010).

The repetition approach for structural analysis is based on time lag representations (Goto, 2003). These methods aim to find repetitions and similar segments with respect to a given feature based on similarity matrices (see further in Chapter 4). To obtain similarity matrices, first the dissimilarity (obtained via Euclidean distance) between all possible pairs of points of a multidimensional, frame-decomposed feature (see 4.3.2) is computed (Foote, 1999). The result is a dissimilarity matrix, which is then inverted. Points in the main diagonal represent similarity between adjacent frames; other diagonals compare similarity between frames that are temporally more distant with each other, such that the further the diagonal is from the main diagonal, the larger the temporal distance between pairs of frames will be. While all musical pieces tend to show clear stripes in the main diagonal (due to high similarity between adjacent frames), similarity matrices of pieces with repeated segments exhibit stripes in non-main diagonals, provided that the extracted features yielded similar descriptions for repeated parts. The position and duration of the stripes can be used to easily indicate which parts are being repeated and when the repetition occurs. However, extracting these paths is computationally not so simple because they are placed over diagonals. Time lag representations offer a simpler representation of the duration and temporal difference between repeated parts, because all the diagonals of the similarity matrix are transformed into horizontal lines. These representations are especially useful for music with a stable tempo, because music with tempo changes will result in curved lines which are more difficult to trace; however, methods for tracing these paths from similarity matrices have been proposed (Müller & Kurth, 2006).

The third approach to audio-based structural analysis, which is related to the previous approach, is based on analysis around the main diagonal of the similarity matrix (Foote, 2000). Novelty-based approaches work under the assumption that structural segments can be found via the detection of high dissimilarity between

subsequent regions, provided that each of these regions is characterized by high self-similarity. Detection of novelty points is obtained via convolution of a checkerboard kernel with Gaussian taper along the main diagonal of the similarity matrix (or equivalently, cross-correlation between these two). Often, from the resulting novelty curves, segmentation points are selected via a peak detection algorithm; these points are evaluated against perceptual boundary indications (ground truth data). A variety of implementations based on novelty curves has been proposed, for instance using summation of spectral and chroma features (Paulus & Klapuri, 2009; Eronen, 2007; Peeters, 2007) and different kernel sizes (e.g., Gaudefroy, Papadopoulos, & Kowalski, 2015). The original implementation, however, still yields satisfactory results for segmentation prediction when compared to newer methods (Aljanaki et al., 2015), even for challenging music (Bohak & Marolt, 2016).

Although similar aims apply to both symbolic and audio-based approaches to segmentation, segmentation in the symbolic domain can be considered to have a more straightforward relationship with music-theoretical studies. The computational validation of musicological rules is more difficult in the audio domain, since these are essentially focused on on analytical readings of the musical score. Audio-based approaches are often oriented towards real-world stimuli, which are often polyphonic, whereas theoretical approaches often focus on monophonic representations such as melodies. Also, studying the relationship between, for example, note onsets, offsets and rests in audio depends on developments in the area of audio-based transcription.

However, audio-based approaches to segmentation prediction that incorporate musicological models have been proposed (e.g. Pauwels et al., 2013). One could also claim that novelty detection implicitly relates with similarity and proximity principles. To give an example of audio-based rules that somehow resemble these models, Bohak and Marolt (2016) applied the assumption that any time point that follows a region of loudness below a given threshold is a segment boundary candidate, and the larger the preceding region of low loudness is, the higher the likelihood of placing a boundary at that time point will be. It could be asked which of these two conditions (low loudness or relatively constant loudness) is more important for boundary perception. Regardless, although loudness can be a relevant segmentation cue, it should be noted that the types of change that determine segmentation can be stimulus dependent; for instance, McFee and Ellis (2014) point out that structure in pop and rock is frequently determined by harmonic change, whereas in jazz it is often based on instrumentation.

2.4 Limitations and challenges in music segmentation

A problem to highlight in segmentation of perceived streams is the difficulty of studying how subjects represent a structure that evolves through time; experimental approaches may often face issues of test validity. Related to this, we should mention another validity issue related to the relevance of boundary indication data in the

study of perceptual segmentation. Music segmentation can be understood as a passive, automatic process, but explicit segmentation tasks involve attention and decision making processes (Zacks & Swallow, 2007; Bigand & Poulin-Charronnat, 2006). Because of this, it is important to discuss how can music segmentation be explored from a more naturalistic point of view. We believe that it is problematic to use explicit tasks to study segmentation, as they might provide information only about some of its aspects.

An issue associated with the use of explicit tasks is that task instructions or other factors could influence the time scale of the segmentation used by listeners. It is important to study the contribution of the operational definition used to define musical boundaries ('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, McKinney, and Kohlrausch, 2006) to the resulting boundary profiles. Also, it is important to better understand the link between musical segment boundaries and attentional processes: segment boundaries can sometimes be understood as points that capture our attention, although not all points in the music that capture our attention may be considered boundaries. It might be interesting to compare segment boundaries obtained via explicit tasks with points of increased attention to the music, and further understand how these points relate to realization of goals (Peebles, 2011; Baldwin, Baird, Saylor, & Clark, 2001; Zacks & Swallow, 2007).

We should also remark that even if numerous stimuli are used in segmentation tasks, not all kinds of music will be represented: this illustrates the need for standard data collection methods for music segmentation. Indeed, using a single method to collect a large amount of compatible data through different experiments may better approach the issue of music segmentation. A tradeoff between participant and example size could be obtained from social tagging (e.g., in the likes of listeners' comments over a representation of the audio waveform), but this would have an impact on the reliability of the data.

Another important question concerns the "semantic gap" problem: how fair is it to "jump" from basic audio features to perceptual structure? Lerdahl and Jackendoff (1983) mention that only partial representations of musical structure can be obtained from musical scores and sound waves. To put it differently, listeners derive a complex representation of structure from the acoustic signal; this operation should not be understood as mere mnemonic processing of multiple musical examples (Lerdahl & Jackendoff, 1983). Other more complex aspects regarding cognition, culture, and our embodiment, to mention a few, must play a role here. This problem extends to musicological analyses of segmentation, which also fail to yield the whole picture. Lerdahl and Jackendoff (1983) mention that although their set of rules had been constantly revised, there are always examples that do not conform to predictions. We also mention another related problem, which is the assumption that listeners' segmentation can lead to a better understanding of perceptual structure. It is important to understand that

segment boundaries involve a reduction of dimensionality with regards to more complex and abstract representations, so in this sense it is problematic to yield interpretations on how musical structure is represented based on findings from segmentation studies.

It should also be highlighted that some musical dimensions (rhythm, tonality) can be understood as being more hierarchical than others (timbre, loudness). GTTM chooses not to formalize "non-hierarchical" dimensions, although they are regarded to contribute to the hierarchical structure that listeners perceive (Lerdahl & Jackendoff, 1983). To our view, studying the contribution of other features besides metrical and tonal hierarchy might help to understand other possible hierarchical organization factors in music: for instance, one could imagine a timbral structure in pop music, where changes in percussion and bass would involve larger time scales than changes in piano, guitar or vocals.

Also, some questions regarding generalizability remain to be asked. It would be important to understand to what extent the results of segmentation studies are representative of the population, especially considering differences between cultures regarding their understanding of what is an instant of significant change in the music. In addition, it would be interesting to compare the results obtained with those from other groups of participants, including listeners from different cultures, experienced listeners, amateur musicians, and amusics.

Further, we mention problems in audio-based structural analysis that concern the evaluation of algorithms (Paulus et al., 2010). For example, multiple ground truths can be obtained for a single piece and may be equally valid. In addition, models often yield higher accuracy for specific musical styles (e.g., pop-rock music), but may fail in other cases. Also, comparisons between models are hard to establish due to the use of multiple evaluation metrics (see Lukashevich, 2008); some measures might be inappropriate from a perceptual viewpoint, because listeners prefer segmentations that score higher on some measures (i.e., *precision*) than on others (Nieto, Farbood, Jehan, & Bello, 2014). Further, retrieval of boundary detection systems is based on the use of thresholds (e.g., 0.5 s and 3 s, see Ehmann, Bay, Downie, Fujinaga, & De Roure, 2011), a problem that relates to the issue of temporal scales. To address this problem, multi-scale models have been proposed in audio music description, for instance, to study tonality (Martorell Dominguez, 2013; Gómez, 2006a; Gómez & Bonada, 2005). Our work on multi-scale modelling of segmentation is based on these efforts.

Other issues worth mentioning concern investigation of structure analysis in the audio domain. First, dealing with a mixture of sounds that overlap in frequency and in time may result in very general representations of structure. Second, audio musical features that are typically used for structural segmentation are far from ideal. That is, Mel-Frequency Cepstral Coefficients (MFCCs Logan, 2000) are still utilized for extraction besides their problems regarding perceptual interpretation: MFCCs describe energy at broad frequency ranges for low coefficient values (e.g., MFCC 1 yields higher values if there is energy at low frequencies), but as the coefficient value increases, the frequency ranges described are so narrow that it is hard to understand what is being described from a perceptual point of

view. Also, MFCCs are sensitive to many different types of changes, including changes of instrumentation, register, voicing, articulation, and loudness. Similarly, changes in chroma features can be due to pitch steps, pitch jumps, or changes of chords. This is particularly problematic when it comes to the contribution of different features to segmentation, since some features focus on more dimensions of musical change than others, and there might also be overlapping between the dimensions that different features describe. Further, rhythmic descriptors are not often extracted; effects of rhythm on musical structure are only accounted by the use of beat synchronous features. Loudness descriptors are not prevalent either: a reason might be that they do not fit the novelty-based approach since they are not multidimensional; MFCCs and chromagram may also implicitly describe amplitude to some extent. The third problem is that deeper analyses of structural aspects in polyphonic music are very complex to achieve; polyphonic audio transcription and pattern discovery based on relationships between notes in the polyphonic signal are still at a relatively early stage. In sum, many problems in structural analysis are associated to general limitations in audio-based music information retrieval regarding audio feature description, transcription, and pattern discovery that will hopefully be solved in the near future.

3 AIMS OF THE THESIS

The study of music segmentation can help us understand the complex perceptual mechanisms involved in the organization of streams in time. Furthermore, it may contribute to solve various challenges regarding automatic prediction of musical structure. Our work examines a particular understanding of music segmentation: the indication of points or instants in which the music changes significantly. These instants are assumed to relate to the perceptual structure of music. Other components related to segmentation, including repetition (within the piece and with respect to other pieces), symmetry, variation, and the influence of schematic knowledge upon e.g. perception of cadential closure, need to be regarded for a thorough understanding of listeners' segmentations. These components are especially important because they can partially or completely change the likelihood that a given local change evokes boundary perception. However, for the most part we excluded these higher level factors from analysis in favour of a greater focus on identification of local musical changes.

A central issue in this endeavor is the relationship between musical boundaries indicated by listeners and changes in the music such as contrasts in instrumentation. Since similarity, repetition and other complex aspects also prompt boundary perception, not all boundaries should be considered as points of significant change; this is the reason why, in our experiments, listeners were asked to focus on significant changes. For the sake of simplicity, we regard all indicated points of significant change as musical boundaries, even though this may not necessarily apply to all indications; in any case, the notion of musical boundaries is rather fuzzy.

It may be useful at this point to recapitulate our main research problems. One can start with an ambitious question: are segmentation strategies universal, or conversely, would different levels of cultural exposure or experience influence these strategies (Drake & Bertrand, 2001)? If universality holds, then listeners of different cultural background, and musical expertise should indicate segment boundary indications at similar time points while listening to music. This work focuses on one aspect of this question, namely on the effect of musicianship upon segmentation. As stated in the previous chapters, the current evidence is insufficient to reach

conclusions, but studies have shown that non-musicians tend to indicate more segment boundaries than musicians, and also other differences were reported. In this respect, finding out whether or not musical training shapes our perception and our appreciation of music might help us gain deeper insights about possible universals in music listening. To add more complexity to this question, music is a dynamic process that evolves over time, so our understanding of a piece of music changes as we listen to it. In this sense, we can distinguish segmentation in real-time from non real-time segmentation, a more refined annotation of indicated boundaries which involves having heard the piece as a whole. However, to our knowledge no studies have closely examined possible differences between these two modes of perceptual segmentation. In this sense, a better understanding of listeners' representation of musical structure requires to isolate the characteristics of online processing of musical material from those related to offline processing.

The second problem that is investigated in this thesis concerns the way actual musical changes point to phenomena of boundary perception. State of the art models in automatic audio segmentation are often based on timbral and harmonic changes, suggesting that these musical dimensions may be useful to predict boundaries (Serrà et al., 2014). However, these studies are usually based on a single annotation of the music that is obtained from one or few listeners with a given level of musical experience. To this date, we do not know about possible effects of musical training and experimental task on prediction accuracy of segmentation. In this respect, it would be relevant to know what musical elements (as described by acoustic features) would different listeners pay attention to in different scenarios.

The third main issue that we examined is the role of particular musical content of a piece on the accuracy of prediction. According to studies, segmentation accuracy depends on the particular characteristics of musical pieces (Peiszer et al., 2008). However the relationship between model accuracy and the musical content of a piece remains unclear. It would be important to understand what manifest characteristics of musical stimuli determine their prediction accuracy. Understanding this would allow to select a prediction strategy, for example tonal or rhythmic, that would better suit a particular piece.

This thesis examined various aspects of music segmentation regarding perceptual modelling, computational modelling and characteristics of individual musical works. We can summarize this work into three main aims:

- To investigate the role of musical training on segmentation and explore possible differences between real-time and non real-time segmentation
- To devise audio-based computational approaches to model perceptual segmentation for different tasks and groups of listeners
- To understand what characteristics of musical pieces that are described by audio musical features would be associated with segmentation model accuracy

Figure 1 shows a general overview of the main topics covered in this thesis and their corresponding studies. Boxes with rounded corners refer to the conducted

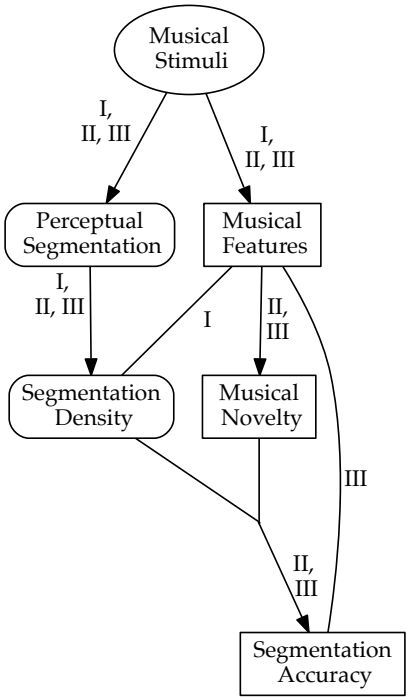

FIGURE 1 Schematic overview of topics and studies included in this thesis.

listening experiments and to the perceptual modelling of segmentation, whereas boxes with sharp corners relate to computational modelling of segmentation based on extracted musical features. There are commonalities between the studies regarding their procedures: boundary data obtained from listeners' segmentations of musical stimuli is used to generate perceptual segmentation boundary density, and musical features are extracted from the audio signal in order to understand different aspects related to segmentation. In both PII and PIII, we investigated the use of novelty detection based on musical features to obtain computational models of segmentation for prediction of segmentation boundary density; different models were compared based on their accuracy. In PI, we explored the relationship between different aspects related to the segmentation boundary density curves and global musical features extracted from the audio stimuli. PIII focused on global characteristics of audio musical features and their potential as predictors of segmentation accuracy.

4 METHODS

This section focuses on a selection of the methodological approaches utilized in our studies regarding modelling of segment boundaries and audio-based computational prediction of perceptual segment boundary density. We cover methods related to data collection and analysis of perceptual segment boundaries, audio-based modelling of segmentation, similarity between different sets of boundary data or data derived therefrom, and finding suitable segmentation time-scale parameters.

4.1 Boundary data acquisition

As mentioned in previous chapters, perceptual segment boundary data can be acquired in real-time and in non real-time contexts. In our experiments, both tasks involved the indication of boundaries during music listening, but the non real-time segmentation (annotation task) involved a higher familiarity with the stimuli, and the possibility to modify the location of indicated boundaries or to remove them after the segmentation.

4.1.1 Participants

We obtained segmentation data responses from 18 non-musicians (11 males, 7 females) and 18 musicians (10 females, 8 males). The mean age was similar across groups: 27.28 years (SD = 4.64) for non-musicians and 27.61 years (SD = 4.45) for musicians. All the participants were local or foreign students and graduates from the University of Jyväskylä and Jyväskylä University of Applied Sciences. Musicians had an average of 14.39 years (SD = 7.49) of training, and most (12 participants) played classical styles; the rest of the musicians (6 participants) played non-classical musical styles such as rock. They played piano (5), guitar (4), flute (2), bass guitar, clarinet, saxophone, cello, violin, viola and voice as main instrument. All the musicians considered themselves either as semiprofessional

(12 participants) or professional (6 participants); they also reported having 6 or more years of training. Non-musicians reported being musically untrained (except for compulsory musical training during school); none of the participants reported skills in dance or sound engineering.

4.1.2 Musical stimuli

For the listening tasks, we selected 6 multi-instrumental pieces and 3 polyphonic piano pieces comprising various styles. The examples were mainly excerpts extracted from longer pieces, with a duration ranging from 2 to 8 minutes. For some analyses of PI and all analyses of PII and PIII, we used 6 of the mentioned pieces; two of them lasted 2 minutes and the other four were trimmed down to around this length for a total experiment duration of around one hour. These pieces considerably differ from one another in terms of musical form and emphasize aspects of musical change of varying nature and complexity:

Genesis: experimental pop-rock instrumental piece that is characterized by the use of different types of electronic percussion and synthesizers within rather long and homogeneous sections in terms of melody, harmony, and in some cases also loudness and number of instruments.

Smetana: extract from a romantic symphonic poem including a long cantando theme played by clarinets and horns that is followed by a stringed fugato theme (De Lisa, 2009); these two parts later return as reprises in different keys and with larger instrumentation.

Morton: fox-trot piano piece with a 4-bar introduction that is followed by variations over a 12-bar blues progression; its rather catchy melodic line is often intruded by sudden breaks, which are responded with rhythmic chordal clusters (Trythall, 2002).

Ravel: impressionist piano composition that experiments with whole-tone and pentatonic scales and is characterized by its high technical virtuosity due to the constant use of arpeggio, glissando and tremolo (Sonntag, 2011); the piece includes changes in dynamics, register, and tempo that may be heard as highly unexpected.

Couperin: piano rendition of a baroque piece for harpsichord that is characterized by highly ornamented semiquaver melodies, which are accompanied by a quaver or semiquaver bass lines; closed cadences are common in the piece but the use of triads is very rare (the piece only contains four triads).

Dvořák: symphonic piece that is mainly based on a traditional Czech folk dance but also suggests a polonaise rhythm; its main theme is introduced by winds and horns, and is later played by the whole ensemble and transposed up a fourth (Šupka, 2013).

In addition, 3 longer pieces that were segmented only in the real-time task were included in PI: *Piazzolla*, a modern Argentinian tango piece, *Dream Theater*, a progressive metal song, and *Stravinsky*, an avant-garde ballet and orchestral concert work. These 8-minute examples were trimmed into chunks of approximately 2

minutes to avoid fatigue of participants; their boundary indications were later concatenated across chunks for analysis.

The main reason for focusing on polyphonic musical pieces was to prompt segmentation relying on processes of texture change; these types of changes are most noticeable in *Genesis*, *Stravinsky*, and *Dream Theater*. These musical pieces, however, did contain repetitions or similar sections (e.g. *Genesis*, *Smetana*, *Couperin*, *Dvořák* and *Dream Theater*); although the effect of repetition was not our focus of analysis, we were interested in employing ecologically valid stimuli, and repetition is a common characteristic of most music. Similarly, we aimed to obtain more generalizable results by including musical pieces with varying degrees of structural complexity. For example, *Genesis*, *Smetana* and *Dream Theater* may induce high level structural boundaries whereas the opposite may be the case for *Ravel*, *Morton*, or *Piazzolla*. Other criteria for selecting the stimuli are mentioned in PI; web links for listening to the musical pieces (using the music streaming service *Spotify*) were included in the *Abbreviations* at the beginning of this thesis.

4.1.3 Experiment I: Real-time task

To obtain real-time segmentation responses, we devised a Max/MSP computer patch that presented the stimuli through headphones and involved the use of keyboard and mouse to record listeners' responses (see Figure 2, top). The interface included a play bar to give listeners an idea of the relative duration of the stimulus and to indicate the current time position; a visual feedback was triggered by each boundary indication. The main task for listeners was to indicate instants of significant change by pressing the space bar key of the computer. The stimuli were presented in randomized order. The boundary data was recorded in a single pass, meaning that listeners neither had the possibility to listen to the stimuli before the segmentation nor were able to modify their boundary indications. The following instructions were given to participants: "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"'.

4.1.4 Experiment II: Annotation task

The second experiment aimed towards obtaining more comprehensive and precise boundary data. The same group of musicians that participated in Experiment I was recruited for this experiment, because they all reported experience in basic audio editing. For this experiment, we prepared an interface in Sonic Visualiser (Cannam, Landone, & Sandler, 2010) that recorded boundary time points and boundary strength ratings for six two-minute stimuli (see Figure 2, bottom). For each stimulus, a waveform was shown as a visual guide; it was possible to zoom the time scale of the waveforms (horizontal zoom). The interface offered the possibility

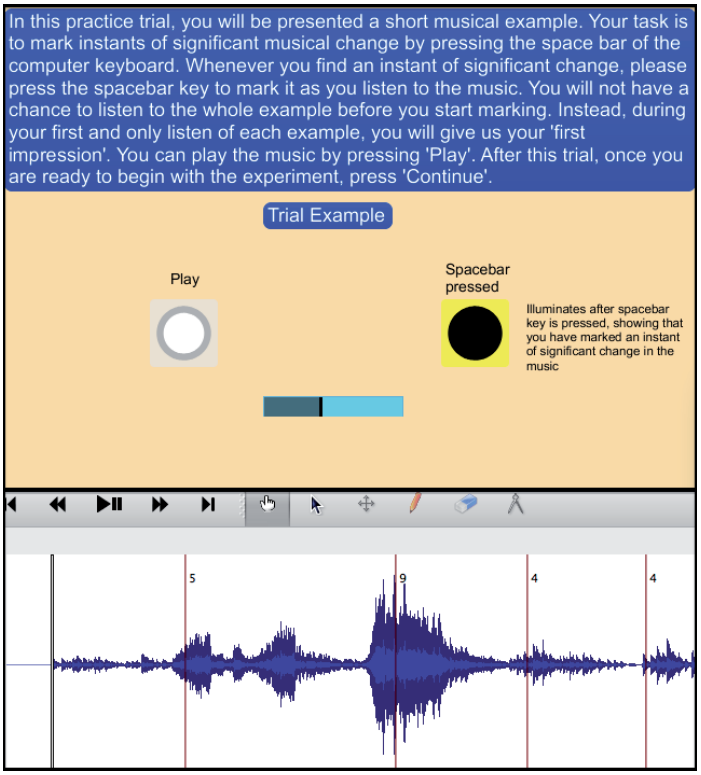

FIGURE 2 Top: Trial instructing subjects to indicate instants of significant change during music listening (Experiment I). Bottom: Part of an annotation segmentation (Experiment II) for stimulus *Ravel*; red vertical bars refer to boundaries indicated by the listener, and numbers situated next to the bars are perceived boundary strength ratings.

to play back the music, add indications, reposition them, remove them, and rate their strength using keyboard and mouse. Listeners were asked to focus on the music rather than on visual content to avoid a bias towards boundary indications based on amplitude changes. Headphones were used to playback the music at a comfortable listening level.

Each of the target stimuli was first presented for listening from beginning to end before segmentation in order to prompt more deliberate indications. Subsequently, subjects were asked to indicate instants of significant change while listening to the music. Finally, listeners could freely play the music from any time point in order to correct the time location of the indications for higher precision or remove any unwanted boundaries, and were asked to rate the perceived strength of each indication. They were asked not to add any new indications, as we assumed that they would tend to "over-segment" the stimuli while focusing on musical excerpts of short duration (see Krumhansl, 1996).

In the written instructions, we included a presentation of the different tools in the interface (also, the experimenter ensured that the interface was understood) and a description of the task, which consisted of five steps:

- Listen to the complete musical example.
- Listen to the complete example, and at the same time mark instants of significant change by pressing the Enter key.
- 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.
- Mark the strength of the significant change for each instant with a value ranging from 1 (not strong at all) to 10 (very strong).
- Move to the next musical example and start over from the first step.

4.2 Perceptual segment boundary data

The aforementioned experiments were used to obtain different boundary data sets based on participant groups and experimental tasks. We then generated models with the collected boundary location data and ratings of boundary strength to obtain intermediate results regarding segmentation across participants.

4.2.1 Boundary data comparison

Analyses of indicated boundaries often involve making comparisons between sets of boundary data. Most commonly, these are done to evaluate performance of algorithms; for instance, MIR studies compare actual and predicted segmentation using performance measures, such as precision, recall and F-measure. In some cases, however, it may be necessary to compare different actual segmentations with each other, for instance if two or more listeners yield rather dissimilar boundary indications for a same musical piece. To this end, one can regard the problem of boundary data comparison as a particular case of the more general problem of comparison between point processes (Dauwels, Vialatte, Weber, & Cichocki, 2009). A point processes a sequence of discrete binary events that occur over time, such as the firings of a neuron. Analogously to boundary indication data from different participants, a group of neurons may or may not send nerve impulses in relative synchrony, although there may be some delays between neurons.

A number of metrics have been suggested to analyze the similarity between pairs or groups of point process data directly. Depending on the problem, one similarity measure may be more appropriate than the other. For example, it could be that two point processes are identical except for one event, which appears in one process but not in the other one; some similarity metrics, such as Victor-Purpura, would penalize each *event deletion* (or conversely, insertion), whereas other metrics may allow a relative number of event deletions . A common issue among most measures is the requirement of a time constant or time scale parameter: how proximal (e.g., in time) would two events need to be in order to define these as related with each other (or conversely, how distant from each other would events need to be to consider these as isolated events)? Finally, some of these measures

are characterized by the ability to control for delays between point processes: using these measures, two identical point processes should yield high similarity even if one of them was time shifted. This is a potentially interesting characteristic considering perceptual segmentation of music, as different segmentation contexts may be associated with higher indication delays.

4.2.2 Modelling boundary data

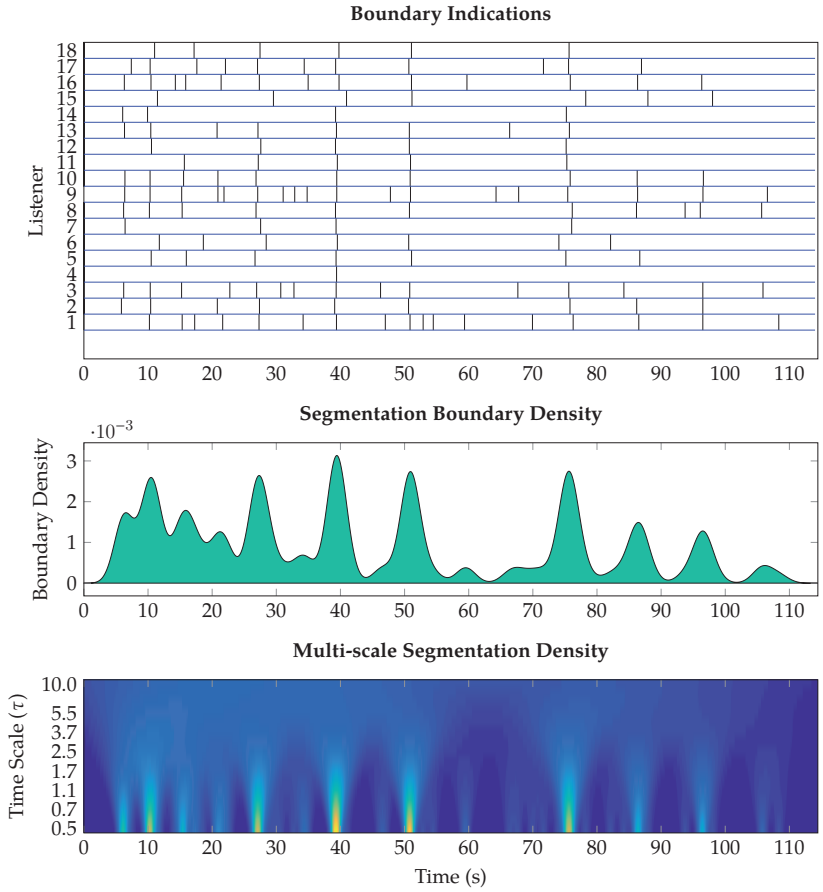

FIGURE 3 Top: Boundary Indications from each listener for stimulus *Ravel*. Middle: Perceptual segment boundary density at a time scale of 1.5 s. Bottom: Multi-scale modelling of density at 16 time-scales ranging logarithmically between 0.5 s and 10 s. Warm colors denote high density (i.e., simultaneous boundary indications by multiple listeners), whereas cool colors denote low density.

Some scenarios might benefit from an alternative approach, which consists of aggregating multiple point processes into a density curve; this method might be useful for comparison between different sets of point processes. For instance, it may be relevant to compare boundary data from a group of musicians with boundaries

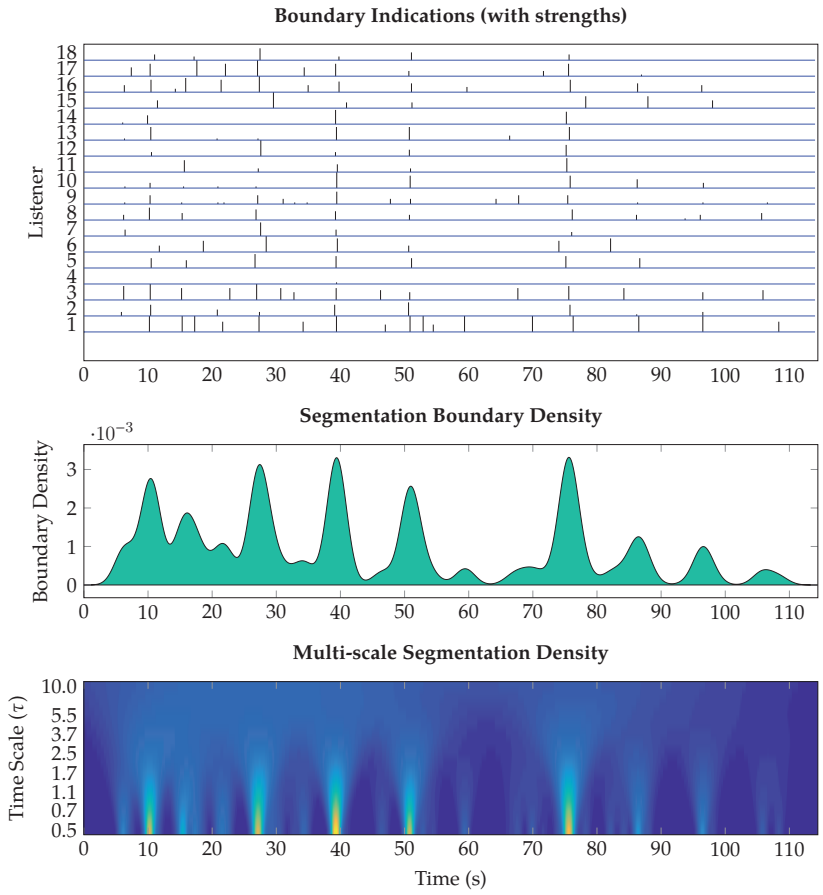

FIGURE 4 Top: Boundary Indications from each listener for stimulus *Ravel*, weighted according to perceived boundary strength ratings. Middle: Perceptual segment boundary density at a time scale of 1.5 s. Bottom: Multi-scale modelling of density at 16 time-scales ranging logarithmically between 0.5 s and 10 s.

obtained from a group of non-musicians. In this case, Kernel Density Estimation (KDE, Silverman, 1986) can be used to obtain a value estimating the probability of boundary indication for each time point. Figure 3 (middle, top) illustrates the modelling of boundary data collected from different listeners as perceptual segmentation density via KDE. High segmentation activity from multiple listeners for a given temporal region yields peak values at the middle of that region: for instance at around 40 seconds the curve reaches its maximum density, because all listeners indicated a change at about the same time. One of the advantages of this method is that the obtained model offers a tradeoff between listeners who indicate more often (listener 1) and those who indicate seldom (listener 4). The model also yields a balance between listeners who tend to segment relatively early with respect to other listeners (listener 14) and those who seem to be more delayed (listener 15).

