Explaining Listener Differences in the Perception of Musical Structure

by

Jordan B. L. Smith

A thesis submitted to the University of London for the degree of
Doctor of Philosophy

Department of Electronic Engineering & Computer Science
Queen Mary, University of London
United Kingdom

September 2014

To my friends and family

Statement of Originality

I, Jordan B. L. Smith, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Jordan B. L. Smith

Date: 1 September 2014

Details of collaboration and publications:

Three published articles have been incorporated into the thesis, and form the basis of Chapters 3–5. Portions of each article also appear in Chapters 1–2, the Introduction and Literature Review. I am the first author on each of these published articles. The citations for these articles (and some additional refereed presentations on the same subjects) are listed here, as well as the major contributions of the other authors.

Chapter 3

- Smith, J. B. L., I. Schankler, and E. Chew. 2014. Listening as a creative act: Meaningful differences in structural annotations of improvised performances. *Music Theory Online*. 20: 3. [SSC14]
- Smith, J. B. L., I. Schankler, and E. Chew. 8 August 2013. "Why do listeners disagree about large-scale formal structure? A case study." Talk at the biennial meeting of the Society for Music Perception and Cognition, Toronto, ON, Canada.

Schankler performed the music which was the foundation of the study, participated as the other subject of the case study, and wrote portions of article related to the circumstances of the performance and the functioning of Mimi. He also edited the rest of the text with Chew.

Chapter 4

- Smith, J. B. L., C.-H. Chuan, and E. Chew. 2014. Audio properties of perceived boundaries in music. *IEEE Transactions on Multimedia* 16: 5. 121928. [SCC14]
- Smith, J. B. L., C.-H. Chuan, and E. Chew. 9 August 2013. "Learning about structural analysis from structural analyses." Talk at the biennial meeting of the Society for Music Perception and Cognition, Toronto, ON, Canada.

- Smith, J. B. L., C.-H. Chuan, and E. Chew. 18 December 2012. "Boundaries and novelty: the correspondence between points of change and perceived boundaries." Presentation at Digital Music Research Network One-day Workshop, London, UK.

Chuan provided the centre of effect computation for the corpus.

Chapter 5

- Smith, J. B. L., and E. Chew. 2013. Using Quadratic Programming to estimate feature relevance in structural analyses of music. *Proceedings of the ACM International Conference on Multimedia*. 113-22. [SC13b]

Abstract

State-of-the-art models for the perception of grouping structure in music do not attempt to account for disagreements among listeners. But understanding these disagreements, sometimes regarded as noise in psychological studies, may be essential to fully understanding how listeners perceive grouping structure. Over the course of four studies in different disciplines, this thesis develops and presents evidence to support the hypothesis that attention is a key factor in accounting for listeners' perceptions of boundaries and groupings, and hence a key to explaining their disagreements.

First, we conduct a case study of the disagreements between two listeners. By studying the justifications each listener gave for their analyses, we argue that the disagreements arose directly from differences in attention, and indirectly from differences in information, expectation, and ontological commitments made in the opening moments. Second, in a large-scale corpus study, we study the extent to which acoustic novelty can account for the boundary perceptions of listeners. The results indicate that novelty is correlated with boundary salience, but that novelty is a necessary but not sufficient condition for being perceived as a boundary. Third, we develop an algorithm that optimally reconstructs a listener's analysis in terms of the patterns of similarity within a piece of music. We demonstrate how the output can identify good justifications for an analysis and account for disagreements between two analyses.

Finally, having introduced and developed the hypothesis that disagreements between listeners may be attributable to differences in attention, we test the hypothesis in a sequence of experiments. We find that by manipulating the attention of participants, we are able to influence the groupings and boundaries they find most salient. From the sum of this research, we conclude that a listener's attention is a crucial factor affecting how listeners perceive the grouping structure of music.

Acknowledgments

I am immeasurably grateful to my advisor, Dr. Elaine Chew, and to my secondary advisor, Dr. Marcus T. Pearce.

Many thanks to my thesis examiners, to reviewers of previous papers, to colleagues past and present, and to my friends and family.

This research was supported by: the Social Sciences and Humanities Research Council; a PhD studentship from Queen Mary University of London; a Provost's Ph.D. Fellowship from the University of Southern California. This material is also based in part on work supported by the National Science Foundation under Grant No. 0347988.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Social Sciences and Humanities Research Council, Queen Mary University of London, University of Southern California, or the National Science Foundation.

Table of Contents

Abstract ii

Acknowledgments iii

Table of Contents iv

List of Figures viii

List of Tables xi

List of Abbreviations xiii

1 Introduction 1

1.1 Grouping structure and listener disagreements 1

1.2 Thesis scope 3

1.3 Thesis outline 4

2 Structure in music theory, psychology, and information retrieval 8

2.1 Music theory, music psychology and MIR 9

2.2 Models of the perception of grouping structure 12

2.2.1 Rule-based generative models 12

2.2.2 Models of expectation 13

2.3 Structural analysis in MIR 15

2.3.1 Listener disagreements in MIR 17

2.3.2 Evaluation and annotation 18
2.4 Bottom-up versus top-down 20
2.5 Listener disagreements 23
2.5.1 Accounting for listener disagreements 24

3 Causes of variation among listeners in boundary and grouping decisions 27

3.1 Procedure 29
3.1.1 Choice of materials 30
3.1.2 Annotation procedure 33
3.2 Differences between annotations 34
3.2.1 Performance no. 1 34
3.2.2 Performance no. 2 38
3.2.3 Performance no. 3 42
3.3 Discussion 46
3.3.1 Factor 1: Attention to musical features 46
3.3.2 Factor 2: Opening moments 48
3.3.3 Factor 3: Difference in information 49
3.3.4 Factor 4: Difference in analytical expectations 51
3.3.5 Factor 5: Analysis method 52
3.4 Conclusion and future work 52

4 An analysis of boundary perception and musical features 58

4.1 Introduction 59
4.1.1 Background 59
4.1.2 Proposed experiment 62
4.2 Materials and methods: the SALAMI dataset 63
4.2.1 Participants and apparatus 64
4.2.2 Stimuli 64
4.2.3 Procedure 65

4.3 Data analysis . . . 67
4.3.1 Audio processing . . . 68
4.3.2 Generating novelty functions and picking peaks . . . 70
4.3.3 Random baseline . . . 71
4.3.4 Analysis metric . . . 71
4.4 Results and statistical analysis . . . 72
4.4.1 Are boundaries points of novelty? . . . 72
4.4.2 Do any boundaries not match a novelty peak at all? . . . 79
4.4.3 Can boundary salience be estimated by annotation concurrence? . . . 81
4.5 Conclusion and future work . . . 84

5 Relating grouping structure to musical features 88
5.1 Introduction . . . 89
5.1.1 Previous methods in SSM calculation . . . 89
5.1.2 Limitations of combining SSMs . . . 90
5.2 Proposed method . . . 92
5.2.1 Self-similarity matrix calculation . . . 92
5.2.2 Combining SSMs . . . 96
5.2.3 Reconstructing an annotation SSM from audio SSMs . . . 101
5.3 Visualizing structural differences . . . 108
5.3.1 Investigating a single difference . . . 109
5.3.2 Reconstructing dissimilar analyses . . . 111
5.4 Conclusion and future work . . . 113

6 The effect of attention on grouping decisions 117
6.1 Introduction . . . 118
6.1.1 The role of attention . . . 118
6.1.2 Proposed experiments . . . 120
6.2 Method . . . 123
6.2.1 Participants . . . 123

6.2.2 Material 124
6.2.3 Procedure 127
6.3 Results 134
6.3.1 Experiment no. 1: Change identification 134
6.3.2 Experiment no. 2: Salience judgements 139
6.3.3 Experiment no. 3: Pattern detection & grouping preference 145
6.3.4 Experiment no. 4: Analysis continuation 158
6.4 Discussion 160
6.5 Conclusion and future work 166

7 Conclusion **168**
7.1 Summary 168
7.2 Limitations 170
7.3 Future work 171
7.4 Summary of key contributions 174

Bibliography **176**

Appendix A Gallery of Web Pages from Experiments in Chapter 6 **191**

Appendix B Computed Models for Experiments in Chapter 6 **195**

List of Figures

3.1 Isaac Schankler performing on a Yamaha Disklavier with Mimi 32
3.2 Analyses of Performance no. 1 35
3.3 Notation Example 1. Performance no. 1 36
3.4 Notation Example 2. Performance no. 1 37
3.5 Analyses of Performance no. 2 38
3.6 Notation Example 3. Performance no. 2 40
3.7 Notation Example 4. Performance no. 2 41
3.8 Analyses of Performance no. 3 42
3.9 Notation Example 5. Performance no. 3 43
3.10 Notation Example 6. Performance no. 3 43
3.11 Notation Example 7. Performance no. 3 44
3.12 Screenshot of Variations Audio Timeliner 53
3.13 Screenshot of Sonic Visualiser 53

4.1 Two annotations for the song "Ain't Too Proud To Beg" 66
4.2 Distribution of f-measure scores for boundaries and non-boundaries . . . 73
4.3 Distribution of f-measure contrast across annotators 76
4.4 Distribution of f-measure contrast across genres 77
4.5 Distribution of f-measure contrast across timescales 78
4.6 Distribution of f-measure contrast across features 79
4.7 Comparison of histograms for boundaries and non-boundaries 81

4.8 Comparison of boundary profiles from annotations and novelty functions for two songs 83

5.1 Self-similarity matrices derived from annotations for “Yellow Submarine” 93
5.2 Self-similarity matrices derived from recording of “Yellow Submarine” . . 95
5.3 Illustration of matrix components. 102
5.4 Optimal reconstruction coefficients of annotation of “Yellow Submarine” . 103
5.5 Optimal reconstruction of annotation of “Yellow Submarine” 105
5.6 Optimal reconstruction coefficients of annotation of “Yellow Submarine” (smaller scales) 107
5.7 Box plot of reconstruction costs 108
5.8 Annotation-derived SSMs by two listeners for “Garrotin” 109
5.9 Feature matrices for the reconstruction of “Garrotin.” 110
5.10 Optimal reconstruction coefficients for two annotations of“Garrotin.” . . . 111
5.11 Feature matrices for the reconstruction of “As the Bell Rings the Maypole Spins.” 112
5.12 Optimal reconstruction coefficients of two annotations of “As the Bell Rings the Maypole Spins” 114

6.1 Voice parts for the “HT-MR” environment 125
6.2 Example two-part stimulus 126
6.3 Example three-part stimulus 126
6.4 Example four-part stimulus 126
6.5 Example targets for Experiment no. 3 132
6.6 Example prompt for Experiment no. 4 133
6.7 Experiment no. 1: Main effect of musical training on accuracy 136
6.8 Experiment no. 1: Interaction plot of feature and music on accuracy . . . 136
6.9 Experiment no. 1: Interaction plot of exposure and music on accuracy . . 137
6.10 Experiment no. 2: Main effect plots of match condition and changing feature on salience 141

6.11 Experiment no. 2: Interaction effects of training, match condition and changing feature on salience 142
6.12 Experiment no. 2: Interaction effects between changing feature, match condition and musical environment on salience 144
6.13 Experiment no. 3: Main effect of focal feature on grouping agreement . . 148
6.14 Experiment no. 3: Main effects of age and musical environment on grouping agreement 149
6.15 Experiment no. 3: Interaction effect between feature and musical environment on grouping agreement. 150
6.16 Experiment no. 3: Main effect of relevance and presence on grouping confidence 152
6.17 Experiment no. 3: Main effects of training, age and focal feature on grouping confidence. 153
6.18 Experiment no. 3: Main effect of training on pattern detection accuracy . 154
6.19 Experiment no. 3: Main effects of focal feature, presence, musical environment and relevance on pattern detection 155
6.20 Experiment no. 3: Interaction effects between presence, musical environment and relevance on pattern detection 156
6.21 Experiment no. 3: Interaction effects between focal feature, musical environment and gender on pattern detection 157
6.22 Experiment no. 4: Main effects of training and age on continuation accuracy 161
6.23 Experiment no. 4: Main effects of musical environment, static feature and focal feature on continuation accuracy 162
6.24 Experiment no. 3: Main effect of training on grouping agreement 166

List of Tables

2-A The scope of different disciplines studying music structure 11

4-A Number of recordings analyzed according to genre and number of annotators 65

6-A Musical training survey questions 123
6-B Summary of experiment sequence 129
6-C Summary of variables in Experiment no. 1 134
6-D Significant effects in linear model for Experiment no. 1 135
6-E Significant effects in updated linear model for Experiment no. 1 138
6-F Summary of variables in Experiment no. 2 139
6-G Significant effects in linear model for Experiment no. 2 140
6-H Summary of variables in Experiment no. 3 145
6-I Answers provided in the pattern-detection task 146
6-J Significant effects in linear model for Experiment no. 3: Grouping preference 147
6-K Significant effects in linear model for Experiment no. 3: Answer confidence 150
6-L Significant effects in linear model for Experiment no. 3: Pattern recognition 151
6-M Summary of variables in Experiment no. 4 159
6-N Significant effects in linear model for Experiment no. 4 159

2-A Experiment no. 1, full model results, table 1 196
2-B Experiment no. 1, full model results, table 2 197
2-C Experiment no. 2, full model results 198

2-D Experiment no. 3, full model results: grouping preference 199
2-E Experiment no. 3, full model results: pattern detection 200
2-F Experiment no. 3, full model results: answer confidence (all trials) 201
2-G Experiment no. 4, full model results 201

List of Abbreviations

CE	Centre of Effect
FP	Fluctuation Pattern
GTTM	Generative Theory of Tonal Music
HMM	Hidden Markov Model
IDyOM	Information Dynamics of Music
IR	Implication Realization [model]
LBDM	Linear Boundary Detection Model
LMA	Live Music Archive
MFCC	Mel-frequency cepstral coefficient
MIDI	Musical Instrument Digital Interface
Mimi	Multimodal interface for musical improvisation
MIR	Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
QP	Quadratic Programming
SALAMI	Structural Analysis of Large Amounts of Music Information
SSM	Self-Similarity Matrix

Chapter 1

Introduction

In this chapter, we introduce the subject of music structure analysis, we describe the problem posed by disagreements between listeners, and we suggest that attention may have an important role in explaining them. This chapter outlines the aims and objectives of the rest of the thesis, and summarizes how the other chapters advance its overall goals. Since this work is very interdisciplinary, we take the time to explain the contrasts between each chapter.

1.1 Grouping structure and listener disagreements

The perception of grouping structure in music is one of the most fundamental and yet poorly understood aspects of listening. Grouping structure refers to how a listener divides a sequence of sounds into segments, and groups these segments together. At the shortest timescales, this process is automatic: a listener does not need to think consciously in order to perceive the beginning and end of each note in a piano piece. But these notes may be grouped into distinct gestures or motives, and these assembled into longer phrases, sections and so forth. Explaining how the mind accomplishes this feat of musical analysis is a formidable challenge, but could lead to a better understanding

of how humans detect patterns of arbitrary kinds, or help us to endow computers with the same abilities.

A listener's perception of structure is partly based on sensory information, as when a change in some musical feature, such as timbre, melody, or rhythmic pattern, signals a section boundary, and it is partly reliant on short-term memory, which allows one to notice when a sequence of musical events repeats. To the extent that these abilities are shared by most listeners, the perception of musical structure can be quite similar across listeners. However, it is also clear that individual factors—such as one's past musical training, one's familiarity with the given piece, and what one chooses to pay attention to—play a role. How significant are these factors, and how do they shape one's structural understanding of a piece? When, if ever, does their influence dominate over directly perceived features of the music, such as vivid timbre changes or proximate repetitions? Answering these questions is the original impetus for this thesis.

The act of perceiving groupings is sometimes referred to as *chunking*. Godøy, Jensenius and Nymoen [GJN10] encourage the terms *exogenous* and *endogenous* to refer to two contrasting influences on how a listener chunks what they hear: exogenous chunking is based on information outside the listener—namely, obvious changes in the music that imply boundaries between chunks; endogenous chunking originates from the listener's own knowledge and expectations, such as the expectation that a metrical pattern will continue. Crucially, each listener has some control over endogenous chunking, in that they may choose to focus on one aspect of the music or another. This terminology is synonymous with the dichotomy between passive, perceptual processes and active, analytical processes. Although exogenous influences are of primary importance at short timescales (less than 5 seconds), the longer the timescale, the more endogenous influences begin to dominate. Endogenous chunking is the less well understood kind, the kind more likely to lead to disagreements between listeners, and the kind that this thesis aims to investigate. In particular, we will study the importance of the intentional focus of the listener.

1.2 Thesis scope

The importance of music's structure is reflected in how its study spans many disciplines. In music theory, of course, form has been studied for centuries, and music theorists have developed many specialized ways of describing how music is composed and how it is heard—and, in some cases, prescribing how it ought to be heard to be best understood. The field of music psychology (encompassing, for our purposes, music perception and cognition and even some neuroscience), on the other hand, is concerned mainly with the hearing part, and seeks to explain how the brain makes sense of musical ideas. Whereas music theory mainly focuses on the hearings of educated listeners, music psychology often seeks to understand how music is understood by regular people, and how musical training and enculturation affect perception. Recently, engineers in music information retrieval (MIR) have taken up music structure analysis as an algorithmic challenge: the goal is to process an audio recording and estimate the structure that a human would perceive. Towards this end, some have developed new methods of transcribing structure—methods that resemble, in some ways, an extension of the original work by music theorists.

Although these fields all have separate aims, the concepts in each are related and illuminate each other. This thesis will therefore engage with all these fields to pursue the same central question: given a piece of music, how will a listener divide it into segments and group these segments into categories? And, in particular, how can we account for the differing perceptions of different listeners?

The four main chapters of this thesis (3–6) vary widely in scope (from case studies to moderate-sized experiments to large-scale corpus analyses), and each chapter investigates listener disagreements and attention in different ways. They also contribute to three different fields: music theory, music psychology and music information retrieval. The objective of each chapter, and their main contributions, are summarized below.

1.3 Thesis outline

Chapter 2: Structure in music theory, psychology, and information retrieval

Chapter 2 is a literature review that describes the kinds of models of grouping structure that have been developed and studied in three fields: music theory, music perception and cognition, and music information retrieval. It discusses the similar issues faced in each field, and shows the importance of listener disagreements in understanding these issues.

Chapter 3: Causes of variation among listeners in boundary and grouping decisions

How do listeners come to disagree about structural analyses? Chapter 3 presents a case study of the structural analyses of two listeners. The disagreements between them were studied to produce hypotheses about how the disagreements originated. The study's small scale allows an unusually deep analysis of the justifications each listener gave for their analyses. One of the main conclusions is that the focus of the listeners accounted most directly for the disagreements; as a result, much of the rest of the thesis deals with the focus of listeners. A version of the chapter was published in *Music Theory Online* [SSC14].

The contributions of this chapter include:

- The study generates a set of hypotheses about how listener disagreements originate.
- The psychological depth of the study is unique in the literature; other studies do not probe the reasoning behind the segmentation decisions as carefully.

Chapter 4: An analysis of boundary perception and musical features

Does acoustic novelty determine boundary salience? Chapter 4 presents a corpus analysis of the acoustic properties of points interpreted as boundaries by listeners. This study's larger scale enables us to draw confident conclusions about how acoustic novelty relates to boundary salience. The chapter tests a hypothesis incorporated into many segmentation algorithms that the moments of greatest novelty cause the perception of boundaries. Our results are especially pertinent since the corpus in question, SALAMI, is used primarily for validating such algorithms. A version of the chapter was published in a special issue of *IEEE Transactions on Multimedia* on Music Data Mining [SCC14].