KDEs are similar to histograms, which are also density estimators. However, Kernel Density Estimation does not separate data points into bins, but yields a smooth distribution via the application of a (usually normal) kernel function to each data point. Kernels of larger variance allow us to describe the contribution of each segment boundary to a larger temporal context, and vice versa. In this sense, this method is optimal for the analysis of perceptual segment boundary data. In our case, it allows for a perceptual modelling of each participant group and experimental task, which simplifies the analysis because it is based on a representative estimate of the segmentation across all listeners. Also, assessing similarity between time series instead of between point processes can be more convenient because comparable time series would share the same number of points, whereas this is not necessarily the case for point processes.

In our analyses, Kernel Density Estimation (Silverman, 1986) was applied to a vector of superimposed boundary indications from multiple participants. The method consists of centering a Gaussian kernel at each boundary, and then summing the kernels together. The result is a smoothed curve where each peak indicates that multiple listeners indicated a boundary around that point; this curve is further normalized to sum 1 for further comparison between different density curves. A set of weights associated with boundary indications can also be included in the computation, so that boundaries that are perceived as stronger would yield higher density, and vice versa. A smoothing parameter, the bandwidth of the Gaussian kernel, needs to be defined in Kernel Density Estimation; this time scale parameter is studied in PI, where different bandwidths are used to obtain multiple curves of perceptual segment boundary density for the same boundary data in order to find an optimal parameter value. It is possible to obtain a representation of the segmentation density at multiple time scales by organizing the perceptual segment boundary density curve into an array according to their bandwidth. The result is a multi-scale model of perceptual segmentation density, where each column refers to density at a given time point for different time scales, and each row corresponds to the perceptual segmentation density curve at a given time scale. Figure 5 shows multi-scale models for non-musicians and musicians in the

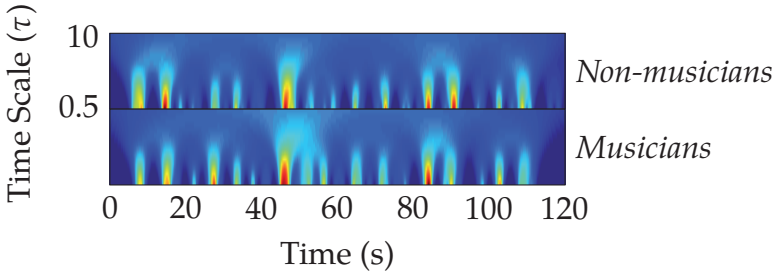

FIGURE 5 Modelling of indicated boundaries via a multiple time scale approach. The kernel density over time (stimulus *Morton* is represented for 16 time scales. Top: Non-musicians in the real-time task (*NMrt*). Bottom: Musicians in the real-time task (*Mrt*).

real-time task (stimulus *Morton*).

4.2.3 Boundary density and boundary strength

Figure 3 illustrates how boundary indications from different listeners are represented as a single-scale curve of boundary density and as a multi-scale representation. Boundary indications by 18 musicians in the annotation task are aggregated into a single-scale model using KDE; 16 single-scale models, each at a different KDE bandwidth, are aggregated to form a multi-scale representation. This could be considered an interesting representation of the musical structure: segments of around 11 s are recurrent (26-39 s, 39-51 s, 76-87 s, 87-97 s, 97-106 s), although perceptually it seems to be very difficult to estimate that these are of similar duration since the rhythmic characteristics of the music vary considerably due to changes in tempo and pulse. The single-scale model, in turn, allows for a visual comparison against a representation that includes boundary strength ratings. Figure 4 shows how these three graphs change when boundary strength is taken into account. In particular, the single-scale perceptual boundary density is different to some degree: for instance, the highest peak in Figure 3 (40 s) becomes the second highest in Figure 4. Although all participants indicated a boundary at this point, 10 participants (3, 5, 6, 7, 8, 12, 15, 16, 17, and 18) rated lower strength for this boundary than for a boundary at around 27 seconds that was indicated by less listeners. It is also clear that some participants (1, 3, 6, 15, and 16) indicated strong boundaries (Figure 4) for time points that do not correspond with high density in Figure 3, although in some cases this seems to be due to lack of agreement regarding the exact location of changes. Even though the density curves in Figure 3 and Figure 4 correlate almost perfectly ($r = .97$), according to these differences the strength of boundary indications does not necessarily relate with the boundary density for the corresponding time points; PI elaborates further on this finding.

4.3 From audio-based features to computational segmentation models

An important part of the conducted analyses in our studies aimed towards finding possible relationships between several characteristics of the musical structure and musical features extracted from the audio signal. Audio musical features and estimates derived therefrom allowed us to understand the obtained segmentation data as responses to acoustic phenomena related to the musical signal. We investigated possible relationships between: I) musical features and optimal time-scales for correlation between segmentation density for different experimental tasks; II) musical features and optimal time lag between multi-scale models for different segmentation tasks; III) changes regarding frame-decomposed musical features (obtained via novelty detection) and segmentation density; IV) musical feature characteristics and computational segmentation model accuracy.

4.3.1 Global musical features

Acoustic musical descriptors of global properties can offer an overall estimate of musical characteristics across the entire duration of a piece. This approach to musical feature extraction cannot account for dynamic evolution (e.g., tempo changes in musical pieces), but can offer useful summaries of the musical pieces under study. In PI, we extracted global rhythmic features in order to understand whether rhythmic characteristics of the stimuli would relate, for example, to listeners' perceptual delays or to the optimal time scale of the segmentation of each piece. For this purpose, we used features dependent on onset detection (Bello et al., 2005): *event density*, *tempo*, and *pulse clarity* (Lartillot & Toiviainen, 2007a; Lartillot, Eerola, Toiviainen, & Fornari, 2008). We chose the default strategies implemented in MIRToolbox 1.6.1 (Lartillot & Toiviainen, 2007a) for their computation. Event density is computed as the number of note onsets per second, for note onsets obtained from an amplitude envelope via peak picking. Tempo is estimated by obtaining the lag of the highest peak from the positive half of the autocorrelation of the amplitude envelope. Pulse clarity basically refers to the coefficient of the highest peak of the aforementioned positive half of the autocorrelation. From these computed features we also obtained *average note duration* (inverse of event density) and *beat length* ($\frac{60}{tempo}$).

4.3.2 Frame-decomposed musical features

In order to describe the dynamic evolution of different musical dimensions, frame decomposition is applied as a previous step to feature extraction. This procedure consists of dividing an audio signal into short chunks (i.e., frames) in order to later perform feature extraction over each frame separately and obtain a feature time series. The duration of each chunk is an important parameter: for instance, the description of tonal features requires larger temporal contexts than in the case of timbral features. The process of frame decomposition actually involves "sliding" a short-term window of a specified duration (called window length) along the signal. The idea of *sliding* involves that two subsequent positions of the window may either correspond to subsequent chunks of the signal or be overlapped with each other by a specified amount; in other words, the size of the window "hop" is a parameter in frame decomposition. Overlapping is used to partially smooth the representation, which is important taking into account that the time points at which the signal is divided are rather arbitrary (they do not necessarily coincide with perceptible signal changes).

In our work, we extracted five frame-decomposed features: a spectral feature, a rhythmic feature, a chroma feature and two tonal features. These were used for novelty detection in PII and to derive a global feature estimate in PIII. To describe timbre we used *subband flux* (Alluri & Toiviainen, 2010), a 10-dimensional feature that describes spectral fluctuations at octave-scaled subbands of the audio signal. The first step to compute this is to divide the signal into subbands using ten second-order elliptic filters (to achieve a sharp cutoff). For each frequency

channel, a spectrogram is computed using a window length of 25 ms and 50\Finally, dissimilarity between successive spectral frames is computed via pairwise normalized Euclidean distance (spectral flux). Unlike the commonly used MFCCs, subband flux features have been specifically designed to model musical polyphonic timbre perception; also, they have shown higher separation ability between genre classes than MFCCs in the context of musical genre classification (Hartmann, Saari, Toiviainen, & Lartillot, 2013).

For rhythmic description, we extracted *fluctuation patterns* (Pampalk, Rauber, & Merkl, 2002), a psychoacoustics-based representation of rhythmic periodicities in the audio signal that is obtained via estimation of spectral energy modulation over time at different frequency bands. First, a spectrogram in dB scale with frequencies bundled into 20 critical bands is computed using a window length of 23 ms and a hop rate of 80 Hz. Following an outer ear model (Terhardt, 1979), frequencies between 2000 Hz and 5000 Hz are emphasized, whereas energy outside this range is attenuated. Further, the spectrogram is weighted based on a perceptual model of spectral masking that, given a high-energy frequency band, attenuates energy at a region of frequencies below that band. Subsequently, for each separate critical band, a second spectrogram is computed (window length 1 s, hop rate 10 Hz) where the highest frequency taken into consideration is 10 Hz (which corresponds to 600 beats per minute). This yields, for each critical band and each frame, a description of loudness modulation. Each modulation coefficient is then weighted based on a psychoacoustic model of fluctuation strength sensation (Fastl, 1982) in order to give emphasis to modulation frequencies that are optimal for the perception of strong fluctuation (e.g., a steady beat). Finally, for each frame, the modulation coefficients are summed together; the result is a description of the dynamic evolution of periodicity for each modulation frequency.

We also computed *chromagram* (pitch class profile, see Fujishima, 1999; Gómez, 2006b), a 12-dimensional feature describing the energy distribution of each pitch class per spectrogram frame. To obtain this feature, a spectrogram for the highest energy over a range of 20 dB and for frequencies ranging between 100 Hz and 6400 Hz is first computed. Then, frequency bins are combined into chroma, corresponding to the different absolute pitches. To each chroma a central frequency cl is associated; it is calculated as $cl = 12 \times log_2(\frac{f}{cf})$, where cf is the central frequency related to C4 (set to 261.6256 Hz). The audio waveform is normalized before the spectrogram computation, and each frame of the resulting chromagram is also normalized by the maximum local value. The chromagram is then wrapped into one octave, by summing together chroma values of same pitch classes, leading to a 12-dimensional feature. We computed the spectrogram using a 3 s window length and 100 ms overlap to obtain a sufficiently high temporal resolution (see Hartmann, Lartillot, & Toiviainen, 2015, for comparisons between different window length parameters).

For description of tonality, we estimated *key strength* (Krumhansl, 1990; Gómez, 2006b), a 24-dimensional feature that represents how well the chromagram fits the different tonal profiles for major and minor keys. The key profiles are based on the probe-tone experimental method and represent the contribution of each of

the 12 chromatic tones to a given key. The key strength values of each frame are estimated via correlation between the pitch class profile and each of the 24 key profiles.

We also extracted *tonal centroid* (Harte, Sandler, & Gasser, 2006) to describe tonality. This 6-dimensional feature describes a projection of the pitch class profile onto interior spaces of the circle of fifths, the circle of minor thirds and the circle of major thirds, which are based on a toroidal representation of the harmonic network (*Tonnetz*) and are used in the Spiral Array model (Chew, 2002) for key boundary detection. At each frame, the chromagram is multiplied with the basis of a 6-dimensional pitch space in order to obtain three co-ordinate pairs, one per circularity inherent in the harmonic network.

Finally, in PIII we proposed a global estimate of amount of local variation (*Feature Flux*) of a given frame-decomposed musical feature. Our goal was to gain a better understanding of the relationship between characteristics of musical features, especially with regard to the suitability of novelty curves for segmentation prediction of a given stimulus. Feature Flux is the mean of the Euclidean distance between successive feature frames; it is calculated by obtaining the squared difference between successive time points for each feature dimension, then computing the squared root of the sum across dimensions, and taking the mean across time points. This estimate allowed us to investigate whether segmentation accuracy of different stimuli would relate to differences in feature local variation.

4.3.3 Novelty detection

Our approach for computational modelling of segmentation consisted of the extraction of frame-decomposed musical features for the computation of novelty curves. These curves describe, for each time point t, the amount of dissimilarity between two consecutive groups of feature frames (t is located in between these groups); this amount of dissimilarity is penalized by the amount of similarity within each of these groups. For instance, subsequent segments in the music that are in different keys would show high novelty for the tonal features at the point in which the key changes (which would delimit these two segments). Stark novelty peaks would be exhibited for high similarity within subsequent segments and high dissimilarity between subsequent segments.

To compute novelty curves, a dissimilarity matrix is first obtained from the audio feature of interest by computing the Euclidean distance between all possible pairs of points in the time series. This matrix is inverted element-wise into a similarity matrix, where important local contrast around the main diagonal represents high dissimilarity between neighboring events. A novelty curve is subsequently obtained via convolution with a checkerboard kernel across the main diagonal of the similarity matrix (see Foote, 2000; Lartillot & Toiviainen, 2007a; Paulus et al., 2010, for detailed explanation). This is illustrated in Figure 6; for each time point t, a novelty score is determined based upon the similarity between the checkerboard kernel (centered at t) and the portion of the similarity matrix

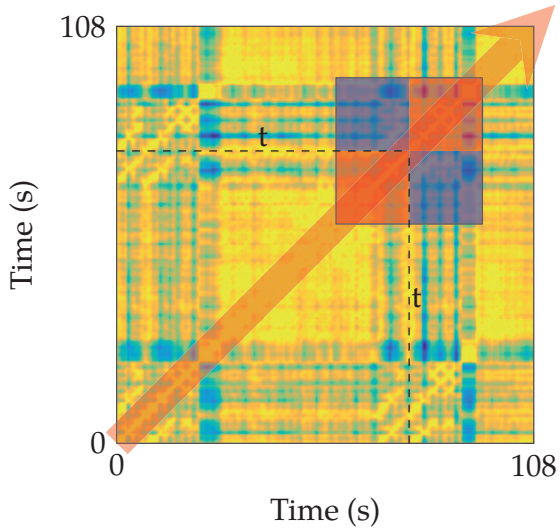

FIGURE 6 Convolution of a checkerboard kernel along the main diagonal of a chromagram-based similarity matrix (stimulus *Genesis*)

that is covered by the kernel. The width of this kernel is a crucial parameter in novelty detection; in PII, we proposed an optimal width for prediction of segment boundary data.

4.3.4 Interaction of novelty features

Besides the computation of novelty features such as chromagram-based novelty (henceforth called *basic* novelty features), we also investigated the possibility of computing new novelty features based on interactions between different novelty curves. We created interactions of basic novelty features via point-by-point multiplication between each pair of basic novelty features; this process is illustrated in Figure 7. For instance, a spectral novelty feature and a tonal novelty feature can be used to generate a spectral-tonal novelty feature: *subband flux* $\circ$ *tonal centroid* would be the result of the interaction of *subband flux* novelty and *tonal centroid* novelty, and would only exhibit novelty peaks for simultaneous change in both spectral and tonal dimensions.

4.3.5 Novelty-based modelling of segmentation density

One of the main goals of our investigation was to compare actual segmentation density based on listeners' responses with predicted segmentation density yielded via novelty detection. We used two approaches to compare perceptual segment boundary density with novelty curves. The first one consisted of computing correlations between these two, and the second one involved generating a model from an optimal subset of novelty features. To perform correlations between perceptual segmentation density and novelty curves, each novelty curve was first

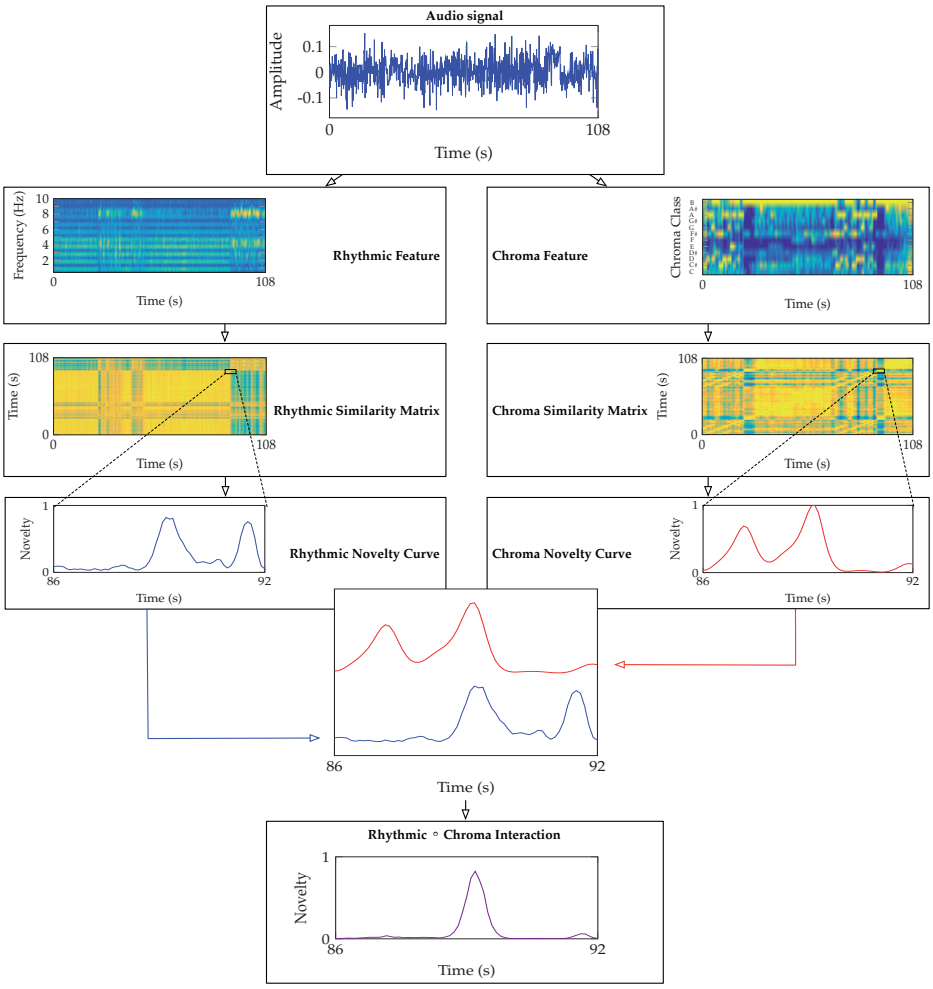

FIGURE 7 Method used to obtain interaction features via pairwise multiplication between novelty curves (stimulus *Genesis*).

normalized to sum 1. Besides the correlation coefficient (r) we obtained p values via Fisher's z transformation of r, with standard scores adjusted for effective degrees of freedom (i.e., corrected for temporal autocorrelation, see Pyper & Peterman, 1998; Alluri et al., 2012). Effective degrees of freedom are calculated based on part of the positive half of the autocorrelations of the novelty curve and the density curve: a high slope means that these are very noisy (i.e., the correlation of the time series with itself would only yield high values at zero lag), whereas a more gradual slope at around zero lag means that the time series are smoother. Since smoother time series have fewer independent time points, these would yield lower effective degrees of freedom than noisier time series.

The second approach was also based on correlation, but involved generating a model based on a combination of an optimal subset of novelty curves. We used combinatorial optimization for this problem. A Genetic Algorithm (Popov, 2005;

Eiben & Smith, 2003) was applied to avoid getting trapped in local minima in the search space; one advantage of this over other search algorithms (such as hill climbing) being that Genetic Algorithms are more likely to find the global optimum (i.e., the solution with the highest fitness value) instead of converging on local optima (a solution that will be the best among solutions that are very similar to each other, but not the best one out of all possible solutions). This is because of mutation, which prevents older and newer generations from being too similar from each other, and thus from stalling the evolution.

Genetic Algorithms can greatly differ in their approach, so here we focus on the particular functions and parameter values used in this study. The optimization was initialized with a set of random subsets of normalized novelty curves, which can be called a first generation population of candidate solutions; each generation had a population of 20 candidate solutions. In our execution, each subset of novelty features was a candidate solution, and was represented as a bit string: for example, 10100 means that for a set of 5 novelty features, feature 1 and 3 were included in the subset, whereas features 2, 4 and 5 were excluded.

At each step, the Genetic Algorithm calls a cost function to obtain the fitness values of each candidate solution, selects candidate solutions in the population (based on fitness values), and generates a new population based on them. In our case, each fitness value is the negative of the correlation between a model derived from a candidate solution and the perceptual segment boundary density; lowest values yield maximum fitness.

New candidate solutions are generated for the next population via uniform mutation or scattered crossover between the selected candidate solutions. In uniform mutation, some of the bit positions from a candidate solution are chosen (e.g., Feature 1 and Feature 4) and a random value (either 0 or 1) is assigned to each position; this mutated solution is passed to the new population. Scattered crossover is applied to a pair of candidate solutions to generate a candidate solution that will be included in the new population; the value of each bit position of the new candidate solution will be the value that either one of the pair has for that position. To ensure that the fitness obtained by at least some of the candidate solutions will not decrease from generation to generation, elitist selection is also applied: two of the candidate solutions are passed on to the next generation based on their fitness values. The Genetic Algorithm stops either after it has iterated 100 times the number of novelty features or if the average relative change in the fitness value over 50 generations is below 1×10^{-6}.

We used a cost function that finds an optimal value of the correlation coefficient y by minimizing the negative of the correlation (i.e., maximiizing the correlation) between actual and predicted segment boundary density,

$$y_{opt} = \underset{y}{\operatorname{argmin}} - \operatorname{corr}(x, p_\alpha)$$

where x is the segmentation density and p_α is the α percentile along features of a given subset. The approach consisted in finding a subset of novelty curves whose α percentile would optimally correlate with the segmentation density curve. By

taking the percentile, we perform a non-linear aggregation of novelty features that, for each time point, ranks the features based on their values. In soft computing, the percentile aggregation involves a monotonically increasing mapping that follows a continuous logic function called conjunction/disjunction function (Dujmović & Larsen, 2007). The 0th percentile (equivalent to the min function) can be roughly understood as a pure logical **AND** conjunction ("all criteria are satisfied") because if the minimum among features is high, then all features should have high values; conversely, the 100th percentile (max function) represents pure **OR** disjunction ("at least one criterion is satisfied") because a high maximum value among features implies that at least one of the features has a high value. Following this logic, 1th-99th percentiles lie between **AND** and **OR**, exhibiting varying levels of *orness* (closeness to maximum). The 50th percentile across features would in this sense be comparable to a *majority judgement*, because it would only result in high values if at least half of the features exhibited high values. We found $\alpha = 50$, which is the median ordinal position, to yield the highest accuracies with the Genetic Algorithm when compared to $\alpha = 25$ and $\alpha = 75$.

4.3.6 Novelty-based predictors of segmentation accuracy

We also investigated whether musical feature and novelty curve characteristics would serve as predictors of segmentation accuracy. Our aim was to understand whether stimuli whose novelty curves of a given feature exhibited certain characteristics would yield high prediction rates for that feature. PIII examines the use of a global estimate derived from novelty curves as an indicator of prediction accuracy with respect to segment boundary data in the annotation task. This approach might help to clarify which characteristics are common for novelty curves that yield high accuracy, and also better explain how the choice of musical feature(s) used to compute novelty curves depends on the stimulus under investigation.

Figure 8 serves as illustration of this method for the tonal centroid feature. For novelty features derived from tonal centroid, *Smetana* exhibited high segmentation accuracy, whereas *Couperin* yielded relatively lower accuracy. The figure allows us to identify three reasons why the segmentation accuracy is higher for *Smetana*. First, the boundary density (bottom) yields a lower number of stark peaks for *Smetana*, whereas for *Couperin* there seem to be more highly agreed significant changes: some of the multiple significant peaks of *Couperin* do not correspond only to changes in tonality, but also by clear rhythmic changes (e.g. 37 s and 84 s). Second, the novelty curves (middle) show a larger temporal distance between highest novelty score peaks for *Smetana* than for *Couperin*: it is possible that novelty peaks that are more distant between each other correspond to instants of change that are perceived as stronger. Third, tonal centroid representations (top) show more distinct sections in *Smetana* than in *Couperin*, probably because many dimensions tend to remain rather constant over time and change simultaneously at few time points; in this respect, *Smetana* shows smoother transitions of tonal features whereas *Couperin* exhibits higher feature fluctuation. In PIII, we systematically investigated these two last possibilities by computing two global descriptors: the aforementioned Feature Flux,

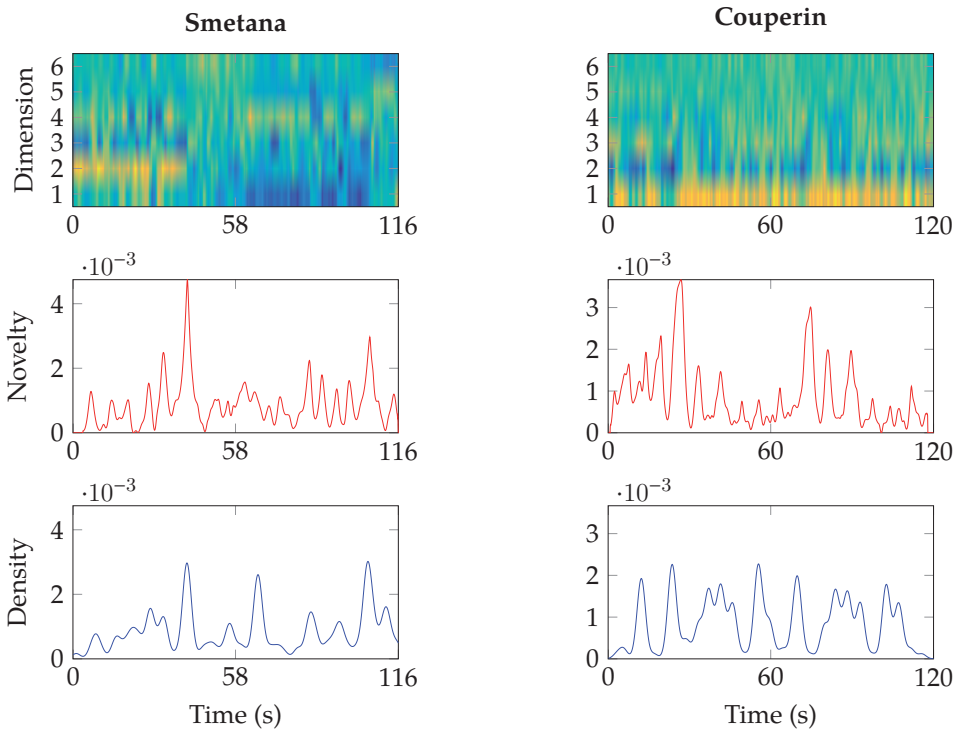

FIGURE 8 Left: Tonal centroid, novelty of tonal centroid and perceptual boundary density in the annotation task for stimuli *Smetana* and *Couperin*.

which characterizes musical features, and the *Mean Distance between Subsequent Peaks* (MDSP), which characterizes novelty features. MDSP is based on peak detection and describes the peak-to-peak duration (in seconds) of novelty curves. We investigated whether, for a given stimulus, the MDSP of a feature can be used to estimate its segmentation accuracy; our underlying hypothesis was that higher mean inter-peak distance would be associated with higher prediction accuracy. We also explored, for each musical piece individually, how the changes suggested by perceptual boundary density, musical feature, and novelty curve representations corresponded with noticeable changes of specific musical dimensions or patterns. Our aim was to better explain the accuracies obtained by clarifying both the types of changes that were attended by the listeners and the musical changes that the algorithms were sensitive to.

4.4 Comparisons between groups and experimental tasks

A number of methods have been utilized in the studies to compare listeners' segmentation for different groups and experimental tasks. Some of the analyses involved indirect comparison between point processes; for instance, in PI we

calculated the mean number of boundary indications for each segmentation task and participant group. In other cases, we estimated how adequate the perceptual density curves were with respect to perceptual point process data: in PI we assessed the goodness of model-to-data fit to find which kernel density estimation time scale would offer the best fit to the data. On multiple occasions we used Pearson correlation coefficient to perform comparisons between time series or between points corresponding to boundary indications in the time series. For instance, in PII we correlated perceptual segmentation density curves and novelty curves for each stimulus at different checkerboard kernels to find an optimal novelty kernel width for prediction of perceptual segmentation density. Also, in PI, we correlated the boundary strength ratings of the indications from the annotation task with density values in the real-time task for time points that corresponded with boundary indications in the annotation task. In addition, in the same study we computed individual multi-scale models (one model per participant) in order to obtain mean inter-subject correlations based on the correlation between each pair of models.

A potential factor that was taken into account in some of the comparisons was the possible indication delay for different groups and tasks (PI and PII). We therefore focused on the estimation of response delays for different participant groups and experimental tasks, following the hypothesis that controlling for such delays would increase correlations between perceptual density curves corresponding to different groups and tasks, and between perceptual density curves and novelty curves. In PI, we compared different groups and tasks via two-dimensional cross-correlation between multi-scale models. For each time scale, we obtained a time lag at which the cross-correlation was maximum. Subsequently, we computed the mean time lag across time scales. This method helped us to determine whether there were delays between boundary data of different segmentation tasks and different participant groups, and to understand the possible relationship between indication delays and rhythmic characteristics of the stimuli. In PII, we applied time shifting to each single-scale model, and obtained a computational model based on a subset of novelty curves that would be optimal for a given time shift of the single-scale model. This procedure allowed us to investigate the segmentation accuracy obtained at different time lags, and find the time lag that would yield maximum accuracy in order to compare prediction for different segmentation sets.

4.5 Segmentation and temporal scales

We also examined the issue of temporal scales in segmentation, which is a problem that concerns not only the comparison between groups and tasks, but also the way segment boundary likelihood is represented. In this sense, appropriate temporal scales need to be specified in the modelling of perceptual segment boundary data via KDE (PI) and also for musical novelty detection (PII). In fact, this is a more general problem related to the analysis of multiple point processes (Dauwels et al.,

2009). For instance, MIR studies on segmentation often need to evaluate model performance via the comparison of two point processes: a predicted segmentation and an actual segmentation or "ground truth". As previously mentioned, to solve the time constant issue, rather arbitrary tolerance windows are used: for instance, a predicted boundary is correct if it is at least half a second apart from an actual boundary (Turnbull, Lanckriet, Pampalk, & Goto, 2007). Our approach is different in this regard, as it prioritizes number of listeners over number of stimuli (cf. Nieto, 2015, for a MIR approach to the multiple human annotations problem), but also needs to address the same issue.

4.5.1 Time scale in kernel density estimation

One of the challenges of our investigation was to find a way to aggregate the boundary indications obtained from multiple listeners that would allow for a comparison between different segmentation sets. To this end, it was necessary to estimate a time scale parameter that would do justice to the segmentation profiles from all or at least most of the listeners. In PI we proposed methods to estimate an optimal segmentation time scale for each stimulus and to find a time-scale for comparison between multi-scale models of perceptual segmentation density.

To find an optimal time scale for each stimulus, we studied the probability of the boundary indication data given a single-scale model of perceptual segmentation density. One option was to compute, for each time scale, the log-likelihood of each boundary indication; that is, the natural logarithm of the density value corresponding to each boundary indication was divided by the total number of boundary indications. Subsequently, we summed the log likelihoods. However, this method leads to results that are biased towards the shortest time scale, which are due to overfitting: at the shortest time scales there tends to be a perceptual boundary density peak for each boundary indication, which results in high log-likelihood for all boundary indications. To avoid overfitting, the estimates were obtained with a leave-one-out procedure, such that for each subject we computed a single-scale model that did not include the boundary indications from that subject (see Duin, 1976, for a similar approach). This procedure is able, for instance, to yield relatively low likelihood at the shortest time scales, because the indications of a given listener may not correspond exactly with other listeners' indications. For each subject, we obtained the log-likelihood between a single-scale model and individual data. Then, we summed for each time scale the individual log-likelihoods together, and subsequently selected the time scale that yielded the maximum sum of log-likelihoods.

To find a time scale at which segmentation densities from different groups and tasks would be most similar, we examined the relationship between groups and between tasks at different time scales. We determined which time scales yielded the highest similarity between models by computing correlations between single-scale models. Again, we found that this approach leads to a bias, in this case favoring the longest time scales, due to the smoothing of boundary indications, which reduces the number of effective degrees of freedom. To solve this issue,

we debiased the correlation via a Monte Carlo simulation (10000 iterations). This

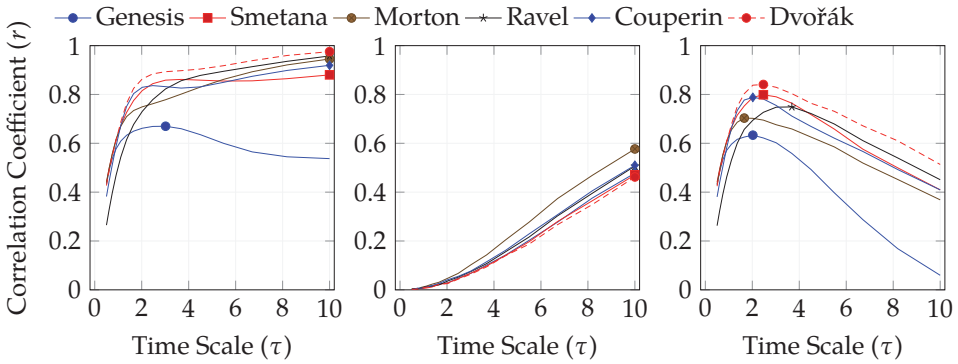

FIGURE 9 Left: Correlation between real-time (*Mrt*) and annotation (*Ma*) task boundary density at each segmentation time scale. Middle: correlation baseline obtained via Monte Carlo simulation. Right: Debiased correlation obtained via subtraction of correlation baseline from the original correlation. Maximum points for each curve are highlighted with markers.

approach is illustrated in Figure 9: we computed a correlation baseline for each combination of example and time scale, and subsequently subtracted it from the original correlation. For instance, to compare real-time and annotation task perceptual segmentation density for a given stimulus and time scale, we first obtained the (biased) correlation coefficient between these two curves. The next step was to obtain a correlation generated by two random density curves with the same number of boundaries as the density curves and the same duration. For this purpose, we generated two random sets of boundary indications, each one with identical number of boundaries to that of the actual boundary data, and with a duration between 0 s and the total stimulus duration. We generated density curves from these random sets and correlated them. The process of creating random boundary sets, perceptual segmentation density curves and obtaining a correlation coefficient between curves was done 10000 times to obtain a distribution of random correlation coefficients. Finally, the mean of the random correlation distribution was subtracted from the actual correlation coefficient between real-time and annotation task perceptual segmentation density to obtain a debiased estimate of the correlation. After computing a Monte Carlo simulation for each time scale, we could determine the optimal time scale for comparison between single-scale models by selecting the time-scale with maximum debiased correlation coefficient.

4.5.2 Time scale in novelty detection

A similar constraint regarding temporal scales applies to musical novelty detection. The size of the checkerboard needs to be specified, at least for the original Foote approach, which is most commonly used. Based upon our results regarding perceptual modelling of boundary indications, we chose a single-scale model

that would be compared against novelty curves obtained from different musical features. Once an optimal time scale of the segmentation was obtained, it was necessary to find an optimal novelty kernel width for the purpose of prediction of segmentation.

The width of the Gaussian kernel bandwidth used for kernel density estimation is not homologous to the width of the checkerboard kernel used for novelty detection. One reason for this is that musical features are often very sensitive, so novelty curves can be relatively noisy. Another reason is that, unlike segment boundary density peaks, novelty peaks are only situated between homogeneous regions, because a checkerboard kernel is cross-correlated along the main diagonal of a self-similarity matrix. In contrast, in kernel density estimation, a normal kernel function is usually used; a Gaussian kernel is situated at each boundary and then the kernels are summed together. Segment boundary density peaks can be situated between heterogeneous regions, leading to areas covering larger temporal regions than novelty curves.

In PII, we employed a systematic approach to find an optimal novelty kernel width based upon correlation between novelty curves and perceptual segment boundary density. We also obtained 26 novelty-based computational models at varying novelty widths in the search for an optimal kernel width value.

5 STUDY SUMMARIES

This chapter presents a concise description of the main problems addressed, methodologies used and results obtained in the three studies that are included in this thesis. The most relevant analyses have been described in the previous chapter; the reader is encouraged to refer to the studies for an explanation of other analyses. The general designs of Study I, Study II and Study III are shown in Figures 10, 18, and 21, respectively.

5.1 Study I

– Hartmann, M., Lartillot, O., & Toiviainen, P. (2016). Multi-scale modelling of segmentation: effect of musical training and experimental task. *Music Perception*, 34(2).

5.1.1 Introduction

While listening to music, people, often unwittingly, break down musical pieces into constituent chunks such as verses and choruses. Music segmentation studies and music-theoretical work suggests that people share a common sense of the instants at which the music in a piece changes in a significant way (Clarke & Krumhansl, 1990; Lerdahl & Jackendoff, 1983), despite varying frequency of indications among individuals (Koniari et al., 2001; Bruderer, 2008). However, neither the effects of experimental task (i.e., real-time vs annotated segmentation), nor of musicianship on boundary perception are clear. This study investigates the contribution of musical training and experimental task in phrase-level segmentation, estimates optimal time scales for segmentation modelling, and proposes an approach for modelling perceptual boundary density. Our work was guided by the following research questions:

– What is the effect of musical training on the indication of musical segment boundaries by listeners in a real-time type of experimental setup?

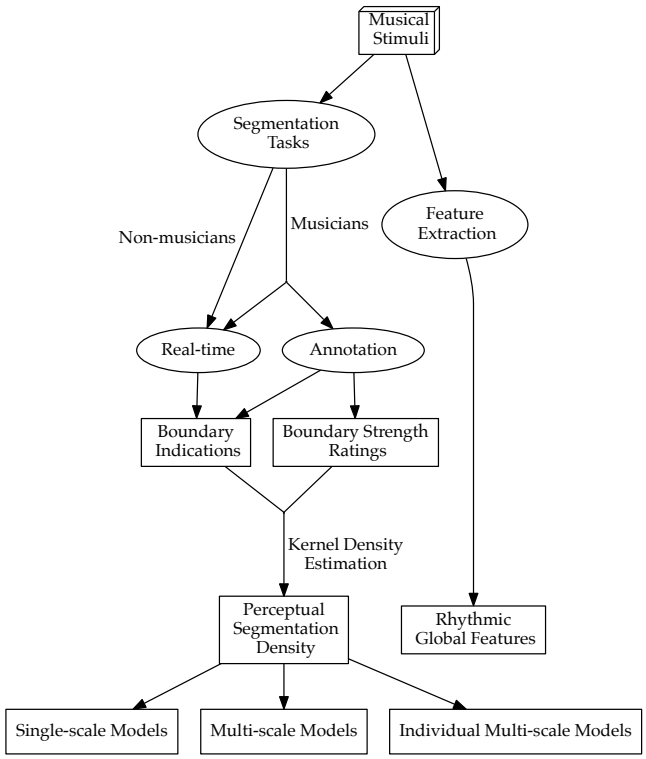

FIGURE 10 General design of Study I.

– What are the differences between a first impression of musical structure as it unfolds over time and an offline, more knowledge-driven music segmentation?
– Which global characteristics of musical stimuli modulate the optimal time scale for modelling perceptual segmentation?

5.1.2 Methods

Two listening experiments on perceptual segmentation were conducted. In the first experiment, we collected real-time segmentation responses from 18 musicians and 18 non-musicians for 9 audio musical stimuli comprising various styles. Listeners were asked to indicate instants of significant musical change by pressing a key on a computer keyboard; they did not have the possibility neither to listen to the example beforehand nor to revise their indications. The second experiment involved the same 18 musicians as in Experiment I and was based on 6 of the 9 stimuli used in that experiment. Participants were asked to listen to the stimulus, perform the same task as in Experiment I, and playback the example from different time points to correct the position of the indications, or remove indications that

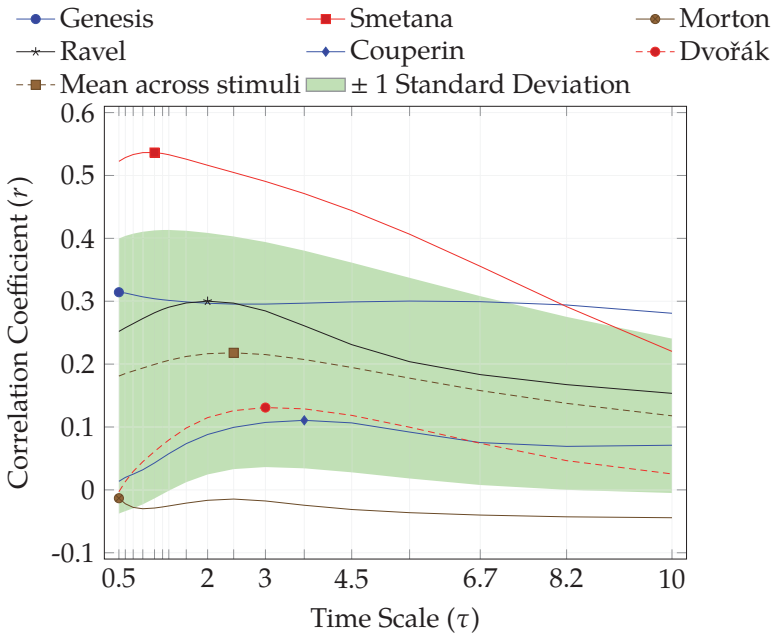

FIGURE 11 Correlation between perceived boundary strength ratings (*Ma*) and real-time task perceptual boundary density (*Mrt*) at the respective time points. Maximum points for each curve are highlighted with markers.

were added by mistake. Finally, they were asked to indicate the strength of each instant of significant change with a value ranging from 1 (not strong at all) to 10 (very strong).

The data was organized into three segmentation sets based on listeners' musical training and on the performed task: non-musicians in the real-time task (*NMrt*), musicians in the real-time task (*Mrt*), and musicians in the annotation task (*Ma*). We then estimated the mean number of indicated boundaries across examples for each participant.

Subsequently, we computed segment boundary probability curves using Kernel Density Estimation (Silverman, 1986) to obtain smooth distributions across participants. We obtained a multidimensional representation of smoothness (multi-scale model) by utilizing varying kernel bandwidths; these corresponded to 16 time scales logarithmically ranging from .5 s to 10 s. In addition, a fourth segmentation set of perceptual density curves was obtained by weighting the annotation task boundaries based on listeners' boundary strength ratings. This weighted version of the annotation task set was called *Maw*.

The possible relationship betweeen perceived boundary strength and perceptual segment boundary density was analyzed by correlating the perceived boundary strength values with the real-time task model values at the respective time points (these two sets were time-aligned with each other). Also, the degree of cohesion in each segmentation set was studied by obtaining mean inter-subject correlations. Another aim was to find a time scale that provided an optimal

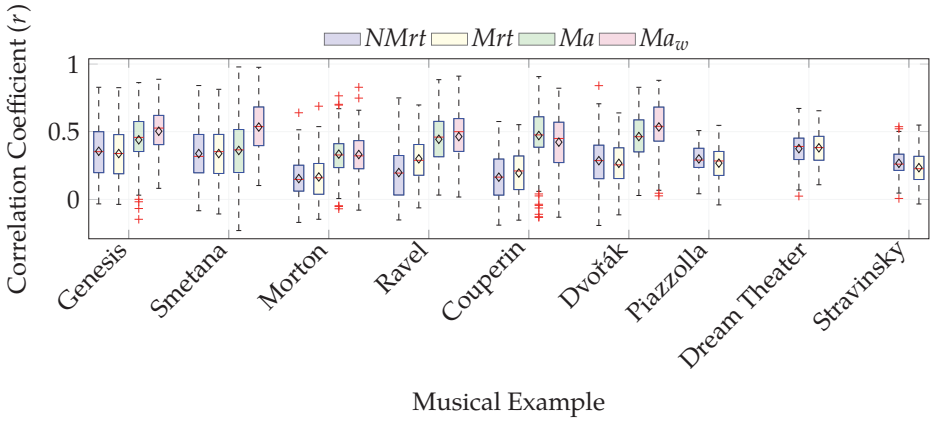

FIGURE 12 Inter-subject correlation coefficient per stimulus for each segmentation set; p-values ($***$: $p < .001$) obtained via 10000 Monte Carlo replications and adjusted using Benjamini-Hochberg correction ($q = 0.05$).

model-to-data fit and to compare segmentation sets. We computed maximum likelihood time scales based on comparison between listeners' boundary data and perceptual boundary density. After this, we looked for possible lags between tasks and groups; to this end, we examined the degree of alignment between segmentation models via a two-dimensional cross-correlation. We investigated whether these possible lags would be associated with rhythmic characteristics of the stimuli. Also, we conducted similarity analyses based on correlations between models: we correlated each multi-scale model column-wise and also at each time scale separately to find an optimal time scale to compare perceptual segment density curves. In addition, we investigated possible links between the optimal time-scale of each stimulus and rhythmic characteristics of the stimuli.

5.1.3 Results

Regarding boundary strength, we did not find a relationship with perceptual boundary density (Figure 11), which suggests that the frequency of indications of a boundary does not necessarily relate to its perceived strength. This result is contradictory with respect to previous findings (Bruderer, 2008); we should mention, however, that Bruderer restricted the analysis to a subset of boundary peaks with different indication frequencies, whereas we analyzed complete boundary data. Also, analysis on cohesion revealed similar inter-subject correlations for each group (Figure 12), suggesting no effect of musicianship, and higher inter-subject correlations for the annotation task, indicating effects of task on between-subject agreement. Analyses regarding optimal time scales showed that bandwidths of around 1.5 seconds are optimal for comparison between segment boundary densities of different groups and tasks provided that the tasks are time aligned by delaying the real-time task by approximately 1 second (Figure 13). We also found effects of task on the mean number of boundary indications across musical

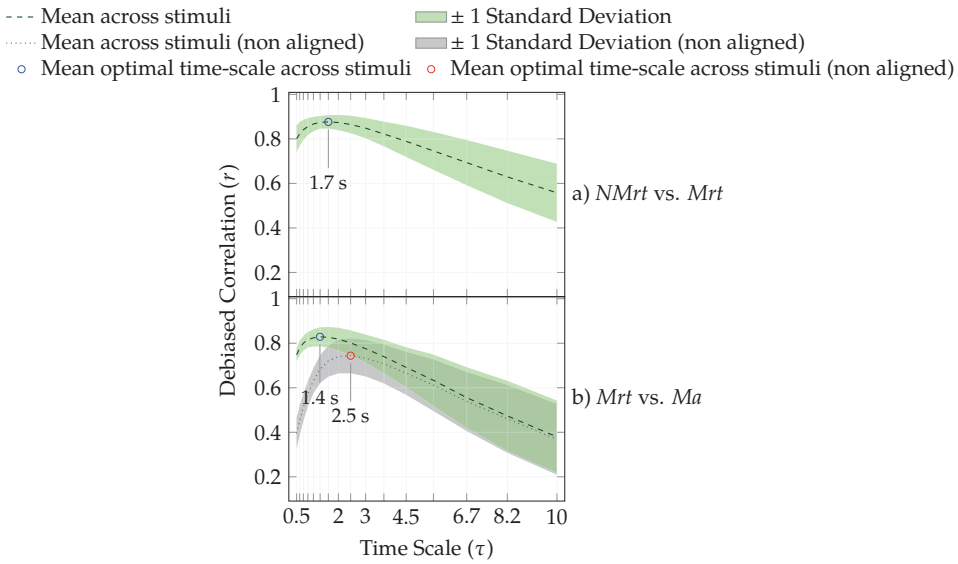

FIGURE 13 Mean correlation across stimuli between perceptual densities from different tasks and groups for each of the 16 time scales studied. Green and gray areas: 1 standard deviation from the mean. Figure *a* compares participant groups (real-time task). Figure *b* presents mean comparisons between segmentation tasks across stimuli for each alignment strategy.

examples from each participant (Figure 14); participants indicated nearly double the mean number of boundaries in the annotation task ($\overline{x}$ = 11.33 boundaries per example, SD = 8.1) compared to the real-time task ($\overline{x}$ = 5.8 boundaries per example, SD = 4.1) for the six musical examples that were common to both. We additionally found that the annotation tasks yielded lower optimal time scales (Figure 15) compared to the real-time task. Also, the time lag between tasks increased for stimuli with higher beat length, and vice versa (Figure 16). This result suggests that the real-time segmentation lag might stem from a recognition delay of around $\frac{3}{4}$ of a beat and a response delay of about $\frac{2}{3}$ of a second. Further, we found a negative link ($r(4) = -.83$, $p < .05$) between pulse clarity and optimal time scale to compare real-time task and annotation task, and a negative relationship ($r(4) = -.82$, $p < .05$) between frequency of events and optimal time scale for task comparison (Figure 17); the optimal time scale for comparison between tasks increased when the pulse clarity or event density of the music decreased.

5.1.4 Conclusion

Overall, musicians tended to segment less (and at a higher time scale) than non-musicians, but our findings did not provide support for the hypothesis of an effect of musicianship upon segmentation. With the caveat that boundary indication data is not enough to understand structural representation, one could interpret that musical training may not have an effect on how structure is mentally represented. However, it is possible either that the real-time task or an analysis oriented towards

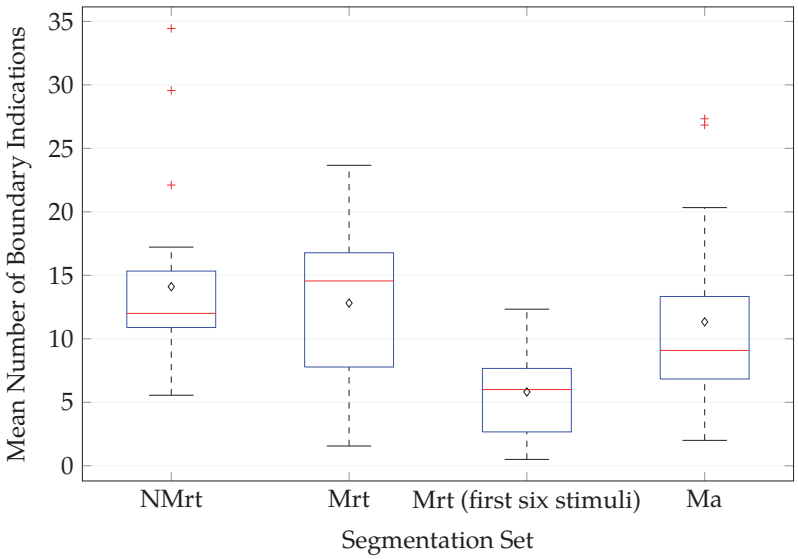

FIGURE 14 Comparison between segmentation sets with respect to the mean number of indicated boundaries across examples from each participant.

global estimates could not reveal actual differences between groups. Our findings do support the view that the experimental task used to collect boundary data has an effect on perceptual segmentation. Significant differences between tasks were found regarding number of boundary indications: listeners segmented more in the annotation task, suggesting that they may have failed to indicate some of the boundaries in real-time. Real-time task responses exhibited a latency with respect to those in the annotation task; the magnitude of this lag was found to depend on rhythmic characteristics (global beat length) of the music. This finding 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. Also, the correlation between tasks increased after the tasks were aligned; this contribution of the real-time task lag to the difference between tasks illustrates the importance of accounting for lags in real-time segmentation. In addition, time scales for optimal fit of models to data were shorter for the annotation task, suggesting that this segmentation was more hierarchical, whereas real-time task boundaries tended to be indicated at large time-scales.

Another interesting finding was that perceived strength of a boundary might not be equivalent to its frequency of indications. This suggests that not all changes indicated by few listeners are weak, probably because the perceptual salience of a change depends on what particular musical dimensions and time scales are different listeners attending to. Similarly, not all significant changes indicated by many listeners are strong, as musical context might deter participants from indicating high strength for certain noticeable changes. Finally, we found that optimal segmentation time scales may depend on global rhythmic pulsation, amount of events, and duration of events. In this respect, the time scale for

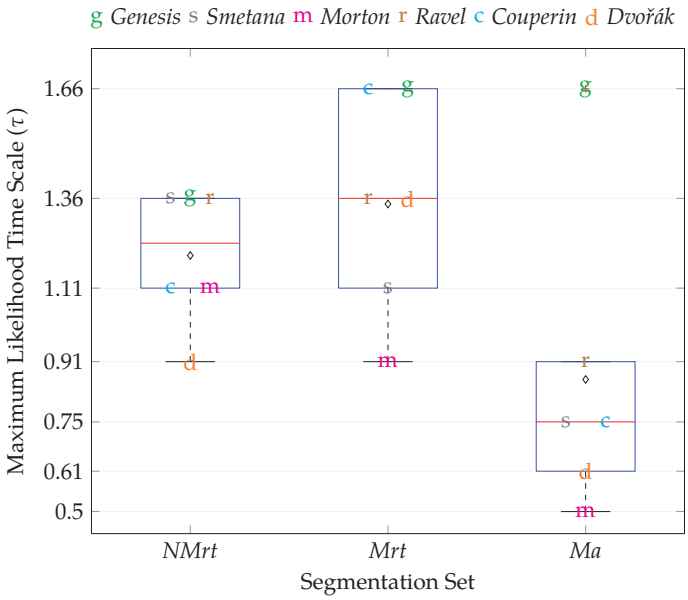

FIGURE 15 Box plot of maximum likelihood time scales for each segmentation set. Each time scale is represented with initials of its corresponding stimulus.

modelling perceptual segmentation could be measured in terms of rhythmic characteristics rather than in seconds; for instance, music with unclear pulse may optimally be modelled at larger time scales, probably because rhythmic cues for segmentation may become less salient for listeners.

5.2 Study II

– Hartmann, M., Lartillot, O., & Toiviainen, P. (in press). Interaction features for prediction of perceptual segmentation: effects of musicianship and experimental task. *Journal of New Music Research*.

5.2.1 Introduction

As music unfolds in time, listeners are able to mentally represent different aspects related to its structure, regardless of their level of musical expertise. A number of studies have proposed alternatives to model the perception of segment boundaries between structural sections, and found for example, that rhythmic and spectral changes in the music to predict listeners' indications (Jensen, 2007). However, the effects of musical expertise and experimental task on computational modelling of structure are not yet well understood. These issues need to be addressed to further understanding of how listeners perceive the structure of music and to improve automatic segmentation algorithms. This study investigated computational pre-

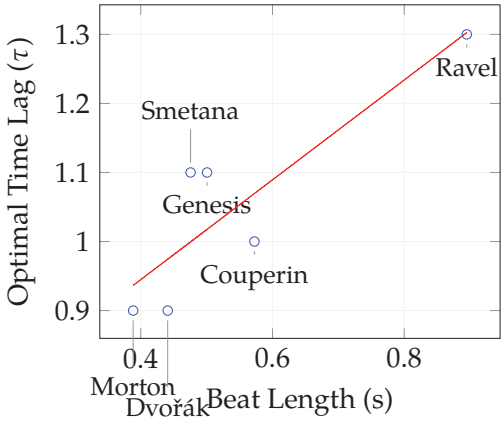

FIGURE 16 Optimal time lag for alignment between real-time and annotation tasks as a function of stimuli beat length (BL). Red line: simple linear regression equation $\tau = .72 \times BL + .66$.

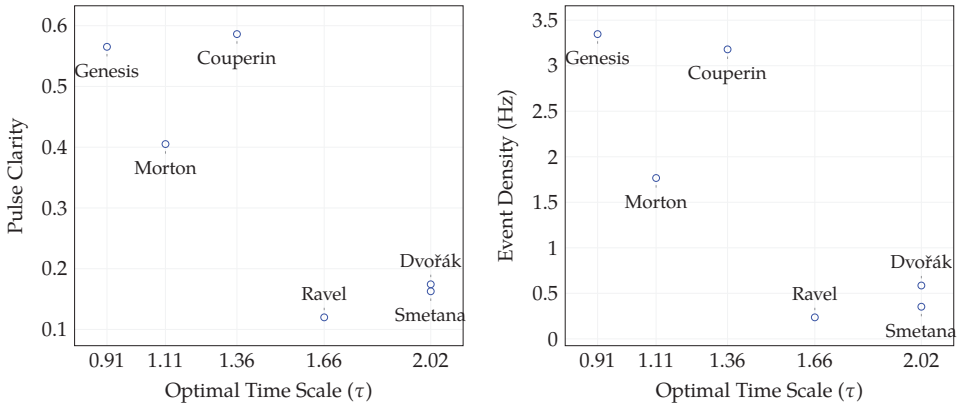

FIGURE 17 Relationship between global rhythmic descriptors and optimal time scales for correlation between perceptual density of real-time and annotation task.

diction of perceptual segmentation density via novelty detection (Foote, 2000). We attempted to shed light on the following research questions:

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

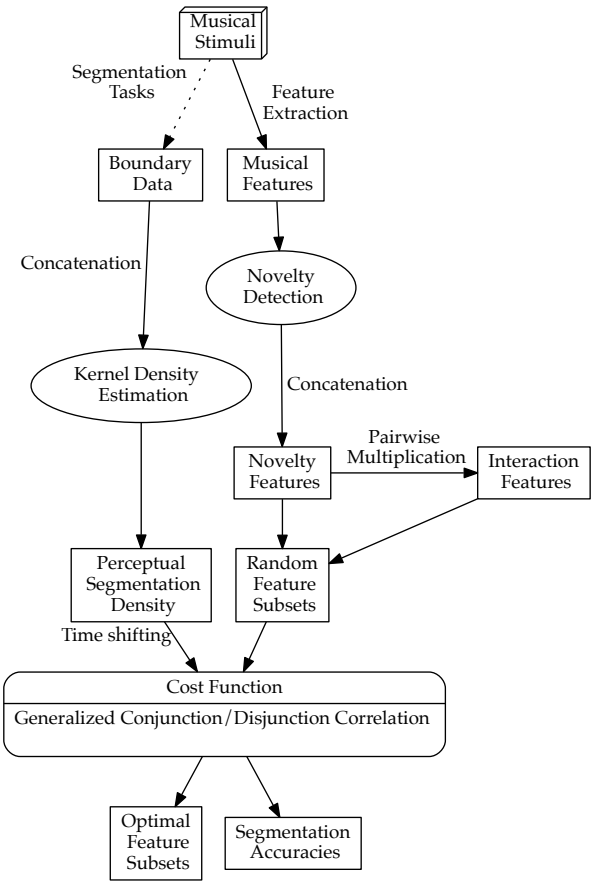

FIGURE 18 General design of Study II.

5.2.2 Methods

We focused on boundary indication data for six of the stimuli that were utilized in PI due to their similar duration; these responses were concatenated across musical stimuli. A perceptual segment boundary density curve was computed for each participant group and segmentation task; we chose a time scale parameter of 1.5 s following results from PI and other work (Befus, 2010; Bruderer, 2008). Subsequently, these density curves were computationally modelled via an approach based on novelty detection. First, for each stimulus, we extracted 5 frame-decomposed musical features describing timbre, rhythm, pitch class and tonal context. Novelty curves were computed from these features and concatenated across stimuli; interaction features were also derived from these curves via point-by-point multiplication between each pair of novelty features. We conducted various systematic explorations to find a novelty kernel width that would yield

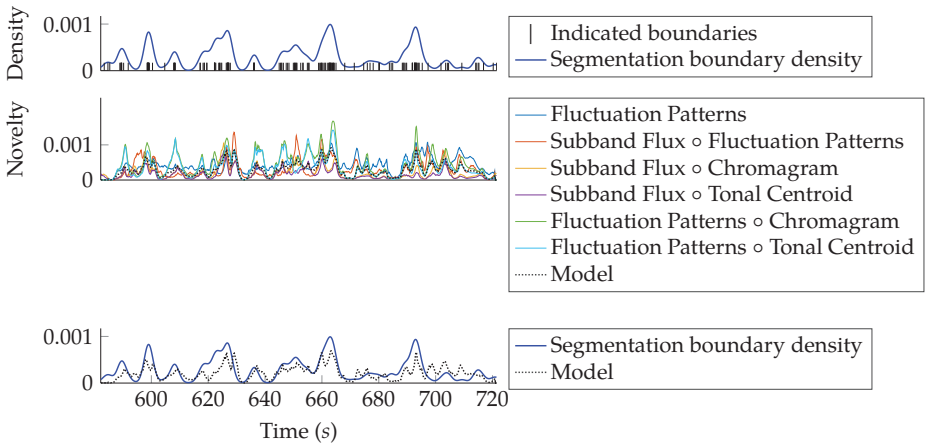

FIGURE 19 Perceptual segment boundary density and computational segmentation model for non-musicians in the Real-time task (stimulus *Dvořák*). Upper graph: Boundary indication data and segmentation boundary density. Middle graph: Model predictors and computational model prediction. Lower graph: Perceptual segmentation boundary density and computational model prediction. Model computed using a time lag of 1.7 s.

optimal prediction with respect to the perceptual segment boundary density curves. Combinatorial optimization was used to find subsets of novelty curves whose derived model prediction would optimally correlate with the perceptual segmentation density curves. Novelty feature subsets were aggregated by taking the 50th percentile (median ordinal position) across features. Figure 19 (middle) shows how the percentile model relates to the feature subset it is based on; since the number of feature subsets is even in this case, the model is the mean of the two middle feature values for each time point. This is a non-linear modelling approach because it assigns weights to features for each time point based on ranked values, so the contribution of each feature to the prediction varies over time. In particular, the median operator offers a 'majority judgement' since prediction values will be high if at least half of the features yield high values. We obtained models for 26 different novelty kernel widths. In addition, the segmentation data was time shifted with respect to novelty curves in order to study the effect of indication delay upon prediction rates.