The contributions of this chapter include:

- It is the largest corpus analysis to date on the subject of structure analysis.
- Confirms and qualifies previous findings (that novelty is a prime motivator for boundary placement) on a much larger corpus than before, using recorded audio from many genres rather than monophonic music within a single genre.
- Finds that degree of novelty is correlated to the likelihood of being perceived as a boundary.
- Demonstrates a new methodology: taking data sets intended for evaluation and repurposing them as the subject of an analysis of perception.

Chapter 5: Relating grouping structure to musical features

Can we deduce what a listener paid attention to based on their analysis? Chapter 5 poses a new MIR problem intended to address the needs of music psychologists: is it possible to estimate what a listener paid attention to in a piece of music based on their annotation? If so, it would be possible to understand the cause of disagreements between listeners. We describe an algorithm that addresses this problem by minimizing

the distance between an annotation and several decomposed self-similarity matrices. Although a large-scale validation of the algorithm is not possible, we demonstrate its ability to analyze differences between annotations on a small set of songs. A version of this chapter was published in the proceedings of *ACM Multimedia* [SC13b].

The contributions of this chapter include:

- Poses a new problem for MIR: music structure analysis explanation.
- Introduces an algorithm to solve this problem using Quadratic Programming.
- Demonstrates usefulness of the algorithm for visualizing and explaining listener disagreements.

Chapter 6: The effect of attention on grouping decisions

Can attention affect the analytical decisions of a listener? Chapter 6 describes a psychology experiment that tests the hypothesis developed throughout the preceding three chapters: that when a listener directs their attention, whether consciously or unconsciously, towards a particular aspect of the music, this affects their perception of grouping structure. While the correlation between these factors is supported by previous work, this experiment tests the causal relationship. We also test whether this effect depends on factors within the music (e.g., the musical features being attended to and manipulated) or individual factors (e.g., level of musical training). A version of this chapter is being prepared for submission to a journal of music psychology.

The contributions of this chapter include:

- Creates a set of artificial but realistic stimuli for use in this and other studies of the perception of structure.
- Reaffirms that listeners perceive music in a multi-dimensional way.

- Discovers that drawing a listener's attention to one feature of the music influences them to analyze the music according to that feature.
- Confirms that listeners are able to accurately continue the analysis of another listener.

Chapter 2

Structure in music theory, psychology, and information retrieval

This chapter presents an overview of previous approaches to structural analysis in three relevant fields: music theory (or musicology), music perception and cognition, and music information retrieval (MIR). We examine how each field has posed and answered these two questions: "How do listeners divide pieces of music into categorized segments?" and "How can we explain disagreements among listeners?"

We first compare these three fields and their approaches to understanding and modeling grouping structure. In Section 2.2, we review landmark and state-of-the-art models of the perception of structure that have been developed in music theory, the cognitive reality of which has been tested by music psychologists. Next, in Section 2.3, we describe how these models have been implemented as algorithms in MIR, and how MIR researchers have sought to manage disagreements among listeners. In Section 2.4, we discuss some ways in which the models have all struggled with similar problems, and finally in Section 2.5 the ways listener disagreements have been addressed.

2.1 Music theory, music psychology and MIR

Musicology is a branch of the humanities; music psychology, of science; and MIR, of engineering. While comparing these fields, we must keep in mind their different aims. In an essay on the gap between musicology and music psychology, Clarke [Cla89] offers a succinct definition of each:

> Broadly speaking, the aim of musicologists and composers in tackling issues of musical structure can be characterized as the attempt to formulate theories of the structural relations within and between musical works, and their origins, development and effectiveness as formal devices. A correspondingly brief summary of the aim of psychologists of music is the development of theories of the mental processing of musical events, or the relationship between the listener, performer or composer and the musical environment.

Twenty five years later, it is necessary to add a brief summary of the aim of MIR research: the development of algorithms that deduce structural analyses from audio recordings, simulating how listeners interpret the structure of pieces of music.

Within music theory, we recognize a division between composer- or text-based approaches and listener-based approaches. The former category includes Schenkerian analysis and set theory, both of which are most concerned with musical relationships that may be apparent in the score but may be very difficult to discern—much less perceive spontaneously—in a performance.[1] This thesis focuses instead on the latter category, listener-based theories of grouping, which includes Lerdahl and Jackendoff's Generative Theory of Tonal Music (GTTM) [LJ83] and Narmour's Implication-Realization (IR) theory [Nar90]. These seek to account for the listener's experience of music, and hence are closely related to music psychology.

[1] Take, for example, Tymoczko: "I am primarily interested in the idealized composer's point of view: my goal is to describe conceptual structures that can be used to *create* musical works, rather than those involved in perceiving music." [Tym11] (22)

One primary way that these fields differ is in how they regard the importance of individuals and populations, and of specific genres or multiple genres of music. First, in music theory, it is common to compare the hearings of individuals (often, the hearings of the authors themselves) and to debate their merits with respect to individual pieces. The merit of a particular grouping could be how well it explains or reflects the compositional principles of the piece in question. Pieces are almost always discussed in the strict context of a genre or style.

In contrast, as a science, music psychology investigates general principles that can explain the way populations of listeners perceive groupings. Towards this end, it is much more common to experimentally test how groups of listeners respond to stimuli. An individual's response is not of special interest, since it is the trend observable in the group that is pertinent. For practical reasons, it is most common to run experiments using either a very limited set of pieces or using artificial stimuli, short music excerpts composed specially for the experiment. Hence, music psychologists are interested in general principles that apply across listeners and across genres, but due to operational constraints usually test these principles in carefully constrained scenarios.

Finally, structural analysis is treated as an engineering problem in MIR, in which and any and all pieces of music are relevant; individual algorithms may target specific genres of music, but only as specific as "classical" or "popular." The goal is not necessarily to understand how humans analyze music, but to replicate this ability algorithmically. (Of course, an understanding of how humans do this seems essential in practice.) MIR is only concerned with specific listeners and pieces at the stage where algorithms are evaluated. To evaluate an algorithm's success, it is executed on a large number of pieces, and its output is compared to structural annotations created by human listeners. Although individual listeners are needed to create the annotations, these annotations are treated as absolutely correct "ground truth" for the purposes of evaluation. Individual pieces may be examined in order to characterize an algorithm's shortcomings.

	Music Theory	Music Psychology	MIR
Relevant listeners	individuals	populations	populations
Typical listeners actually studied	individuals	many listeners	individuals' annotations
Types of listeners	experts	all	experts
Relevant pieces	individual pieces and styles	all pieces	all pieces
Typical number of pieces studied	individual pieces	individual pieces and artificial stimuli	large numbers of pieces
Overall aim	generate theories	test theories	implement theories

Table 2-A: The scope of different disciplines studying music structure

Despite these broad differences, there has been significant overlap between the fields, and it can be difficult to place individual contributions firmly in one field or another. Music theories inform and are sometimes directly implemented as MIR algorithms (e.g., [Cam01] is inspired by GTTM, and [HHT06] implements GTTM); music psychology seeks to confirm or disprove the mental reality of music theoretic models of grouping (e.g., [CK90], [FC04]); and, most recently, some developments in MIR regarding grouping annotation procedures could inspire new theories of music (e.g., [BDSV12b]).

The typical scope of each discipline is summarized in Table 2-A. Since this thesis includes contributions to each of these fields, this table illustrates the broad scope of the thesis. In our effort to better understand how listeners perceive structure, we will: compare the hearings of pairs of listeners (Chapters 3 and 5), examine a small group of listeners (Chapter 4) and eventually look at a very large number of listeners (Chapter 6). We will look at disagreements among listeners for individual pieces (Chapters 3 and 5), for large collections of pieces (Chapter 4), and for artificial stimuli (Chapter 6).

With this preamble, we now turn to the central question: how has the perception of structure been modeled in these fields?

2.2 Models of the perception of grouping structure

2.2.1 Rule-based generative models

The problem of modeling grouping structure has often been approached in a constructive, ground-up manner: a theory seeks to explain how the tiniest sonic units (e.g., notes[2]) are identified by a listener, how these are chunked into larger units (e.g., triplets), and how this chunking procedure continues at higher hierarchical levels (to melodic motives, phrases, and sections).

For example, in Tenney and Polansky [TP80] and later in Lerdahl and Jackendoff's [LJ83] Generative Theory of Tonal Music (GTTM), simple gestalt rules are proposed to describe how listeners perceive and group sounds. GTTM's grouping rules include, among others: the proximity rule (a boundary is likelier to be heard when a longer note, or a rest, sits between two shorter notes) and the change rule (a boundary is likelier to be heard when some parameter of the musical surface—register, loudness, duration—changes).

GTTM's grouping rules are explained with simple stimuli and short melodies, and the full theory (including rules for describing metrical and hierarchical structure) is demonstrated on pieces of moderate size and complexity [LJ83] (250–278). Subsequent studies have confirmed the perceptual validity of some of the rules postulated in GTTM. In an experiment by Clarke and Krumhansl, participants listened to entire pieces and indicated where they heard boundaries between segments, and afterward freely explained their choices for each boundary; most of the reasons offered related to the grouping preference rules of GTTM (i.e., they pertained to changes and to parallelism) [CK90]. Frankland and Cohen showed that quantified versions of the rules of GTTM could be used to predict how listeners segmented short melodies, although not all of the preference

[2]We are strictly concerned with horizontal groupings, rather than vertical groupings; that is, we set aside the listener's task of segregating simultaneous but independent streams in the music. This subject is reviewed well by Deutsch [Deu99].

rules tested were shown to be equally effective [FC04].

The success of GTTM has inspired similar theories: Cambouropoulos' Linear Boundary Detection Model (LBDM) simplifies the rule structure of GTTM considerably by measuring changes in all musical factors relative to their own scale [Cam01]. Temperley's Grouper algorithm combines GTTM's proximity and parallelism preference rules with an *a priori* preference for segments of roughly 10 notes [Tem01].

However, the aforementioned works all engage primarily with the first two grouping rules of GTTM: the proximity rule and the change rule. GTTM also specifies more complex preference rules which have been much harder to implement in practice: an intensification rule states that simultaneous changes lead to higher-level grouping boundaries, and rules for parallelism and symmetry state that similar and similar-sized segments tend to be grouped at higher levels. The application of these rules is not made precise in [LJ83], and the rules have been implemented algorithmically more rarely. When they have, they have required refinements to the model (e.g., [HHT04] defined weights for the relative importance of the more complex rules) or strict limits on its application (e.g., [Cam06] is designed to handle only a limited range of parallel situations).

2.2.2 Models of expectation

According to Narmour's Implication-Realization (IR) theory, groupings derive not from psychological gestalts but from the dynamic way expectations are established and then realized or denied. The two most basic expectations are that a repetition will beget another repetition, and that a change will beget another change; from these, Narmour derives a taxonomy of melodic types. The connection to grouping structure is that Narmour posits that groupings are bounded by points of greater closure, with the perception of closure being induced by a set of melodic conditions. Closure is most emphatic when a large interval is followed by a small interval in the opposite direction, but other conditions, such as resolution to a consonance and note lengthening play a role.

Like GTTM, IR theory was explicated mostly using short melodies, and has since been tested in listener studies. For example, Krumhansl had participants listen to melodic fragments whose last tones varied systematically [Kru96]. Participants' ratings of how well the tone continued the fragment were able to be fit to a linear model based on the IR criteria. However, Schellenberg et al. performed a similar experiment and found that a more parsimonious model of expectation could model listener responses better; they also found that expectations changed as a function of age, which was significant because Narmour had posited that his rules of expectation were universal [SAPM02].

Also like GTTM, the IR model has been criticized for giving "short shrift" to the top-down influences on expectation, such as expectations generated by stylistic knowledge or prior listenings of a piece, compared to the bottom-up influences [Roy95]. Pearce's Information Dynamics of Music (IDyOM) model is a descendant of IR that seeks to remedy this [Pea05, PW06, PMW10b]. Like IR, IDyOM is, at its core, a model of expectation, with groupings predicted as the result of changes in expectation. Unlike IR, expectedness in IDyOM is not computed from a set of rules; rather, it models the information theoretic properties of the melody (the expectedness of each note and the certainty with which the next note is predicted) with unsupervised learning. IDyOM draws on comparable models of how infants learn to segment speech, but is improved in two important ways: first, it has the ability to incorporate many different melodic attributes at once; for example, the model can consider the surprisingness (and relative informativeness) of a pitch sequence along with its pitch-class sequence, interval sequence, sequence of durations, and many other viewpoints. Second, the model combines expectations from both a short-term model (which learns expectancies based on the local piece only) and a long-term model (trained on a corpus of previously-heard melodies). In this way, the model successfully integrates bottom-up and top-down influences.

For both the gestalt-based and expectation-based theories of grouping structure, we have seen that the most persuasive evidence comes from studies of listening-based segmentation of modest-length sequences, usually monophonic melodies [MOG00, HTS02,

Kru96, FC04, PMW08], although some have used full-length and full-textured pieces of music [CK90, BMK06]. At a short enough timescale (e.g., the size of a phrase or shorter), where listeners' responses are most consistent, these models may offer an adequate explanation of grouping judgements. But groupings at larger scales seems to involve either a complex combination of preference rules based on parallelism, tonal stability/instability, caesuras, and countless other sonic features summing together, or a complex combination of expectations based on a similarly wide array of attributes.

2.3 Structural analysis in MIR

In the wake of GTTM, with the advent of more powerful computers in the 1990s, there was great interest in implementing it and other theories of music as algorithms. Stammen and Pennycook adapted the preference rules of GTTM for a real-time segmentation system [SP94], and an algorithmic implementation of GTTM has been developed by Hamanaka et al. [HHT04] and Hirata et al. [HTH07]. GTTM's successors, LBDM and Grouper, were conceived as algorithms, and the IDyOM model has also been implemented as an algorithm [PMW10a].

Earlier grouping structure algorithms operated on symbolic representations of music, such as MIDI (e.g., [SP94]), MusicXML (e.g., [HHT04]), or some other abstract representation (e.g., [Cam01]). Since 1999, grouping structure research in MIR has come to focus much more on audio representations. Instead of dealing directly with abstract musical parameters such as notes and instrument parts, researchers extract musical features from audio, such as chroma (a vector expressing the relative strength of each pitch or pitch class) or Mel-frequency cepstral coefficients (MFCCs, a vector that characterizes the shape of the sound's frequency spectrum, and hence timbre). These features are then processed to estimate the grouping structure of the music.

Much of the work in this area draws on Foote's seminal paper on the use of self-similarity matrices (SSMs) to visualize repeated patterns in music [Foo99]. An SSM

displays the computed similarity between all points in a recording, and is useful for discovering homogenous passages of music (which appear as blocks on the main diagonal), strong discontinuities (which appear as sharp corners between blocks), and repetitions (which appear as strong diagonal lines off the main diagonal). Since Foote, SSMs have been refined and used by countless others for structural analysis. (A formal description of SSMs and a more extensive review of SSM techniques developed in MIR appear in Section 5.1.)

Using the SSM, bottom-up and top-down factors may be taken into account. A bottom-up algorithm begins with a search for local discontinuities: Foote proposed convolving the diagonal of an SSM with a checkerboard kernel for discovering the sharp corners between blocks [Foo00], and this approach has been used many times since (e.g., [PK08b]. [Pei07]). On the other hand, one may begin by searching for the longest repetitions and deducing a finer-scale segmentation using several such observations, as in [Got06] or [MK07]. Such approaches are top-down in the sense that they prioritize a search for parallelism over a search for local discontinuities. Seen this way, we may consider all block-based structural analysis approaches to be bottom-up, and all sequence-based approaches to be top-down, in the sense that they define structure primarily with regard to discontinuities or to parallelism, respectively. The differences between these types of approaches were articulated by Peeters [Pee04], and since then some have aimed to combine the insights of both approaches (e.g., [PK09] and [GCJM13]).

Another way to incorporate a top-down view of structure is to build in stylistic expectations; for example, Shiu, Jeong and Kuo's algorithm filters the SSM to reinforce repeated four-measure sequences [SJK06]. Another class of algorithms uses clustering or Hidden Markov Models (HMMs) to obtain a description of a piece as a sequence of states. For example, Abdallah et al.'s algorithm [ASRC06] first computes an HMM with many states to obtain a very fragmented representation of the audio, in a manner comparable to both [LC00] and [AS01]. Next, they cluster histograms of HMM states to estimate large-scale structure. The approach is bottom-up in the sense that it builds a

representation of a piece's structure iteratively from the frame level to the segment level, but it is top-down in the sense that the frame-level judgements are based on comparisons between frames across the entire piece, not on local comparisons.

2.3.1 Listener disagreements in MIR

The previous section illustrated the range of approaches pursued in MIR for structural analysis, but the state-of-the-art in this field is not the focus of this review. (Technical advances have been made the past five years, but the methods of the field are reviewed in [PMK10] and [Smi10].) We are most interested in two questions: first, how have researchers sought to account for differences among listeners in their algorithms; and second, how do they account for different modes of attending—for example, a focus on local discontinuities or on large-scale groupings?

The answer to the first question is simple: they have not. The algorithms all take a single input (an audio recording) and produce a single output (a structural analysis), and the possibility of multiple interpretations is generally considered irrelevant. Even those models that take advantage of complex top-down factors are deterministic. Discovering multiple plausible structural descriptions of a piece is simply not part of the problem definition.

Regarding the second question, the field has effectively addressed different modes of attending by dividing the broadly-defined "structural analysis" task into subtasks. For example, segmentation (boundary detection) and segment labeling (grouping analysis) are considered separate but related tasks, and are usually evaluated separately. Some researchers narrow the problem further, focusing only on recognizing choruses or the most repeated part (e.g., [BW01], [Ero07], [MGJ11]) or motive recognition, sometimes called intra-opus pattern discovery (e.g., [CAFW13]).

2.3.2 Evaluation and annotation

Ignoring the possibility of different interpretations and different modes of listening is perfectly reasonable, given how algorithms are evaluated. Evaluation consists of executing an algorithm on a large collection of recordings, and comparing their output to a matching collection of ground-truth annotations. The similarity between the output and the ground-truth is appraised using a variety of metrics (see [Luk08] and [SC13a] for a review). Crucially, most corpora of annotations have a single annotation for each song, reflecting the operational assumption in this field that there is a single, best analysis of each song's structure—which, since there is no agreed-upon way to define the "average" analysis, is just one person's single hearing.

Although musical analyses are often hierarchical—each segment has subsegments and is part of a supersegment—in MIR evaluations, only one timescale is used. There is an informal understanding in the community of which timescale is most relevant—in pop music, the length of verses and choruses is roughly the unit size—but this timescale is not defined precisely. An analysis of several algorithms' performance at the Music Information Retrieval Evaluation eXchange (MIREX) suggests that algorithms often fail to target the level of detail encoded in the annotation [SC13a].

This status quo has been criticized, and newer corpora have pursued two remedies: first, clarifying the type of structure the annotations describe; and second, including several annotations per song.

In 2009, Peeters and Deruty observed that existing annotations conflate many aspects of structure—namely, musical function, similarity, and instrumentation [PD09]. Their critique was incorporated into the SALAMI dataset, which has a different annotation for each of these three aspects of structure; additionally, musical similarity is described separately at both a short and long timescale [SBF$^+$11]. Another attempt to refine the definition of structure has been pursued by Bimbot et al. [BLBSV10]. They defined a method for obtaining a segmentation that, in addition to setting an optimal segment size,

defines criteria for considering a span a segment, such as interchangeability and similarity. They also proposed a "System and Contrast" model which describes the typical ways that segments are composed and related to each other [BDSV12a], and proposed a system of labeling segments based on this model [BDSV12b]. The model resembles descriptions of Classical formal structures (such as Caplin's account of sentences and periods [Cap98]) and represents an important contribution of the theory of popular music—a unique contribution to music theory from MIR. A taxonomy of common transformations of segment lengths is presented in [BLBSV10], and a taxonomy of typical deviations and hybrids of segments in [BDSV12b].

The approaches of Peeters and Deruty and of Bimbot et al. both aim to minimize inter-annotator disagreement by being more precise about how the annotations should be written. However, as reported in [BLBSV10] and [SBF+11], inter-annotator disagreement persists despite clearer instructions and is significant. Although these approaches clarify what structure is, they still rely on annotators deciding for themselves and for each piece what patterns in the music are most relevant for the analysis, and what patterns are part of the uninteresting background.

One way to respond to this is to include several annotations per song. To consider such listener disagreements, the SALAMI corpus includes two annotations for most songs. Considering each annotation and both timescales, there are up to four different ground-truths per song—four different ways of hearing the structure. The recently published JAMS specification is designed to manage several independent annotations per song, and additionally preserves information about the origin of each analysis, such as the rules followed by the annotator and the annotator's musical background [HSN+14].

Yet another possibility, still speculative, is to build annotations in a probabilistic manner. Bruderer et al. devised a method of merging the boundary indications of multiple listeners by convolving each listeners' sequence of boundaries with a Gaussian kernel, and summing the results across listeners [BMK09]. (The size of the kernel had been optimized for maximum separation of boundary indications within each response.)

Although it may be impractical to collect the quantity of data they used on a larger scale, a collection of such annotations would be highly valuable. Bruderer's approach is a good system for annotating boundaries in a fuzzy way, but no one has yet devised a comparable approach for assigning fuzzy labels to segments, which are themselves fuzzily bounded.

2.4 Bottom-up versus top-down

In the previous two sections, we have witnessed a common struggle in all three disciplines to develop a theory or model of analysis that balances the influence of *bottom-up* and *top-down* factors. This pair of terms has been shared among the disciplines, but sometimes with different meanings. The proximity and change rules of GTTM are bottom-up in the sense that local discontinuities have an effect on higher-level groupings, while the parallelism rule is top-down in the sense that long-term similarity has an influence on shorter timescales. In the IR model and in IDyOM, expectations may either be bottom-up (i.e., originating within the local context of the piece), or top-down (originating from knowledge beyond the piece). In MIR, a focus on blocks in an SSM, which is a focus on homogenous states in music, is related to the bottom-up approach in GTTM, while a focus on stripes in an SSM, or on repeated sequences in music, is like the top-down approach. Algorithms can base their estimates entirely on knowledge derived from the acoustic signal, or can apply top-down constraints based on prior expectations: constraints on the size of segments, the number of unique segments, and so forth.

"Bottom-up" and "top-down" are similar to the terms "exogenous" and "endogenous," used by Godøy, et al. to describe, respectively, influences that originate outside and inside the listener [GJN10]. The music exerts exogenous influence over the listener, in the form of sharp local changes or clear, verbatim repetitions that are automatically perceived; the listener exerts endogenous influence over the music, in the form of expectations and knowledge (as in IR and IDyOM), but also in the form of conscious and

perhaps deliberate efforts to pay attention to particular aspects of the music. For example, depending on the listening situation, one may pay more attention to the melody (perhaps while listening in a car) or to the beat (while exercising).

Clarifying the different goals that a listener may have in analyzing a piece of music is central to Hanninen's theory of analysis [Han12]. Hanninen writes that a listener may adopt one of three fundamental orientations: a focus on discontinuities in the acoustic signal (sonic orientation), a focus on associations between passages (associative orientation), and a focus on how a particular theory of music applies to the piece at hand (theoretical orientation). The sonic and associative orientations are further synonyms for the bottom-up and top-down approaches of GTTM-like approaches, respectively, while the difference between the theoretical orientation and the others is similar to the difference between endogenous and exogenous influences. (With the growing set of synonyms, the analogies get muddied; depending on the context, the associative orientation could be seen as bottom-up or as endogenous—that is, top-down.) However, Hanninen further points out that the orientations are interdependent: sonic and associative attending nearly always happen simultaneously to some extent, and observations made from one perspective form the basis of new observations in the other. Jones and Boltz also argued that there are two contrasting modes of listening: an analytic mode that is focused on tracking local events (bottom-up), and a future-oriented mode that is more expectation-driven, in which attention is drawn to longer timescales [JB89].

Bottom-up and top-down; states and sequences; sonic and associative orientations; novelty and repetition; exogenous and endogenous; automatic and conscious; analytic and future-oriented. That we have accrued such a wealth of synonymous dichotomies hints that, regarding grouping analysis, the difference between bottom-up and top-down is difficult to define precisely, incredibly complex, and yet crucial for understanding how groupings are made. Each of these binaries refers to a competing set of influences in the mind of the listener. And yet, the listener does mediate between them; bottom-up *and* top-down processes coexist.

Perhaps this is the nugget of the problem: grouping structure seems to be, at once, a passive *perception*, something that a given listener is compelled to do, and an active *analysis*, a willful interpretation of the listener. The fact that a listener exerts some conscious control over their perception of structure may be the main cause of the listener disagreements noted in the previous section.

The ability of listeners to control their perception is apparent in the case of ambiguous musical stimuli that resemble, in some ways, optical illusions. There is a famous optical illusion where a drawing appears to be both a rabbit and a duck. A viewer will at first perceive the "dubbit" to be one animal, but then their perception may spontaneously switch to the other animal, and at this point, the viewer can deliberately choose to see the dubbit however they please. The lines of the drawing constrain, in a bottom-up way, what interpretations are most stable—it is not equally a duck, a rabbit, or a horse—but the imagination of the viewer allows them to control their perception.

Comparable situations in music are plentiful, especially with regard to meter and rhythm. The hemiola pattern, in which a single unit can be stably subdivided into two or three sub-units, can be used as an occasional device or as a meter; both uses appear in musics across the globe. In a discussion of ambiguity in music, Karpinski cites several examples where the downbeat, as indicated by the barlines, does not match how it is likely to be initially perceived; however, with effort, a listener can choose to hear each passage one way or the other [Kar12]. A vivid example that is a personal favourite: Paul Simon's song "Gumboots" has a bistable meter nearly throughout, with two possible downbeats a beat apart. Each of these examples is like the dubbit in that multiple interpretations are possible, but the range of interpretations is still constrained by the music. Musical dubbits thus lie at the centre of the tension between bottom-up and top-down factors: a listener can choose how they perceive notes to be grouped, but their choices are limited by options permitted by bottom-up considerations.

2.5 Listener disagreements

Such ambiguous situations might reasonably be identified as a main source of listener disagreements, although Agawu has argued that, given a theoretical context, ambiguities should never be irresolvable and analyses ought not diverge [Aga94], and Francès found that listeners who were given clear expectations about the music to come agreed with each other's analyses much more than did listeners who were not guided [Fra58]. Lewin attributed differences in listeners' interpretations to a deficiency in analytic discourse that fails to account for the fact that listeners are in fact analyzing many phenomena at once. But as annotators in MIR can attest, clarifying an analytical procedure does not make it perfectly repeatable. One issue is that many musical situations are not merely ambiguous in the sense that they support bistable percepts; if this were so, listener disagreements could be reduced to a set of conscious, "duck-or-rabbit"-type decisions. In fact, listeners often disagree without necessarily being aware of the alternative hearings; disagreements can arise without listeners believing that they've made any conscious "decisions" at all. It is difficult to identify all the decisions one has made and top-down factors one has felt with mere introspection.

Listener disagreements are noted in all studies on listening-based segmentation. For example, for every song studied by Frankland and Cohen, while the average agreement between listeners' boundary segmentations was high, and in some cases perfect, there were always pairs of negatively correlated analyses [FC04]. Bruderer, Martin, and Kohlrausch had listeners segment full pieces, and found that of all the boundaries indicated by participants, only a few were agreed upon by all listeners [BMK09]; similar results were observed in [MOG00]. What's more, listeners may even disagree with themselves: in [FC04] and [BMK09] within-subject agreement across trials was sometimes low, and Margulis found that after hearing a piece multiple times, a listener indicates different boundaries, in a way that suggests their attention has been drawn to repetitions of greater length [Mar12].

These studies attest to the fact that listener disagreements are real and often substantive, and that focusing on the nature of these disagreements can lead to insight. As another example of this: by studying the rate of agreement in perceiving boundaries, [BMK09] showed that boundary salience is correlated with the likelihood of being perceived as a boundary.

While such disagreements seem natural and commonplace, music theories such as GTTM and IR do not necessarily account for them. Narmour argued that the fundamental expectations in IR were universal, which was disputed by [SAPM02]. And consider the simplifying assumption made by GTTM's authors: the first sentence of the book limits its scope to "a listener who is experienced in a musical idiom" (1). In a world of identically experienced listeners, differences in interpretation would not exist. The authors recognize this but point out that while "the 'experienced listener' is an idealization, [....] there is normally considerable agreement on what are the most natural ways to hear a piece" (3). This assumption aids in their project of deducing a set of gestalt rules for generating analyses of tonal music that resemble the perceptions of humans. Several of the research studies cited so far certainly do corroborate the observation that considerable agreement among listeners exists as far as musical structure is concerned. That evidence may even extend to the neuroscientific literature: Abrams et al. found brain activity to be synchronized across listeners of the same piece of music [ARC+13]. But confidence that agreement among listeners is "considerable" does not explain the origin of the disagreements. One aim of this thesis is to investigate precisely that.

2.5.1 Accounting for listener disagreements

We are not the first to observe that listeners can disagree, and other research has suggested refinements to existing theories that could account for such differences. For example, regarding GTTM, Deliège suggested that listeners might apply the same gestalt rules as one another, but with slightly varying weights depending on their musical experience

[Del87]. Her evidence was that musicians in her study produced segmentations more often concurrent with GTTM's Grouping Preference Rules than non-musicians. On the other hand, [FC04] found that listeners with varying musical backgrounds parsed melodies quite similarly to one another.

A second refinement could be to model how listening-based segmentation is affected by a listener's familiarity with the piece. As noted briefly earlier, Margulis found that an individual's attention was drawn to longer repetitions after hearing a piece multiple times [Mar12], and Frankland and Cohen found that a listener's second and third hearings of a piece agreed more closely with one another than their first hearing did to the second or third [FC04]. Palmer and Krumhansl found that the more familiar a listener was with a piece, the closer the agreement was between how they segmented two different simplified versions of the piece, each retaining only the rhythm or the melody [PK87]. The evidence collectively suggests that listeners refine and crystallize their interpretations of a piece as it becomes more familiar.

The problem of explaining listener disagreements is hypothetically sidestepped (or at least reduced) when prior probabilities are used to train the model. For example, Grouper and IDyOM are both segmentation models whose parameters may be set according to the statistical properties of a corpus of music; by feeding them a variety of corpora to represent different listeners' experiences, these models could output a range of grouping predictions for different listeners. Hansen, Vuust and Pearce have shown that providing IDyOM with corpora of jazz or general melodies, one can predict the different expectancy ratings of jazz and classical musicians [HVP13]. In another (also probabilistic) view, listener disagreements could arise due to perception being a stochastic process: for example, individuals may perceive boundaries with a probability proportional to the boundary's intrinsic salience. This view aligns with the findings of [BMK09], who found that of all the boundaries indicated by participants in a segmentation task, the few that were agreed upon by all listeners happened to be those with the highest rated salience, and even those who did not indicate a popular point as a boundary tended to agree it was

salient. All of these probabilistic interpretations of human perception allow differences among listeners to be explained as the variance in the input and output of a perceptual mechanism that is common to everyone.

While a probabilistic interpretation of listening is appealing, it might not be a satisfying description of the conscious experience a listener has when they interpret the structure of a piece of music. This brings us to a second simplification admitted by Lerdahl and Jackendoff: they are concerned only with the "final state of the [listener's] understanding," and not the "mental processing" that precedes it [LJ83] (4). While it is true that the structural description provided by a listener is the most concrete evidence that can be examined, in order to understand how these descriptions deviate from one another, we must interrogate the listeners about their mental processing. The study presented in Chapter 3 seeks to do exactly that.

Chapter 3

Causes of variation among listeners in boundary and grouping decisions

In chapter 2, we saw how differences among listeners were not well accounted for by models and algorithms for determining grouping structure. Important theories of music cognition, such as the Generative Theory of Tonal Music, posit an archetypal listener with an ideal interpretation of musical structure, and many studies of the perception of this structure focus on what different listeners have in common. However, we also saw that previous experiments have revealed interesting differences in how listeners perceive structure, showing a dependence on a listener's familiarity with the piece, and on their musical background. The impact of these and other endogenous factors (that is, factors that depend on the listener) is not understood in detail. Determining their impact, and determining which other factors are important, may be essential to developing more advanced models of music perception.

In this chapter, we embark on a case study of the structural analyses of two listeners with very different perspectives on the music: one is the performer, the other only

a listener. Our study has two goals: to identify the differences between the listeners' analyses and to explain why these differences arose. For this study, the listeners analyzed the structure of three improvised duets, which were performed by one of the listeners and Mimi (Multimodal Interaction for Musical Improvisation), a software system for human-machine improvisation. The ambiguous structure of the human-machine improvisations, as well as the distinct perspectives of the listeners ensured a rich set of differences for the basis of our study.

We compare the structural analyses and argue that most of the disagreements between them are attributable to the listeners paying attention to different musical features. Following the chain of causation backwards, we identify three more ultimate sources of disagreement: different commitments made at the outset of a piece regarding what constitutes a fundamental structural unit, differences in the information each listener had about the performances, and differences in the analytical expectations of the listeners.

A case study similar to ours was reported by Bamberger [Bam06]. She conducted interviews of three listeners with different musical backgrounds, and their hearings of a Beethoven minuet were compared in an effort to understand musical development and how people learn to have more complex hearings of pieces of music. Although the focus of her study differs from ours, she touched on issues relevant to us here. For instance, she discussed how the differences between hearings of a piece could be understood as what she terms "ontological differences" (a musical ontology being a determination of what musical ideas count as genuine abstract entities or units). She also suggested that a listener's musical knowledge can influence which musical features and relationships they deem relevant. We will see in Section 3.3 how these factors—listeners' differing musical knowledge, beliefs, and ways of attending—led, in our case, to diverging musical ontologies and differing interpretations of musical structure.

The justification for our choice of material and the method for collecting the annotations are described in Section 3.1. Referring both to the annotations and to the listeners' written accounts of why they analyzed the music as they did, the differences between the

analyses for each piece are studied in Section 3.2. The results of these comparisons are summarized and discussed in Section 3.3, and our conclusions are presented in Section 3.4.

3.1 Procedure

Our goal in this study is to develop a better understanding of how and why listener disagreements occur. To do so, we compare the different listeners' analyses of pieces of music. In this section, we describe the compromise we struck between the size of our experiment and the level of detail of the responses gathered, and justify our choice of materials, procedures, and participants.

Most significantly, we have opted to limit the "participants" of our study to two listeners: myself and Isaac Schankler, the composer/performer of the music in question. While this precludes the possibility of drawing unbiased or statistically powerful conclusions from our observations, our approach facilitates a deeper examination of the differences between our analyses. As will be explained in this section, our choice of methodology is intended to maximize the number and diversity of listener disagreements observed, while allowing as deep an investigation as possible into the causes of these disagreements.

Studies of listeners' analyses usually tout their large size as an advantage: with increased size comes increased statistical power and greater generalizability. Indeed, with many participants (e.g., [BMK09]) or many pieces of music (e.g., the corpus analysis in Chapter 4), it is possible to observe broad patterns in how listeners perform chunking, or in how chunking decisions relate to the music that was heard. However, when studying listener disagreements, increased size can be a liability. Firstly, the information we are most interested in—the listeners' justifications for their responses—is information that is difficult to quantify or categorize, and hence difficult to interpret in large quantities. Secondly, we would like to have the participants reflect on each other's analyses and explain why they did not respond in the same manner, and this information can only

be collected after the first part of the analysis takes place. By using only ourselves as participants, we simplify this process.

Still, it cannot be denied that including the responses of more participants could improve this case study. However, this study was originally conceived and published (see [SSC14]) as a contribution to the field of music theory, in which it is common to reflect on the implications of just a single listener's analysis, and where single-author articles are the norm. In this context, our use of two listeners, while the bare minimum required to make comparisons across listeners, may be less unsatisfying.

3.1.1 Choice of materials

The choice of music to study was guided in part by my experience collecting the dataset for the Structural Analysis of Large Amounts of Music Information (SALAMI) project [SBF+11]. The SALAMI dataset consists of over 2,400 annotations of nearly 1,400 recordings in a wide variety of musical styles, ranging from Renaissance motets to Dixieland jazz to electronica. It was observed that some categories of music, such as song-form popular music, inspired far fewer disagreements than others, such as avant-garde jazz. Through-composed and improvised works in particular seemed to demand more willful interpretation from the listener.

Since we wanted the music in our case study to elicit as many and as diverse a set of listener disagreements as possible, we chose to focus on a human-machine improvisation scenario, described below, that presents unique challenges for grouping and segmentation.

Mimi (Multi-modal Interaction for Musical Improvisation) is a software system designed for human–machine improvisation [FCT07, Fra09]. Using a MIDI keyboard, an improviser's performance is recorded into a buffer (called the "oracle") and modeled by Mimi. Mimi then walks through the oracle, recombining parts of the improviser's performance into new musical material; in this manner, Mimi and the musician are able to perform concurrently in an overlapping, improvised duet. The performer retains

control over some aspects of Mimi's behavior, including the content of the oracle (which can be added to or deleted altogether), the recombination rate (which controls how likely Mimi is to juxtapose fragments of the oracle), and whether Mimi is generating music or not (naturally, this control must be used in order to end a piece). A visualization accompanying the performance gives the performer information about what Mimi is about to play and has just played, as well as a display of all the musical material currently in Mimi's memory.

Performances with Mimi provide interesting challenges for the listener seeking to understand musical structure; first of all, there is the improvised nature of the performance, which is, in the words of George Lewis, characterized by a "refreshing absence of the moral imperative concerning structure" [Lew09]. Put simply, improvisation is not necessarily bound to formal structures traditional in popular, classical, or other music.[1]

A second and perhaps more intriguing challenge is interpreting the actions of Mimi: Mimi has no knowledge of how, nor the ability, to intentionally create an ending of a phrase, a section, or the entire piece. Any perceived structure could be said to be partly derived from the creativity of the human improviser, whose performance provides the basis for Mimi's material and whose decisions in response to Mimi may reinforce previous patterns or introduce new material. It may also be partly and serendipitously due to the probabilistic connections Mimi makes between similar note material of disjoint sections. But in the absence of these chance connections or the improviser's interventions (as when the improviser clears the oracle or tells Mimi to stop generating music), the material Mimi generates tends to be structurally amorphous, especially at larger scales.

The third and final challenge is that of integrating the improviser's musical ideas and Mimi's concurrent, perhaps not compatible, layers of musical material. For instance, at any given moment, the listener must decide who is in the foreground, Mimi or the improviser. But, as in an Escher drawing, there may be more than one interpretation

[1] And yet, whether through the latent tendencies of the performer or by the constraints of Mimi's programming, traditional formal structures may still emerge from performances with Mimi. This is discussed in other articles related to these same three performances and some others; see [SCF14, SSFC11].

of the same lines. The focus of the listener's attention—whether they are concentrating on the improviser, on Mimi, or on both—may thus have a significant impact on the perception of structure. This task is further complicated by the fact that, depending on the instrument patches chosen for the improviser and Mimi, the two voices may not always be distinguishable.