5.2.3 Results

We found that musicians' segmentation yielded lower accuracy (Table 1), and involved a more varied set of features for prediction (key strength did not appear in *NMrt* model) and more feature interactions than non-musicians'. Prediction of the annotation task yielded higher rates than for the real-time task, which required time shifting of the segmentation data for optimal modelling (Figure 20); in fact, time shifting reverted the result, resulting in higher accuracy for the real-time task when compensated for delays. Also, this prediction involved more novelty

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

***$p < .001$

TABLE 1 Correlations between perceptual segmentation density and computational models' predictions obtained via percentile optimization. *P*-values adjusted for effective degrees of freedom.

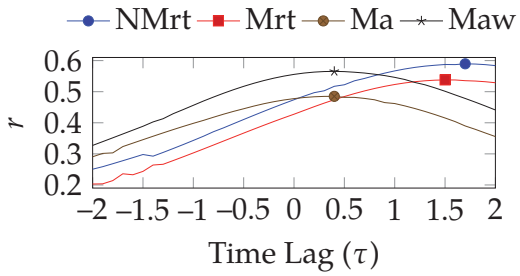

FIGURE 20 Correlation between perceptual segment boundary density and models' predictions obtained after time lags ranging from –2 s to 2 s, incremented by steps of 100 ms. Positive time lags refer to delay of novelty with respect to perceptual segment boundary density, and vice versa. Maximum points for each curve are highlighted with markers.

features, particularly rhythmic and tonal features (Table 1), than the prediction of the real-time task. Further, annotation task models that were weighted based upon boundary strength ratings yielded maximum accuracy, and involved more interaction features. Another finding was that the use of relatively large novelty kernel widths (half checkerboard kernel width of 11 s) was required to obtain novelty curves that would optimally predict perceptual segmentation density; similar kernel widths have been previously used to overcome high levels of detail in novelty curves (Pauwels et al., 2013; Liem, Bazzica, & Hanjalic, 2013; Klien, Grill, & Flexer, 2012). Finally, an interesting methodological result was that, for all segmentation sets, interaction novelty features yielded higher correlations with perceptual segment boundary density than basic novelty features. For instance, the highest correlation between a basic novelty feature and the density in the annotation task with added boundary strength weights was $r = .39$, $p < .001$ (fluctuation patterns), whereas in the case of interaction novelty features the highest correlation was $r = .49$, $p < .001$.

5.2.4 Conclusion

Comparing groups, musicians' segmentation may rely more on schematic knowledge, involve more dimensions of musical change and levels of the structural hierarchy, and result from faster musical structure processing. Regarding the tasks, real-time segmentation was associated with larger response delays and a focus on fewer musical dimensions than in the case of annotation segmentation. Also, the increase in accuracy found for perceptual density curves with added boundary strength weights was associated with higher emphasis given to acoustically stark musical changes, which are the ones that novelty curves would better predict. Models that include boundary strength weights might yield a clearer representation of a hierarchy of high-dimensional musical change, because adding strength weights increased the number of interaction features selected; it may be, for example, that rhythmic-tonal novelty could often be perceived as more perceptually salient than spectral-rhythmic novelty. Finally, we should highlight that both the proposed interaction features and the percentile optimization modelling approach yielded a correlation increase with respect to the use of basic novelty features only.

5.3 Study III

– Hartmann, M., Lartillot, O., & Toiviainen, P. (submitted). Musical Feature and Novelty Curve Characterizations as Predictors of Segmentation Accuracy.

5.3.1 Introduction

Novelty detection is a well-established method for analyzing the structure of music based on acoustic descriptors. Work on novelty-based segmentation has mainly concentrated on enhancement of these descriptors and similarity matrices derived therefrom (Paulus et al., 2010). Studies have also focused on improving novelty detection via alternative novelty kernels (Kaiser & Peeters, 2013) and peak detection (Gaudefroy et al., 2015) approaches. Less attention, however, has been paid to possible global characteristics of musical features and novelty curves that could determine segmentation accuracy. This is an important issue as it may help unearth acoustic cues prompting perceptual segmentation and find predictors of accuracy. This study focused on estimating segmentation accuracy from a characterization of the musical features themselves, as well as a characterization of the novelty curves. For perceptual density curves obtained from six individual musical examples via an annotation segmentation task, we investigated spectral, rhythmic and harmonic predictors of accuracy. Specifically, our aim was to understand whether local variability of musical features and distance between novelty peaks would be associated with segmentation accuracy. Two research questions guided this investigation:

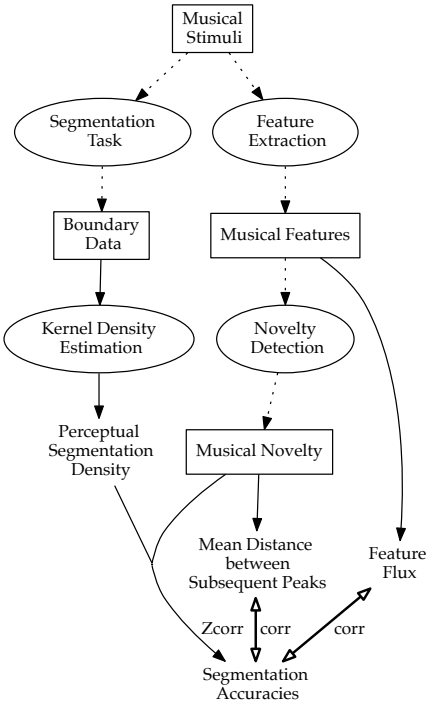

FIGURE 21 General design of Study III.

– What specific aspects of musical stimuli that account for segmentation accuracy can be directly described from musical features?
– What stimulus-specific attributes of novelty curves determine optimal segmentation accuracy?

5.3.2 Methods

For each stimulus, segmentation accuracies were obtained by correlating perceptual segment boundary density in the annotation task with 5 novelty curves describing timbre, rhythm, pitch class, and tonal context; the resulting coefficients were *z*-transformed for further comparisons. Two global characterizations derived from musical features and novelty curves were computed: amount of local variation (mean *Feature Flux*) in the musical feature, and mean distance between subsequent peaks (*MDSP*) in the novelty curve. Subsequently, we correlated across stimuli and for each musical feature the aforementioned estimates with the obtained segmentation accuracy. We then focused on the correlation ranks to better understand possible relationships between segmentation accuracy and feature estimates. Explorations of perceptual segment boundary density, musical features and novelty curves were conducted for each stimulus to offer an in-depth analysis

of the results.

5.3.3 Results

Segmentation accuracies varied both depending on stimulus and on type of feature used for novelty detection; no single novelty feature yielded maximum accuracies for all stimuli. Regarding the global characterizations computed (Table 2), we found

Musical Feature	**Feature Flux**	**MDSP**
Subband Flux	.54	–.03
Fluctuation Patterns	–.30	.11
Chromagram	–.68	.27
Key Strength	**–.74**	.47
Tonal Centroid	–.66	**.50**

TABLE 2 Correlation between z-transformed accuracy and characterizations of musical features (Feature Flux) and novelty curves (MDSP).

that, for rhythmic and tonal features, segmentation accuracies tended to increase for stimuli with milder local changes and fewer novelty peaks. Spectrum yielded a less clear trend: stimuli with high local variability for subband flux exhibit increased accuracy, and no relationship was found between MDSP and accuracy. Also, an exploration of musical and perceptual factors that may have contributed to accuracy showed that novelty detection accuracy may drop due to temporal imprecision of perceptual boundaries, and to knowledge-driven or other biases of attention from listeners. In some cases, they seem to focus on style specific, 'decorative' aspects of the music if other features change too often, for instance, in *Couperin* tonal centroid novelty peaks can detect the chord changes, but these changes occur very frequently, so listeners' boundary placements seem to be prompted by ornamentation (e.g. mordents) instead. Additionally, musical changes occurring in a single dimension may mask other prominent changes, e.g., musical changes triggering peaks for multiple novelty features may not prompt boundary perception if these occur in the middle of a ritardando (*Ravel*, 60 s) or a melodic phrase (*Morton*, 114 s, and *Dvorak*, 75 s). Further, large temporal gaps between endings and beginnings of melodies may cause a disjunction between actual and predicted segmentation, for instance in *Couperin* listeners tend to segment melodic endings probably because these are evoked by cadences (89 s), whereas novelty peaks tend to appear at beginnings due to clear local discontinuities (90 s).

5.3.4 Conclusion

Overall, the findings suggest that, at least for rhythmic and pitch-based features, stimulus-dependent characteristics derived from musical features and novelty curves can be utilized as predictors of segmentation accuracy. High segmentation accuracy seems often to be associated with novelty profiles that yield high scores for few temporal regions; that is to say, listeners might pay more attention to musical

dimensions that change significantly but not too frequently during segmentation. Regarding the results for spectrum, it is possible that high local variability was related to high accuracy because only a small fraction of the spectral changes achieved stark novelty peaks; these may have corresponded with actual boundaries. Another interpretation is that the results for spectrum were influenced by music with high local spectral change (due, e.g., to vast and diverse percussion) but few structural sections of long duration. In this respect, a larger sample size is needed to better understand how the possible relationships between characterizations and accuracy are affected by musical styles.

6 DISCUSSION

The perceptual organization processes that are involved in perception of musical structure are not well understood; for instance, the association between musical dimensions and listeners' interpretations of musical structure remains unclear. Understanding how listeners parse musical pieces and represent their structure can be useful in gaining deeper insights on the perceptual and cognitive aspects of listening that are intrinsic to music, and for the development of technological applications such as automatic structural segmentation systems. For instance, a systematic investigation of what types of acoustic changes in the music to which listeners would attend could be used to predict listeners' segmentation and help towards development of real-world applications, including smart music players (e.g., that could find guitar solo parts from an album or collection), editing aids for audio engineers, DJ and remixing applications, and interactive music visualizations. Another concrete benefit would be the potential to overcome, via audio music segmentation, the tape recorder paradigm (Tzanetakis & Cook, 1999) in musical playback, which currently makes it difficult to find interesting points of change in the music. Meaningful information related to the structure of musical pieces could be at listeners' disposal in the context of digital audio, just like modern musical notation offers a regular segmentation scheme with barlines providing reference points for easier score reading, or like the grooves of vinyl records can offer visual descriptions of structural changes in musical works.

One of the hypotheses that were formulated in this investigation was that the conducted experimental task does have an effect on perceptual segmentation. The results supported this assumption: real-time and annotation segmentation data yielded different number of boundaries (Figure 14), inter-subject correlations (Figure 12), temporal scales (Figure 15), and time lags. According to this, listeners' offline annotations of musical structure should not be compared with real-time responses to music listening, which in practice means that annotation segmentation data may not be comparable to descriptions related to physiological (skin conductance, heart rate variability, pupil dilation) or psychological (valence and arousal) determinants of emotional state, spontaneous movement to music, or brain responses. These descriptions are possible correlates of perceptual musical

structure and should be compared against real-time segmentation data (compensation for boundary indication delays may be required), but not with non real-time data such as "expert" annotations (Mungan et al., in press) and ground truths used in MIR structural analysis tasks. One may ask, at this point, whether or not a real-time segmentation task with multiple indication (and possibly also *listening only*) trials per participant would lead to segmentation density data that would highly correspond to the responses obtained in the annotation task. If this happened to be the case, it could be argued that real-time segmentations of unfamiliar pieces are snapshots of the gradual understanding of the structure of musical pieces. In this respect, a study of real-time segmentation of an acousmatic music piece reported that the similarity between subsequent trials neither increased monotonically nor reached its global maximum at the last trial (Mendoza Garay, 2014).

With regard to musical training, we assumed that there would be an effect on boundary perception: for instance, musicians would exhibit longer indication delays. This result, however, remains unclear because we observed high similarity between segmentations from musicians and non-musicians (Figure 13), few differences between groups regarding inter-subject correlations (Figure 12), and relatively similar mean number of boundary indications (Figure 14); for example, the optimal time lag for alignment between groups was negligible. In this respect, it is likely that the real-time task does not offer the chance to display higher level knowledge in listeners' responses; in other words, both musicians and non-musicians may have focused on the detection of local changes, which is a low level bottom-up segmentation process. In contrast, an annotation task would possibly have revealed differences between groups, as top-down processing may have been utilized, for instance, to remove inappropriate boundaries; in offline contexts, it may be that repeated listening and experience would have played a role in the placement of boundaries. Another possibility is that a direct comparison between perceptual segmentation density curves cannot reveal differences between groups, whereas more detailed analyses of boundary data for specific musical excerpts may be needed to discern these.

In contrast with the previous result, musical training was found to have an effect on model prediction rates and musical features involved in novelty-based segmentation models (Table 1), and on optimal lag of models (Figure 20). This suggests that differences between groups seem to be clearer after their segmentations are compared against computational models based on musical features: effects of musicianship upon segmentation may be observable if acoustic musical changes are taken as a reference point, but not if the segmentations are compared directly. If this was the case, then this would point to an influence of explicit schematic knowledge on musicians' segmentation. In this respect, non-musicians seemed to stumble more on local surface discontinuities, elicited for example by embellishments, instead of grouping short sequences of events together into larger sequences; this phenomenon was also observed by Mungan et al. (in press). This brings us back to results obtained via comparison between perceptual density curves, as differences in optimal time scales and number of indicated boundaries might be related with the model prediction rates found.

According to analyses regarding maximum likelihood time-scales for each group and stimulus (Figure 15), non-musicians seem to segment at lower time-scales than musicians. This finding corresponds to the observation that, altogether, non-musicians indicated nearly double the amount of boundaries than musicians (note that both groups are similar, however, regarding the mean number of boundaries indicated by each participant). The results from PI might explain why novelty curves yielded higher prediction rates for non-musicians than for musicians in PII (Table 1). Taking into account that novelty curves are highly sensitive to local discontinuities in the music (because the approach is "bottom-up"), lower prediction rates for musicians could imply that they are able to anticipate future changes that might be much more important than ongoing ones (possibly in some cases by relating to past changes or patterns).

We also found that the experimental task seems to have an impact on prediction and on features involved in segmentation modelling (Table 1). This implies that annotation segmentation data is not suitable for testing computational models tailored to real-time processes, such as segmentation and expectation models that focus on short-term memory or preceding musical context (Pearce, 2005; Lartillot, Cereghetti, Eliard, & Grandjean, 2013). It is noteworthy that, the result was inverted when the real-time task was compensated for delays (Figure 20), suggesting that the main assumption involved in novelty detection may apply to real-time perception of boundaries: stark discontinuities occurring between self-similar regions are optimal boundary candidates. This view is also supported by the observation that annotation task models with added strength weights yielded maximum accuracies (Table 1). The main difference between non real-time segmentation and real-time segmentation compensated for delays could be the indication of local discontinuities that are less prominent as they would serve secondary functions, exhibit lower intensity, or evoke changes in a shorter temporal scale; the contribution of these 'extra' indications to the annotation task models is diminished when boundary strength weights are included. Regarding the lag found between tasks, we highlight that two different approaches showed that real-time segmentation is delayed with respect to annotation segmentation by approximately 1 second. In PI, we used two-dimensional cross-correlation to compare multi-scale models from different tasks and found a mean optimal time lag for alignment of 1.05 s; in PII we obtained segmentation models for perceptual boundary density curves at different time shifts (Figure 20) and found optimal prediction rates based on a delay of about 0.5s for *Ma* and 1.5s for *Mrt* (from this we can assume that *Mrt* is delayed by about 1 second with respect to *Ma*, although different optimal models were obtained for each task).

We also highlight that the differences found between tasks with respect to boundary indications (Figure 14) and to time scales (Figure 15) are particularly interesting. The result is surprising because one would expect that listeners would segment less often as they get familiarized with the music, due to increased knowledge regarding goals and intentions (Zacks et al., 2007). There are several ways to interpret this finding. One of them is to point to actual differences between the characteristics of the conducted tasks: the real-time task may be more difficult

so listeners failed to indicate some of the boundaries, or perhaps in the annotation task they tried to indicate boundaries as often as possible so as to be able to assign different ratings of boundary strength. Another interpretation is to associate it with an increase of familiarity: as listeners apprehend the general structural aspects and get familiar with the music, they may discover other changes that may be deemed significant. In this respect, previous findings suggest that listeners tend to indicate more boundaries as they focus on shorter excerpts of the same stimulus (Krumhansl, 1996). We believe that, due to the effect of alignment between tasks, listeners did not fail to indicate very many boundaries in the real time task, and that since participants tended to indicate a large amount of low strength indications in the annotation task, both of the remaining explanations are possible. First, we cannot rule out the possibility that listeners were biased to include more indications of low strength due to the task instructions. Second, it is likely that higher familiarity was associated with more frequent segmentation. If this was the case, it might be that knowledge regarding future goals does not increase with familiarity, but can be better associated with schematic knowledge instead. Alternatively, it is possible that knowledge about goals leads to fewer indications at a given time scale, but also that increased familiarity with the music prompts representations that involve more levels of the structural hierarchy, resulting in a higher number of boundaries at multiple time scales.

Regarding the role of musical stimuli characteristics on segmentation modelling, we found that rhythmic and pitch-based changes seem to generally be relevant for modelling of segmentation (Table 1); timbre-based features, in contrast, may yield less clear descriptions regarding structure as they might be more sensitive to various types of acoustic changes (e.g. instrumentation, register, voicing, articulation, loudness). Furthermore, we found that accuracies obtained via novelty detection can be foreseen from extracted acoustic features: music characterized by low local discontinuity of a feature would yield high segmentation accuracy for that feature, at least in the case of rhythmic and pitch-based features (Table 2). From a perceptual viewpoint, the aforementioned finding seems to support the assertion that listeners tend to segment at time points that introduce stark discontinuity with respect to a rather homogeneous context; in other words, few rare changes in the music are more prominent perceptually. This relates to a result reported in PI: music with a clearer pulse is optimally modelled at shorter time scales (Figure 17). It could interpreted that if the pulsation in the music is unclear, this frequent lack of regularity may hinder the perception of strong boundaries. In other words, music with unclear pulse could be associated with the proposed theoretical constant $A + B \rightarrow C$ (Narmour, 1992), according to which expectation is not violated for repeated discontinuity (because discontinuity should occur in between homogeneous contexts to be perceived as a strong boundary). Similarly, music with high local discontinuity of a feature would yield low predictability for that feature as listeners may tend not to indicate too many boundaries for features that change often (e.g., for varying rhythmic pulsation).

We also found in PI that music with higher frequency of events tends to be modelled at shorter time scales (Figure 17). A plausible interpretation is that

music with very frequent events is often likely to exhibit more local discontinuity, so listeners will tend to indicate boundaries at shorter time scales, and these boundaries will tend to be relatively less salient. We additionally found a positive relationship between beat length and the optimal time lag for alignment between real-time and annotation tasks (Figure 16); this result suggests that the indication delay could be understood as the summation of a delay dependent on stimulus beat length (recognition delay) and a constant time lag among stimuli (response delay). The nonzero intercept of the regression line indicates a constant time lag, suggesting an identical response delay for all stimuli: listeners may have required over half a second to respond to the recognized change by indicating a boundary. The recognition of a perceived change as significant probably varied depending on the stimuli beat length: listeners possibly required less than a beat (between 0.4 s and 0.9 s, depending on the stimulus) to pass in order to recognize a musical change as significant.

6.1 Summary of contributions to music segmentation

- We utilized an interdisciplinary approach, lying in between the fields of music perception and cognition and music information retrieval, that relied on both statistical analysis and on a careful exploration of the musical stimuli and indicated boundaries.
- Our work examined various materials involved in the audio-based modelling and prediction of segmentation: boundary data from multiple listeners, its perceptual modelling using kernel density estimation, computational modelling of segmentation density via novelty detection, and characteristics of musical features and novelty curves with regards to this modelling.
- For the first time, two boundary indication tasks that are commonly used for segmentation were experimentally compared; a deep understanding of the differences between different segmentation tasks is key to gain new insights about the representation of musical structure by listeners.
- The issue of temporal scales was systematically studied for the first time in the context of perceptual segmentation, addressing the problem of temporal scales which is an important contribution to the study of similarity in music segmentation.
- We contributed to the understanding of a perceptual interpretation regarding novelty detection, since we examined, in the light of listeners' responses, the contribution of different musical features to novelty-based prediction of structure and the relationship between acoustic characteristics of the stimuli and segmentation accuracy.
- We proposed the generation of *interaction novelty features* for segmentation, which yielded an increase of accuracy with respect to basic novelty features.
- We suggested an optimization modelling approach that exhaustively looks for optimally performing sets of features for segmentation, allowing a better

understanding of the relative contribution of different musical dimensions and the effect of time lag upon segmentation modelling.

6.2 Main findings and implications

Regarding the role of musicianship in perceptual segmentation, we found high similarity between segmentation densities from musicians and non-musicians, as well as similar inter-subject correlations and high mean alignment between their segmentation densities. We did not find clear differences neither in mean number of boundary indications, nor in time scales of the segmentation. We observed, however, that altogether musicians segmented less and at slightly higher time scales than non-musicians. In regards computational modelling of segmentation, we found that musicians' segmentation yielded lower accuracies, and involved more musical features and less time shifting of data for optimal prediction. In sum, we lack sufficient evidence for an effect of musicianship on perceptual segmentation.

Examining the influence of experimental tasks upon perceptual segmentation, we found differences regarding number of boundary indications, inter-subject correlations, time scale of the segmentation, and a time lag between tasks. Furthermore, in computational modelling analyses the annotation task density yielded higher accuracies, required smaller time shifting for optimal prediction, and involved more features for modelling than the real-time task. However, data time shifting led to higher accuracy for the real-time task than for the annotation task. An associated finding was that listeners' ratings of perceived boundary strength in the annotation task were not related to segmentation density at boundary time points, and that annotation task models that were weighted based on these ratings yielded maximum segmentation accuracy. These findings support the view that the experimental task used (real-time vs. non real-time) has an effect on perceptual segmentation.

With regard to musical features, the optimal time scale for comparison between segmentation tasks was found to be associated with rhythm characteristics of the music, as it increased for stimuli with lower pulse clarity or event density, and vice versa. Also, we found that the optimal time lag for comparison between segmentation tasks was shorter for stimuli with shorter beat length. We observed that simultaneous changes in rhythm and pitch-based features contributed to the prediction of both groups and tasks. In addition, for pitch-based and rhythmic features, we found that music with lower local discontinuity and larger distance between subsequent novelty peaks for a given feature yielded higher segmentation accuracy for that feature. According to these findings, rhythmic and pitch-related aspects of the stimuli are relevant for computational modelling of segmentation and can be used as predictors of segmentation accuracy; rhythmic characteristics also contribute to time scale and time lag of boundary indications.

Results regarding segmentation prediction and musical training suggest that musicians' segmentation would involve more schematic knowledge, attention

paid to a higher number of dimensions of musical change and more levels of the structural hierarchy, and higher speed of musical structure processing. However, it is also possible that both groups of participants focus generally on similar instants of change, since the examination of boundary data and perceptual segmentation density derived therefrom did not provide evidence supporting the hypothesis of an effect of musical training.

In contrast, our analyses suggested a clear effect of experimental task upon segmentation; this finding implies that different experimental paradigms can lead to clear differences (i.e., beyond occasional "missed" or delayed indications) regarding segmentation responses, such as segmentation at different time scales and systematic time lags. Furthermore, according to the results obtained via computational modelling, real-time segmentation may be associated with higher response delays and attention to fewer dimensions of musical change than annotation segmentation; compensation for delays revealed interesting similarities between novelty detection and real-time perception. Further, our results imply that weighting of the density in the annotation task is associated with more emphasis towards stark musical changes and clearer representation of a hierarchy of high-dimensional musical change.

Regarding musical features, the results imply that the time scale of the segmentation can be measured in number of events rather than in seconds. The findings also suggest that listeners pay attention to interactions between musical dimensions, particularly between rhythm and pitch-based changes; these may be associated with higher levels of the structural hierarchy; listeners may also focus on musical dimensions that change less frequently. In addition, the suitability of novelty curves for prediction of a particular stimulus can be directly estimated from acoustic features, that is, without the requirement of a novelty detection step, which is much more efficient taking into account the computational cost of novelty detection.

6.3 Methodological considerations

One of the problems of our investigation is the lack of a set of indications from non-musicians in the annotation task (*NMa*), which creates difficulties for the analysis of some of our results. Based on the obtained results, one could expect that a *NMaw* set would yield maximum prediction rates, as it would offer a precise representation of the most salient changes from non-musicians, which seem to be more accurately predicted than musicians. Unfortunately, we cannot test this hypothesis with the current material.

It is also important to mention an issue regarding the multi-scale approach conducted to derive a representation of musical structure from listeners' indications. The different levels of representation obtained by multi-scale models can be used to represent perceptual musical structure with regards to a given temporal context, but they differ in their precision regarding the location of musical change. This

leads to a representation of structure that fails to inform the exact location of structural boundaries for large temporal scales. While smoothing is beneficial to aggregate indications by different listeners that refer to the same perceived instant of change, it is problematic for localization of hierarchically superior boundaries. Since boundaries yield very broad density peaks for larger kernel bandwidths, it is difficult to locate them in the music. This is because a large time scale parameter for Kernel Density Estimation reduces the bias of the estimation, but increases its variance, which results in higher similarity in density between neighboring regions. To illustrate this, at large time scales it becomes difficult to estimate whether listeners' indications of a given musical change are relatively isochronous.

Regarding the analysis of time shifting the perceptual segment density to compute different computational models of the segmentation (Figure 20), we remark that the segmentation sets were compared against different models. For each segmentation set, this approach finds the time lag that would yield the best performance for prediction, but it is not the optimal way to compare segmentation sets with each other. Bruderer (2008) computed the cross-correlation between perceptual segmentation density of participants and perceptual segmentation densities of symbolic musical features. This allowed examination of whether all musical features were either advanced or delayed with respect to the perceptual segmentation density. Bruderer (2008) did not find a trend in this regard. In this respect, it may be challenging to find an appropriate reference point for assessment of time lags in perceptual segmentation.

In regards to the lack of relationship found between boundary strength and boundary density, this result can be seen as counterintuitive and it conflicts with the study by Bruderer (2008). We should highlight, however, that the boundary density corresponding to a given indication time point does not necessarily correspond with the peak density for that boundary; for instance, the density for listeners that placed a boundary earlier or later than the rest of the listeners may be relatively low regardless of their indicated strength. In this respect, comparisons of strength ratings with boundary density peaks would perhaps be more appropriate to investigate the relationship between boundary density and indications of boundary strength.

Another issue is that other musical dimensions that may contribute to segmentation were not covered; for instance, the role of loudness has not been investigated. Novelty curves based on acoustic features often describe contrast between absence and presence of events, which is an aspect that is associated to loudness. Further, subband flux is sensitive to loudness because it describes spectral magnitude changes over time. Loudness should be disentangled as much as possible from other musical dimensions to study its role on perceptual segmentation, although this is not an easy task, as many musical features are derived from spectral energy. We should also add at this point that novelty detection is designed for modelling changes in the spectrogram, so it is suitable for multidimensional features but not for unidimensional ones such as loudness.

An associated problem is about the validity of the extracted musical features with regards to representation of musical dimensions. In particular, rhythmic and

key descriptors may yield unreliable information for certain stimuli. Also, different features may not be entirely independent from each other. This collinearity issue is problematic because it can hinder evaluation of the contribution of each feature to the obtained models. For instance, if the obtained optimal subset is almost equally optimal to other possible subsets, then slightly different segmentation density curves could have a great effect on the feature subsets selected by the models.

We also remark that the number of examples used in this work is insufficient to generalize our results regarding musical features and automatic segmentation. Peaks selected from the obtained optimal models for segmentation, for instance, should be further tested against existing datasets such as SALAMI (Smith et al., 2011). Another possibility in this respect would be to also evaluate other structural segmentation approaches on the stimuli used in our work to examine the reliability of the findings.

Finally, we should also mention that our investigation could have benefited more from more music-theoretical perspectives, especially with regard the analysis of specific musical excerpts and the contribution of higher-level components of segmentation, such as musical parallelism. It is useful to remark, however, that interpretations of the results in finer detail can get rather convoluted when dealing with real-world, polyphonic audio stimuli, when compared to, for instance, sequences of notes.

6.4 Future directions

To better understand the observed differences between tasks, future studies could empirically examine other types of real-time and non real-time tasks. For instance, it may be useful to investigate multiple real-time segmentation trials for the same piece, as well as to record reposition of boundaries in annotation segmentation tasks. Also, it would be interesting to use a between-group design in order to avoid too much familiarity with the stimuli in the annotation task; this might help to better explain the effect of familiarity upon the differences between tasks. Regarding indication delays, the real-time task setting could involve asking listeners to indicate instants of significant change as soon as they are noticed, to make sure that they react as fast as possible to the perceived changes.

It should also be noted that other possible solutions can be applied regarding kernel density estimation; notably, it has been found that adaptive kernel density estimation may yield better results for large samples (Shimazaki & Shinomoto, 2010). In addition, rhythmic characteristics of the stimuli seem to allow for estimation of required time shifts and segmentation time scales for perceptual boundary data modelling of a given stimulus; this possibility could be further explored. Also, regarding novelty detection, and considering that a fixed common time scale should not be appropriate for all stimuli, multi-granular approaches to novelty (Lartillot et al., 2013) could be investigated for prediction of segmentation density. In addition, we believe that interaction features may be useful for further

investigation: it might be interesting to generate interactions between more than two features to help reduce the number of novelty peaks, potentially leading to fewer “false positives”.