Over the course of three weeks, Schankler produced three separate improvisations (hereafter referred to as Performance no. 1, no. 2, and no. 3) with Mimi, all on a Yamaha P90 weighted-action keyboard (see 3.1) in a laboratory setting. The three performances were recorded as MIDI files, from which audio tracks equivalent to the original performances could be made. These were the recordings consulted during the annotation stage.

Figure 3.1: Isaac Schankler performing on a Yamaha Disklavier with Mimi at the People Inside Electronics concert at the Boston Court Performing Arts Center in Pasadena, California, in June 2010.

3.1.2 Annotation procedure

The annotation procedure was inspired by previous work with the SALAMI project. The formal structure of each piece was independently annotated by Isaac Schankler (the improviser, hereafter referred to as Annotator 1) and by me (an independent listener, hereafter referred to as Annotator 2). Using different software tools with similar functionality—Annotator 1 used the Variations Audio Timeliner[2] and Annotator 2 used Sonic Visualiser[3]—the listeners analyzed each piece at two hierarchical levels. In accordance with common practice in formal musical analysis, the large-scale level was annotated with uppercase letters, and the small-scale level with lowercase letters, to indicate which portions of the piece were judged to contain similar musical material. In keeping with Lerdahl and Jackendoff's well-formedness rules for structural grouping, overlapping sections were disallowed, all portions of a piece were labeled, and boundaries at a given hierarchical level were respected at smaller-scale levels.

Each analysis was produced in a single session, each lasting roughly a half hour, although this time was not prescribed beforehand; indeed, aside from producing annotations in the same format, the annotators had total freedom: they were free to listen to the pieces as often as they liked, and to return to particular spots or repeat short excerpts.

In a departure from the procedure used by SALAMI, both listeners also wrote brief notes explaining their choice of boundaries and groupings in a separate session after annotating each piece. The responses were worded freely, but at a minimum the listeners were expected to justify, with reference to the recording, each boundary and the similarity of sections labeled with the same letter.

These justifications did not generate explanations from both participants for every moment where the interpretations diverged. Consider the case in which listener no. 1

[2] `http://variations.sourceforge.net/vat`
[3] `http://www.sonicvisualiser.org`

perceived a boundary where listener no. 2 did not. We may refer to the first listener's explicit justification for this perception, but listener no. 2's remarks may not include an explanation for not experiencing this perception. The process of identifying and explaining differences thus required more than just collating responses. So, after enumerating all the significant differences between our analyses, we (the two listeners) discussed each one, reflecting on our listening experiences and elaborating on our interpretations of the pieces. The next three sections recount the outcome of these conversations for the three pieces.

3.2 Differences between annotations

In this section, we consider the three performances separately. For each, we list the differences between our annotations and offer reasons to account for these differences. We will collect our observations and attempt to generalize from them in Section 3.3. The pair of annotations for each performance are shown in Figures 3.2, 3.5 and 3.8; in each, the upper part is from Annotator 1 (the improviser) and the lower part is from Annotator 2 (the independent listener). Recordings of each performance, set to animated versions of these figures, are available online.[4]

3.2.1 Performance no. 1

In Performance no. 1, Annotator 2 roughly agreed with all of Annotator 1's small-scale boundaries (the smaller bubbles in Figure 3.2), but Annotator 2's version has more small-scale boundaries, and it also differentiates subsections within each main section (e.g., A_1 includes a, b, c, and d subsections). This leads to two compelling divergences in the large-scale segmentation. Setting aside the small deviations in timing (e.g., the few seconds difference in the boundary between Annotator 1's a_2/a_3 and a_3/b_1 transitions,

[4]Performance no. 1: `https://vimeo.com/96662176`
Performance no. 2: `https://vimeo.com/96662177`
Performance no. 3: `https://vimeo.com/96662178`

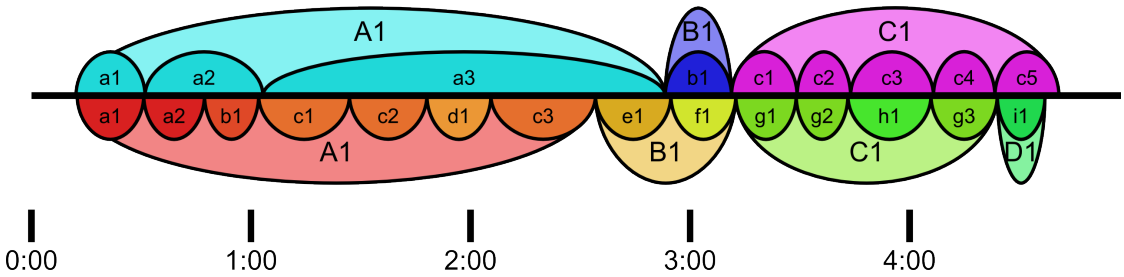


Figure 3.2: Analyses of Performance no. 1 by Annotator 1 (top), and Annotator 2 (bottom).

and the disagreement about when the piano stopped ringing at the end of the piece), the differences that require explanation are:

(1) Why is Annotator 2's A_1 section much more segmented than Annotator 1's?

Both annotators identified the same initial sequence as a single musical idea a_1, but they conceptualized this passage differently because they focused on different musical parameters. To Annotator 1, the idea was defined by its mood—an amorphous, ethereal melody with pedal—and the segments a_2 and a_3 were distinguished by the melody moving to a different voice (Mimi) or to a new register. On the other hand, Annotator 2's hearing was marked by a strong sense of rhythmic phrasing, established when the four-part opening phrase a_1 is answered by Mimi with a similar phrase a_2. This pace is only followed roughly for the rest of the A section, but because the material is very open, containing relatively short gestures with long pauses in between, it is easy to imprint a loose pace of phrases onto the music.

(2) Why does Annotator 2 hear the transition section B as beginning earlier than does Annotator 1?

Both annotators agreed that the material beginning at 2'54" (i.e., at the beginning of Annotator 1's b_1 and Annotator 2's f_1) was wholly different from the material in section A_1. Indeed, Mimi is silent during this section, and it is melodically and rhythmically distinct from all of section A_1. (See Figure 3.3.) However, Annotator 2 perceived a "pre-transition function" in segment e_1, leading him to place the beginning of the section

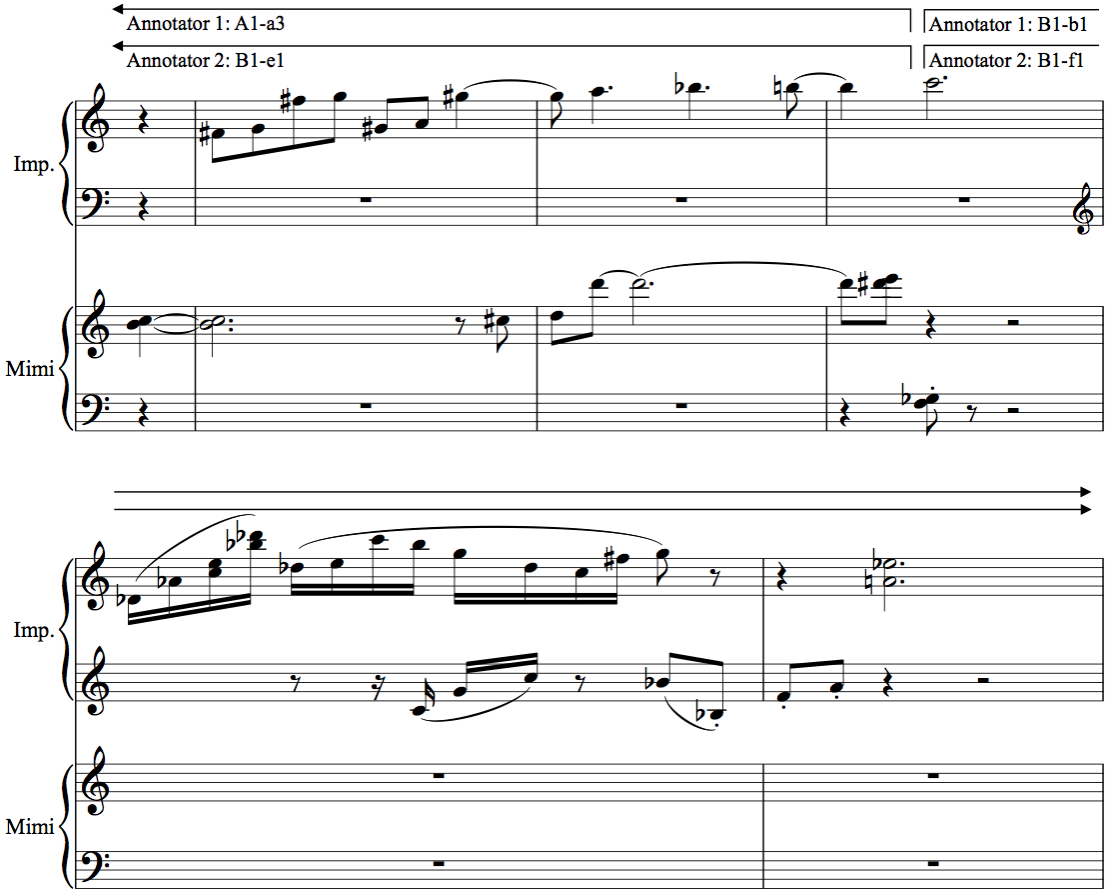


Figure 3.3: Notation Example 1. Performance no. 1, large section boundary (A_1–B_1) for Annotator 1.

earlier than Annotator 1. While the material in e_1 is similar to the rest of section A_1, there are a few cues that arguably distinguish it: a new descending theme from the improviser with a repeated rhythm, and a rising, fading motive that follows, both of which feel like ending material and anticipate the change at 2'54". (See Figure 3.4.)

(3) Why do Annotator 1 and Annotator 2 disagree about the differentiation of musical ideas in section C_1?

While Annotator 2 differentiated between subsegments throughout the piece, Annotator 1 did not; he posits that this is because that option did not occur to him at the time. It is hard to say whether the labeling differences of these subsections of C_1 (or the subsections of A_1) are very meaningful, since Annotator 1 and Annotator 2 also

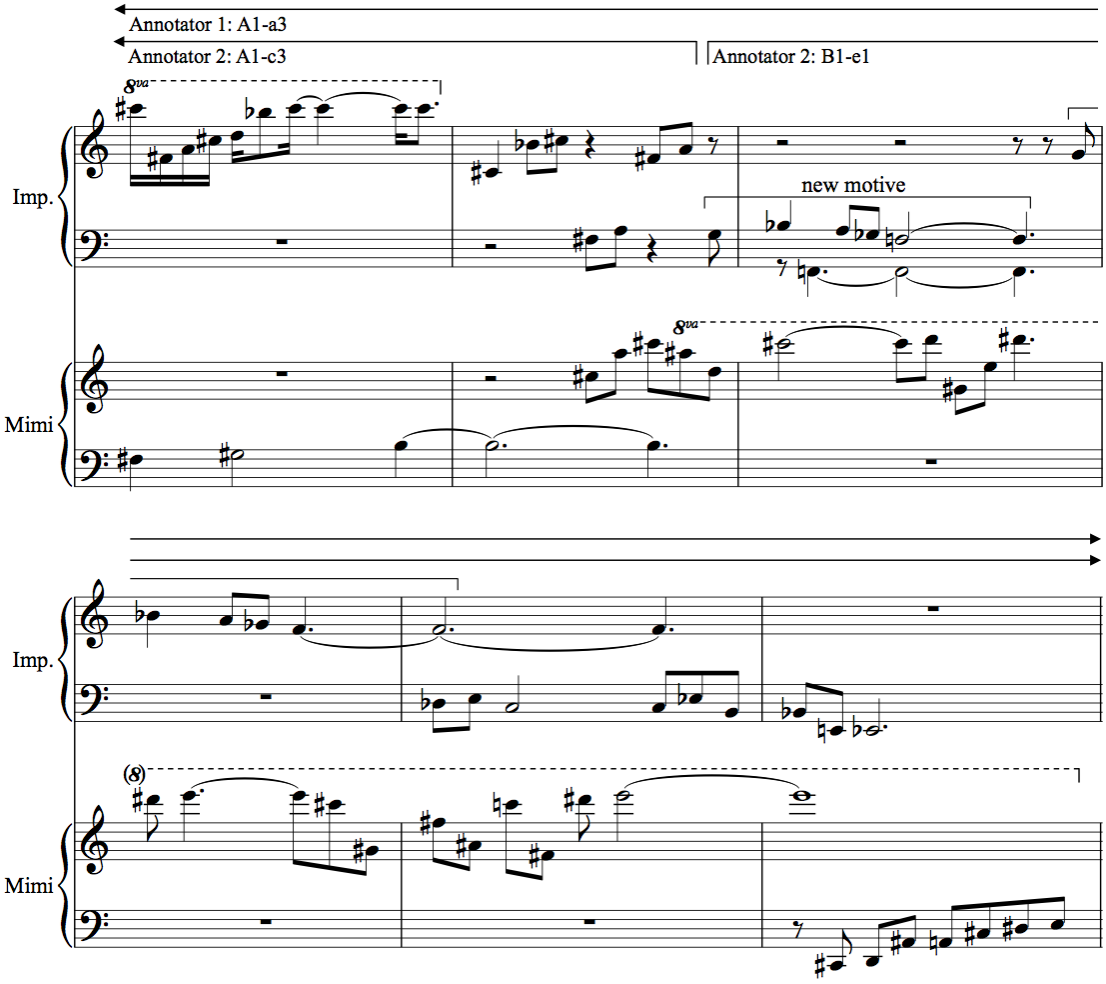


Figure 3.4: Notation Example 2. Performance no. 1, large section boundary (A_1–B_1) for Annotator 2.

initially employed slightly different naming conventions: Annotator 1 used letters and prime notations (e.g. A, A', A''), and Annotator 2 used a combination of letters, subscript numbers and prime notations (e.g. A_1, A_2, A_2'). The analyses shown in Figure 3.2 are adaptations of the original analyses, meant to enable comparison; for the later performances, the annotators used the same format as each other. In the diagrams in this chapter, subscripts are only used to indicate repetitions of musical ideas.

(4) Why do Annotator 1 and Annotator 2 disagree about the labeling of the final section (C vs. CD)?

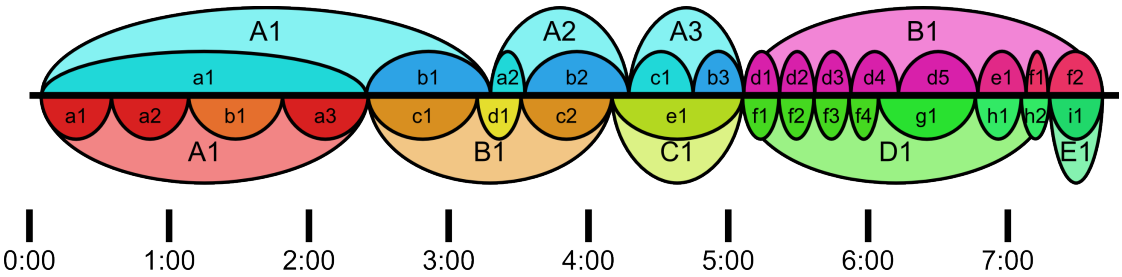


Figure 3.5: Analyses of Performance no. 2 by Annotator 1 (top), and Annotator 2 (bottom).

The improviser's part does change dramatically at Annotator 2's D_1, while Mimi continues in the same vein. Annotator 2 separated D_1 from C_1 because the figure played by the improviser in D_1 was not only musically distinct but, with its descending triads and relatively thin texture, seemed to have a strong sense of ending function, whereas Annotator 1 attended to the continuity in the melodic material in Mimi's voice.

3.2.2 Performance no. 2

The most striking differences between the annotations (see Figure 3.5) are in the grouping and labeling of the first five minutes: Annotator 2's A_1 is subdivided further than Annotator 1's a_1; the placement of the boundaries near 3'11" and 4'17" are disputed; and the larger section that encompasses Annotator 1's b_1 and Annotator 2's c_1 is disputed. There are also some subtle differences in the labeling of the subsections in the last two minutes.

(1) Why does Annotator 2 subdivide Annotator 1's a_1 further?

Annotator 1 attuned to the textural similarities that joined his a sections (their atmospheric quality) and their contrast with the b sections (louder and rhythmic, i.e., with a strong and regular pulse). Annotator 2 identified the same contrast between the material designated a_1 and b_1 by Annotator 1, but, as with Performance no. 1, he made further subdivisions and associations based on recurring melodic motifs: if treating a_1 as the opening theme, both a_2 and a_3 begin with the first part of the opening theme,

and a_2 ends with the ending of the opening theme.

(2) Why does Annotator 1 hear the first five minutes as a series of binary groups (A_1, A_2, A_3), whereas Annotator 2 hears a duo of ternary groups (A_1, B_1) with an additional section (C_1)?

The annotators agree that Annotator 1's b_1 presents a contrast to all that precedes it, but disagree about the larger structural interpretation of b_1. To Annotator 1, section a_1 was internally self-similar, and so the change at 2'25" (at the beginning of his b_1) struck him as the midpoint of a larger grouping. This hearing was reinforced by the subsequent alternation of atmospheric a_1 material and rhythmic b_1 material as repetitions of this binary structure. In contrast, Annotator 2 had already heard a ternary-like structure in the material preceding 2'25" (*aaba*) and so was inclined to hear the material in section B, with the entrance of a new quarter-note triplet motive, as beginning a new section, also ternary (*cdc*). In hearing things this way, he overlooked the similarity of d_1 to the opening material, instead focusing on the broad textural self-similarity of his B_1. (See 3.6.)

(3) Why does Annotator 2 not identify either of sections d_1 or e_1 as being a repetition of previous material?

Annotators 1 and 2 characterized section a_1 differently: Annotator 1 heard a long self-similar span with a particular texture, and hence easily associated the return of this material in his a_2. To Annotator 2, a_1 was a melody, which recurred in varied form in a_2 and a_3. With this in mind, he heard the return of the theme in d_1 as a severe truncation of the theme, a kind of quotation in an otherwise distinct passage.

(4) Why do the annotators disagree about the placement of the boundaries near 3'11" and 4'17"?

In a_2, Annotator 1 heard a return to the opening material, and hence his section begins at the onset of restatement of the theme (see 3.7); in d_1, Annotator 2 heard a

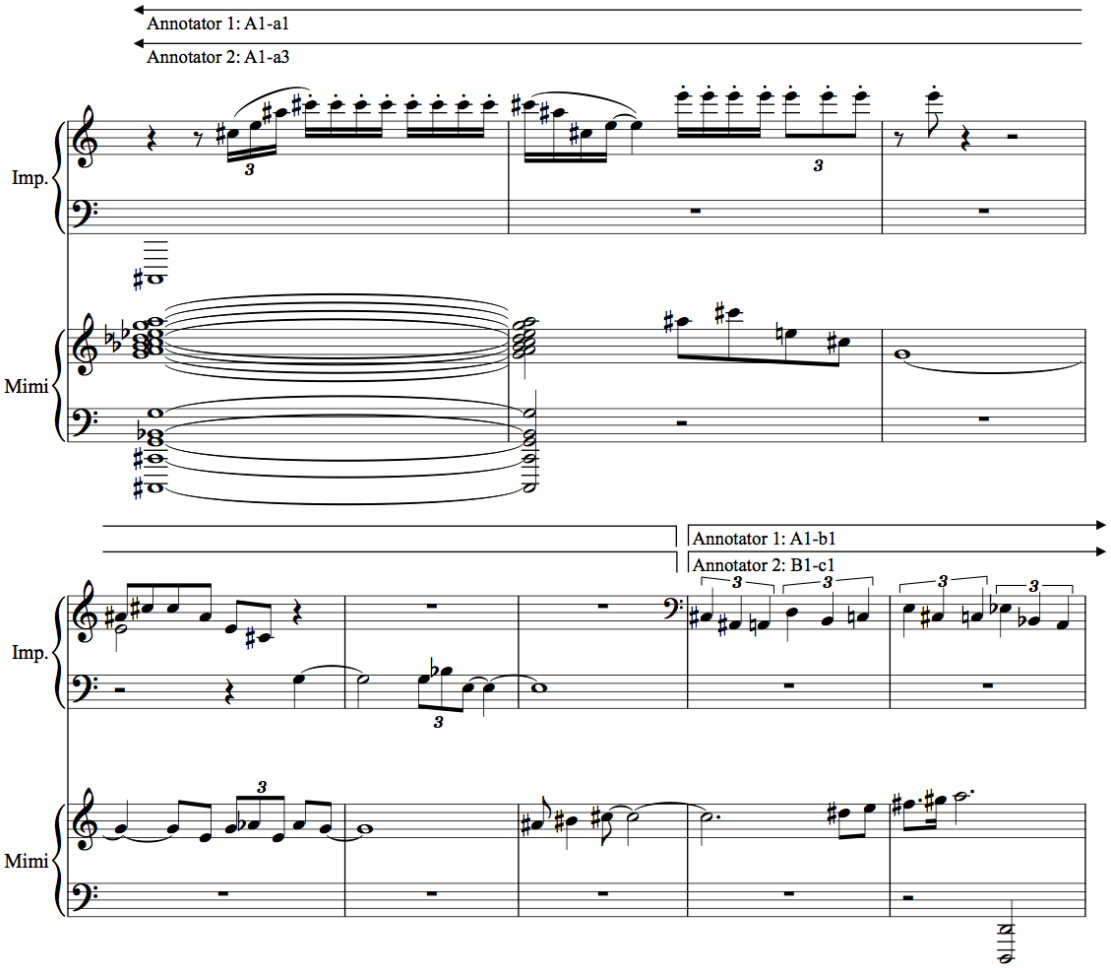


Figure 3.6: Notation Example 3. Performance no. 2, large section boundary (A_1–B_1) for Annotator 2.

brief reprieve between statements of the c_1 material, and hence identified the moment where we deviate from the material of c_1 as the boundary. Both annotators recognized the introduction of new material by the improviser at 4'17", and Annotator 1 placed his boundary (the beginning of c_1) in line with this. Annotator 2 placed the boundary (e_1) earlier, at the onset of a stark register shift at 4'10".

(5) Why do Annotator 1's d_4 and Annotator 2's g_1 overlap (6'04" to 6'13")?

Both listeners perceived that this final section (Annotator 1's B_1) begins with the improviser and Mimi engaging in an approximate canon with a period of about 15 seconds between voices. This pattern breaks down shortly after the 6'00" mark. Here, Annotator

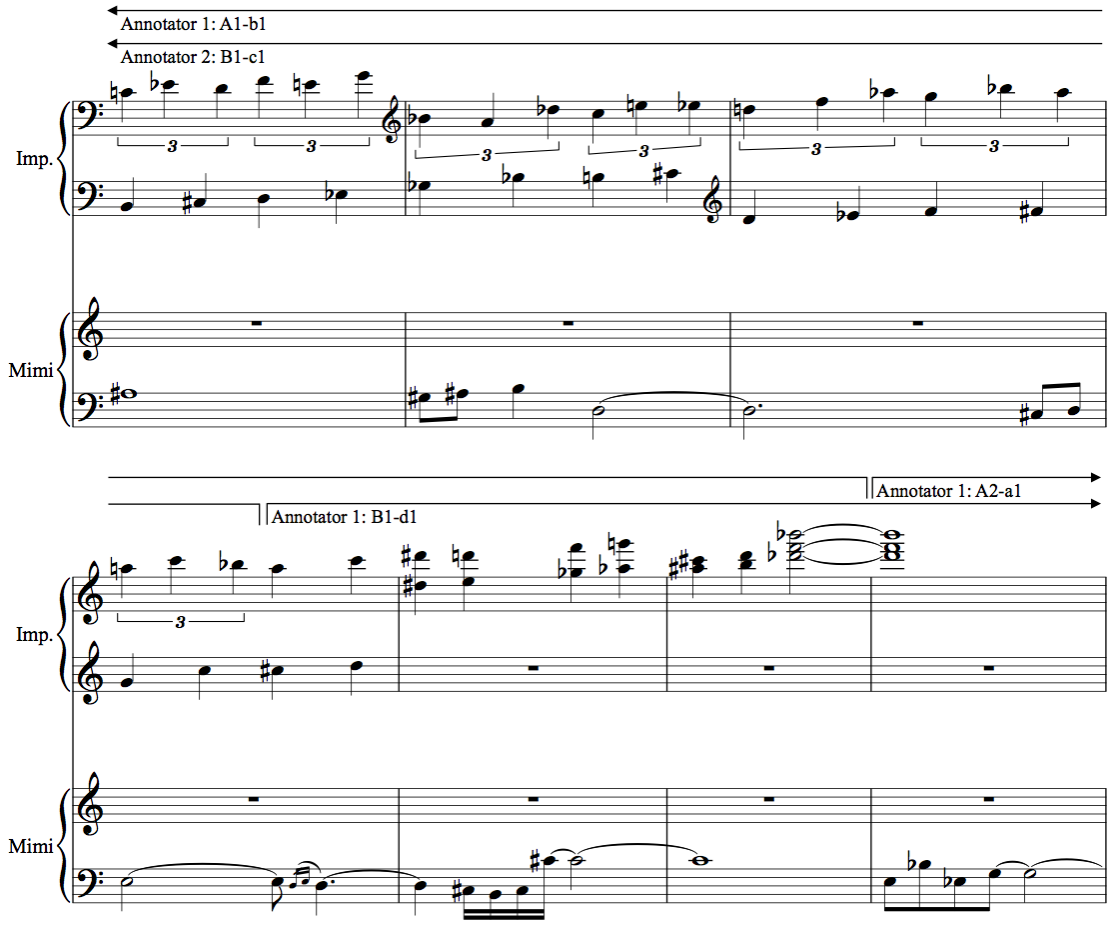


Figure 3.7: Notation Example 4. Performance no. 2, boundary discrepancy at 3'11"–3'16."

1 heard a prolongation of the last phrase (d_4), followed by a new section in which the improviser introduces a new musical idea in the lower register while Mimi continues with the canon material. Annotator 2 did not focus on the new theme, and instead heard at g_1 an accelerated continuation of the canon between the voices. This canon has a much shorter period of a few seconds, the improviser and Mimi now trading gestures rather than phrases.

(6) Why is the span from 7'05" to 7'17" (Annotator 1's f_1) grouped with the subsequent material (f_2) by Annotator 1, and with the preceding material by Annotator 2 (h_1/h_2)? And why is Annotator 1's f_2 given its own large-scale section by Annotator 2?

From 7'05", the improviser introduces two contrasting ideas: a loud, downward-

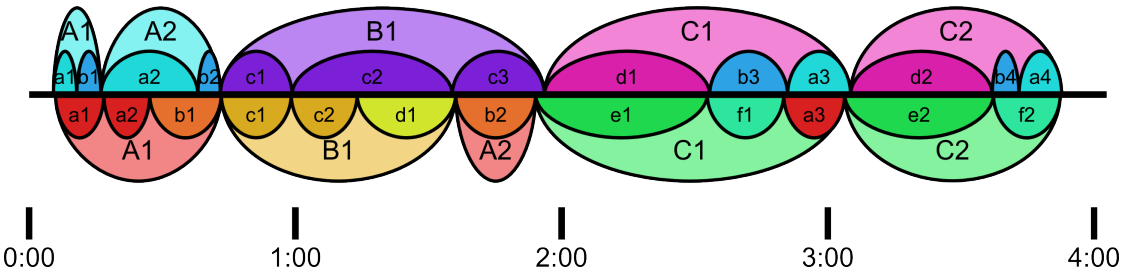


Figure 3.8: Analyses of Performance no. 3 by Annotator 1 (top), and Annotator 2 (bottom).

leading progression (Annotator 1's e_1), and an ethereal theme played sparsely in a high register (Annotator 1's f_1). These are repeated by Mimi in canon in Annotator 1's f_1 and f_2; in the latter of these, the improviser also provides sparse accompaniment. Since he marked e_1 and f_1 as distinct, it can be seen that Annotator 1 focused on the difference between the themes introduced by the improviser. On the other hand, Annotator 2 focused on the repetition of the louder, more prominent musical idea in sections h_1 and h_2. This meant that he heard a greater degree of discontinuity between h_2 and i_1 than did Annotator 1. This abrupt change to a sparse texture, suggestive of a concluding function, also led Annotator 2 to indicate a higher-level boundary between large-scale sections.

3.2.3 Performance no. 3

In contrast to Performances no. 1 and no. 2, Annotator 1's and Annotator 2's analyses of Performance no. 3 (which were created before the listeners had conferred on Performance no. 2) are largely in agreement, especially with regard to the larger sections (i.e., the uppercase letters). Most of the differences can be understood in terms of attending strategies: Annotator 1 paid the most attention to motivic recurrence, while Annotator 2 paid the most attention to surface qualities (e.g., register and texture). However, thematic segmentation also played a role in differentiating the interpretations: Annotator 1 segmented the opening theme into individual motives, while Annotator 2 did not segment the theme. This had implications for the final section of the performance, when

this thematic material returns.

(1) Why does Annotator 1 subdivide Annotator 2's opening section a_1 into two subsections? (This also applies to the subdivision of the last section, Annotator 2's f_2.) Further, why does Annotator 1 further subdivide Annotator 2's A_1?

Annotator 1 heard the opening 10 seconds (from 0'05" to 0'16") as ab and Annotator 2 heard it as simply a. The difference may hinge on a matter of metrical interpretation, and since there is no "ground truth" set of intended note lengths, the preferred interpretation is a creative choice. In Figure 3.9, the notes at the boundary between a_1 and b_1 are notated as triplets, suggesting rhythmic continuity. However, if the notes are instead heard as quarter notes, as shown in Figure 3.10, Annotator 1's boundary now falls between gestures (instead of in the middle of a triplet), emphasizing the shift in register at the proposed boundary. (In both of these examples, barlines are chosen to emphasize certain patterns and divisions; no particular meter is implied.) This is a vivid example of how a structural analysis can depend on how the listener has made sense of the fundamental units of the piece, an issue discussed by [Bam06] to which we will return in Section 3.3.

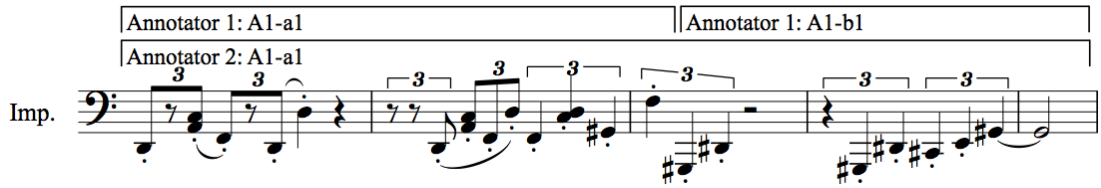


Figure 3.9: Notation Example 5. Opening of Performance no. 3, notated as triplets.

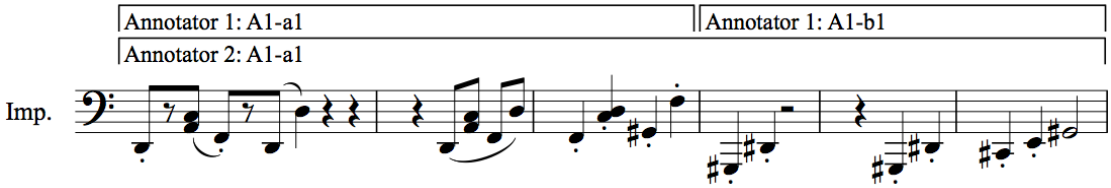


Figure 3.10: Notation Example 6. Opening of Performance no. 3, notated as straight eighth notes.

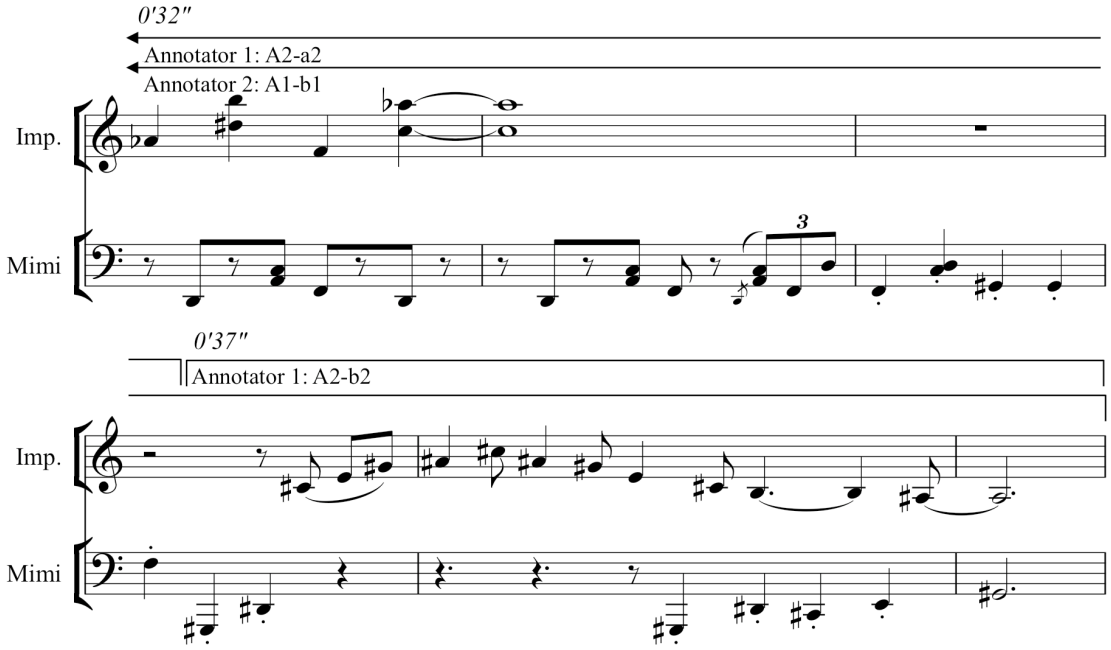


Figure 3.11: Notation Example 7. Large section boundary (A_1–A_2) for Annotator 1.

This initial discrepancy meant that Annotator 1 was more focused on segmenting the rest of the opening A section according to the recurrence of these separate a and b ideas. For instance, after the end of the initial idea, Annotator 1 placed his next boundary at b_2, where Mimi repeats the material of his b_1 section and the improviser introduces a new gesture (see Figure 3.11). Annotator 2 placed his next boundary earlier, at his b_1, citing a significant change in register and accompaniment; his b_1 section is united by Mimi's use of a_1 material and the improviser's presentation of contrasting, non-thematic material. Annotator 2's conflation of the two parts of the first idea also led to his fusing the two last segments in Annotator 1's analysis (b_4 and a_4) into one (f_2).

(2) Why do the annotators disagree about the grouping of the material from 0'27" to 0'37"?

Annotator 2 perceived a strong change at 0'27" (b_1) as the improviser introduced new accompanying material. In contrast, Annotator 1 heard a continuation of the a_1 material in Mimi's voice, and Mimi eventually returns to the b_1 material at 0'37".

(3) Why does Annotator 2 mark boundaries at 0'27" (b_1) and 1'13" (d_1), when Annotator 1 does not?

As stated before, Annotator 2 heard a discontinuity at 0'27" in the improviser's material. But Annotator 2 partly attributes both boundaries to the pauses that precede them. In both cases, the thematic continuity of the section led Annotator 1 to forego an additional boundary.

(4) Why does Annotator 2 recognize a return of material from the opening section at his b_2 when Annotator 1 does not?

To Annotator 2, b_1 was characterized by Mimi's playing fragments of the original motive, with the improviser adding novel accompaniment. Thus b_2 represented a return to this configuration. In contrast, Annotator 1 felt that this section continued the chaotic, fragmented feel of section B_1.

(5) Why does Annotator 2 further subdivide Annotator 1's B_1?

Given that Annotator 1 did not identify c_3 as a return to material from the previous section A_1, his choice of the large-scale grouping (AB) is no surprise. Annotator 2 did identify a return to the previous section at b_2, and the significance of this return led him to hear a larger-scale ternary grouping, ABA.

(6) Why does Annotator 2 not recognize a return of material from A at his f_1?

At Annotator 1's b_3, Mimi repeats the pattern played at b_1 by the improviser, who then responds with a melodic inversion of the material. The counterpoint is repeated at Annotator 1's b_4, with the parts swapped: the improviser plays the original ascending b_1 motif, and Mimi repeats the inverted theme from b_3. As the improviser, Annotator 1 recalls these imitations being deliberate, and hence was aware of their relationship to the earlier material at the time of performance. However, Annotator 2 was not aware of the repetition until it was pointed out to him! This oversight can possibly be explained by Annotator 1's b idea having less primacy in Annotator 2's analysis. Since Annotator

2 did not hear it as a "head" of any section, he was less apt to hear just the "tail" of the opening theme return, either at 2'33" or 3'37" (Annotator 1's b_3 and b_4)—even though he heard these as repetitions of each other.

3.3 Discussion

The questions we ask in this chapter are: in what ways may two listeners disagree about the structure of a piece of music, and what factors cause or explain these differences? In the previous section, we presented the analyses produced by two listeners of three improvised pieces, and enumerated the differences between them. We also sought to explain how each difference arose by referring to the listeners' introspective notes on why they made the decisions they did. We are now interested in following the chain of causation backward, first considering the proximate causes of the disagreements—the circumstances that explain the disagreements most immediately—and extrapolating from these possible ultimate causes. In this section, we discuss these causes in a loose progression from most to least proximate. As the causes get deeper, they become more speculative but also, we suggest, more important and illuminating.

3.3.1 Factor 1: Attention to musical features

The simplest and most expected explanation for why the two listeners disagreed is that they paid attention to different musical features. For example, in Performance no. 1, the annotators segmented A_1 differently because Annotator 1 found the shifts in register more salient, while Annotator 2 paid attention to the pauses and melodic gestures that supported a regular phrase rhythm. They also gave these subsections different labels because the former focused on the textural similarity between them, and the latter on the slightly different motives in each.

Both annotators reported attending to a similar set of musical parameters at various

times: melodic themes and their repetition; rhythm, texture, and register; and whether Mimi or the improviser were playing a particular part (recall that these two voices had different timbres). Still, sometimes annotators attributed their decisions to parameters not mentioned by the other; for example, Annotator 2 invoked the perceived function of a section to justify some of his decisions, but Annotator 1 never indicated that this was an important attribute. (This occurs with the concluding sections that Annotator 2 heard at the end of Performance no. 1 and no. 2, and in the preparatory e_1 section that he heard in Performance no. 1.) On the other hand, in Performance no. 3, Annotator 1 identified a melodic inversion at b_3, which Annotator 2 did not attend to.

The annotators did not seem to consistently prefer one musical attribute over another: in the disagreement over the labeling of the final three subsections of Performance no. 2 (eff vs. hhi), it was Annotator 2 who found the overall texture salient, whereas Annotator 1 paid attention to the different themes being played by the performer. But in their analysis of section A_1 in Performance no. 1, the annotators focused on the opposite features.

The instances where the function of a part was cited as a reason to segment or differentiate it recalls the observation of Peeters and Deruty that music structure is multi-dimensional, consisting of attributes that can be independent, such as musical function, similarity, and instrumentation [PD09]. In this view, some disagreements could be attributed to listeners focusing on different dimensions of structure, although it remains to be explained why some people focus on different dimensions to begin with. As explained in Chapter 2, Peeters and Deruty proposed an annotation format that would separate these dimensions, a scheme that was adopted for SALAMI. If musical similarity were similarly decomposed, attention to different musical features could explain disagreements between listeners; this notion is explored in Chapter 5.