7 CONCLUSIONS

The study of music segmentation can offer insights into the way acoustic events in music are organized and understood by listeners as meaningful content involving multiple time granularities. Drawing inspiration from unresolved issues in previous perceptual segmentation studies, we investigated the role of musical training and experimental task on music segmentation. We also focused on the problem of computational modelling of perceptual segmentation as a way to elucidate possible differences between participant groups and segmentation tasks. We were interested in pinpointing the relative contribution of different musical features upon prediction, and understanding the possible relationship between segmentation accuracy of a musical example and characteristics of musical features used for prediction. Our methodological framework for the study of modelling of music segmentation tried to maximize external validity by using a relatively large participant size, and sought to satisfy aspects of ecological validity by including real-world musical stimuli. We also aimed to strengthen the internal validity of the approach: we carefully considered the temporal scales used in the analysis of the magnitude of musical changes, and took into account the possibility of indication delays in perceptual segmentation.

Our findings suggest that a number of aspects related to segmentation depend on the setting used to collect boundary indication responses; in real-time segmentation, listeners indicate less boundaries and at larger time scales, are less consistent with each other, and exhibit higher time lags than in non real-time segmentation. We also observed differences for segmentation prediction based on audio-based descriptors of musical change. Real-time segmentation involved models of lower dimensionality and yielded lower accuracy than non real-time segmentation; compensation of real-time data for indication delays inverted this last result. Furthermore, boundary strength ratings from participants were shown to improve segmentation accuracies when included in non real-time segmentation models, and did not correspond with likelihood of boundary indication based on perceptual data.

The role of musicianship in segmentation remains unclear: we found high similarity between perceptual segmentation models, similar inter-subject correlation

for both groups, similar number of boundary indications, and no trend regarding possible delays between one group and another. However, non-musicians tended to segment at shorter time scales, and optimal computational prediction models for this group yielded higher segmentation accuracy, involved a less varied set of optimal features, and required larger time shifting than for musicians.

We observed interesting results regarding the contribution of musical features on perceptual segmentation and its prediction. According to our findings, features describing simultaneous changes in rhythm and pitch were often involved in optimal computational prediction models. In addition, for features describing rhythm, pitch, and tonality, musical pieces exhibiting lower local fluctuation for a given feature yielded higher prediction accuracy for that feature; a similar trend was found regarding the mean distance between peaks of musical novelty detected from these features. Moreover, we found that the global beat length of the music is associated with differences in optimal time lag between segmentation tasks: the higher the beat length (i.e., the slower the tempo), the higher the lag, and vice versa. Also, we observed a relationship between rhythmic characteristics of stimuli and optimal time scales for modelling of segmentation responses: music with higher pulse clarity and higher frequency of onset events involved shorter optimal time scales for segmentation.

Overall, our work sheds light on the relevance of audio-based musical features for prediction of perceptual segmentation density obtained via aggregation of multiple listeners' responses, as well as other aspects of segmentation derived therefrom. The results provide evidence regarding the perceptual validity of rhythmic and pitch-based audio musical descriptors for prediction of segmentation. We also found that the accuracy of segmentation prediction can be traced by characteristics of acoustic musical features. This finding can be applied to the optimization of automatic segmentation techniques, as it might be possible to predetermine what musical features would yield higher accuracy for a given stimulus in order to implement appropriate segmentation strategies.

Finally, our investigation represents a new step in the study of music segmentation by combining music perception and cognition with state-of-the-art methods in music information retrieval. This interdisciplinary approach incorporated multiple viewpoints, as we examined listeners' responses, different segmentation tasks, musical stimuli and audio-based musical features. Approaches incorporating various strategies to approach segmentation can help to gain a deeper understanding of how the structure of music is perceived and how can it be predicted.

BIBLIOGRAPHY

Agres, K. & Wiggins, G. A. (2015). Schematic processing as a framework for learning and creativity in CBR and CC. In *Workshop proceedings from the twenty-third international conference on case-based reasoning*. Frankfurt: CEUR.

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

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

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*(4), 3677–3689.

Aucouturier, J.-J. & Sandler, M. (2001). Segmentation of musical signals using hidden markov models. In *Audio Engineering Society Convention 110*. Audio Engineering Society; 1999.

Ayari, M. & McAdams, S. (2003). Aural analysis of arabic improvised instrumental music (taqsıım). *Music Perception: An Interdisciplinary Journal*, *21*(2), 159–216.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child development*, *72*(3), 708–717.

Barbedo, J. & Lopes, A. (2007). Automatic genre classification of musical signals. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 157–157.

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

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047.

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

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

Bod, R. (2002a). A general parsing model for music and language. In *Music and artificial intelligence* (pp. 5–17). Springer.

Bod, R. (2002b). A unified model of structural organization in language and music. *Journal of Artificial Intelligence Research*, (17), 289–308.

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

Bohak, C. & Marolt, M. (2016). Probabilistic segmentation of folk music recordings. *Mathematical Problems in Engineering*, *2016*.

Bregman, A. S. (1994). *Auditory Scene Analysis: the perceptual organization of sound* (Bradford, Ed.). MIT Press.

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

Bruderer, M., McKinney, M., & Kohlrausch, A. (2006). Perception of structural boundaries in popular music. In *Proc. of the 9th International Conference on Music Perception and Cognition* (1983, pp. 157–162). Bologna.

Cambouropoulos, E. (1998). *Towards a general computational theory of musical structure* (Doctoral dissertation, University of Edinburgh).

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference* (pp. 17–22).

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation. *Music Perception*, *23*(3), 249–268. Retrieved from http://www.jstor.org/stable/10.1525/mp.2006.23.3.249

Cambouropoulos, E. (2010). The musical surface: challenging basic assumptions. *Musicae Scientiae*, *14*(2 suppl), 131–147.

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

Chew, E. (2002). The spiral array: an algorithm for determining key boundaries. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and artificial intelligence* (pp. 18–31). Edinburgh: Springer.

Clarke, E. & Krumhansl, C. L. (1990). Perceiving musical time. *Music Perception*, *7*(3), 213–251.

Conklin, D. & Anagnostopoulou, C. (2001). Representation and discovery of multiple viewpoint patterns. In *Proceedings of the international computer music conference* (pp. 479–485). Citeseer.

Dauwels, J., Vialatte, F., Weber, T., & Cichocki, A. (2009). On similarity measures for spike trains. In *Advances in Neuro-Information Processing* (pp. 177–185). Springer.

De Lisa, G. (2009). Smetana: má vlast. Retrieved from http://genedelisa.com/2009/03/smetana-ma-vlast/

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. Retrieved from http://dx.doi.org/10.1080/09298215.2013.859710

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. (1989). A perceptual approach to contemporary musical forms. *Contemporary Music Review*, *4*(1), 213–230.

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

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.

Deutsch, D. (1999). Grouping mechanisms in music. *The psychology of music*, *28*.

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

Duin, R. P. W. (1976). On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Transactions on Computers*.

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

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

Eiben, A. E. & Smith, J. E. (2003). *Introduction to evolutionary computing*. Springer.

Ellis, D. P. (1996). *Prediction-driven computational auditory scene analysis* (Doctoral dissertation, Massachusetts Institute of Technology).

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

Fastl, H. (1982). Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise. *Hearing Research*, *8*(1), 59–69.

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

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

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

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

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

Gaudefroy, C., Papadopoulos, H., & Kowalski, M. (2015). A multi-dimensional meter-adaptive method for automatic segmentation of music. In *13th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6). IEEE.

Gómez, E. (2006a). *Tonal description of music audio signals* (Doctoral dissertation, Universitat Pompeu Fabra).

Gómez, E. (2006b). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing*, *18*(3), 294–304.

Gómez, E. & Bonada, J. (2005). Tonality visualization of polyphonic audio. In *Proceedings of international computer music conference*. Citeseer.

Goto, M. (2003). A chorus-section detecting method for musical audio signals. In *Proceedings of the 2003 IEEE international conference on acoustics, speech, and signal processing* (Vol. 5, pp. V–437). IEEE.

Handel, S. (2006). *Perceptual coherence: hearing and seeing*. Oxford University Press, USA.

Hansen, N. C. & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in psychology*, *5*, 1052.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: building event schemas. *Memory & cognition*, *34*(6), 1221–1235.

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

Hartmann, M., Lartillot, O., & Toiviainen, P. (2015). Effects of musicianship and experimental task on perceptual segmentation. In J. Ginsborg, A. Lamont, M. Philips, & S. Bramley (Eds.), *Proceedings of the ninth triennal conference of the european society for the cognitive sciences of music*. Manchester.

Hartmann, M., Saari, P., Toiviainen, P., & Lartillot, O. (2013). Comparing timbre-based features for musical genre classification. In *Proceedings of the sound and music computing conference 2013, SMC 2013*. Stockholm, Sweden.

Jensen, K. (2007). Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 159–159.

Johnson, E. K. & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567. doi:10.1006/jmla.2000.2755

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

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

Klien, V., Grill, T., & Flexer, A. (2012). On automated annotation of acousmatic music. *Journal of New Music Research*, *41*(2), 153–173.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace and Company.

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*(3), 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 *12th International Conference on Music Perception and Cognition*. Thessaloniki, Greece.

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

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

Krumhansl, C. L. (2001). *Cognitive foundations of musical pitch*. Oxford University Press.

Kurby, C. A. & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in cognitive sciences*, *12*(2), 72–79.

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. (2011). Cultural impact in listeners' structural understanding of a tunisian traditional modal improvisation, studied with the help of computational models. *Journal of interdisciplinary music studies*, *5*(1), 85–100.

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

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: design, validation and optimization. In *Proceedings of the 9th International Conference on Music Information Retrieval* (pp. 521–526). Citeseer.

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

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

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

Levy, M. & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *Audio, Speech, and Language Processing, IEEE Transactions on*, *16*(2), 318–326.

Liem, C. C. S., Bazzica, A., & Hanjalic, A. (2013). Looking beyond sound: unsupervised analysis of musician videos. In IEEE (Ed.), *Proceedings of the 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*. Paris, France.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for music modeling. In *International Symposium on Music Information Retrieval* (Vol. 28, p. 5). Citeseer.

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

Maddage, N. C., Xu, C., Kankanhalli, M. S., & Shao, X. (2004). Content-based music structure analysis with applications to music semantics understanding. In *Proceedings of the 12th annual ACM international conference on Multimedia* (pp. 112–119). MULTIMEDIA '04. New York, NY, USA: ACM.

Margulis, E. H. (2005). A model of melodic expectation. *Music Perception*, *22*(4), 663–714.

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

Mauch, M., Noland, K., & Dixon, S. (2009). Using musical structure to enhance automatic chord transcription. In *Proceedings of the 10th international society for music information retrieval conference* (pp. 231–236).

McFee, B. & Ellis, D. P. (2014). Learning to segment songs with ordinal linear discriminant analysis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5197–5201).

Mendoza Garay, J. (2014). *Self-report measurement of segmentation, mimesis and perceived emotions in acousmatic electroacoustic music* (Master's thesis, University of Jyväskylä).

Meredith, D. (2015). Music analysis and point-set compression. *Journal of New Music Research*, *44*(3), 245–270.

Meredith, D., Lemström, K., & Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, *31*(4), 321–345.

Müller, M., Chew, E., & Bello, J. P. (2016). Computational Music Structure Analysis (Dagstuhl Seminar 16092). *Dagstuhl Reports*, *6*(2), 147–190. doi:http://dx.doi.org/10.4230/DagRep.6.2.147

Müller, M. & Kurth, F. (2006). Towards structural analysis of audio recordings in the presence of musical variations. *EURASIP Journal on Advances in Signal Processing*, *2007*(1), 1–18.

Mungan, E., Yazıcı, F., & Kaya, M. U. (in press). Perceiving boundaries in unfamiliar turkish makam music: evidence for Gestalt universals? *Music Perception*.

Narmour, E. (1992). *The analysis and cognition of melodic complexity: the implication-realization model*. University of Chicago Press.

Newtson, D. & Engquist, G. (1976). The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, *12*(5), 436–450.

Nieto, O. (2015). *Discovering structure in music: automatic approaches and perceptual evaluations* (Doctoral dissertation, New York University).

Nieto, O., Farbood, M. M., Jehan, T., & Bello, J. P. (2014). Perceptual analysis of the F-measure for evaluating section boundaries in music. In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval*.

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

Parncutt, R. (1998). Listening to music in the real world? a critical discussion of Marc Leman's music and scheme theory: cognitive foundations of systematic musicology. *Journal of New Music Research*, *27*(4), 380–408.

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

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

Pauwels, J., Kaiser, F., & Peeters, G. (2013). Combining harmony-based and novelty-based approaches for structural segmentation. In *Ismir* (pp. 601–606).

Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (Doctoral dissertation).

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.

Pearce, M. T. & Wiggins, G. (2006). The information dynamics of melodic boundary detection. In *Proceedings of the ninth international conference on music perception and cognition* (pp. 860–865).

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

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

Peeters, G. & Deruty, E. (2009). Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proc. of 3rd workshop on learning the semantics of audio signals* (pp. 75–90).

Peiszer, E., Lidy, T., & Rauber, A. (2008). Automatic audio segmentation: segment boundary and structure detection in popular music. In *Proceedings of the 2nd International Workshop on Learning the Semantics of Audio Signals (LSAS)*. Paris, France.

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

Poli, G. D., Rodà, A., & Vidolin, A. (1998). Note–by–note analysis of the influence of expressive intentions and musical structure in violin performance. *Journal of New Music Research*, *27*(3), 293–321. doi:10.1080/09298219808570750

Popov, A. (2005). Genetic algorithms for optimization. *User Manual, Hamburg*, *2013*.

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

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rafailidis, D., Nanopoulos, A., Manolopoulos, Y., & Cambouropoulos, E. (2008). Detection of stream segments in symbolic musical data. In *Proceedings of the 9th international conference on music information retrieval* (pp. 83–88).

Rosner, B. & Meyer, L. (1982). Melodic processes and the perception of music. In *Psychology of music*. Academic Press.

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.

Schellenberg, E. G. (1997). Simplifying the implication-realization model of melodic expectancy. *Music Perception: An Interdisciplinary Journal*, *14*(3), 295–318.

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

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language*, *57*(1), 24–48. doi:10.1016/j.jml.2006.10.004

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

Shimazaki, H. & Shinomoto, S. (2010). Kernel bandwidth optimization in spike rate estimation. *Journal of Computational Neuroscience*, *29*(1), 171–182. doi:10.1007/s10827-009-0180-4

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

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

Sonntag, C. M. (2011). *Jeux d'eau and its colleagues: water and artistic expression at the turn of the 20th century* (Master's thesis, Ball State University).

Šupka, O. (2013). Slavonic dances for orchestra, series I. Retrieved from http://www.antonin-dvorak.cz/en/slavonic-dances1-for-orchestra

Temperley, D. (2001). The cognition of basic musical structures. (Chap. 3). London: MIT press.

Tenney, J. & Polansky, L. (1980). Temporal Gestalt perception in music. *Journal of Music Theory*, *24*(2), 205–241.

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

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

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

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

Trythall, R. (2002, Summer). "Jelly roll blues", observations on performance practice. Retrieved from http://www.richardtrythall.com/33.html

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

Tzanetakis, G. & Cook, P. (1999). Multifeature audio segmentation for browsing and annotation. In *Ieee workshop on applications of signal processing to audio and acoustics* (pp. 103–106). IEEE.

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.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological bulletin*, *133*(2), 273.

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

YHTEENVETO (FINNISH SUMMARY)

Musiikkia kuunnellessa havaitsemme moninaisia äänellisiä tapahtumia. Pystymme esimerkiksi populaarimusiikkikappaleesta erottelemaan säkeen ja kertosäkeen, sekä kykenemme paikallistamaan milloin kappale siirtyy osasta toiseen. Tällaiset tapahtumat ilmenevät monella tasolla; musiikissa voi erottaa esimerkiksi motiiveja, fraaseja ja osia. Tässä työssä käsitellään musiikin rakenteen muutoksia, eli *merkittävän muutoksen hetkiä*.

Useissa musiikin havaitsemista ja kognitiota käsittelevissä tutkimuksissa on pyritty ymmärtämään minkä tyyppiset muutokset musiikkikappaleessa muodostavat kappaleessa havaitun rakenteen. Edelleen ratkaisematta on kuitenkin, miten musiikillinen harjautuneisuus, rakenteen reaaliaikainen ja ei-reaaliaikainen jaottelu sekä eri musiikillisten elementtien suhteellinen merkittävyys vaikuttavat rakenteen havaitsemiseen ja ennustamiseen. Näiden kysymysten tarkastelu voi johtaa parempaan ymmärrykseen niistä mekanismeista, jotka ovat käytössä eri ihmisillä eri tilanteissa, sekä antaa tietoa musiikin havaitun rakenteen ja akustisten piirteiden välisistä yhteyksistä.

Työssä järjetettiin kaksi kuuntelukoetta, joissa koehenkilöiden tehtävänä oli kuunnella musiikkikappaleita ja segmentoida ne kahdessa havaitsemistehtävässä: reaaliaikaisesti ilman kuuntelun keskeytystä sekä ei-reaaliaikaisesti vapaassa tahdissa. 18 muusikkoa osallistui kumpaankin tehtävään, minkä lisäksi 18 ei-muusikkoa osallistui reaaliaikaiseen tehtävään. Tulosten analysoinnissa käytettiin ydinestimointimenetelmää, jolla yhdistettiin koehenkilöiden vastaukset kappaleessa havaittujen muutosten tiheyttä kuvaavaksi ajalliseksi käyräksi, sekä akustisiin piirteisiin pohjautuvaa muutoksen havaitsemismenetelmää.

Tutkimuksen tulokset osoittivat segmentointitehtävätyypin merkityksen rakenteen erottelussa ja ennustamisessa, sekä selvensivät yleisten ja paikallisten musiikillisten elementtien merkittävyyttä tässä prosessissa. Tulokset eivät kuitenkaan selittäneet tutkimuskirjallisuudessa esiintyviä musiikilliseen harjautuneisuuteen liittyviä eroja. Tutkimus lisäksi korostaa musiikin homogeenisten osien väleissä esiintyvien paikallisten muutosten merkitystä musiikin rakenteellisten muutosten havaitsemisessa, kuulijoiden merkintäviiveiden vaikutusta rakenteen segmentoinnissa, sekä segmentoinnin aikaskaalan määrittelystä johtuvia ongelmia mallinnuksessa.

Avainsanat: musiikin rakenne, ydinestimointimenetelmä, muutoksen havaitsemismenetelmä, musiikilliset piirteet, musikillinen harjautuneisuus, segmentoinnin havaitsemistehtävä.

APPENDIX 1 FUNCTIONS USED IN THE STUDIES

Some of the MATLAB functions written for the analyses described in this thesis are available in YouSource: https://yousource.it.jyu.fi/dissertation/appendix

The repository can be cloned using git:

```
git clone git://yousource.it.jyu.fi/dissertation/appendix.git
```

- Paper I
  - `segread.m`: reads in boundary indication data from listeners for further processing
  - `segdensity.m`: computes a normalized density curve of boundary indications
- Paper II
  - `segnovelty.m`: computes a set of novelty features from a musical file
  - `seginteract.m`: creates interactions between novelty features via pairwise multiplication
  - `segcorrelation.m`: correlation between density curve and normalized novelty features
  - `segoptimize.m`: percentile-based model of the density curve based on computed novelty curves
- Paper III
  - `segfeatures.m`: computes a set of frame-decomposed features from a musical file
  - `segcharact.m`: computes global characterizations derived from musical features and novelty curves

ORIGINAL PAPERS

PI

MULTI-SCALE MODELLING OF SEGMENTATION: EFFECT OF MUSICAL TRAINING AND EXPERIMENTAL TASK

by

Martín Hartmann, Olivier Lartillot, and Petri Toiviainen 2016

Music Perception

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

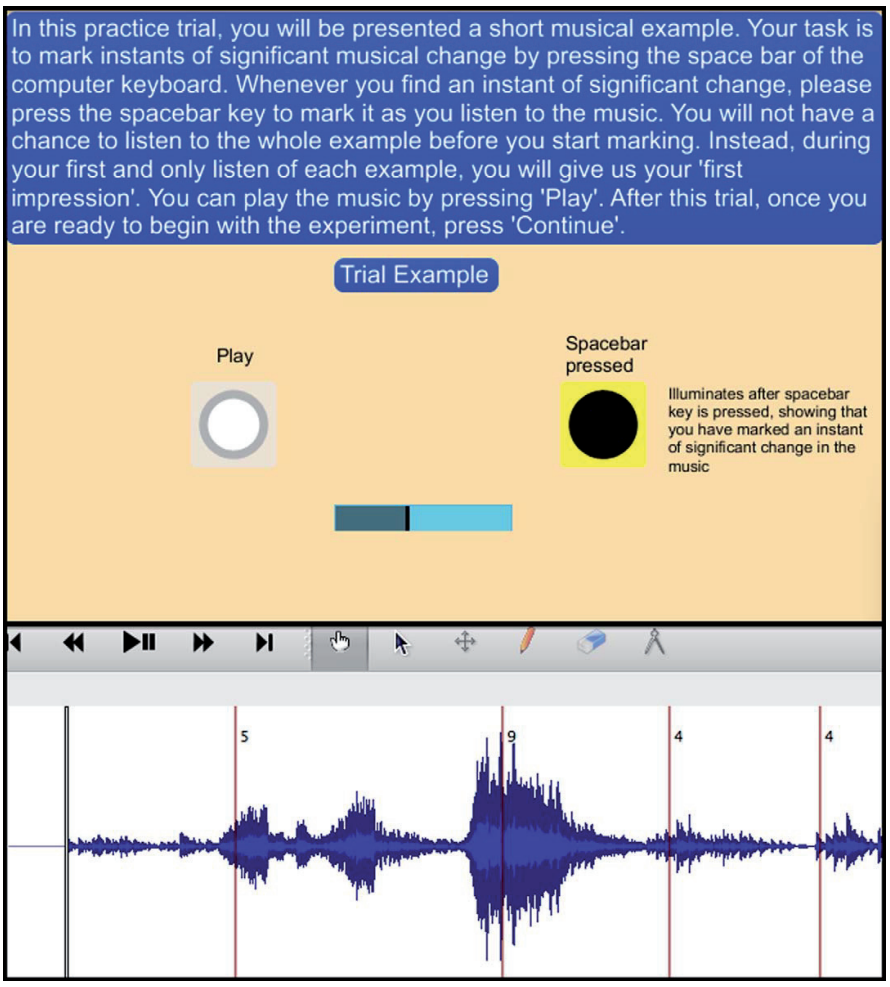

FIGURE 1. Upper image (Experiment 1): Trial instructing listeners to indicate instants of significant change while listening to the music. Lower image (Experiment 2): Part of an annotation segmentation performed by a musician for the stimulus *Ravel*; vertical bars indicate marked boundaries, and numbers 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	$\bar{x}$ age (*SD*)	Range	$\bar{x}$ years training (*SD*)	Range	$\bar{x}$ 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		Ma_w
$Task_{\text{boundary strength weights}}$		

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).

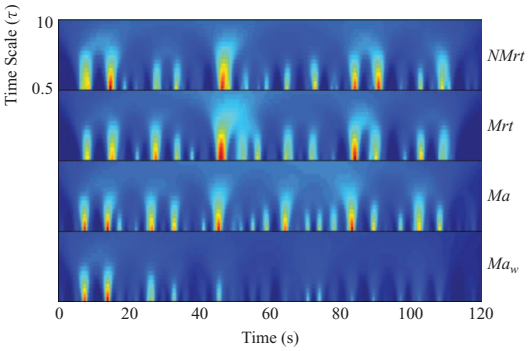

FIGURE 2. Each of the four sets of indicated boundaries was modelled via a multiple time scale approach. The kernel density over time of stimulus *Morton* is represented for 16 time scales.

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 ($\bar{x} = 3.1$, $SD = 1.3$), Piazzolla ($\bar{x} = 2.9$, $SD = 1.6$), and Ravel ($\bar{x} = 2.4$, $SD = 1.4$). Regarding nonmusicians, they were most familiar with Stravinsky ($\bar{x} = 2.7$, $SD = 1.5$), Dream Theater ($\bar{x} = 2.6$, $SD = 1.1$), and Piazzolla ($\bar{x}= 2.4$, $SD = 1.4$).

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

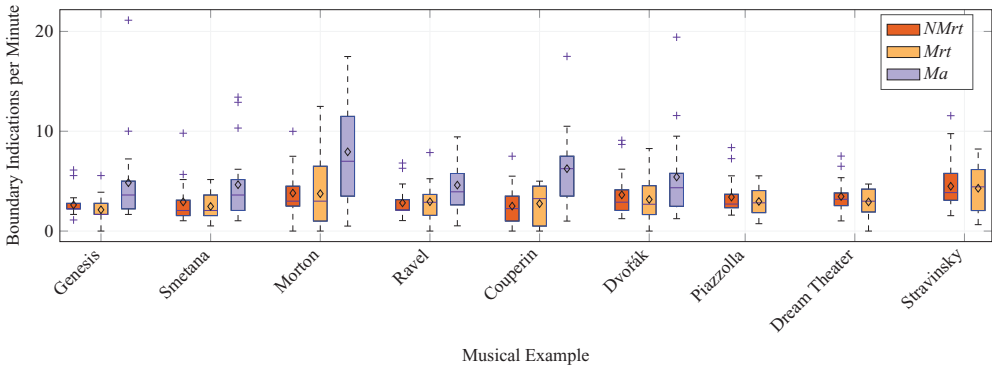

FIGURE 3. Box plot comparing participant groups and segmentation tasks with respect to the number of indicated boundaries per minute for each stimulus.

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

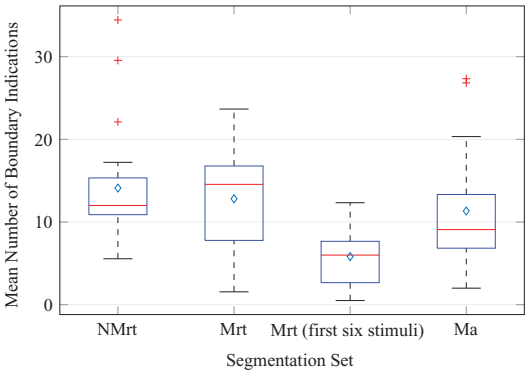

FIGURE 4. Box plot comparing participant groups and segmentation tasks with respect to the mean number of indicated boundaries by each participant.

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

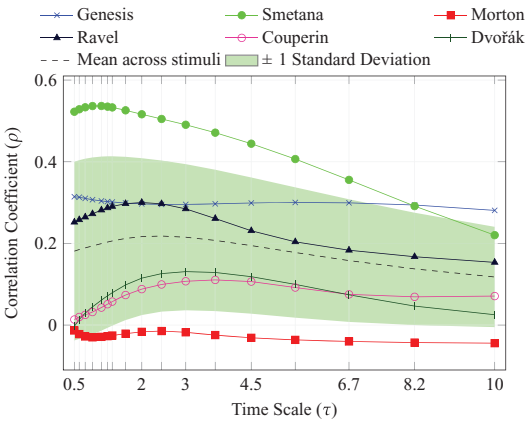

FIGURE 5. Confidence interval plot of correlation between boundary strength ratings in the annotation task and boundary density in the real-time task.

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

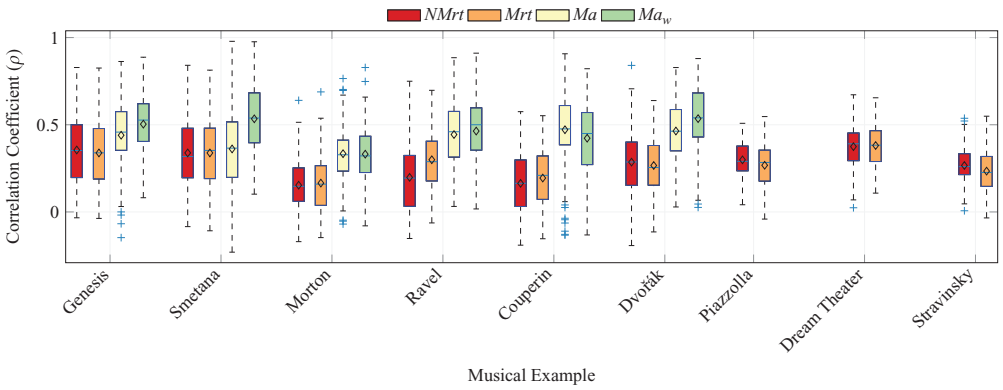

FIGURE 6. Box plot showing inter-subject correlation coefficient per stimulus for each segmentation set; p values (***$p < .001$) obtained via 10,000 Monte Carlo replications and adjusted using Benjamini-Hochberg correction ($q = 0.05$).

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

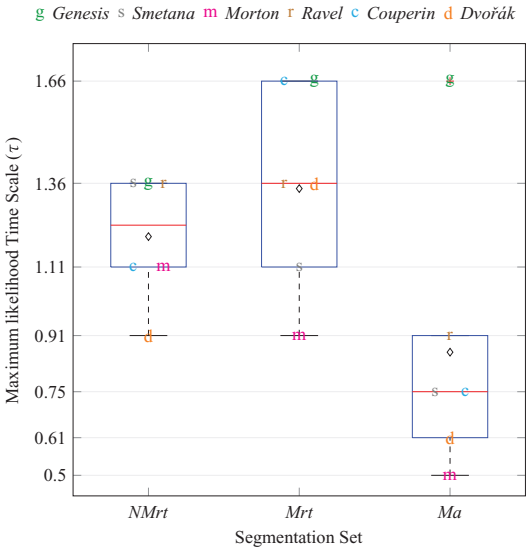

FIGURE 7. Box plot of maximum likelihood time scales for each segmentation set. Each time scale is represented with initials of its corresponding stimulus.

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*
Dream Theater	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

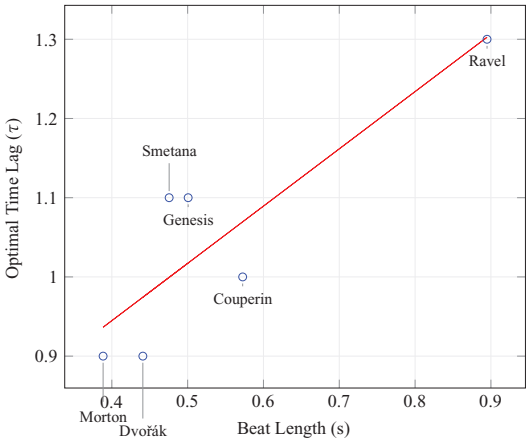

FIGURE 8. Scatter plot of optimal time lag for alignment between tasks as a function of stimuli beat length (BL). Trend line: simple linear regression equation $\tau = .72 \times BL + .66$.