3.3.2 Factor 2: Opening moments

While most of the differences seem well explained by referring to the listeners' attention to different musical features, it is more concise to attribute later differences between two annotations to earlier differences. That is, how the listener happens to perceive the opening moments of a piece—what they initially perceive as the basic units in their chunking, or what they initially call A and B—appears to greatly determine how the rest of the analysis will proceed.

For example, in Performance no. 2, Annotator 2 heard an opening section A_1 as having a basically ternary structure; this may have encouraged him to perceive the following material (B_1) as a ternary grouping as well. Similarly, Annotator 1 heard a binary contrast within the opening section (A_1), which would reinforce the binary interpretation of the next two sections (A_2, A_3).

It makes sense that the opening moments would lay the framework for the rest of the piece, since they would strongly affect one's expectations. In Performance no. 1, Annotator 2 identified a regular four-phrase structure in the first section a_1; this seemed to lead him to expect a similar phrase rhythm in subsequent material, resulting in more regular section lengths. The opening moments establish for the listener what design principles the composer or improviser is using: what contrasts are relevant and what units can be repeated.

The opening moments were clearly crucial in Performance no. 3. Here, the opening 10 seconds crystallized in the mind of Annotator 1 as two distinct themes (a_1 and b_1), but as a single theme to Annotator 2 (a_1). The fact that the material Annotator 1 calls b_1 did not strike Annotator 2 as a distinct theme likely explains why Annotator 2 did not recognize the return of this b material later on as b_2, b_3, and b_4, even though he did recognize that b_3 and b_4 were similar to each other.

3.3.3 Factor 3: Difference in information

Assuming that the perception of the opening moments is crucial in forming an analysis, how is it that listeners differ in how they perceive these opening moments? A deeper explanation ought to include differences between the listeners that are present before the listening begins. Generally speaking, differences in information are anticipated as an important factor in psychological studies; for example, participants are classified as musicians and non-musicians (i.e., people with and without specialized musical knowledge). Here, we consider a more specific difference: a difference in the type and thoroughness of the knowledge each listener has about the piece.

Annotator 1, as the improviser in the performances, had a more intimate understanding of how the piece was constructed than Annotator 2 before each later listened to and analyzed the performances. This difference had an impact on the slightly different procedure used by the listeners: Annotator 1 tended to analyze pieces section by section, nearly finalizing his analysis of the first half before listening to the second half, for example. The ability of Annotator 1 to work through the large sections in series suggests that the large-scale analysis (or at least the large-scale segmentation) may have already been decided at the beginning of the annotation process. In contrast, Annotator 2 tended to work in parallel: he annotated boundaries in real time while listening through the whole piece several times, and in between auditions he re-listened to specific parts to adjust his annotations. This contrast between the listeners suggests an important difference in the initial information each had about the performance. To Annotator 1, the lay of the land was already well known; Annotator 2 had to do more scattered scouting before he could finalize his understanding of the large-scale patterns. While this observation may seem particular to the scenario at hand, comparable situations arise frequently among listeners: some analyze a piece only after becoming familiar with it as a performer or in casual listening, while others do so as new listeners.

Elizabeth Margulis has found that listeners who are less familiar with a piece of

music are more likely to focus on shorter repetitions, while those who are more familiar are likely to focus on longer repetitions [Mar12]. Extrapolating from repetition (which never occurs exactly in the three performances studied here) to similarity, we see the same pattern reflected in the differences between our annotations: in Performance no. 1, Annotator 2, the newer listener, subdivides A_1 more than Annotator 1 on the basis of a perceived phrase rhythm and on local changes in texture, whereas Annotator 1 focuses on the self-similarity of the entire passage. Similarly, in Performance no. 2, Annotator 1 points out what unites sections A_1, A_2, and A_3 at a large timescale, whereas Annotator 2 does not recognize these similarities. Finally, in Performance no. 3, although Annotator 1's conception of the opening moments at first appears more fine-grained than Annotator 2's, it leads to an analysis that recognizes more repetitions and returns globally, requiring only four distinct section types (a to d) compared to Annotator 2's six types (a to f).

The different levels of familiarity with the pieces also seemed to influence the musical features to which the listeners paid attention. Annotator 2 (whose annotations are generally more segmented than Annotator 1's) attributed more of his boundaries to surface features, such as long silences, sudden note clusters, and changes in register, than did Annotator 1. For example, Annotator 2's large-scale boundary in Performance no. 3 between B_1 and A_2 is attributed to a long pause. In the same performance, Annotator 2 starts his section C_1 where Mimi plays some disruptive clusters, whereas Annotator 1 begins C_1 a few moments later, when the improviser takes up the new theme.

One final difference in information is quite specific to the present circumstances but nonetheless bears mentioning: the fact that Annotator 1 was the improvising performer and hence had memories of creating the music. Annotator 1, being thus more aware of details such as what part of the oracle Mimi had access to and when Mimi was active and inactive, may have been less willing to give an analysis that did not reflect these events. For example, in Performance no. 1, his section B_1 exactly aligns with when Mimi was turned off; Annotator 2, however, heard parts of the previous section as being a part

of this transition section. In Performance no. 2, the oracle is cleared and reset only at the boundary between Annotator 1's A_3 and B_1; perhaps Annotator 1, knowing this, was less inclined to differentiate the large-scale subsections of each half with different letters, as Annotator 2 did. Memories of the performance may also have helped ensure that intentional but subtle repetitions, such as the return and inversion of an earlier motif in Performance no. 3 (at b_3), were reflected in Annotator 1's analysis. While the difference in information between the listeners in our case was extreme by design, listeners certainly differ along similar lines: access to the score may radically affect how a listener perceives the structure of a piece, and listeners may differ in their insight into the relevant instrumental practice (e.g., pianists and non-pianists analyzing a piano sonata) or prior knowledge of the specific piece being performed.

3.3.4 Factor 4: Difference in analytical expectations

Beyond the information the listeners had about this specific piece, we consider the role of information about music in general, involved here as analytical expectations. Some of our results suggest that the listeners may have had different *a priori* expectations about what the analyses would look like. Since the two listeners have different backgrounds and experience in music theory, analysis, and musical taste, it is difficult to speculate as to where these expectations would arise. However, the two sets of annotations differ strikingly in one property: the small-scale segments perceived by Annotator 2 tend to have more equal size than those of Annotator 1. For example, in Performance no. 2, Annotator 2's A_1 has 4 subsegments, each roughly one quarter the duration of A_1. Annotator 1's A_1, on the other hand, is subdivided highly asymmetrically. The trend appears to be somewhat consistent across the three performances, although a larger study would be needed to confirm this difference. If it were found to be a consistent trend, it may reflect a strong expectation on the part of Annotator 2 that subsegments will be of equal size. This is not an unreasonable expectation, given that composed music often includes repeating or contrasting sections with similar lengths, and may be shared

by many listeners. It would be interesting to determine whether this expectation affects how music is analyzed.

3.3.5 Factor 5: Analysis method

Finally, we acknowledge that some of the differences between the annotations arose from non-identical analysis methods. This is most important for Performance no. 1, in which Annotator 2 differentiated the subsections by letter, but Annotator 1 did not, saying it did not occur to him as an option. This was noticed immediately after the first analysis, and the issue was corrected before the next pieces were annotated.

However, even for these later analyses, the listeners used different annotation tools—Variations Audio Timeliner (VAT) for Annotator 1 and Sonic Visualiser (SV) for Annotator 2—and this may have affected the analyses more subtly. Both tools allow one to divide a timeline into segments and to label these segments. However, the bubbles in the visual interface of VAT may have emphasized the groupings and hierarchical relationships (see Figure 3.12), while the vertical segment lines in SV may have emphasized the segmentation (see Figure 3.13). SV also displays the audio waveform, showing how loudness varies throughout the piece, and this may have made the changes in loudness more salient. This difference in tools is consistent with one main difference in how the two annotators generally approached the pieces: Annotator 1 more often attended to groupings and to qualities that unified sections, while Annotator 2 attended more to local discontinuities and heard more boundaries.

3.4 Conclusion and future work

We examined two listeners' analyses of three improvised performances and found the differences between these analyses to reveal several insights.

Attention. First, we observed that these differences were often due to the fact

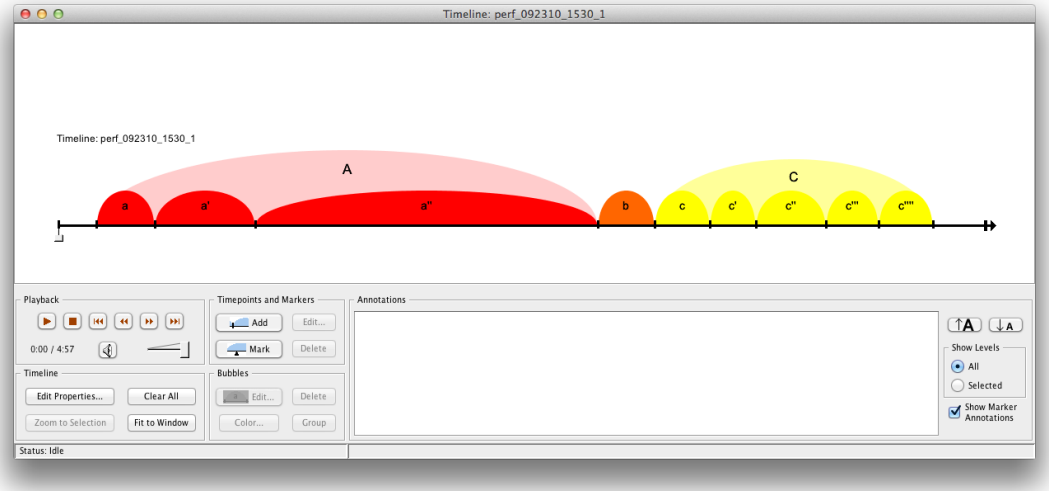


Figure 3.12: Screenshot of Annotator 1's analysis of Performance no. 1 in Variations Audio Timeliner

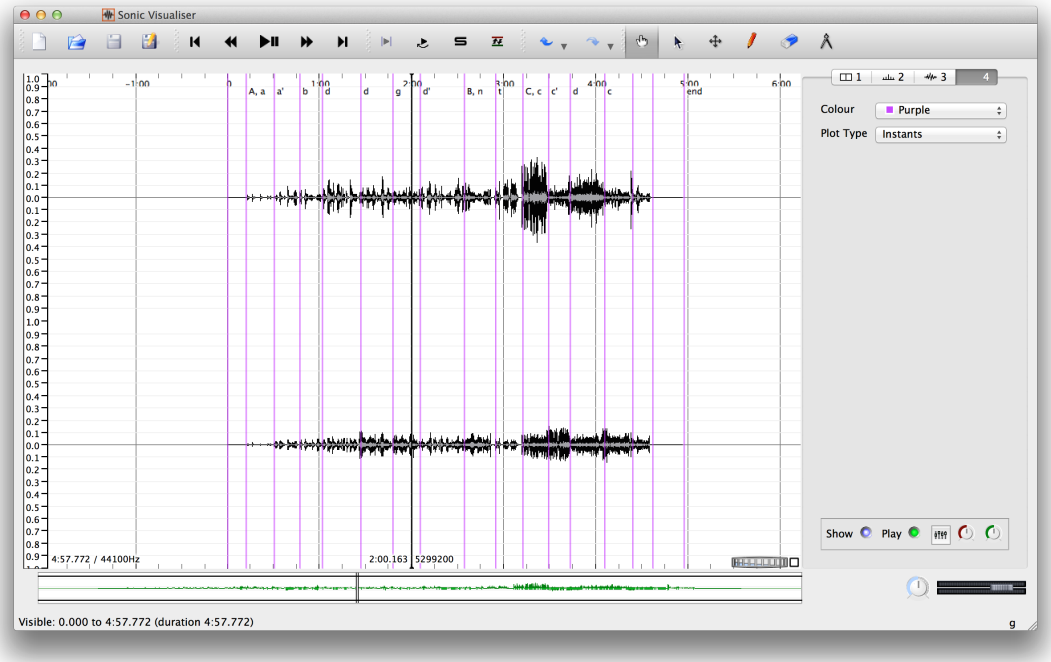


Figure 3.13: Screenshot of Annotator 2's analysis of Performance no. 1 in Sonic Visualiser

that the listeners paid attention to different musical features. Attention itself is already widely studied, but usually only as a global concept: researchers are interested in how much the listener is paying attention to the music, not what the listener attending to

in the music. For example, [AP09] showed that a model of attention and surprise can bear a striking resemblance, in practice, to a theory of musical structure, and attention is a key concept in existing theories of music such as Farbood's model of musical tension [Far12]. Jones and Boltz [JB89] have shown that paying attention to short and long timescales can affect one's perception of time, but it remains to be studied how this can in turn affect one's interpretation of musical structure. Since listeners are able to focus their attention on (or have their attention unwittingly drawn to) particular aspects of a piece of music—patterns in a vocal line, recurrences of a motif, shorter or longer timescales—we recommend following up this research in a way that treats music, the object of attention, as multi-dimensional. The way the attention of the listener wanders between these aspects could become the subject of a new theory of analysis.

Opening moments. We next noted that differences in how two listeners heard the opening moments prefigured most of the remaining differences. It appears to be in these opening moments that listeners decide what will comprise their basic units of analysis and what types of abstraction—melodic, textural, rhythmic—will serve them best. The mental representation formed here serves as a template, allowing the listener to form expectations about how the material will develop in the rest of the piece. One conclusion from this—that knowing how a listener understands the beginning of a piece allows one to predict how the rest will be understood—is a readily testable hypothesis that would form the basis of exciting future work.

Information. On a deeper level, we speculated that access to information could affect the area of attention or focus. In our case, Annotator 1 had more information than Annotator 2 in several ways: he had created and performed the piece, his memory helped him to better disentangle his own and Mimi's contributions, and he had simply heard the piece much earlier than Annotator 2. Several differences in the annotations seemed well explained by these differences in information. While these differences in information may appear to be circumstantial, comparable differences arise between listeners who have access to a score and those who do not, or among listeners who have listened to

a piece different numbers of times and whose familiarity with it varies—a factor whose importance has already been demonstrated in previous studies (e.g., [Bam06, Mar12]).

Analytical expectations. Finally, we found that listeners may bring *a priori* global expectations to the analysis. In the analyses studied here, this was suggested by the conspicuously regular phrase length indications of Annotator 2, which contrasted with the more asymmetric groupings in Annotator 1's analyses. These global assumptions are formed over the entirety of a person's listening history; they are based on familiarity with the style of music at hand, instrumental experience, and exposure, if any, to music theory, or to the piece in question. These analytical expectations may also influence how a listener initially understands the opening moments of a piece.

These four insights resonate with Bamberger's [Bam06] argument that perceptual disagreements among listeners can be ascribed to differences in ontology, which are in turn affected by a listener's values and belief system (which are shaped by the information they have, and reflected in their expectations) and preferences (which influence the features and relations to which they pay attention). Of course, while this system of beliefs appears to be the source of listener disagreements, a listener's analysis of a piece is still predicated on external cues present in the music: for example, prosodic cues (stresses, pauses, and shaping of tempo and loudness as communicated by the performer) or repetitions that guide a listener's attention or expectations.

The last observation, that the **analysis method** (meaning the annotation format and the tools used) will affect the outcome, is interesting but was not the objective of this research. As explained in Section 3.1, the choice of annotators—two listeners with very different prior relationships to the piece—and the choice of musical materials—improvised pieces that lack exact repetitions—were made to elicit a rich set of disagreements. These were deliberate choices, but the difference in software was happenstantial; each listener simply used the software that each was more comfortable with. And since we planned to work with the same analytical constraints (e.g., use Lerdahl and Jackendoff's well-formedness rules; describe the structure at two timescales), we did not think

much of the different tools. As has been observed in the MIR research on annotation methodologies, it can be deceptively hard to constrain the analysis task for listeners to ensure that, notwithstanding all their differences in attention, knowledge and expectations, they are doing the same thing.

Although the conclusions of this case study have been carefully reasoned from our annotations and our accounts of why we heard what we heard, two of our choices in design remain shortcomings. First, the small size of this study means we cannot confidently extrapolate our conclusions to all listeners and all disagreements. It remains a task for future work to demonstrate whether the rough chain of causality that we have proposed—which proceeds from *a priori* expectations, to knowledge about the particular piece, to how the opening moments are heard, down to how one's attention is directed throughout a piece—is in fact a good general description of how listener disagreements arise. Second, the objectivity of our results could be questioned, since we two listeners doubled as the two meta-analysts. This arrangement certainly made it easier to discuss the musical analyses in minute detail and to understand the subtle influences at play (detail which would have been difficult to match by interviewing others), but we cannot say how two observers of our discussion would have made different conclusions.

Thus, future work could take the form of very similar studies with more listeners in which the differences between the listeners are narrower, and the sources of disagreements fewer in number. For example, in a study with two non-performing listeners, neither of whom had heard the music before, the importance of prior information should be reduced. Moreover, if both listeners were explicitly coached to hear the opening moments in the same manner, and then asked to complete the analysis, subsequent disagreements should not have depended on this factor. (The fourth experiment reported in Chapter 6 presents a comparable scenario.) Although the present study was devised as a contribution to music theory, these proposed extensions would be contributions to music psychology.

Notwithstanding concerns about how the present study was conceived, the outcome is a rich set of new questions to explore. For example, how do people's musical backgrounds

affect their perception of structure? This question could potentially be addressed with experiments comparable to Margulis's [Mar12] (in which listeners were asked to identify repetitions), but where listeners' musical background were catalogued in detail.

One question inspired by this study is pursued later in this thesis: what musical features do listeners pay attention to, and does this directly impact their perception of a piece's structure, as we have suggested here? At first glance, it seems that answering this question directly would require an auditory attention-tracking system, some analogue to the eye-tracking systems used to study visual attention. Since none exists, a carefully constructed set of artificial stimuli will be necessary to study this question, and this is the main project undertaken in Chapter 6.

We would also like to know how quickly listeners decide on a set of basic musical ideas when they begin to listen to a piece of music, and how definitively this guides their interpretation of the piece. Supposing a listener devises a running hypothesis of the piece's structure while listening to it, how easily or how frequently is this hypothesis revised? What kinds of musical events are capable of causing this? If listeners are permanently beholden to any aspect of their first impressions, this has wide-reaching implications for those who make music. Although the experiments in Chapter 6 do not address this broader question, Experiment no. 4 begins this work by establishing the plausibility that listeners are able to continue an analysis in the face of new changes after having committed to an initial decision.

Chapter 4

An analysis of boundary perception and musical features

The previous chapter's case study re-demonstrated one fact that was known from the literature: that a listener is likely to attribute their perception of a boundary to a stark change in the music. This assumption has been incorporated into a majority of segmentation algorithms in the fields of computational musicology and music information retrieval. However, it is not clear whether the perception of a boundary is actually attributable to a stark change in the music, or whether listeners merely make this attribution *post hoc* when asked to explain themselves. In a study on the visual cues people used to estimate whether people were intensely happy or sad, [ATT12] found that people tend to attribute their judgements to facial expressions, even when these were non-informative (in their examples, emotion was in fact indicated by body language). They called this the "illusory facial effect." It could be that there is an "illusory novelty effect" in music: a tendency to attribute boundaries to novelty when in fact novelty is uninformative. We approach this question by looking at how acoustic novelty relates to the perception of boundaries in a large and varied corpus.