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

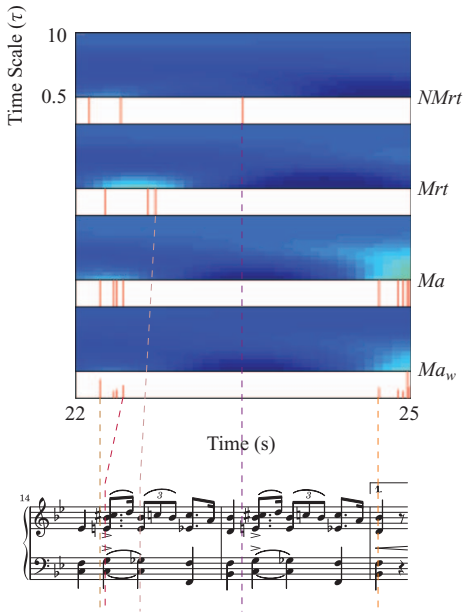

FIGURE 9. Multi-scale analysis for a 3-s extract of the stimulus *Morton*. Solid vertical lines: boundary indications by listeners. Dashed lines: approximate location of boundaries in the score.

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

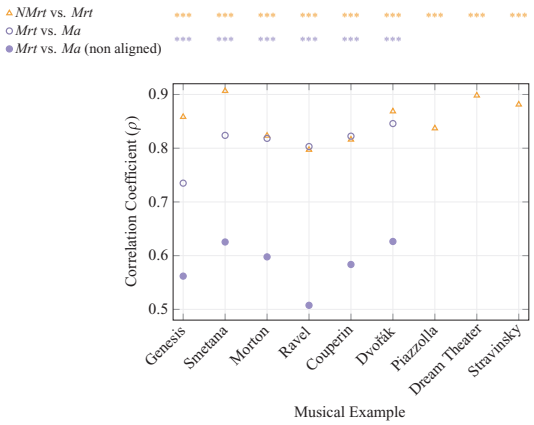

FIGURE 10. Multi-scale model correlation per stimulus comparing participant groups, segmentation tasks and alignment strategies; *p* values (***$p < .001$) obtained using Monte Carlo simulation and adjusted via Benjamini-Hochberg correction ($q = 0.05$).

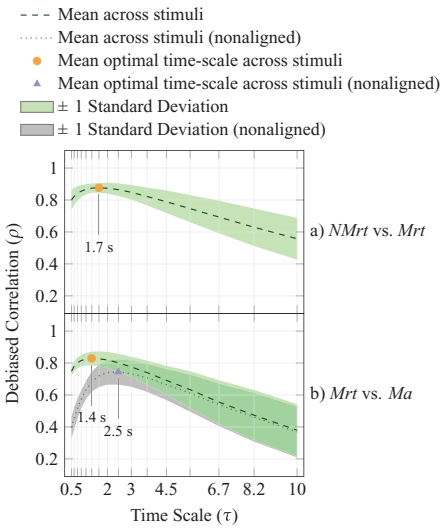

FIGURE 11. Mean correlation across stimuli between sets of indicated boundaries at 16 different time scales. Shaded areas: 1 standard deviation of the mean (estimate of correlation dispersion of different stimuli). Figure *a* compares participant groups in the real-time task. Figure *b* shows mean comparisons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

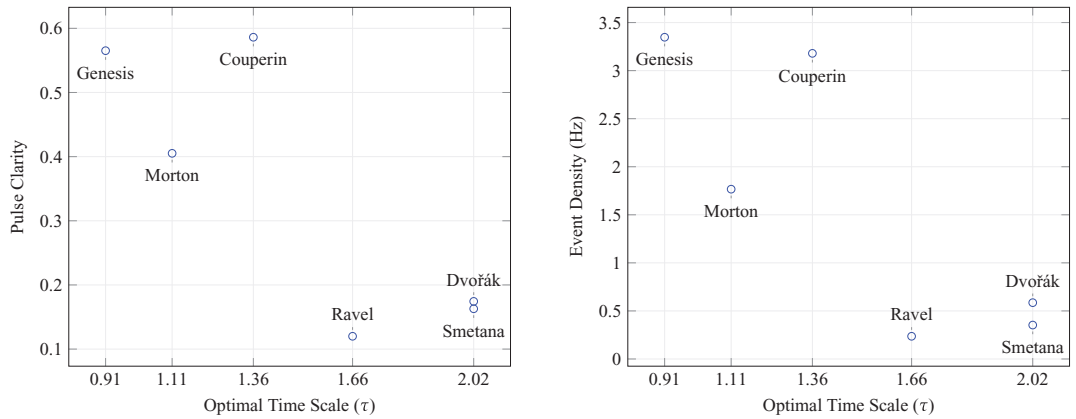

FIGURE 12. Relationship between two rhythmic descriptors and optimal time scales for correlation between tasks. a) Negative link, $r(4) = -.83$, $p < .05$, between pulse clarity and optimal time scale to compare real-time task and annotation task. b) Negative relationship, $r(4) = -.82$, $p < .05$, between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

PII

INTERACTION FEATURES FOR PREDICTION OF PERCEPTUAL SEGMENTATION: EFFECTS OF MUSICIANSHIP AND EXPERIMENTAL TASK

by

Martín Hartmann, Olivier Lartillot, and Petri Toiviainen 2016

Journal of New Music Research

Journal of New Music Research, 2016
http://dx.doi.org/10.1080/09298215.2016.1230137

Routledge
Taylor & Francis Group

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

Martín Hartmann[1]*, Olivier Lartillot[2], Petri Toiviainen[1]

[1]*University of Jyväskylä, Jyväskylä, Finland;* [2]*Aalborg University, Aalborg, Denmark*

(Received 15 March 2016; accepted 23 August 2016)

Abstract

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

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

1. Introduction

Humans possess the ability to perceptually parse ongoing streams into discrete events. This perceptual operation, which is called segmentation, makes it possible to understand activities that involve sound and movement, just like it is possible, in a messy room, to recognise each of its objects (Zacks & Swallow, 2007). It has central importance, for instance, in the area of speech perception, as it is needed for language acquisition: infants exploit different speech segmentation cues to identify words in sequences of syllables and to recognise larger groupings such as clauses (Johnson & Jusczyk, 2007; Seidl, 2014). Similar but specialised psychological processes may apply to music listening, since musical events that share-related characteristics or high temporal proximity are often grouped into sequences, even in passive listening contexts. This temporal psychological process of integrating musical events into larger units, which has been proposed to be universal (Drake & Bertrand, 2001), can be inversely formulated: listeners segment long musical streams when they perceive changes and repetitions. Musical feature change is a common cue for segmentation: listeners indicate segment boundaries if they easily perceive that there is a contrast, such as a stark change in dynamics or instrumentation. Multiple strategies are

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

exploited by composers (Deliège, 2001), improvisers (Dean, Bailes, & Drummond, 2014) and performers (Poli, Rodà, & Vidolin, 2007) to induce perception of musical changes, and communicate musical structure to the listener. This paper focuses, however, on musical listeners only, and on a particular conception of segmentation. We refer to segmentation in its broad sense, as we understand perceptual segment boundaries as *significant instants of musical change*; implications of this choice are discussed further.

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

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

The potential of combining different acoustic features for segmentation and structural analysis has been mentioned in MIR studies (Turnbull, Lanckriet, Pampalk, & Goto, 2007; Paulus & Klapuri, 2009). Few novelty-based studies (Paulus & Klapuri, 2009; Eronen, 2007; Peeters, 2007) have yielded enhanced automatic structural analyses via the summation of spectral and chroma features; this operation can be considered as a logical disjunction (**OR**), because changes of either or both spectral and chroma features would result in novelty peaks. To our knowledge, no studies in this area have implemented logical conjuction (**AND**) operations, which would yield novelty peaks only after concurrent change of both features. For example, an interaction feature resulting from a spectral novelty curve and a chroma novelty curve would not register a given spectral change unless it was accompanied by a simultaneous chroma change, and vice versa. From a computational perspective, such a novelty feature interaction approach seems appropriate because it can reduce the effect of spurious novelty peaks derived from high feature sensitivity; it may also be relevant from a perceptual viewpoint, since listeners probably pay most attention to changes that are evoked by more than one musical dimension (see Smith, Schankler, & Chew, 2014).

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

Musical experience seems to have an impact on listeners' focus of attention during music listening and on their representation of structure. Non-musicians are often considered to pay more attention to aspects related to the musical surface; they often tap with the fastest pulse during finger tapping tasks (Martens, 2011), and tend to place more boundary indications than musicians in segmentation studies (Hartmann et al., forthcoming-b; Bruderer, 2008; Deliège, 1987), suggesting that non-musicians focus more on changes in timbre, fast rhythmic layers, and pitch jumps. Most research has found that non-musicians focus less on harmonic functions than musicians, for instance in a task that consisted in rearranging musical segments, non-musicians paid more attention to rhythmic and metric aspects than to tonality (Deliège, Mélen, Stammers, & Cross, 1996). Moreover, a rhythm identification study showed that musicians' perception of rhythmic patterns for temporal sequences with harmonic accompaniment was more influenced by location of chord changes than non-musicians', whose answers were less consistent, and biased

towards responses that fitted the inferred meter (Dawe, Platt, & Racine, 1995). Based on these findings, it could be posited that non-musicians' segmentation can be more accurately predicted from the audio signal than musicians'; musicians would pay also attention to deeper aspects such as tonal context, which cannot be accurately modelled since they are rooted on implicit knowledge of Western tonal hierarchies. Other studies on processing and perception of musical structure (see Tillmann & Bigand, 2004) however suggest that schematic knowledge (see Justus & Bharucha, 2001) is built through mere exposure to music, as both groups focused on musical surface and deeper aspects of structure during tasks involving harmonic priming and manipulation of global organisation of pieces. Hence, it becomes unclear if musically trained listeners are more influenced by schematic expectancies during segmentation than untrained listeners or, conversely, if for both groups few musical events suffice to generate accurate forecasts about mode or upcoming chords in the music (Tillmann & Bharucha, 2002). Thus far, no studies have investigated the prediction of musicians' and non-musicians' segmentation, nor systematically examined whether or not these groups pay attention to same or different acoustic features during segmentation tasks. A deeper understanding on how musical training shapes our perception and understanding of structure and an examination of what musical dimensions listeners are attending to are needed in order to gain further insights on how musical structure is processed.

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

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

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

- To what extent does musicianship affect segmentation, and more specifically, how does computational prediction of segmentation for musicians differ from that of non-musicians?
- What is the effect of experimental task on segmentation, particularly on the modelling of real-time and non real-time segmentation tasks?

- Related to the previous question, what is the contribution of perceived boundary strength ratings on prediction of non real-time segmentation?

As a first hypothesis, we expected to find an effect of musicianship on model prediction, as non-musicians should be more accurately predicted by the segmentation models: they would focus more on perceived local acoustic changes, which could be accurately detected via novelty-based methods. Musicians would instead segment more based on other aspects, such as learned musical schemata, and find relatively irrelevant surface events to be context and cues for ulterior changes that may be much more significant. Also, more features were expected to be involved in musicians' prediction, particularly more interaction novelty features, because musicians would pay attention to more musical dimensions and to co-occurring feature changes at multiple levels of the musical structure. In addition, we expected smaller response delays for musicians than for non-musicians due to extensive training on sense of timing cues.

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

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

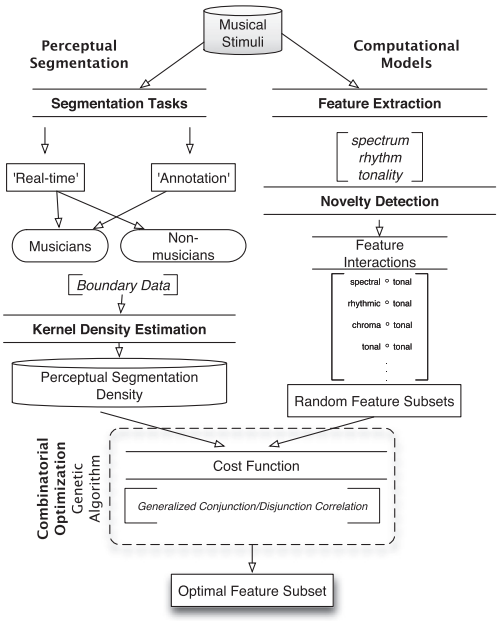

Fig. 1. General design of the study.

2. Method

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

2.1 Experiment I: Real-time task

2.1.1 Subjects

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

2.1.2 Stimuli

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

2.1.3 Apparatus

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

2.1.4 Procedure

Participants were asked to indicate significant instants of change while listening to the music by pressing the space bar key of the computer; the stimuli were presented in random order. For each participant and stimulus the boundary data was recorded in a single pass: they neither had the chance to listen to the stimuli before the segmentation nor were able to modify their boundary indications after the task. The task instructions were as follows: 'Your task is to mark instants of significant musical change by pressing the space bar of the computer keyboard. Whenever you find an instant of significant change, please press the 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'.

2.2 Experiment II: Annotation task

2.2.1 Subjects

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

2.2.2 Stimuli

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

2.2.3 Apparatus

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

2.2.4 Procedure

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

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

2.3 Perceptual segment boundary density

For each stimulus, the collected boundary indication data from different listeners was aggregated into a perceptual segmentation density curve for each participant group and segmentation task. First, we organised the segmentation data into three groups: musicians in the real-time task, non-musicians in the real-time task, and musicians in the annotation task. Next, we aggregated boundary indications from all participants so as to obtain a single profile of indications per stimulus. Subsequently, we concatenated the boundary data from each stimulus to obtain three boundary profiles spanning a duration of 12 min 5 s each. For each profile we derived a time series of density of segmentation. These segment boundary probability curves were obtained via KDE. This approach is illustrated in Figure 4 (upper plot), where segmentation density peaks in

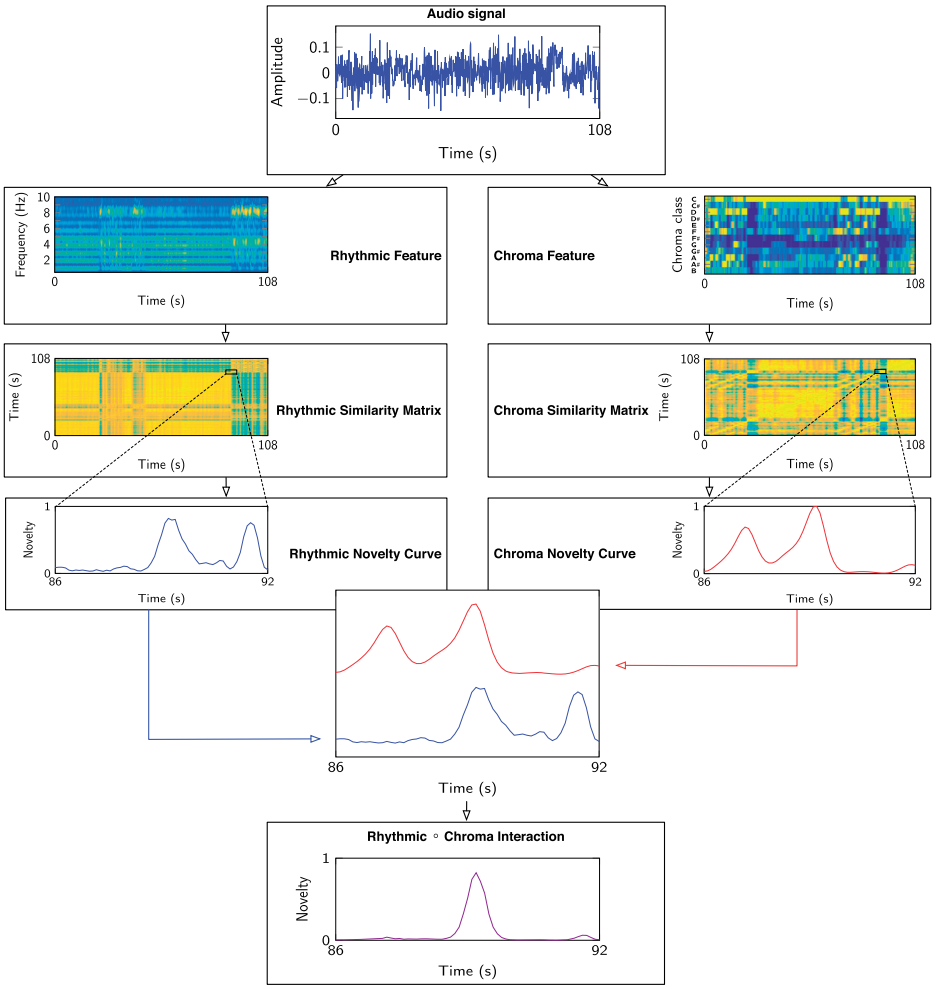

Fig. 2. Method used to obtain interaction features via pairwise multiplication between novelty curves.

the curve imply that multiple participants indicated boundaries at relatively close time points. The amount of closeness required between two boundary indications for them to be represented by the same density peak is defined by the time scale parameter τ, which corresponds to the bandwidth of the KDE Gaussian kernel; in other words, this parameter determines the degree of smoothness of the KDE. We chose a segmentation density time scale of $\tau = 1.5$ s following previous studies that focused on the Gaussian kernel bandwidth for modelling perceptual segmentation (Befus, 2010; Bruderer, 2008); particularly, Hartmann et al. (forthcoming-b) found a mean optimal time scale for comparison between real-time and annotation task boundary density curves at 1.4 s and a mean optimal time scale for comparison between musicians' and non-musicians' boundary density at 1.7 s. The sample rate of the KDE was set to 10 Hz since it was deemed sufficiently accurate for point process data of this nature. Besides the three obtained segmentation density curves, the annotation task data was also modelled taking into account listeners' boundary strength ratings, yielding a weighted boundary density curve. In total, we obtained four curves describing probability density estimates of the boundary data: boundary density for non-musicians in the real-time task (*NMrt*), musicians in the real-time task (*Mrt*), musicians in the annotation task (*Ma*) and musicians in the annotation task with added boundary strength weights (Ma_w).

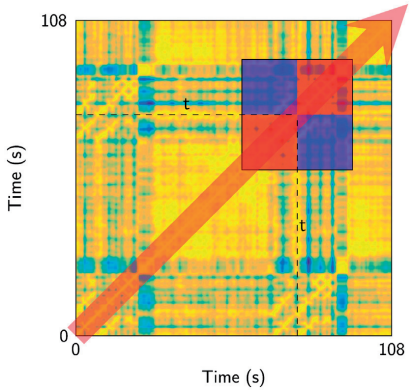

Fig. 3. Convolution of a Gaussian checkerboard kernel along the main diagonal of a chromagram-based similarity matrix.

2.4 Computational modelling

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

2.4.1 Feature extraction

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

2.4.2 Novelty detection

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

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

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

2.4.3 Optimal checkerboard kernel width

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

2.4.4 Non-linear modelling

We investigated the prediction of perceptual data from combinations of novelty curves via a non-linear modelling approach. The approach consisted in finding a subset of novelty curves whose 50th percentile (median ordinal position) would optimally correlate with the segmentation density curve. This procedure involves a non-linear aggregation of novelty features that assigns weights to features for each time point based on ranked values. From the perspective of soft computing, the percentile aggregation involves a monotonically increasing mapping that follows a continuous logic function called conjunction/disjunction function (Dujmović & Larsen, 2007). Roughly, the 0th percentile (equivalent to the *min* function) can be understood as a pure logical **AND** conjunction (‘all criteria are satisfied’) because if the minimum among features is high, then all features should have high values;

conversely, the 100th percentile (*max* function represents pure **OR** disjunction ('at least one criterion is satisfied') because a high maximum value among features implies that at least one of the features has a high value. Following this logic, 1–99th percentiles lie between **AND** and **OR**, exhibiting varying levels of *orness* (closeness to maximum). Hence, taking the 50th percentile across features would be comparable to a 'majority judgement', because it would only result in high values if at least half of the features exhibited high values. Several statistics, including arithmetic mean, median, min, max and percentile belong to the family of ordered weighted averaging operators (Yager, 1988, 2006), but have different characteristics; for example in arithmetic mean aggregation all data elements get equal weights, whereas percentiles use only one argument to determine the aggregated value (for an odd number of arguments).

2.4.5 Combinatorial optimisation

In order to find an optimal subset of features for computational modelling we performed discrete combinatorial optimisation. Via this approach we searched for a combination of novelty features whose percentile-based model would yield highest prediction rates, i.e. maximum correlation with the perceptual segmentation density. A generalised conjunction/disjunction correlation was used as a cost function criterion within a combinatorial optimisation routine. The cost function finds the optimal value of the correlation coefficient y by minimising the negative of the correlation between actual and predicted density, $y_{opt} = \text{argmin}_y - \text{corr}(x, p_\alpha)$, where x is the segmentation density and p_α is the α-percentile along features of a given subset. The reason for using combinatorial optimisation was the high number of possible feature combinations per perceptual segmentation density curve (2^{15}). We used a Genetic Algorithm search heuristic to find an optimal feature subset for each perceptual segmentation density curve. The optimisation cost function was initialised with a random subset of features. Since the algorithm employs a stochastic selection at each iteration, it tends to avoid local optimal solutions, i.e. subsets that are only best within the context of neighbouring combinations. As a result, we obtained for each segmentation density curve an optimal percentile model, the correlation between these two curves and an optimal subset of features for computing the model. Correlation p-values (H0: *no correlation between observed and predicted segmentation density*) were obtained via Fisher's z-transformation of r, with standard scores adjusted for effective degrees of freedom (i.e. corrected for temporal autocorrelation, see Pyper & Peterman, 1998; Alluri et al., 2012).

3. Results

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

3.1 Novelty kernel width

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

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

Comparing density curves, Figure 5 shows that the annotation task density curve with added weights tended to yield the highest correlations for most features. Adding weights to the annotation task lead to an increase in correlation (with respect to *Ma*) for all but three features when using a novelty width of 11 s (A.3). A possible reason for this correlation increase could have been the larger variance of the boundary density in the annotation task with added weights, which might have increased similarity with novelty curves due to their high variance. If the increase in correlation was the result of

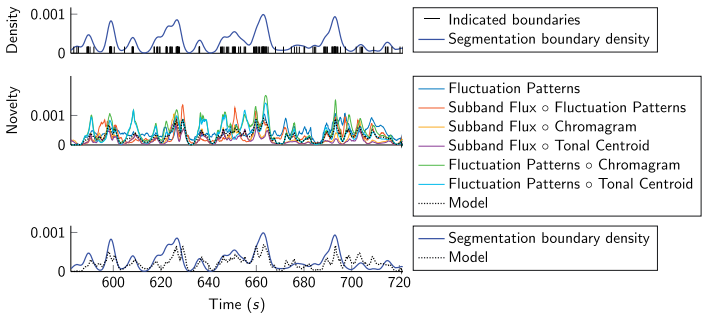

Fig. 4. Perceptual segment boundary density and computational segmentation model for non-musicians in the Real-time task (stimulus Dvořák). Upper graph: Boundary indication data and segmentation boundary density. Middle graph: Model predictors and computational model prediction. Lower graph: Perceptual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

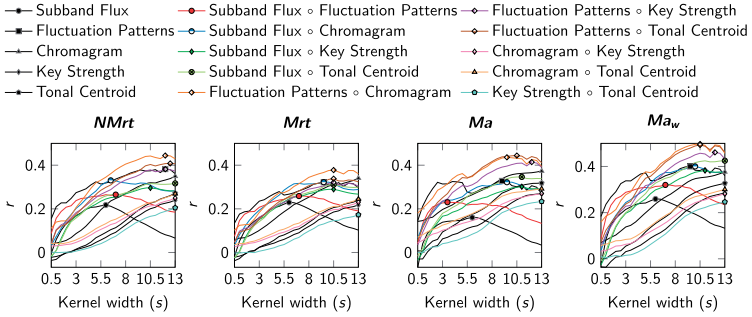

Fig. 5. Correlation between perceptual segment boundary density and novelty curves at novelty widths ranging from 0.5 to 13 s. Maximum points for each curve are highlighted with markers.

simply adding variance to the boundary density via addition of weights, then a random set of weights would be likely to yield a density curve that would result in increased correlation with respect to the weighted annotation task density. To test this possibility, we performed a Monte Carlo permutation (20,000 iterations). At each iteration, (1) a random vector of boundary weights (between 1 and 10) of length equal to the number of boundary indications in the annotation task was generated, and a kernel density curve of the annotation task that included the random vector of weights was correlated with each of the 15 novelty curves. This resulted in a correlation distribution per novelty feature; for each distribution, the sum of the values that were equal or higher than the correlation reported in the study (A.3) for *Maw* was divided by the length of the distribution. Features that showed an improvement after adding weights to the annotation task tended to yield correlations for *Maw* that were unlikely to be reached by using a random set of strength weights ($p < .001$ for eight features; $p < .01$ for one feature; $p > .05$ for key strength, chromagram ∘ key strength, and key strength ∘ Tonal centroid). This suggests that higher variance of the boundary density was probably not an important factor in the correlation increase obtained from listeners' boundary strength ratings; in other words, boundary strength ratings from listeners include relevant information that lead to an increase of segmentation prediction accuracy.

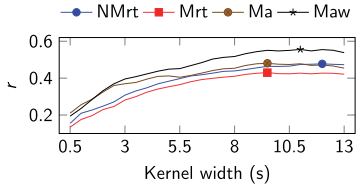

Fig. 6. Correlation between perceptual segmentation density and computational model prediction obtained via percentile optimisation for novelty widths ranging from 0.5 to 13 s. Maximum points for each curve are highlighted with markers.

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

	NMrt	**Mrt**	**Ma**	**Maw**
Subset	Fluctuation patterns Chromagram Tonal centroid Subband flux ∘ Fluctuation patterns Fluctuation patterns ∘ Chromagram	Fluctuation patterns Key strength Subband flux ∘ Fluctuation patterns Subband flux ∘ Tonal centroid Fluctuation patterns ∘ Chromagram	Subband flux Fluctuation patterns Tonal centroid Subband flux ∘ Tonal centroid Fluctuation patterns ∘ Chromagram Fluctuation patterns ∘ Tonal centroid	Fluctuation patterns Tonal centroid Subband flux ∘ Fluctuation patterns Subband flux ∘ Tonal centroid Fluctuation patterns ∘ Chromagram Fluctuation patterns ∘ Tonal centroid
Category	Rhythmic Chroma Tonal Spectral ∘ Rhythmic Rhythmic ∘ Chroma	Rhythmic Tonal Spectral ∘ Rhythmic Spectral ∘ Tonal Rhythmic ∘ Chroma	Spectral Rhythmic Tonal Spectral ∘ Tonal Rhythmic ∘ Chroma Rhythmic ∘ Tonal	Rhythmic Tonal Spectral ∘ Rhythmic Spectral ∘ Tonal Rhythmic ∘ Chroma Rhythmic ∘ Tonal
r	.47***	.43***	.48***	.56***

***$p < .001$.

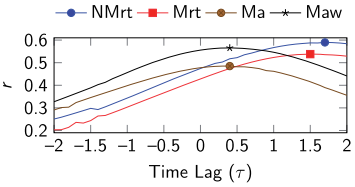

Fig. 7. Correlation between perceptual segment boundary density and models' predictions obtained after time lags ranging from −2 to 2 s, incremented by steps of 100 ms. Positive time lags indicate delay of novelty curves with respect to perceptual segment boundary density, and vice versa. Maximum points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

3.2 Model prediction rates

We next examined the prediction obtained from novelty-based computational models for different participant groups and segmentation tasks. To achieve this, we performed combinatorial optimisation using generalised conjunction/disjunction correlation as a cost function. We further investigated at this stage the novelty kernel width parameter by obtaining 26 computational models at varying novelty widths. Figure 6 shows that prediction rates tend to increase as a function of novelty but gradually reach a plateau; novelty curves based on a kernel width of 11 s yielded the highest overall prediction rates. Table 1 shows the correlation between optimal models and segmentation density for each participant group and segmentation task at a novelty width of 11 s; interaction features include the symbol ∘, which indicates pairwise multiplication between two basic features. Notably, prediction rates were lower for musicians than for non-musicians in the real-time task. This result suggests that musicians' segmentation relies more on schematic knowledge than in the case of non-musicians. Comparing experimental tasks, we found higher prediction rates for the annotation task. This suggests that some boundaries are difficult to anticipate in real-time segmentation, and are hence either indicated after longer delays or not indicated at all, leading to more noisy segmentation data. Related to this finding, the effect of experimental task was clearer for the annotation task density curve with added weights, which yielded the highest prediction rates. This indicates that the strength attributed by listeners to boundaries aids to the computational prediction and suggests a positive relationship between musical novelty and perceived boundary strength.

3.3 Selected feature subsets

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

3.4 Time lag between actual and predicted density

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

4. Discussion

This section will discuss three hypotheses that have been formulated with regard to prediction of music segmentation. It is important to highlight at this point that the approach

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

4.1 Musicianship

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

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

4.2 Experimental task

Three results were found to support our second hypothesis, which states that the conducted experimental task has an effect on model prediction. First, prediction rates for the annotation task were higher than for the real-time task, but controlling for delays inverted this result. This suggests that listeners' delayed indications are responsible for the relatively lower prediction rates in the real-time task, and that once these are compensated, this task yields higher similarity with 'bottom-up' novelty-based predictions since listeners neither know with certainty about the unfolding patterns and developments of a piece of music, nor can clearly estimate the relative

significance of a given musical change. The second result concerning segmentation tasks is that prediction of annotation task involved more novelty features, particularly rhythmic and tonal features. This result suggests that listeners pay attention in this task to more levels of the structural hierarchy. The third result with respect to this hypothesis is that the annotation task exhibited a shorter optimal time lag for segmentation prediction than the real-time task. This result is highly expected mainly because the annotation task allowed participants to modify the position of boundaries, but it is noteworthy that the alignment between segment boundaries and instants of musical novelty leads to an increase in prediction rates for offline segmentation tasks as well.

4.3 Boundary strength weights

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

4.4 General discussion

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

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

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

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

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

4.5 Considerations for future research

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

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

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

Focusing further on listeners' response delays, our findings showed that segmentation data can often exhibit up to 1.7 s delays with respect to musical changes; this compensation for response delays increased prediction rates in all models except for the annotation task without added weights. In this regard, retrieval evaluation of boundary detection systems is commonly based on both 0.5 and 3 s thresholds (Ehmann, Bay, Downie, Fujinaga, & De Roure, 2011), however according to our findings, 3 s would yield overly optimistic results, especially considering that the segmentation ground truth data for these evaluations is collected via annotation tasks; future research on MIR should consider hit rate evaluation only at a time threshold of 0.5 s.

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

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

can be circumvented; for instance, a recently proposed multi-granular method (Lartillot, Cereghetti, Eliard, & Grandjean, 2013) detects novelty by considering both the amount of contrast between neighbouring homogeneous passages and the temporal scale of the preceding passage.

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

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

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

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

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

Acknowledgements

The authors would like to thank Alan Marsden and an anonymous reviewer for giving us helpful comments on an earlier version of this paper. Thanks also to Emily Carlson for proofreading the paper.

Funding

This work was supported by the Academy of Finland [project numbers 272250 and 274037].

ORCID

Martín Hartmann http://orcid.org/0000-0002-2897-0610
Olivier Lartillot http://orcid.org/0000-0002-3294-2719
Petri Toiviainen http://orcid.org/0000-0001-6962-2957

References

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

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

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage, 59*, 3677–3689.