Our study asks two questions. The first is relevant to MIR: is the connection between

novelty and boundaries a general one that holds over all genres, musical features, and timescales? The second is relevant to music psychology: does the connection between novelty and boundaries hold in full audio contexts, in diverse corpora? Together, these questions are relevant to our study of listener disagreements, since so far we have relied on listeners' own justifications for their analyses to understand their differences. We would like to know if there is a gap between listeners' analyses and the standard explanation of novelty, and whether this depends, for example, on the genre or type of acoustic change being considered.

Towards this end, we present in this chapter a statistical analysis of a large corpus of recordings whose formal structure was annotated by expert listeners. From each recording, we compute several novelty functions, which measure the rate of change of acoustic properties at different timescales. Our findings corroborate those of previous perceptual experiments: nearly all boundaries correspond to peaks in novelty functions. Moreover, most of these boundaries match peaks in novelty for several features at several timescales. We observe that the boundary-novelty relationship can vary with listener, timescale, genre, and musical feature. Finally, we show that a boundary profile derived from a collection of novelty functions correlates with the estimated salience of boundaries indicated by listeners.

4.1 Introduction

4.1.1 Background

In Chapter 2 we noted some drawbacks among existing musicological models of musical structure. Chief among them was that most algorithmic implementations of these theories, such as GTTM (e.g., [HHT06, FC04]), LBDM [Cam01], Grouper [Tem01] and IDyOM [PMW10a], target the simplest musical context: monophonic melody. This is because as the musical context expands from monophony to polyphony, or from monotim-

bral to multitimbral music, the estimates of the bottom-up models become less reliable as top-down influences, such as stylistic expectations or the recognition of repetitions, guide the perception of boundaries more strongly. The same happens as the musical context expands from short phrases to full pieces: at longer timescales, it is more likely that the rules governing the analysis may conflict, and the predictions will be muddied by factors that are hard to model, such as parallelism. Models that do deal with parallelism, such as [Cam06], use only short, immediate repetitions. A second drawback is that while the models can all claim some degree of generality, the focus on melodic segmentation hints that they mainly apply to Western tonal music, even though IDyOM, for example, uses unsupervised learning and has succeeded in a variety of contexts, from chorale melodies to folk music [PW06].

Hence there is a need for general models of the perception of structure in full-textured, polyphonic music. Since gathering listener responses to full pieces of music is expensive and impractical for a reasonable-sized experiment, we propose to take advantage of comparable resources that the MIR community has created: ground truth collections of structural annotations. A ground truth annotation is a description provided by a listener that is assumed to be the sole correct formal analysis. Of course, no such absolute truth exists, as the copious evidence of listener disagreements presented in Chapter 2 attests. However, even those studies that found listeners marking boundaries at different times (e.g., [CK90], [BMK09]), it was mainly observed that despite this variation, listeners will agree on many boundaries, and especially on the most important ones. Thus, we may hope that ground truth provided by one listener represents how many listeners might hear a piece of music.

Collections of annotations are mainly used to measure the effectiveness of algorithms. What if we treated these annotations not as tools, but as objects of study in themselves? Since each annotation reflects a listener's perception of a piece of music, we can analyze the annotation to test basic assumptions about how music is heard. Thinking of annotations as objects of study rather than tools for studying algorithms, we may actually

derive some interesting conclusions about music cognition from the existing MIR literature. For example, [TLPG07] used machine learning to classify points in a recording as either boundaries or non-boundaries, and found that of over 800 feature dimensions considered, all three timescales and all four feature classes (harmony, melody, timbre and rhythm) were represented among the 20 most informative. This suggests that listeners are likely to integrate information from many musical parameters at many timescales when judging the location of boundaries. Paulus and Klapuri [PK08a] found that, when searching for similar sequences in music, it was optimal to calculate audio features over short time windows, but when searching for similar homogenous sections, a longer window was preferable. This may be evidence that when listeners judge two sections to be similar based on repeated sequences, the sequences they attend to are relatively short, whereas when listeners judge two sections based on their having an overall similar sound, this has been determined over a longer timescale.

One drawback of using ground truth collections to investigate perception is that collections of annotations tend to include just one listener's analysis per piece. In listening studies in music psychology, it is more common to collect twenty or more responses per piece. This drawback is offset by the opportunity to study far more pieces: compared to the six songs heard by 21 listeners in [BMK09], the corpus studied in this chapter has two listeners per piece, but 746 pieces.

An important question about the perception of boundaries is why listeners make the boundary indications they do. Both [CK90], who studied 20th-century and Classical music, and [BMK09], who studied popular music, collected free responses from participants about what cues they were attending to when they indicated a boundary. In both cases, listeners mostly indicated that a change in a particular parameter, such as timbre, rhythm, melody, register, articulation or harmony, motivated the response, while several indications were also attributed to parallelism or to a pause or break. Deliège [Del87], in testing the applicability of GTTM's grouping rules to perception, found that the salience of the rules differed with regards to implying boundaries. Sanden, Befus

and Zhang [SBZ12] asked listeners to indicate boundaries while paying attention only to a single musical feature, such as timbre or harmony, and found that the resulting segmentations varied in how well they related to the overall perceived structure of the songs, and hence that the features were of varying importance to the listeners.

The cumulative evidence points to novelty being important to every genre. However, none of these studies looked at genre effects systematically. (Sanden et al. [SBZ12] did study differences in genre but only used one exemplar song per genre.) Second, although common sense tells us that novelty certainly plays a large role in the perception of boundaries, it could still be that there is an "illusory novelty effect" in music: a tendency to exaggerate the importance of novelty in accounting for boundaries. A third concern is that the studies cited above are limited in the number of pieces considered. Their evidence is supported by hand-picked, often very short stimuli that present exactly the musical contrasts being investigated. This is a result of the unfeasibility of collecting listener's responses to large numbers of long stimuli.

These three concerns are addressed in the present study. By using corpora of structural annotations created by the MIR community, we test the conclusions of previous research on a much larger scale than previous work, with many more pieces and more pieces per genre. Also, by comparing annotations not to the output of algorithms but to the acoustic novelty of the songs, we test how well the presumed explanation of listeners accounts for actual analyses of music.

4.1.2 Proposed experiment

We conduct a series of analyses of the relationship between structural annotations—records of how structure was perceived by listeners—and features of the recordings, a record of what they heard. We first test the hypothesis that boundaries correspond to moments in the recording at which relevant musical features change greatly. This is done by computing novelty functions with respect to many features for a large corpus,

and evaluating the match between the most novel points and the annotated boundaries. Second, we investigate how the outcome of this comparison depends on the listener, the genre of the piece, and the musical features considered. Finally, we examine how the consensus novelty among musical features correlates with the consensus boundary indications of listeners. Our approach is similar to [SBZ12], in which listeners were asked to segment eight pieces while paying attention to a single musical feature. In their case, the responses were compared to the perceived structure of the pieces; in our case, we compare the separate acoustic properties of the signal to the perceived structure.

The present work stands out from previous research in some important respects. First, our musical stimuli are complete, full-textured recordings, rather than short excerpts or simplified stimuli such as melodies or MIDI renditions. Second, our study does not focus on a narrow genre of music; since the present investigation spans a wide range of genres, our observations may be more generalizable. Both of these differences lend our analysis an ecological validity that can be difficult to achieve in an experiment using few or artificial stimuli. Finally, our methodology is notable since, rather than collect data from an experiment, we mine insight from a large dataset developed for other applications.

The rest of the chapter proceeds as follows. Section 4.2 describes the corpus of annotations used as if it were an experiment in music psychology rather than a study in MIR. Section 4.3 describes how the features and novelty functions were computed from the recordings. The experiments are described in Section 4.4, with the results and discussion of each presented in turn in Sections 4.4.1 through 4.4.3. The implications of the results are discussion in the conclusion, Section 4.5.

4.2 Materials and methods: the SALAMI dataset

The data analyzed were originally created for the Structural Analysis of Large Amounts of Music Information (SALAMI) project. The SALAMI project's goal is to use automatic

structural analysis algorithms to analyze several hundred thousand musical recordings, which would allow musicologists interested in form to pursue research on a scale that was previously impossible. The project funded the creation of the largest ever corpus of human-generated structural annotations in order to demonstrate the effectiveness of these algorithms [SBF+11]. This corpus contains descriptions of nearly 1400 recordings, nearly 1000 of which were each analyzed by two independent listeners. Annotations for half of the total collection have been released to the public domain; the private half, which was not used in this study, will be released after serving for a few years as a benchmark dataset for evaluations at the Music Information Retrieval Evaluation Exchange. The SALAMI data are described briefly in this section; a complete account of its design and its properties can be found in [SBF+11], and the "Annotator's Guide" used as a reference by the participants is available on the SALAMI website.[1]

4.2.1 Participants and apparatus

The nine annotators (four men, five women) hired to provide annotations were all in their 20s and pursuing an advanced degree (Master's or PhD) in either music theory or composition. They were trained to use Sonic Visualiser, a powerful software package that allows quick data entry and navigation of the recording, and they could use any means to listen to the music.

4.2.2 Stimuli

The SALAMI collection contains roughly one quarter each of popular, jazz, classical, and world music. An additional portion was drawn from the Live Music Archive (LMA), consisting mostly of popular and jazz recordings. Of the public half of SALAMI, 761 recordings were considered: 498 were annotated by two listeners and 263 by one listener. A breakdown of the number of annotations within each genre is given in Table 4-A.

[1] `http://salami.music.mcgill.ca/`, `http://www.music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf`

Genre	One annotator	Two annotators
Popular	51	101
Jazz	10	112
Classical	44	65
World	30	78
Live Music Archive	113	143

Table 4-A: Number of recordings analyzed according to genre and number of annotators

All recordings were mp3s with 44.1 kHz sampling rates. The sound quality varied somewhat between files—while most mp3s had a bit rate between 128 and 192 kbps, some had variable bit rates and others had bit rates as low as 96 kbps—but none of these differences were expected to affect listeners' perceptions of structure, and this is not investigated here. Indeed, the poor sound quality of the original recordings was often a greater concern: the LMA includes some audience recordings of live concerts, which may include background noise or clipping. SALAMI's annotations do not record the listeners' familiarity with the music. It is unlikely that any annotator had heard much of the corpus before given the extreme breadth of the corpus, but it is also unlikely that the occasional hits in the collection, such as Michael Jackson's "Thriller," were unknown to the annotators.

4.2.3 Procedure

The annotators' descriptions were multi-dimensional in that three kinds of information were indicated separately: musical similarity (which was annotated at short and long timescales), formal function (e.g., "chorus" or "transition" labels), and lead instrumentation. Only the long timescale of the musical similarity layer was considered in the present research. In this layer, annotators indicated boundaries and provided uppercase letter labels ("A", "B", etc.) to indicate which sections were similar or shared the same fundamental musical idea. Annotators decided for themselves whether the unifying idea was primarily harmonic or melodic, or due to some other musical attribute. Labels could be inflected with a prime symbol to indicate substantial variation. Annotators

were encouraged to indicate on average five distinct uppercase letters per song, and to align their analyses with the metrical grid of the piece, if applicable, so a section beginning with a pickup would be annotated as beginning on the down beat. An example pair of annotations is shown in Figure 4.1.

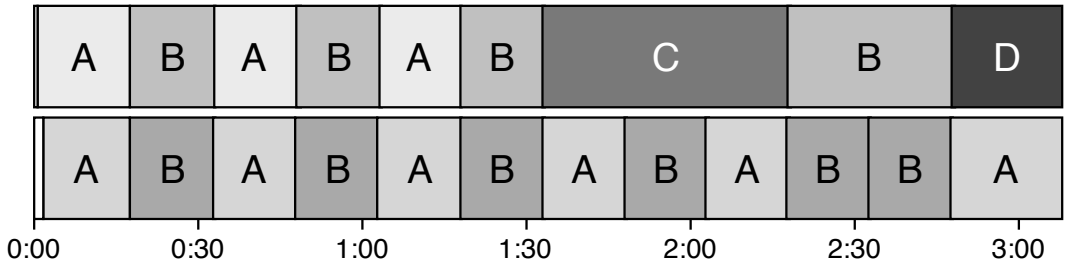


Figure 4.1: Two annotations for the song "Ain't Too Proud To Beg" by The Lost (salami_id: 1420). The shading of the segments emphasizes the labels within each annotation separately.

It should be noted that the annotators did not indicate any written justifications for the perceptions, such as "this point is a boundary because it marks a change in harmony." Thus, we are taking it for granted that the SALAMI annotators, like the participants in previous studies (including [CK90] and [BMK09]), would have justified many of the boundaries with a change in some musical attribute. Certainly, it seems that something *ought* to have changed when a boundary lies between two sections with different labels, and hence with different fundamental musical ideas.

Annotators used Sonic Visualiser according to the following workflow: first, listen through the full piece and indicate section boundaries in real time by pressing a key. On a second listening, pause to correct or adjust the position of boundaries as necessary. Next, provide labels for each section in each of the three layers—similarity, function, lead instrument. Finally, after skipping around to make corrections or resolve ambiguities as necessary, listen through the song a final time to confirm. The number of times each recording was fully heard is not known, but was requested to be at least three.

4.3 Data analysis

Structural analysis algorithms are commonly evaluated by executing the algorithm on a recording and grading the result against a ground truth annotation. This grade is difficult to interpret in isolation, so to fairly assess the significance of the result, a baseline approach, such as an algorithm that outputs random analyses or that makes predictions according to some naïve approach, should be executed on the same corpus.

In contrast to a typical evaluation, our goal is to study the annotations themselves, and not the effectiveness of an algorithm. Thus our analysis proceeds in an inverted manner: instead of comparing how well a given algorithm and a naïve baseline approach can predict the boundaries in an annotation, we compare how well the annotated boundaries and a random baseline set of points (non-boundaries) can predict the output of an analysis algorithm. In our case, this algorithm is based only on the rate of change of selected musical features. Our approach will effectively measure the amount of information that the annotations contain regarding these changes.

The following subsections describe the audio features used to characterize the music, the steps used to estimate the points of greatest musical change, the manner in which random non-boundaries were selected, and the comparison metrics used in the evaluation. The outcome of the experiment depends upon all of these choices: different changes in the music might be captured by different features and novelty-detection algorithms, and different evaluation metrics could tilt the results in different ways. These choices limit the scope of the experiment and should be kept in mind as the results are interpreted. However, at the core of the experiment is a fair and straightforward comparison between how well the true boundaries and the random non-boundaries each match a set of changepoints.

4.3.1 Audio processing

Five audio features were used to encapsulate information from the following musical parameters: timbre, harmony, key, rhythm, and tempo. The object was to select features that would differ when these musical parameters differed, and be stable when the parameters did not differ. None of the audio features chosen are totally independent of each other, but each was designed to efficiently encapsulate information about a particular parameter while minimizing input from other information. Designing effective audio features is an area of ongoing research, and we do not attempt to contribute to this endeavour here. Instead, we have selected well-known features with readily available implementations, and in most cases used reasonable default values for the feature parameters.

None of these features is alleged to represent how the listener processes these musical attributes; the listener certainly perceives the music more holistically, basing their analysis on the properties not of frequency bands but of notes and other discrete events. The novelty-seeking approach tested here could be applied to more abstract representations using automatic beat tracking, transcription and source separation. However, these remain areas of active research: we lack robust tools with known error rates for these tasks that have been tested on a corpus as varied as the SALAMI data used here. Rather than employ intermediate and imperfect transcription efforts, we choose to estimate features directly from the audio. Assuming that changes in musical parameters are in fact reflected by changes in our audio features, our study will test how these musical changes relate to the perception of boundaries.

For timbre we chose Mel-frequency cepstral coefficients (MFCCs), widely regarded as an acceptable representation of the timbre of a short audio snippet [APS05]. The values in an MFCC vector indicate the strength of different periodicities in the Mel-scaled spectrum and hence characterize the shape of the spectrum with a minimum of harmonic information. MFCCs were calculated using windows of 0.19 seconds and a

hop size of half that. The lowest coefficient was discarded, since it relates specifically to overall loudness, and the next 12 coefficients were used.

For harmony we used the chromagram, which gives the strength in the signal of each pitch class from A to $G\sharp$. The method used takes the constant-Q transform of the signal, which scales the spectrum so that each bin corresponds to a single pitch, and then sums the contributions of each pitch class. Our window size was 0.1 seconds with a hop size of half that. Both MFCCs and chromagrams were calculated using Queen Mary's Vamp Plugin set [LGC+11].

The center of effect (CE), which refers to a music segment's estimated tonal center within Chew's Spiral Array model of tonality [Che00], was used to provide information on the key. While the center of effect generator (CEG) algorithm finds the key itself, we use only the CE as a proxy for the key, which facilitates rate of change computations. The CE was calculated using the audio key-finding system from [CC07], which uses a fuzzy analysis scheme to extract the pitches sounded from the spectrum, maps the pitches to their letter names, then calculates the CE, i.e. the geometric mean of their representations in the Spiral Array. Window size was 0.37 seconds with a hop size of one quarter of that.

The remaining two features are derived from a sonogram, which applies a model of the ear to estimate the perceived loudness in each of the twenty Bark-scale frequency bands. A fluctuation pattern (FP), also called a rhythmogram, measures the strength of loudness fluctuations between 0 and 10 Hz in each frequency band [PRM02]. A 1200-element vector, giving the strength of 60 modulation frequencies in each of the 20 Bark-scale frequency bands used, describes each window of the FP. The periodicity histogram gives the estimated strength of periodicities over the tempo range of 40 to 240 beats per minute (0.6 to 4 Hz) in a version of the signal that has been filtered to emphasize sharp attacks [PDW03]. The strength of a period is the number of times its amplitude (estimated using a comb filter approach) exceeds a given threshold over a short series of windows. FPs and periodicity histograms were calculated with the MA

Toolbox [Pam04] using a window size of 3 seconds and a hop size of 0.37 seconds.

4.3.2 Generating novelty functions and picking peaks

From each feature, we calculate a novelty function. Novelty functions were first proposed for segmenting audio by Foote [BM08], who estimated the amount of novelty at a point as the sum of the self-similarity of the passages that preceded and followed that point, and the dissimilarity between the two. Our novelty function ignores the internal similarity of the windows and focuses on the dissimilarity distance: we calculated at each point the Euclidean distance between the average feature vector before and after that point. It is essentially the same as the function used by the Argus algorithm for segmentation by tonal center [Che05], and can be seen as a continuous-time version of the difference features successfully employed by [TLPG07].

Varying the window size over which to take this average allows one to look at how the musical parameters evolve at different timescales; we used values starting at 0 (i.e., the first derivative of the feature vectors) and up to 30 seconds at 5 second intervals, meaning 7 different timescales altogether. Given that listeners have indicated that they usually perceive boundaries in response to a changing musical feature, difference features are a natural physical measure to use. In both [Che05] and [TLPG07], using difference functions at multiple timescales has been shown to be effective at predicting boundaries.

Peaks in the novelty function are hypothesized to indicate likely positions for boundaries. Of course, if the novelty function is sufficiently noisy, then there will be peaks throughout, and all boundaries and non-boundaries will be found to lie near peaks. We thus want to select only the tallest peaks. Our chosen peak-picking method first applies a smoothing filter to the novelty function that averages each value with the 10 previous and subsequent values (hence, a window of 20 times the hop size of that feature, or less than 2 seconds for MFCCs, chroma and CE, and roughly 7 seconds for the rhythm and tempo features). Then we pick the top 10 peaks with the following heuristic: once a peak

was added, any other peaks within 6.5 seconds were made ineligible. These choices were made to match the properties of our collection of annotations: the median number of segments per recording was 10 (the number of peaks we selected), and the smallest average segment length for a recording was greater than 6.5 seconds (the buffer we imposed around selected peaks).