Befus, C. (2010). *Design and evaluation of dynamic feature-based segmentation on music* (PhD thesis). Lethbridge: Department of Mathematics and Computer Science, University of Lethbridge.

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

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

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

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

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

Chew, E. (2002). The spiral array: An algorithm for determining key boundaries. In Anagnostopoulou, C., Ferrand, M., & Smaill, A. (Eds.), *Music and artificial intelligence* (pp. 18–31). Edinburgh: Springer.

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

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

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

Dean, R. T., Bailes, F., & Drummond, J. (2014). Generative structures in improvisation: Computational segmentation of keyboard performances. *Journal of New Music Research, 43*, 1–13. doi: 10.1080/09298215.2013.859710

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

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.

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

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

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

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

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

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

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

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

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

Harte, C., Sandler, M., & Gasser, M. (2006). *Detecting harmonic change in musical audio*. In *Proceedings of the 1st ACM workshop on Audio and Music Computing Multimedia* (pp. 21–26), Santa Barbara.

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

Hartmann, M., Lartillot, O., & Toiviainen, P. (Forthcoming-a). Musical feature and novelty curve characterizations as predictors of segmentation accuracy.

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

Jensen, K. (2007). Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing, 2007*, 159–159.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: when speech cues count more than statistics. *Journal of Memory and Language, 44*, 548–567. doi:10.1006/jmla.2000.2755

Justus, T. C., & Bharucha, J. J. (2001). Modularity in musical processing: The automaticity of harmonic priming.

Journal of Experimental Psychology: Human Perception and Performance, 27, 1000–1011.

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

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

Lartillot, O., Cereghetti, D., Eliard, K., & Grandjean, D. (2013). A simple, high-yield method for assessing structural novelty. In G. Luck & O. Brabant (Eds.), *Proceedings of the 3rd International Conference on Music & Emotion (ICME3)*. Finland: Jyväskylä.

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

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

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

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

Martens, P. A. (2011). The ambiguous tactus: Tempo, subdivision benefit, and three listener strategies. *Music Perception, 28*. 433–448.

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

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

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

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

Poli, G. D., Rodà, A., & Vidolin, A. (1998). Note-by-note analysis of the influence of expressive intentions and musical structure in violin performance. *Journal of New Music Research, 27*, 293–321. doi: 10.1080/09298219808570750

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

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

Seidl, A. (2007). Infants' use and weighting of prosodic cues in clause segmentation. *Journal of Memory and Language, 57*, 24–48. doi: 10.1016/j.jml.2006.10.004

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). Boca Raton, FL: CRC Press.

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

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

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

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

Tillmann, B., & Bharucha, J. J. (2002). Effect of harmonic relatedness on the detection of temporal asynchronies. *Perception & Psychophysics, 64*, 640–649.

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

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

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

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

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

Appendix A. Appendices

A.1. Musical stimuli—List of abbreviations

Genesis Banks, T., Collins, P. & Rutherford, M. (1986). The Brazilian. [Recorded by Genesis]. On *Invisible Touch* [CD]. Virgin Records. (1986)

Spotify link: http://open.spotify.com/track/7s4hAEJupZLpJEaOel5SwV

Excerpt: 01:10.200-02:58.143. Duration: 01:47.943

Smetana Smetana, B. (1875). Aus Böhmens Hain und Flur. [Recorded by Gewandhausorchester Leipzig—Václav Neumann]. On *Smetana: Mein Vaterland* [CD]. BC—Eterna Collection. (2002)

Spotify link: http://open.spotify.com/track/2115JFwiNvHxB6mJPkVtbp

Excerpt: 04:06.137-06:02.419. Duration: 01:56.282

Morton Morton, F. (1915). Original Jelly Roll Blues. On *The Piano Rolls* [CD]. Nonesuch Records. (1997)

Spotify link: http://open.spotify.com/track/6XtCierLPd6qg9QLcbmj61

Excerpt: 0-02:00.104. Duration: 02:00.104

Ravel Ravel, M. (1901). Jeux d'Eau. [Recorded by Martha Argerich]. On *Martha Argerich, The Collection, Vol. 1: The Solo Recordings* [CD]. Deutsche Grammophon. (2008)

Spotify link: http://open.spotify.com/track/27oSfz8DKHs66IM12zejKf

Excerpt: 03:27.449-05:21.884. Duration: 01:54.435

Couperin Couperin, F. (1717). Douziéme Ordre / VIII. L'Atalante. [Recorded by Claudio Colombo]. On *François Couperin : Les 27 Ordres pour piano, vol. 3 (Ordres 10-17)* [CD]. Claudio Colombo. (2011)

Spotify link: http://open.spotify.com/track/6wJyTK8SJAmtqhcRnaIpKr

Excerpt: 0-02:00.863 Duration: 02:00.863

Dvořák Dvořák, A. (1878). Slavonic Dances, Op. 46 / Slavonic Dance No. 4 in F Major. [Recorded by Philharmonia Orchestra—Sir Andrew Davis]. On *Andrew Davis Conducts Dvořák* [CD]. Sony Music. (2012)

Spotify link: http://open.spotify.com/track/5xna3brB1AqGW7zEuoYks4

Excerpt: 00:57.964-03:23.145. Duration: 02:25.181

A.2. Extracted musical features

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

Spectral

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

Rhythmic

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

Chroma

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

Tonal

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

A.3. Correlations between perceptual segment boundary density and novelty features

Table A1. Correlations between perceptual segmentation density and basic features. Maximum coefficients of each set are indicated in boldface. Coefficients from features selected via optimisation are highlighted. *P*-values adjusted for effective degrees of freedom, and for multiple comparisons via Benjamini–Hochberg correction ($q = 0.05$).

Feature type	Basic feature	NMrt	Mrt	Ma	Maw
Spectral	Subband flux	.10	.14*	.07	.17**
Rhythmic	Fluctuation patterns	**.38*****	**.32*****	.31***	**.39*****
Chroma	Chromagram	.32***	.31***	**.36*****	.35***
Tonal	Key strength	.21***	.19**	.25***	.26***
	Tonal Centroid	.23***	.21***	.31***	.30***

$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

Table A2. Correlations between perceptual segmentation density and feature interactions. Maximum coefficients of each set are indicated in boldface. Coefficients from features selected via optimisation are highlighted. *P*-values adjusted for effective degrees of freedom, and for multiple comparisons via Benjamini–Hochberg correction ($q = 0.05$).

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

$^{**}p < .01$; $^{***}p < .001$.

PIII

MUSICAL FEATURE AND NOVELTY CURVE CHARACTERIZATIONS AS PREDICTORS OF SEGMENTATION ACCURACY

by

Martín Hartmann, Olivier Lartillot, and Petri Toiviainen

Manuscript submitted for publication

Musical Feature and Novelty Curve Characterizations as Predictors of Segmentation Accuracy

Martín Hartmann · Olivier Lartillot · Petri Toiviainen

Abstract Novelty detection is a well-established method for analyzing the structure of music based on acoustic descriptors. Work on novelty-based segmentation prediction has mainly concentrated on enhancement of features and similarity matrices, novelty kernel computation and peak detection. Less attention, however, has been paid to characteristics of musical features and novelty curves, and their contribution to segmentation accuracy. This is particularly important as it can help unearth acoustic cues prompting perceptual segmentation and find new determinants of segmentation model performance. This study focused on spectral, rhythmic and harmonic prediction of perceptual segmentation density, which was obtained for six musical examples from 18 musician listeners via an annotation task. The proposed approach involved comparisons between perceptual segment density and novelty curves; in particular, we investigated possible predictors of segmentation accuracy based on musical features and novelty curves. For pitch and rhythm, we found positive correlates between segmentation accuracy and both local variability of musical features and mean distance between subsequent local maxima of novelty curves; this suggests that segmentation accuracy increases for stimuli with milder local changes and fewer novelty peaks. According to the results, novelty tends to be concentrated on few temporal regions for accurately predicted stimuli; implications regarding prediction of listeners' segmentation are discussed in the light of theoretical postulates of perceptual organization and musical expectation.

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

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

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

Keywords music segmentation, musical structure, novelty detection, kernel density estimation, musical features

1 Introduction

Musical segments are a representation of the perceived structure of music, and hence carry its multifaceted, interwoven, and hierarchically organized nature. Transitional points or regions between segments, which are called *perceptual segment boundaries*, can emerge from temporal changes in one or more musical attributes, or from more complex configurations involving e.g. repetition and cadences. Exhibited acoustic change can be measured to yield an estimate of *novelty* with respect to previous and upcoming musical events, with the aim to delineate perceptual segment boundaries: for each instant, its degree of acoustic novelty is expected to predict the likelihood of indicating a boundary. The success of predicting musical segment boundaries from acoustic estimates of novelty depends on the level of structural complexity of a musical piece. For instance, music that unambiguously evokes few sharp segment boundaries and clear continuity within segments to listeners should exhibit high accuracy, whereas pieces that prompt many boundaries and a rather heterogeneous profile might be more challenging for prediction. Music is however multidimensional, which implies that accuracy for a given stimulus could depend on the musical feature or features under study.

Novelty detection (Foote, 2000) approaches can be used to obtain segmentation accuracies for different musical features, such as timbre and harmony descriptors. This work focuses on the assessment of possible predictors of segmentation accuracy that could be obtained directly from musical features or from novelty points derived from these features. We explored musical features of different types, because listeners rarely focus on a single dimension of music. To illustrate possible applications of the proposed approach, one could imagine a segmentation system that could extract candidate features from a particular musical piece, and discard any features that would not seem to be informative of musical changes in order to focus only on those changes that would be deemed relevant for a listener. This would result in more efficient prediction, as the system would not require to compute subsequent structure analysis steps for irrelevant features. Peiszer, Lidy, and Rauber (2008) envisioned a similar scenario; according to their view, future segmentation systems should automatically select optimal structural analysis parameters separately for each individual musical piece. Our endeavor is motivated by the direct impact of music segmentation on other areas of computational music analysis, including music summarization, chorus detection, and music transcription, and its relevance for the study of human perception, as it can deepen our understanding on how listeners parse temporally unfolding processes.

MIR (Music Information Retrieval) studies on segmentation and structure analysis typically require perceptual data for algorithm evaluation. Often, data collection involves the indication of segment boundaries from one or few listeners for a large amount of musical examples; this results in a set of time points for each piece (e.g. Turnbull, Lanckriet, Pampalk, & Goto, 2007). In contrast, approaches on music segmentation within the field of music perception and cognition commonly involve the collection of perceptual boundaries from multiple participants in listening experiments; the collected data is often aggregated across listeners for its analysis. Kernel Density Estimation (*KDE*, Silverman, 1986) has been used in recent segmentation studies (e.g. Bruderer, 2008; Burunat, Alluri, Toiviainen, Numminen, & Brattico, 2014; Hartmann, Lartillot, & Toiviainen, 2016b) to obtain a precise aggregation across boundary data. This approach consists of obtaining a probability density estimate of the boundary data using a Gaussian function; in a KDE curve, temporal regions that prompted boundary indication from many listeners are represented as peaks of perceptual boundary density. This continuous representation of perceptual segment boundary probability has been used to compare different stimuli, groups of participants, and segmentation tasks (Bruderer, 2008; Hartmann et al., 2016b). Prediction of perceptual boundary density based on rules has been applied to symbolic data representations such as LBDM (Local Boundary Detection Model, see Cambouropoulos, 2001) and GTTM (Generative Theory of Tonal Music, see Lerdahl & Jackendoff, 1983; Frankland & Cohen, 2004), showing that rhythmic and timbre-based changes may contribute to perceptual segmentation (Bruderer, 2008).

Prediction of perceptual boundary density in the audio domain often involves the computation of novelty curves (Foote, 2000), which roughly de-

scribe the extent to which a temporal context is characterized by two continuous segments separated by a discontinuity with respect to a given musical feature. Usually, novelty curves are compared against segment boundary data for tasks related to structural analysis, but they have also been applied to other problems such as onset detection (Lartillot, Eerola, Toiviainen, & Fornari, 2008; Bello et al., 2005) and the study of rate of stylistic musical evolution over 50 years (Mauch, MacCallum, Levy, & Leroi, 2015). As structural features, novelty curves have also shown to contribute to models of expressed emotions in music: Eerola (2011) computed statistics derived from novelty curves and other musical features to predict listeners' perceived emotions via regression models, and found that pitch-based novelty tends to be low for music with high valence ratings. Recent work on structural analysis devised alternatives to Foote's novelty kernel and suggested methods combining multiple novelty kernel sizes (Gaudefroy, Papadopoulos, & Kowalski, 2015; Kaiser & Peeters, 2013). However, the original implementation still yields satisfactory results for segmentation prediction when compared to more recent methods (Aljanaki, Wiering, & Veltkamp, 2015), even for relatively challenging datasets (Bohak & Marolt, 2016). Other recent studies have compared novelty curves with perceptual boundary density curves (Hartmann, Lartillot, & Toiviainen, 2016a) and compared peaks derived from both curves (Mendoza Garay, 2014), showing that novelty detection can predict segmentation probabilities derived from numerous participants.

A preliminary step in novelty detection and other segmentation frameworks consists of the extraction of a musical feature, which will determine the type of musical contrast to be detected. Timbre, tonality and to some extent rhythm (Jensen, 2007; Klien, Grill, & Flexer, 2012) have been considered to be important features for structural analysis. Relatively high prediction of musical structure has been found for two musical features: MFCCs (Mel-Frequency Cepstral Coefficients) for timbre description (Peiszer, 2007; Foote, 2000) and Chromagram (Bartsch & Wakefield, 2001) or similar features (Serrà, Muller, Grosche, & Arcos, 2014) for description of pitch changes; also combined approaches have been proposed (Gaudefroy et al., 2015; Eronen, 2007; Turnbull et al., 2007). The segmentation accuracy achieved by novelty curves seems to highly depend on the musical feature that is used, and on the choice of temporal parameters for feature extraction (Peeters, 2004). Differences between musical pieces are another factor to consider regarding accuracy; as a general rule it could be agreed that music with a clearly defined structure, that is, characterized by discontinuity of musical features at segment boundaries, should yield higher segmentation accuracy; the novelty approach is likely to succeed if distinguishable segments of a musical piece are delimited by contrasting feature values. In addition, different musical pieces might require different musical features for optimal prediction (Peiszer et al., 2008); as pointed out by McFee and Ellis (2014), structure in pop and rock is frequently determined by harmonic change, whereas jazz is often sectioned based on instrumentation. No single feature can optimally predict all musical examples; certain features are more appropriate than others depending on particular aspects of musical pieces.

This mechanism is however not well understood at present: it is unclear what characteristics of musical features contribute to the segmentation accuracy for a given musical piece. Addressing this issue would be important since it would enable the possibility to select optimal musical features for further novelty detection, avoiding the computation of novelty curves that would not yield satisfactory results; it would also help to develop better alternatives to the novelty detection approach, with the aim of reducing computational costs.

To analyze the impact of different factors associated to novelty curves upon segmentation and compare different segmentation algorithms, a number of performance measures have been proposed, such as precision, recall, and F-measure (McFee, Nieto, & Bello, 2015; Nieto, 2015; Nieto, Farbood, Jehan, & Bello, 2014; Lukashevich, 2008; Peiszer et al., 2008); also correlation between time series has been applied for this purpose (Hartmann et al., 2016a; Lartillot, Cereghetti, Eliard, & Grandjean, 2013). One of the factors that has been shown to highly contribute to the segmentation accuracy is the width of the novelty kernel (e.g. Hartmann et al., 2016a), which roughly refers to the temporal context with respect to which novelty is estimated for each time point. It also appears that the relative accuracy for a given stimulus may depend on the segmentation approach used: for the same music collection, different novelty-based algorithms may find different examples to be most challenging (Bohak & Marolt, 2016;

Peiszer et al., 2008). Several studies have examined which steps within novelty detection would result in lower accuracy for 'challenging' stimuli; for instance, threshold-based peak picking has been found to be problematic for stimuli with boundaries of varying degrees of salience (Gaudefroy et al., 2015), and self-similarity matrices have been deemed to yield inaccurate curves for stimuli characterized by repeated segments or tempo change (Paulus, Müller, & Klapuri, 2010).

However, to the best of our knowledge, no study has systematically investigated what specific aspects of novelty curves contribute to their accuracy for a given stimulus. It would be relevant to investigate what characteristics of novelty curves relate to their accuracy, as this could allow to predict the relative suitability of a novelty curve for a given stimulus without the need of direct comparison against ground truth, and to bypass computation of novelty curves that would be assumed not to deliver satisfactory performance with regard to a particular stimulus. From the viewpoint of music perception, it would be useful to better understand the extent to which musical characteristics perceived by listeners are directly apparent from novelty curves, and to gain more knowledge on the types of musical changes that prompt both boundary perception and high novelty scores.

Recently, Hartmann et al. (2016a) studied segmentation accuracy achieved for concatenated musical pieces using different novelty curves. It was found that optimal prediction of perceptual segment boundary density involves the use of large kernel widths; the study also highlights the role of rhythmic and pitch-based features on segmentation prediction. The present study is a follow-up to the paper by Hartmann et al. (2016a), as it focused further on prediction of perceptual segmentation density via novelty detection, and examined the same musical pieces; in particular, we investigated one of the perceptual segmentation sets (an *annotation* segmentation task performed by musician listeners) studied by Hartmann et al. (2016a), and explored perceptual segmentation density and novelty curves for individual musical stimuli. The goal of this study was to understand whether or not local variability of musical features and distance between novelty peaks are related with the accuracy of segmentation models. The following research questions guided our investigation:

1. What specific aspects of musical stimuli that account for segmentation accuracy can be directly described from musical features?
2. What stimulus-specific attributes of novelty curves determine optimal segmentation accuracy?

As regards the first research question, we expected to find an inverse relationship, dependent on musical stimulus, between magnitude of local feature variation and accuracy obtained via novelty detection. For instance, musical stimuli displaying minimum local tonal contrast would yield optimal segmentation accuracy via tonal novelty curves. The rationale behind this hypothesis is that if there is not much local change in a feature, then the local changes that occur in that feature should be more salient.

One of the most basic Gestalt principles, the law of *Prägnanz*, relates to this rationale, because it states that, under given conditions, perceptual organization will be as "good" as possible, i.e. percepts tend to have the simplest, most stable and most regular organization that fits the sensory pattern (Koffka, 1935). Maximum and minimum simplicity are also characteristic in "good" perceptual organizations, as in a soap bubble, which has the highest possible volume for its surface and the lowest possible surface for its volume. Ahlbäck (2007) applied this property to melodic segmentation, stating that a higher contrast between boundaries and group events should lead to an increase in likelihood of recognizing groupings based on both sameness and difference. To illustrate maximum-minimum simplicity in the context of musical features, a simple organization from the viewpoint of perception would be characterized by the highest magnitude of local change that is possible for the duration of this change and the lowest possible duration of change for a given magnitude of local change.

Our rationale also follows general notions within the area of musical expectancy (Narmour, 1992): one of the ways in which expectation is generated follows the theoretical constant $A+A \rightarrow A$, which roughly means that successive events that are identical or similar with each other generate expectation of identical or similar events. Another way to generate expectation would correspond to the theoretical constant $A+B \rightarrow C$, which denotes that dissimilarity between successive events leads to the expectation of another event that is dissimilar to the previous ones. In contrast, a sense of closure would be perceived by listeners in the case of

$A + A \rightarrow B$; here the expectation is violated because $\rightarrow A$ is not present (one could also consider $A + B \rightarrow B$, in which retrospective boundary perception would occur due to the unexpected repetition of B). In a nutshell, sameness generates the expectation of more sameness, and discontinuity leads to expected discontinuity; expectation violation and boundary perception would generally occur instead when a temporal sequence of musical events characterized by homogeneity is followed by stark discontinuity. This scenario is similar to the aforementioned one regarding the *Prägnanz* law; the main difference is that expectation theory focuses on previous context to determine the likelihood of a boundary.

Regarding the second question, we expected that stimuli yielding higher accuracy for a given feature would exhibit a relatively large temporal distance between novelty peaks for that feature. Although this relationship could depend on other factors such as tempo and duration, we believed that the absolute time span between peaks would serve as a rough predictor of accuracy for the reasons above mentioned regarding, e.g., maximum and minimum simplicity. To give an example, optimal accuracy was expected for stimuli characterized by long and uniform segments with respect to instrumentation that would be delimited by important timbral changes, whereas music characterized by more frequent timbral change and gradual transitions would yield lower accuracy for a timbre-based novelty measure.

2 Method

Figure 1 illustrates the research design utilized in our investigation. The upper part of the figure concerns computational prediction of segmentation via novelty detection and a perceptual modelling of collected segmentation data via perceptual segmentation density estimation. These two topics were more thoroughly covered in Hartmann et al. (2016b) and Hartmann et al. (2016a), respectively. The bottom part of the figure, specifically the solid line connections, refer to investigation of correlates of perceptual boundary density prediction, which is the main focus of this study.

2.1 Segmentation Task

To obtain perceptual segment boundary density of the stimuli, we collected boundary data from

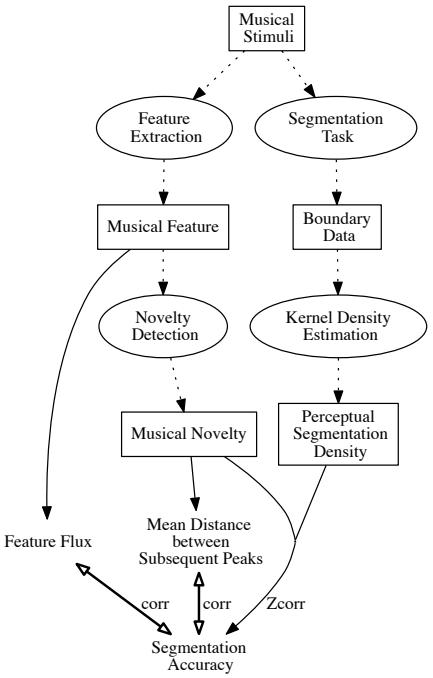

Figure 1: General design of the study.

participants via a listening task that involved an offline annotation. This non real-time segmentation task, called *annotation task*, is described with more detail in Hartmann et al. (2016b). The reason for analyzing prediction of non real-time segmentation was to reduce the number of intervening factors: compared to a real-time segmentation, boundary placements obtained via offline segmentation are probably better aligned in time with respect to musical changes; also, inter-subject agreement has been found to be higher for offline than for online tasks (Hartmann et al., 2016b).

Subjects

18 musicians (11 males, 7 females) with a mean age of 27.61 years (SD = 4.45) and an average musical training of 14.39 years (SD = 7.49) took part of this experiment. They played classical (12 participants) and non-classical musical styles (6 participants) such as rock, pop, folk, funk, and jazz; their main instruments were piano (5), guitar (4), flute (2), bass guitar, clarinet, saxophone, cello, vi-

olin, viola and voice. All participants were familiar with basic audio editing software, and also with the stimuli used for this task since they all had previously completed a real-time segmentation task (which was not analyzed in this study) involving the same stimuli.

Stimuli

We selected 6 instrumental music pieces; two of them lasted 2 minutes and the other four were trimmed down to this length for a total experiment duration of around one hour. The pieces (see Appendix I) comprise a variety of styles and considerably differ from one another in terms of musical form; further, they emphasize aspects of musical change of varying nature and complexity. *Genesis* is a pop-rock instrumental piece that combines multiple experimental sounds and effects (using electronic percussion and synthesizers) within relatively long, homogeneous sections in terms of melody, harmony, and to some extent loudness and instrumentation. *Smetana* is an extract from a romantic symphonic poem that consists mainly of a long cantando theme played by clarinets and horns followed by a fugato theme played by strings (De Lisa, 2009); these two sequences later return as reprises in different keys and including more instrumentation. *Morton* is a fox-trot piano piece that consists of a 4-bar introduction followed by variations over a 12-bar blues progression; its melodic line is clear despite being often intruded by sudden breaks, which are responded with rhythmic chordal clusters (Trythall, 2002). *Ravel* is an impressionist piano composition that demands technical virtuosity from the performer as it is characterized by a heavy use of arpeggio figuration, glissandi and tremolo; the piece includes very abrupt changes of dynamics, register, and tempo, and uses whole-tone and pentatonic scales (Sonntag, 2011). *Couperin* is a piano rendition of a baroque (rococo) piece for harpsichord, roughly characterized by highly ornamented semiquaver melodies that are accompanied by a quaver or semiquaver bass lines; although cadences is common, triads only appear four times throughout the piece. *Dvořák* is a symphonic piece that is mainly based on the sousedská Czech folk dance but also suggests a polonaise rhythm in several parts; the piece is in major key, but one of its motifs is also repeated several times in minor mode (Šupka, 2013). Its main theme is introduced by winds and horns but is later played by the whole ensemble and transposed up a fourth.

Apparatus

An interface in Sonic Visualizer (Cannam, Landone, & Sandler, 2010) was prepared to obtain segmentation boundary indications from participants. Stimuli were presented in randomized order; for each stimulus, the interface showed its waveform as a visual-spatial cue over which boundaries would be positioned (subjects were asked to focus solely on the music). Participants used headphones to play back the music at a comfortable listening level, and both keyboard and mouse were required to complete the segmentation task.

Procedure

Written instructions were given to participants; these included a presentation of the interface tools and a task description, which consisted of the following steps:

1. Listen to the whole musical example.
2. Indicate significant instants of change while listening to the music by pressing the Enter key of the computer.
3. Freely play back from different parts of the musical example and make the segmentation more precise by adjusting the position of boundaries; also removal of any boundaries indicated by mistake is allowed.
4. Rate the perceived strength of each boundary (ratings of boundary strength were collected for another study). Start over from the first step for the next musical example.

2.2 Perceptual Segment Boundary Density

We obtained a perceptual boundary density estimate across the segmentation data collected from musician participants; this estimate would be further compared against novelty curves to assess their accuracy. The perceptual boundary data of all participants was used to obtain a curve of perceptual segmentation density using a KDE bandwidth of 1.5 s; values around this bandwidth were found optimal for comparison between perceptual segmentation densities (Hartmann et al., 2016b) and were also utilized in other segmentation studies (Befus, 2010; Bruderer, 2008). From each tail of the perceptual density curves, 6.4 s were trimmed for more accurate comparisons with novelty curves (see below).

2.3 Feature Extraction and Novelty Detection

We computed novelty curves from 5 musical features describing timbre (Subband Flux), rhythm (Fluctuation Patterns), pitch class (Chromagram) and tonality (Key Strength, Tonal Centroid) using MIRtoolbox 1.6.1 (Lartillot & Toiviainen, 2007); the features used for novelty detection are described in Appendix II. For each feature, a self-similarity matrix was obtained by computing the Euclidean distance between all possible pairs of feature frames. Novelty for each time point was computed via convolution between each self-similarity matrix and a Gaussian checkerboard kernel (Foote, 2000) with half width of 11 s. Large kernel sizes have been previously used to overcome high levels of detail in novelty curves (Hartmann et al., 2016a; Liem, Bazzica, & Hanjalic, 2013; Klien et al., 2012).

As done with the perceptual density curves, we truncated the novelty curves by trimming 6.4 s from each extreme to avoid edge effects. We chose the smallest value that would eliminate, for all stimuli, any novelty spikes caused by the contrast between music and silence in the beginning and end of tracks; trimming the extremes of the novelty curves also increased the number of novelty points that derive from a full checkerboard kernel. Once the novelty curves were computed, we also trimmed 6.4 s from the extremes of each dimension of the musical features in order to obtain the predictors of accuracy described below.

2.4 Characterization of Musical Features and Novelty Curves

From each musical feature matrix $\mathbf{F}$ we calculated mean *Feature Flux*, an estimate of the amount of local variation; Feature Flux is the Euclidean distance between successive feature frames. First, for each time series $\mathbf{F}_d$, where d corresponds to a feature dimension, the squared difference between successive time points is obtained. Next, a flux time series $\mathbf{v}$ is obtained as the squared root of the sum across dimensions:

$$\mathbf{v_t} = \sqrt{\sum_{n=1}^{N} (\mathbf{F}_d(t) - \mathbf{F}_d(t-1))^2}$$

Finally, mean Feature Flux is obtained by averaging the flux time series $\mathbf{v}$ across time points:

$$\textit{Feature Flux} = \frac{1}{K} \sum_{t=1}^{K} \mathbf{v}_t$$

From each novelty curve, we obtained *Mean Distance Between Subsequent Peaks* (MDSP), which describes the peak-to-peak duration (in seconds) of novelty curves. To compute this estimate, we first obtained from the novelty curve a vector of novelty peak locations $\mathbf{v}$, where $\mathbf{v}_i$ corresponds to the i^{th} peak; MDSP was calculated as follows:

$$MDSP = \frac{1}{N} \sum_{i=1}^{N} (v_i - v_{i-1})$$

2.5 Segmentation Accuracy and its Correlates

We compared perceptual segmentation density and novelty curves to obtain segmentation accuracy. To this end, we performed correlations between novelty and perceptual segmentation density for each stimulus and musical feature.

We focused on the possible relationship between accuracy and the aforementioned characterizations of musical features and novelty peaks. Hence, for each feature we correlated across stimuli the accuracy with Feature Flux and with MDSP. In order to perform these correlations, the accuracies of each feature required to follow an approximately normal distribution, so we subsequently transformed accuracies via Fisher's z transformation of r. The normalization of Z_r involved the calculation of effective degrees of freedom to correct for temporal autocorrelation (Alluri et al., 2012; Pyper & Peterman, 1998).

3 Results

3.1 Prediction of Perceptual Segmentation Density

Figure 2 shows the correlation between novelty curves and perceptual boundary density for each stimulus; Appendix III includes graphs comparing perceptual density curves and novelty curves for each musical feature and stimulus. The prediction accuracies were found to vary depending on stimulus; for instance, *Smetana* yielded very high correlations, whereas these were rather low for *Couperin*. Also, for any given stimulus, accuracy differed according to the feature and feature type used for novelty detection; for instance, *Smetana* yielded higher accuracy for pitch-based features than for rhythmic and spectral-based features. Interestingly, no single novelty feature successfully predicted perceptual boundary density for all stimuli.

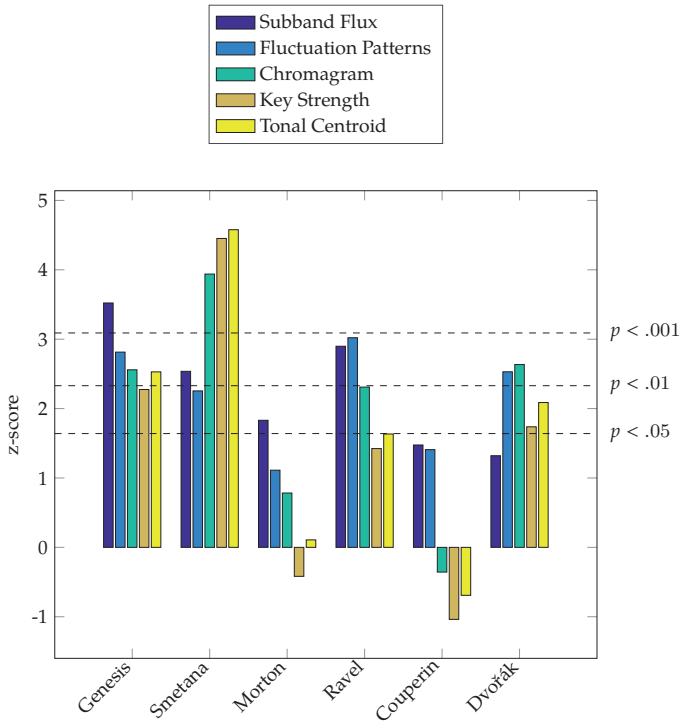

Figure 2: Correlation between novelty curves and perceptual boundary density.

At this point, it might be relevant to illustrate the data analyzed in this study and at the same time explore two musical stimuli that were found to be contrasting with respect to their prediction. Figure 3 visually compares representations for two stimuli that yielded optimal and worst accuracies for Tonal Centroid, *Smetana* and *Couperin*, respectively. The upper graphs show the Tonal Centroid time series for each stimulus; dimensions 1-2, 3-4 and 5-6 correspond to pairs of numerical coordinates for the circle of fifths, minor thirds, and major thirds, respectively. The middle graphs present novelty curves for Tonal Centroid, which result from the representations in the upper graphs. The lower graph shows perceptual segmentation density based on listeners' segmentation. Some clear differences can be seen between the profiles for *Smetana* and *Couperin*. The three most contrasting peaks of perceptual density for *Smetana* correspond to changes of key: at 40 s the music changes from an F key to a fugue in G min, at 66 s the F melody is reprised in C♯, and at 104 s begins a similar fugue transposed to F min. These changes are quite clear in the Tonal Centroid representation, yielding a relatively adequate novelty prediction, although the rather gradual transition at around 66 s does not yield a stark novelty peak.

In contrast, the Tonal Centroid time series of *Couperin* does not allow for a straightforward acoustic interpretation of the perceptual segmentation density profile. According to the peaks of perceptual density, listeners seem to base their indications primarily on the ending of cadences; these are characterized by the use of heavy ornamentation (e.g. mordents) and chords, which are rather salient because the piece is almost exclusively two-voiced. Due to these cues, listeners seem to place more segment boundaries on endings of melodies than on beginnings, for instance at 23 s. In comparison to this, Tonal Centroid and its corresponding novelty curve would describe slightly delayed musical changes, as harmonic transitions become apparent only after enough development of subsequent melodic material. Another reason behind the highly inaccurate prediction is that listeners placed boundaries for changes of register, rests and durational changes, which

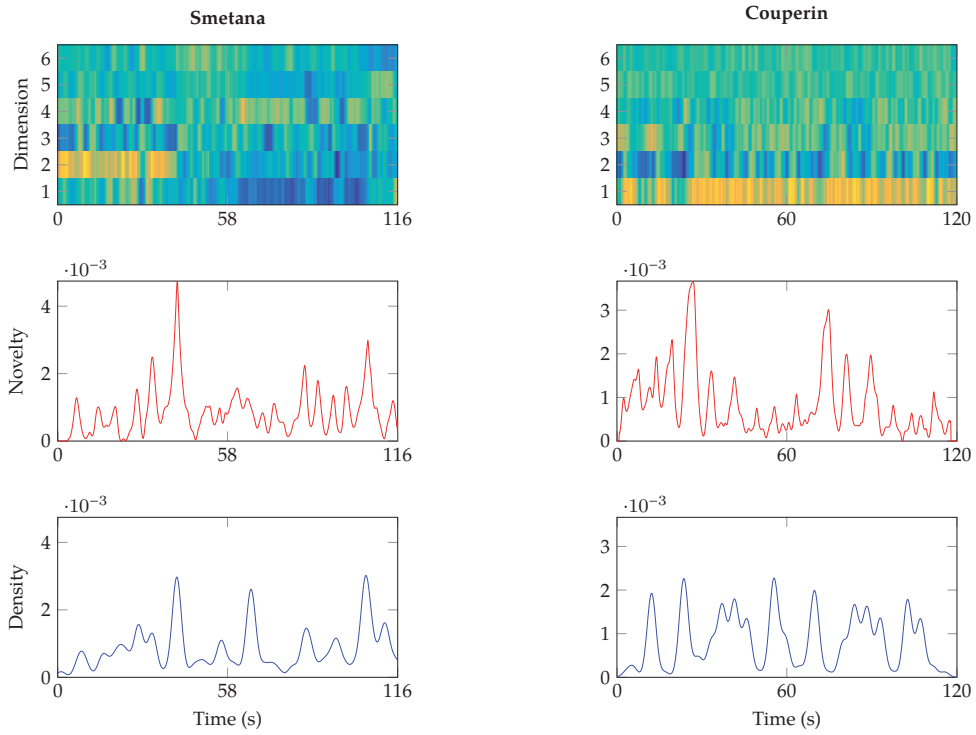

Figure 3: Tonal Centroid, novelty of Tonal Centroid and perceptual boundary density for stimuli *Smetana* and *Couperin*.

Musical Feature	Feature Flux	MDSP
Subband Flux	.54	–.03
Fluctuation Patterns	–.30	.11
Chromagram	–.68	.27
Key Strength	**–.74**	.47
Tonal Centroid	–.66	**.50**

Table 1: Correlation between z-transformed accuracy and characterizations of musical features (Feature Flux) and novelty curves (MDSP).

might partly explain the higher accuracies obtained for spectral and rhythmic features.

3.2 Finding Predictors of Segmentation Accuracy

Subsequently, we analyzed characterizations of extracted features and of novelty curves derived therefrom, looking for correlates of accuracy. We focused on Feature Flux, a global estimate of the extracted features, and on MDSP, which was obtained from novelty curves, to find whether or not these would be indicative of novelty curve accuracy. Table 1 shows the correlation between segmentation accuracies and the obtained characterizations of musical features and novelty curves. We found a strong negative correlation between accuracy and Feature Flux for pitch-based features; as regards accuracy and MDSP, we obtained moderate to strong positive correlations for tonal features (*strong* and *moderate* mean $|r| > .5$ and $.3 < |r| < .5$ respectively, following Cohen, 1988). Although these results did not reach statistical significance at $p < .05$, some interpretations can still be made. According to the results, accuracy increases for stimuli with fewer local change in pitch content and less peaks in pitch-based novelty curves. A similar pattern of results was found for rhythm; we obtained for Fluctuation Patterns a moderate negative correlation between Feature Flux and accuracy, and a weak positive correlation between MDSP and accuracy. Timbre seemed to yield an opposite trend, at least for Feature Flux; Subband Flux exhibited a strong positive correlation between Feature Flux and accuracy, and no or very weak correlation between MDSP and ac-

curacy. This suggests that high accuracy is associated with more local variability of spectral fluctuation.

Next, we focused on the highest correlations for Feature Flux and MDSP, particularly on which stimuli occupied the first and last ranks for segmentation accuracy and value of feature characterizations. Our aim was to further understand which specific musical characteristics were associated with the accuracy yielded by the features. Figure 4 shows scatter plots of segmentation accuracy (i.e. standard score) as a function of characterizations for each feature. The strongest correlation for Feature Flux was exhibited by Key Strength, suggesting that this feature is a good predictor of segmentation when the local variability therein is low. *Genesis*, which obtained the lowest Feature Flux for Key Strength, is characterized by rather long sections containing few tonal changes, or sections without harmonic accompaniment. Its rather high accuracy may stem from the fact that perceived changes often coincide with chord progressions; this interpretation also applies to *Smetana*, which exhibited maximum accuracy. This suggests that Key Strength yields higher accuracy for music characterized by stable sections with respect to tonal center. In *Couperin*, the piece with highest Feature Flux and lowest accuracy for Key Strength, harmonic changes are frequent, and chords with primary harmonic functions only rarely appear one after another (this would lead to weak cadential effects). Other types of changes seem to guide boundary perception in this piece instead, such as stark durational change, ornamentation, change of register, voicing and repetition. This suggests that music with frequent tonal center changes result in low accuracy for Key Strength because listeners may focus their attention on musical dimensions that change less often.

In addition, we obtained a high positive correlation between accuracy and MDSP for Tonal Centroid. This suggests that Tonal Centroid yields higher accuracy for music with few novelty peaks from this feature, such as in the case of *Dvořák*, which includes relatively few chords of long duration. Conversely, *Couperin*, which is characterized by frequent chord progressions, exhibited low accuracy and low MDSP for Tonal Centroid. This result suggests that high amount of tonal change may lead to a decrease in the perceived strength of this type of change.

We were particularly interested in the correlations obtained for Subband Flux, since an opposite trend was observed for this feature (Table 1). Notably, *Genesis* was the piece with highest Feature Flux, lowest MDSP, and highest accuracy for Subband Flux. High local Subband Flux variability for this stimulus may be due to the vast and diverse instrumentation (especially percussion) used in the piece; these constant, short-term spectral changes lead to multiple local dissimilarities in spectral fluctuation. Low distance between subsequent Subband Flux novelty peaks for *Genesis* seems to result from an overall flat novelty profile characterized by multiple peaks of mild contrast. Multiple spectral changes occurring within a rather short time for this musical example might, in turn, clarify the reason behind the opposite trend exhibited by Subband Flux in the correlation between accuracy and characterizations. Possibly, stimuli with high local variability for Subband Flux exhibit increased accuracy because only a small number of spectral changes yielded stark novelty peaks; these few peaks would correspond with changes at rather large times scales and, hence, with perceptual boundaries. In contrast, the piece with lowest Feature Flux for Subband Flux was *Morton*, possibly due to the low amount of local change as it is a piano piece with relatively dampened sound and low dynamic contrast. This piece however contains contrasting sequences, such as a change from a short monophonic part consisting of triplets to regular rhythm and chordal accompaniment; a stark novelty peak is generated at 75 s due to this change, but it does not correspond with high perceptual density because it is situated in the middle of a phrase. As regards MDSP, *Ravel* was the piece with highest MDSP for Subband Flux but did not yield low accuracy, for instance due to the dynamic changes of this piano piece (27 s and 75 s), which were accurately predicted.

4 Discussion

Understanding which specific aspects of musical pieces influence novelty-based segmentation prediction is a crucial but challenging issue. One possible way to address this problem is to focus on the particulars of this approach and tackle the question of what characteristics of musical features and their respective novelty curves predict segmentation accuracy for different musical pieces. This study tries to fill the gap in this respect, and aims to open a discussion on the possibility of predicting accuracy directly from musical feature

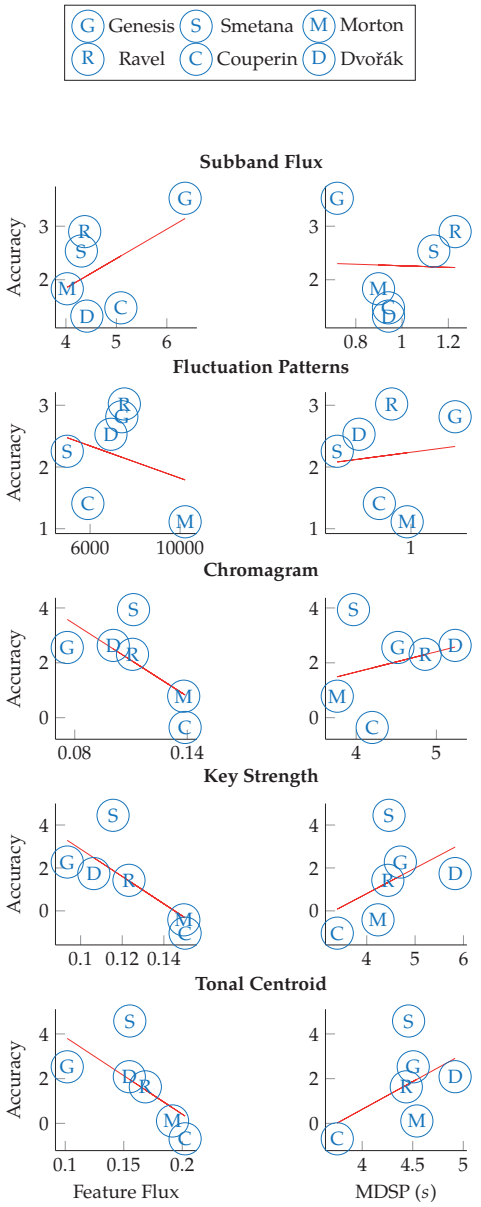

Figure 4: Scatter plots of segmentation accuracy as a function of characterizations of musical features and novelty curves.

characteristics; we believe that automatic segmentation systems could optimize their parameters for stimuli of different idioms and that such an approach would lead to an increase in accuracy.

Regarding the validity of the approach, we also highlight that our method for studying segmentation involves amassing annotations from multiple listeners for the same stimulus; this possibility has been previously explored only by few studies on segmentation prediction (Mendoza Garay, 2014; Hartmann et al., 2016a). In contrast, MIR studies on music segmentation are often based on data

coming from one or few annotators; the computation of a KDE is usually not needed since the set of annotated segment boundaries is directly compared against peaks picked from a novelty curve. In this sense, analyses of perceptual segmentation based on data that is probably more representative of the musician population should be useful for better understanding both perception and prediction of musical structure.

This section examines our research questions in light of the proposed analysis, and assesses the extent to which the stated hypotheses could be supported. Finally, we conclude this article with possible directions for future research.

4.1 Segmentation Accuracy

The first step to address the research questions was to investigate the accuracy to predict perceived boundaries yielded by different musical features for different musical examples. As shown in Figure 2, accuracy seems to highly vary according to musical piece, motivating further analyses on novelty detection that focus on each piece separately. It is also apparent that no single algorithm is robust for prediction of all examples, suggesting the importance of incorporating combinations of multiple features (Hartmann et al., 2016a; Turnbull et al., 2007; Müller, Chew, & Bello, 2016), multiple time scales (Kaiser & Peeters, 2013; Gaudefroy et al., 2015), and other aspects of segmentation (e.g. repetition principles, see Gaudefroy et al., 2015; Paulus & Klapuri, 2006) into novelty approaches. Overall, however, the results seem to support the idea that performance in structural analysis heavily may depend more on musical stimuli than on the algorithm used or the choice of parameters (Peiszer et al., 2008), which could serve to justify subsequent steps of our analysis.

Exploring musical and perceptual factors that contribute to low accuracy, we found that novelty prediction often fails due to issues related to temporal imprecision of perceptual boundaries, and to knowledge-driven and other biases of attention from listeners: i) they seem to focus on style specific, 'decorative' aspects of the music, for instance in *Couperin* they are biased towards ornamentation instead of chord changes when these are too frequent, whereas tonal novelty peaks would only emerge from chords; ii) listeners' selective attention tends to privilege specific musical dimensions, e.g. musical changes triggering novelty peaks may not prompt boundary perception if these occur in the middle of a ritardando (*Ravel*, 60 s) or a melodic phrase (*Morton*, 114 s, and *Dvořák*, 75 s); iii) there may be clear temporal gaps between endings and beginnings of melodies, causing a disjunction between actual and predicted segmentation time points, for instance in *Couperin* listeners often segment melodic endings as they are probably motivated by cadences (89 s), whereas novelty peaks frequently appear at beginnings due to introduction of new material (90 s).

4.2 Feature-based Prediction of Novelty Detection Performance

Our first hypothesis was that accuracy obtained via novelty detection would increase for stimuli with low local variation of musical features. In support with this hypothesis, we overall found Feature Flux to be a good predictor of correlation between perceptual segmentation density and novelty (Table 1), and a same pattern of results for pitch-based and rhythmic features. This suggests, for these features, that increased local feature continuity may be indicative of higher accuracy of novelty curves.

The second hypothesis of this study was that accuracy would increase for stimuli with more distant novelty peaks. Indeed, mean distance between subsequent peaks was found to be somewhat indicative of the accuracy of novelty curves with respect to perceptual segmentation density. This suggests, particularly for tonal features, that longer novelty peak-to-peak duration is associated with higher correlation between detected novelty and perceptual segment density.

Focusing on pitch-based and rhythmic features, the results indicate that high local variability is associated with low accuracy. A possible interpretation is that music characterized by few local changes in pitch or rhythm often involves few, highly contrasting pitch-based or rhythmic structural boundaries. For instance, rhythmically stable melodies are clearly separated by highly discernible rests and long notes in *Dvořák*. This could be interpreted in the light of theoretical approaches to musical expectation (Narmour, 1992), according to which similarity between successive events generates the expectation of another similar event. If eventually this expectation is not satisfied, a sense of closure may be perceived, prompting the indication of a segment boundary; this may explain why, for example, accurate tonal-based segmentation predictions exhibit low vari-

ability at a local level and often involve few, perceptually stark boundaries that delimit homogeneous groupings of events. In addition, the idea that a segment boundary situated after an homogeneous sequence tends to be perceived as strong also resembles an assumption used by Bohak and Marolt (2016) for automatic segmentation: any time point that follows a region of loudness below a given threshold is a segment boundary candidate, and the larger the preceding region of low loudness is, the higher the likelihood of placing a boundary at that time point will be.

Related to our previous result, we found that novelty curve characteristics can be used as predictors of accuracy: larger distance between subsequent novelty peaks was found to result in higher novelty accuracy (Table 1), especially for tonal features. As aforementioned, this relationship relates to the idea of maximum and minimum properties of "good" organizations proposed by Gestalt theorists (Koffka, 1935). In this sense, music characterized by novelty peaks that are clearly isolated should yield higher accuracy as they would relate to perceptually salient musical changes.

We should highlight that the features yielding highest correlations with accuracy were tonal. It is possible that interpretations derived via perceptual organization rules and expectation violation are better applicable to the case of prediction via tonal features because these features focus unambiguously on changes in perceived tonal context (and not on e.g. loudness changes). In contrast, other descriptions used are somewhat more vague: i) Subband Flux discontinuities encompass changes of instrumentation, register, voicing, articulation, and loudness; ii) Fluctuation Pattern changes could be attributed to rhythmic patterns, tempo, articulation, and use of repetition; iii) changes in Chromagram are manifested in pitch steps, pitch jumps, and use of chords. In this respect, tonal features consider a single dimension of musical change, whereas other features analyzed in this study may yield more intricate descriptions.

Accuracy seemed to increase for stimuli with little local change and more distance between peaks (Table 1), however Subband Flux seemed to yield an opposite trend. In this regard, high local variability of changes in instrumentation, register, loudness, etc., seems to be associated with higher accuracy. It could be the case that musical pieces with high local spectral change and multiple novelty peaks are also characterized by few structural sections of long duration, and yield a relatively straightforward prediction; for instance, *Genesis* contains multiple short sounds and effects, yet its sections are clearly delimited by important instrumentation changes, which probably had a positive effect on accuracy. As a matter of fact, Figure 4 shows that *Genesis* might have largely influenced the sign of the Subband Flux correlation results for both Feature Flux and MDSP. Again, stylistic information might help to disentangle these and other problems regarding segmentation accuracy.

4.3 General Discussion

One of the main aims of this study was to find out methods to select musical features that would be efficient in segmentation prediction for a given stimulus; to this end, we investigated the relationship between accuracy and characterizations of musical features and novelty curves. According to the results, for most features there is an inverse relationship between local variability and accuracy, and a direct relationship between mean distance between subsequent novelty peaks and accuracy. This suggests that stimuli whose features are characterized by low variation between successive time points, and whose novelty curves have few peaks, are likely to yield higher segmentation prediction accuracy. A possible reason that explains these results is that music with infrequent musical change often yields perceptually salient boundaries; according to the Gestalt rules of perceptual organization, similar events that are proximal in time are grouped together, creating a strong sense of closure whenever a dissimilar event is perceived. Following this interpretation, if a given musical dimension changes frequently, an increase of listeners' attention towards other dimensions evoking strong closure may occur during segmentation.

4.4 Considerations for Further Research

Since this study focused on the analysis of segmentations of the same musical pieces from multiple listeners, the number of segmented stimuli does not suffice to draw solid conclusions about the correlates of segmentation accuracy; this should be considered a major methodological caveat. More musical stimuli are clearly needed to assess the generalization ability of our results; future studies should in this respect increase the number of musical stimuli used for perceptual segmentation tasks

while maintaining a satisfactory participant sample size, and test whether statistically significant results are obtained for a high number of degrees of freedom. As regards the sample of participants used, this study only focused on musician listeners, mainly following MIR studies, which recruit expert annotators for the preparation of musical structure data sets. Further work should concentrate on annotation segmentation from nonmusicians in order to understand accuracy of novelty curves with regards to the majority of the population.

Another issue to consider is that the novelty detection approach is designed to yield maximum scores for high continuity within segments and high discontinuity at segmentation points, so in this sense it is not surprising that music exhibiting clear discontinuity between large sequences of homogeneity for a given feature will yield higher segmentation accuracy for that feature. In this regard, our results should be further tested using other approaches; for instance probabilistic methods (e.g. Pauwels, Kaiser, & Peeters, 2013; Pearce & Wiggins, 2006) would be suitable as they offer alternative assumptions regarding location of actual boundaries.

Also, the positive relationship between Subband Flux and characterizations could be regarded as an inconclusive outcome, and deserves further examination. As aforementioned, it is possible that this relationship only holds for musical styles characterized by a heavy use of different kinds of percussion and sound effects, and few structural sections.

In addition, other potential predictors of accuracy should be systematically analyzed to better understand how accuracy can be implied from musical features and novelty curves. For instance, the mean 'peakiness' or slope of novelty curves and also their entropy could be studied; based on the obtained results, it would be expected that stimuli yielding higher prediction accuracy would be characterized by large novelty entropy and slope.

We also highlight that novelty curve accuracy could be further explored by estimating false positive rate and false negative rate; for instance, the area between curves shown in Appendix III could be analyzed to assess the tendency towards overestimation or underestimation of perceptual segmentation density. The suitability of this approach can be compared against standard performance measures that are based on peak picking.

Further examination is needed in order to understand the extent to which accuracy decreases due to the challenges imposed to different participants by particular stimuli; level of agreement between listeners and optimal time scale for density estimation are issues to consider because they should have an impact on perceptual segmentation density curves and hence on prediction. These issues are not only relevant to perceptual segmentation density studies rooted on music perception; questions regarding optimal representation of judgments from various listeners and at multiple segmentation time scales also reach MIR studies, which have recently shifted their focus towards the design of new segmentation ground truths (Nieto, 2015; Smith, Burgoyne, Fujinaga, De Roure, & Downie, 2011; Peeters & Deruty, 2009) that include multiple annotators, musical dimensions, and temporal scales.

Finally, as an outcome of this study it can be stated that listeners tend to focus on musical dimensions that do not change often. This interpretation is plausible and highlights the importance of e.g. tonal and tempo stability, as well as the role of repetition and motivic similarity in musical pieces. Future work should test this possibility by conducting listening studies in which listeners would describe what is the most salient dimension for different time points in the music; further, as suggested by Müller et al. (2016), automatic detection of these acoustic description cues should also be a relevant task regarding structural segmentation.

Acknowledgements The authors would like to thank Emily Carlson for proofreading the paper. This work was financially supported by the Academy of Finland (project numbers 272250 and 274037).

References

Ahlbäck, S. (2007). Melodic similarity as a determinant of melody structure. *Musicae Scientiae*, *11*(1 suppl), 235–280.

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

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

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*(4), 3677–3689.

Bartsch, M. A. & Wakefield, G. H. (2001, October). To catch a chorus: using chroma-based representations for audio thumbnailing. In *Signal processing* (Vol. 1001, pp. 15–18). IEEE.

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

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047.

Bohak, C. & Marolt, M. (2016). Probabilistic segmentation of folk music recordings. *Mathematical Problems in Engineering*, *2016*.

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

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

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference* (pp. 17–22).

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

Chew, E. (2002). The spiral array: an algorithm for determining key boundaries. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and artificial intelligence* (pp. 18–31). Edinburgh: Springer.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd). New Jersey: Lawrence Erlbaum.

De Lisa, G. (2009). Smetana: má vlast. Retrieved from http : / / genedelisa . com / 2009 / 03 / smetana-ma-vlast/

Eerola, T. (2011). Are the emotions expressed in music genre-specific? an audio-based evaluation of datasets spanning classical, film, pop and mixed genres. *Journal of New Music Research*, *40*(4), 349–366.

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

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

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

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

Gaudefroy, C., Papadopoulos, H., & Kowalski, M. (2015). A multi-dimensional meter-adaptive method for automatic segmentation of music. In *13th International Workshop on Content-Based Multimedia Indexing (CBMI)* (pp. 1–6). IEEE.

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

Hartmann, M., Lartillot, O., & Toiviainen, P. (2016a). Interaction features for prediction of perceptual segmentation: effects of musicianship and experimental task. *Journal of New Music Research*.

Hartmann, M., Lartillot, O., & Toiviainen, P. (2016b). Multi-scale modelling of segmentation: effect of musical training and experimental task. *Music Perception*, *34*(2), 192–217.

Jensen, K. (2007). Multiple scale music segmentation using rhythm, timbre, and harmony. *EURASIP Journal on Applied Signal Processing*, *2007*(1), 159–159.

Kaiser, F. & Peeters, G. (2013). Multiple hypotheses at multiple scales for audio novelty computation within music. In *Proceedings of the*

International Conference on Acoustics, Speech, and Signal Processing. Vancouver.

Klien, V., Grill, T., & Flexer, A. (2012). On automated annotation of acousmatic music. *Journal of New Music Research*, *41*(2), 153–173.

Koffka, K. (1935). *Principles of Gestalt psychology*. New York: Harcourt, Brace and Company.

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

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

Lartillot, O., Eerola, T., Toiviainen, P., & Fornari, J. (2008). Multi-feature modeling of pulse clarity: design, validation and optimization. In *Proceedings of the 9th International Conference on Music Information Retrieval* (pp. 521–526). Citeseer.

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

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

Liem, C. C. S., Bazzica, A., & Hanjalic, A. (2013). Looking beyond sound: unsupervised analysis of musician videos. In IEEE (Ed.), *Proceedings of the 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*. Paris, France.

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

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), 150081.

McFee, B. & Ellis, D. P. (2014). Learning to segment songs with ordinal linear discriminant analysis. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5197–5201).

McFee, B., Nieto, O., & Bello, J. P. (2015). Hierarchical evaluation of segment boundary detection. In *Proceedings of the 15th conference of the International Society for Music Information Retrieval*.

Mendoza Garay, J. (2014). *Self-report measurement of segmentation, mimesis and perceived emotions in acousmatic electroacoustic music* (Master's thesis, University of Jyväskylä).

Müller, M., Chew, E., & Bello, J. P. (2016). Computational Music Structure Analysis (Dagstuhl Seminar 16092). *Dagstuhl Reports*, *6*(2), 147–190. doi:http : / / dx . doi . org / 10 . 4230 / DagRep.6.2.147

Narmour, E. (1992). *The analysis and cognition of melodic complexity: the implication-realization model*. University of Chicago Press.

Nieto, O. (2015). *Discovering structure in music: automatic approaches and perceptual evaluations* (Doctoral dissertation, New York University).

Nieto, O., Farbood, M. M., Jehan, T., & Bello, J. P. (2014). Perceptual analysis of the F-measure for evaluating section boundaries in music. In *Proceedings of the 15th Conference of the International Society for Music Information Retrieval*.

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

Paulus, J. & Klapuri, A. (2006). Music structure analysis by finding repeated parts. In *Proceedings of the 1st ACM workshop on audio and music computing multimedia* (pp. 59–68).

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

Pauwels, J., Kaiser, F., & Peeters, G. (2013). Combining harmony-based and novelty-based approaches for structural segmentation. In *Ismir* (pp. 601–606).

Pearce, M. T. & Wiggins, G. (2006). The information dynamics of melodic boundary detection. In *Proceedings of the ninth international conference on music perception and cognition* (pp. 860–865).

Peeters, G. (2004). Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach. *Computer Music Modeling and Retrieval*, 169–185.

Peeters, G. & Deruty, E. (2009). Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proc. of 3rd workshop on learning the semantics of audio signals* (pp. 75–90).

Peiszer, E. (2007). *Automatic audio segmentation: segment boundary and structure detection in popular music* (Master's thesis, Vienna University of Technology).

Peiszer, E., Lidy, T., & Rauber, A. (2008). Automatic audio segmentation: segment boundary and structure detection in popular music. In *Proceedings of the 2nd International Workshop on Learning the Semantics of Audio Signals (LSAS)*. Paris, France.

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

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

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

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

Sonntag, C. M. (2011). *Jeux d'eau and its colleagues: water and artistic expression at the turn of the 20th century* (Master's thesis, Ball State University).

Šupka, O. (2013). Slavonic dances for orchestra, series I. Retrieved from http://www.antonin-dvorak . cz / en / slavonic - dances1 - for - orchestra

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

Trythall, R. (2002, Summer). "Jelly Roll Blues", observations on performance practice. Retrieved from http://www.richardtrythall.com/33.html

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

A Appendices

A.1 Musical Stimuli - List of Abbreviations

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

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

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

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

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

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

A.2 Extracted Musical Features

Novelty curves were obtained from similarity matrices of musical features. To this end, the following five musical features describing spectral, rhythmic, pitch-related and tonal attributes were extracted from the musical signal:

Spectral

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

Rhythmic

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

Pitch Class

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

Tonal

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

A.3 Perceptual Density and Novelty Curves

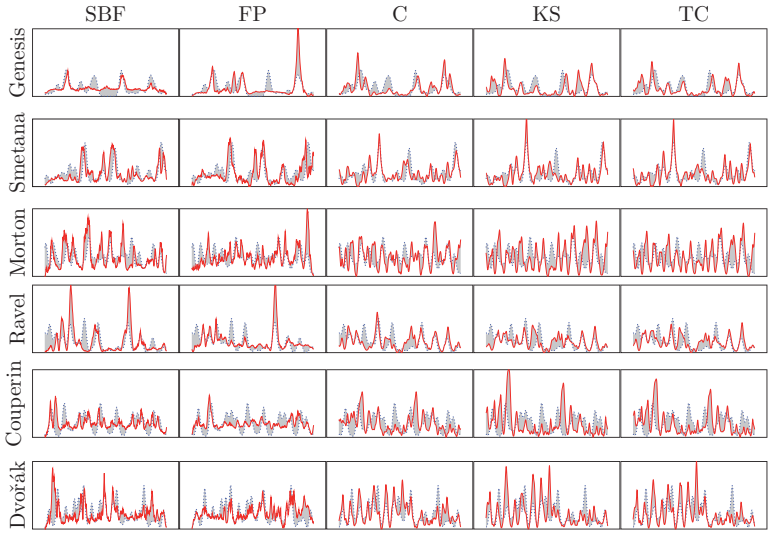

Figure 5: Perceptual boundary density plotted against novelty curves for each stimulus and musical feature. Perceptual segment boundary density curves are indicated with a dotted line; novelty curves are indicated with a solid line. Also the area between curves is colored in the figure.