4.3.3 Random baseline

To properly assess the audio properties of the boundaries, it is necessary to compare them to a set of non-boundaries. We selected random non-boundaries with the following constraints: first, for each recording, there should be an equal number of non-boundaries and boundaries. This ensures that the mean segment lengths are identical. Second, the boundaries should lie a minimum distance from all true boundaries. We set a buffer of 1.5 seconds, ensuring that even in the annotation with the shortest mean segment length, non-boundaries could be drawn from at least half of the recording. (Note that the mean segment length across the entire corpus was over 25 seconds, so this problem was rare.) With these two constraints, non-boundaries were drawn with uniform probability over the eligible portions of the recording.

4.3.4 Analysis metric

The chosen peaks in the many novelty functions now constitute our "ground truth," and we have two sets of points to compare it to: one the annotated boundaries, the other a random set of non-boundaries. We can now calculate how well each set of points predicts the peaks. Although two annotations were available for some recordings, we evaluated each separately.

The evaluation metrics we use are precision, recall, and f-measure. If we designate the set of annotated boundaries as A and the set of novelty function peaks as P, then the set of boundaries in A that 'hit' or are nearer to some peak in P than some given

threshold (we use values of 3.0 and 0.5 seconds) is expressed as $A \cap P$. We can then express precision as the fraction of attempts that are successful $(|A \cap P|)/|A|$ and recall as the fraction of peaks that are found $(|A \cap P|)/|P|$. We are most interested in the f-measure, their harmonic mean. Note that we did not include in our evaluation any trivial boundaries, such as those that indicate the start or end of the recording, or any boundaries occurring in the first or final 1.5 seconds of the piece.

The values for the thresholds, 3.0 and 0.5 seconds, are those standardly used in the literature, and the f-measure is itself the most widely-used metric for boundary evaluation. Nevertheless, one shortcoming to using fixed thresholds is that, since song lengths vary substantially, one value may actually be applying proportionally different standards to different songs. Also, the f-measure has very recently attracted criticism for weighting recall and precision equally, when there is evidence that precision is perceptually more important to listeners [NFJB14]. Despite this, we carry on using it here not only because it is standard but because it is intuitive and well-understood.

4.4 Results and statistical analysis

4.4.1 Are boundaries points of novelty?

We first ask: are boundaries points of novelty? For each of the 761 recordings, we calculated 35 novelty functions, one for each combination of five features and seven timescales, and extracted sets of peaks as described previously. For each set of 35, we calculated the average f-measure between these novelty functions and the boundaries and non-boundaries. We then compared these average scores for each of the 1,253 annotations, resulting in 1,253 paired trials. Of these mean f-measures, the median value for boundaries (0.328) was nearly twice that for non-boundaries (0.178) using a boundary match threshold of 3 seconds (see Figure 4.2). A paired Wilcoxon Signed Rank test confirmed that the difference in medians was significant (U = 771,373.5, p-value $< 10^{-15}$), with

a large effect size (r = 0.59). This indicates that boundaries are a better indicator of novelty peaks than non-boundaries. Indeed, the mean f-measure for boundaries was larger than that of non-boundaries in 93.9% of the annotations.

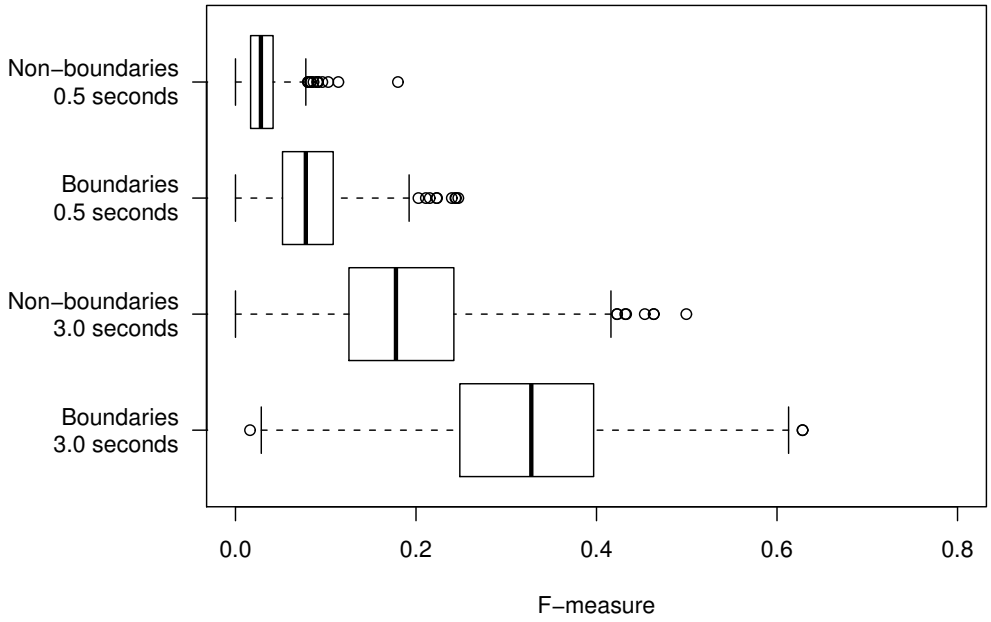


Figure 4.2: Distribution of f-measure scores for boundaries and for random sets of non-boundaries, given a grading threshold of 3.0 or 0.5 seconds. Outliers in a modified box plot are those that lie more than 1.5 times the interquartile range outside the second and third quartiles.

When the boundary match threshold is reduced to 0.5 seconds, the likelihood of being near a boundary shrinks for both sets of points, but the contrast between them grows: the median f-measure for boundaries (0.078) was more than twice that for non-boundaries (0.028). A Wilcoxon test again confirmed that the distributions have a different median (U = 744,216.5, $p < 10^{-15}$). The mean f-measure was greater for boundaries for 90.3% of the annotations. Despite the poorer overall performance, the effect size (r = 0.58) still indicates a large practical significance.

Since the boundaries surpassed the non-boundaries at predicting points of novelty, we can conclude that boundaries indeed tend to be more novel than other points in a

piece. But what do the numbers mean qualitatively? The maximum f-measure possible is 1, indicating perfect recall and precision, but in practice, even two similar listeners are unlikely to replicate each other's analyses with such accuracy. Since we would not expect any algorithm to predict boundaries as well as another listener, we can use inter-annotator agreement as a performance ceiling. Using the subset of 492 pieces in our corpus that were annotated twice, and a threshold of 3.0 seconds, the median f-measure of inter-annotator agreement was 0.769. This is more than twice the median agreement between the novelty functions and the boundaries, which was 0.326 for this subset. This large difference was of course significant according to a Wilcoxon test ($U = 18.9$, $p < 10^{-15}$), and the effect size ($r = 0.60$) reflects that the factor by which points of novelty predict boundaries better than non-boundaries is almost the same as the factor by which boundaries are better predicted by another listener's annotated boundaries than by points of novelty.

On a scale from 0 to 1, we have found the two f-measures we wish to compare (0.178 for non-boundaries-to-novelty and 0.326 for boundaries-to-novelty), as well as a performance ceiling (0.769 for boundaries-to-boundaries) that is less than 1. Is there a comparable performance floor, and is it greater than 0? Although the f-measure between one listener's set of annotated boundaries and the associated set of non-boundaries is 0 by design, it might be greater than 0 if we compare one set of boundaries to the non-boundaries estimated from the other listener's annotation. The median of this measure was 0.118, which differed from the above medians with approximately the same significance and effect size. Hence, using boundaries instead of non-boundaries to predict points of novelty led f-measure to increase from 0.178 to 0.326; a listener attempting to identify instead of to avoid the boundaries indicated by another listener led f-measure to increase from 0.118 to 0.769. This larger increase suggests that although the boundaries relate more to novelty than do the non-boundaries, qualitatively, this is much less significant than the perceptual difference between boundaries and non-boundaries.

If we were to compare our novelty functions to state-of-the-art structural analysis sys-

tems, we would likely find that they surpass our performance. At the 2012 MIREX evaluation, using a corpus of annotations comparable to ours, the mean f-measure achieved by nine algorithms varied between 0.42 and 0.49 using a 3.0-second threshold, and between 0.16 and 0.29 with a 0.5-second threshold. While all of these means far exceed the mean f-measures achieved in this study, this comparison is not fair: the algorithms submitted to MIREX use far more information than novelty (e.g., sequential repetitions, multimodal feature distributions), to estimate structure, and so it is expected that they would fare better. The purpose of this experiment is to investigate how well measures of novelty explain the information contained in the annotations; hence the relevant comparison is between the annotated boundaries and the random sets of non-boundaries.

However the results are parsed, we have observed that boundaries annotated by listeners are more likely than chance to be associated with a peak in novelty, suggesting that annotators do attend to novelty in the signal—and that the annotations, in turn, contain information about acoustic novelty. Does the size of this effect vary according to the listener, to the genre, or to the type of novelty function calculated? In the next four subsections, we address these questions by examining the effect of these factors on f-measure contrast, which we define as the amount by which the boundary f-measure exceeds the non-boundary f-measure for each novelty function, using a threshold of 3 seconds.

Differences among listeners

Among 1,253 annotations, a Kruskal-Wallis test indicated a significant effect of annotator ($\chi^2 = 15.577$, df = 8, p = 0.049) on the f-measure contrast, suggesting that the annotator's responses correlated with boundaries to varying degrees. However, a multiple comparison test (using a Bonferroni correction) found no pairs of annotators for which f-measure contrast differed significantly. The distributions shown in Figure 4.3 show that differences between the annotators are minimal, suggesting that altogether the annotators were similar in the way their annotations reflected musical changes.

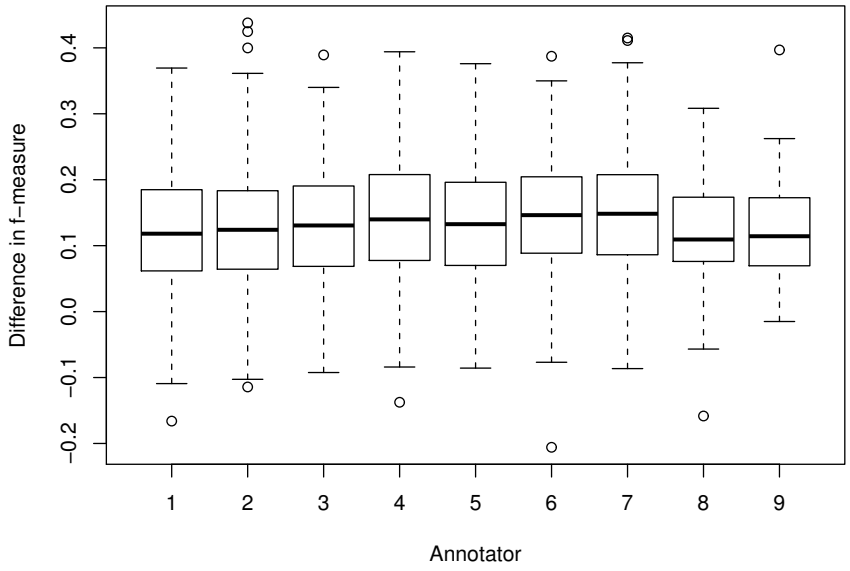


Figure 4.3: Distribution of f-measure contrast (the absolute improvement in f-measure achieved by sets of boundaries over non-boundaries) among different annotators. All results found using a grading threshold of 3.0 seconds.

Differences among genres

The effect of genre (see Figure 4.4) was also significant according to a Kruskal-Wallis test ($\chi^2 = 63.631$, df $= 4$, p $< 10^{-12}$). A multiple comparison test found a difference in the f-measure contrast between five of the ten pairs of genres: four of these indicated that f-measure contrast was smaller in classical than in other genres, with a small to moderate effect size ($0.19 \leq$ r ≤ 0.33); the fifth indicated a small difference between popular and jazz (r $= 0.17$). This could indicate that when annotating classical music, listeners paid more attention to criteria other than novelty, such as parallelism; or, that the transitions between sections in a classical piece tend to be less sudden—that is, there are more elided boundaries than in other kinds of music.

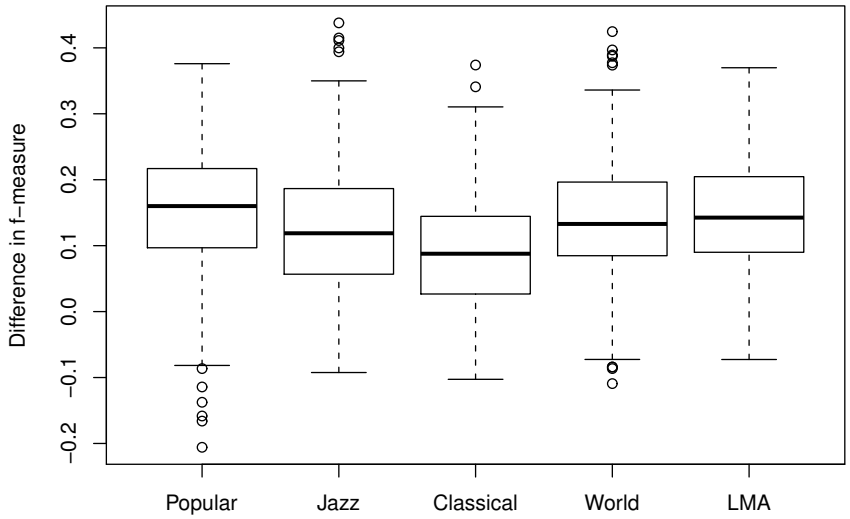


Figure 4.4: Distribution of f-measure contrast among different genres.

Differences among timescales

To evaluate the effect of window size, we averaged the f-measure contrast across features for each of the seven window sizes and for each annotation. A Friedman test found a significant effect of window size ($\chi^2 = 844.94$, df $= 6$, $p < 10^{-15}$), and many pairs of timescales differed. All comparisons between the 0-second window size and another showed a small to moderate effect size ($0.20 \leq r \leq 0.32$). As seen in Figure 4.5, the immediate derivative (timescale 0) did not improve very much on the baseline at all, suggesting that novelty at this timescale was of little relevance to the annotators. Additional comparisons yielded a small difference between the 30-second window size and window sizes between 5 and 20 seconds ($0.11 \leq r \leq 0.19$), and between the 25-second window size and window sizes between 5 and 15 seconds ($0.10 \leq r \leq 0.16$). This suggests that these longer timescales are also less relevant in terms of acoustic novelty. The 10-second timescale improved the f-measure the most, suggesting that it was the most perceptually relevant timescale for establishing section boundaries. It is interesting that although the mean segment length across all pieces was roughly 25 seconds, the

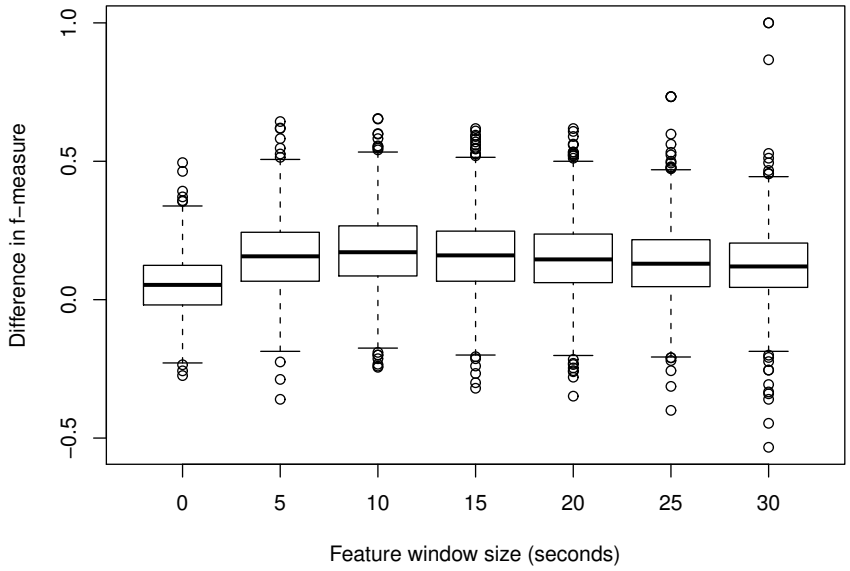


Figure 4.5: Distribution of f-measure contrast among different timescales.

25-second window offered less contrast to the baseline than the 10-second window. This could simply be explained by the fact that the boundaries of short sections risk being obscured by a large window, but a section larger than a shorter window size is less likely to be obscured.

Differences among features

A Friedman test found differences in f-measure contrast among features averaged across timescales to be significant ($\chi^2 = 529.71$, df = 4, $p < 10^{-15}$). A multiple comparison test followed by calculation of effect size yielded small differences between timbre and key ($r = 0.27$), timbre and tempo ($r = 0.21$), as well as rhythm and key ($r = 0.25$) and rhythm and tempo ($r = 0.21$), suggesting that timbre and rhythm were both more reliable indicators of boundaries than tempo or key (see Figure 4.6). The effectiveness of harmony lay somewhere in between: it was found to differ from timbre, tempo and rhythm with a small effect size ($0.10 \leq r \leq 0.12$) and to differ from key with a slightly larger effect size ($r = 0.19$). That tempo should be a less reliable predictor of boundaries

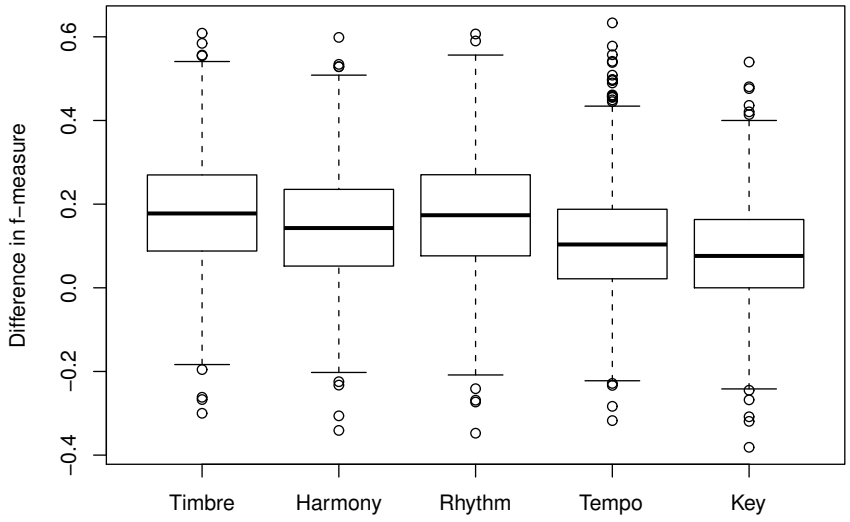


Figure 4.6: Distribution of f-measure contrast among different features.

is a reasonable result, since in most popular and jazz music, which comprise at least half the data studied, tempo does not commonly vary across sections. However, it is a surprise for key. The features for key (center of effect) and harmony (chroma) provide similar information, but while chroma merely provide the raw pitch content, center of effect condenses this information into a single estimate of the tonal center. Our results suggest that for the purpose of locating boundaries, this process filters out more signal than noise.

4.4.2 Do any boundaries not match a novelty peak at all?

The mean f-measure indicates how well the annotated boundaries predict the set of peaks given by a particular novelty function. But we would not expect every boundary to be suggested by every musical feature at every timescale. A further question to ask is if there are any boundaries that do not match any peak at all; this would indicate the minimum extent to which boundaries are not associated with changes in musical parameters.

To answer this question, we produce a histogram showing the number of novelty function peaks associated with each boundary, using a threshold of 3 seconds (Figure 4.7). The comparable histogram for non-boundaries is given below the x-axis. It shows that 7.1% of annotated boundaries do not match a peak in any novelty function, meaning 92.9% match at least one—and most match many more. The median number of novelty functions matched is eleven; since there are five features and seven timescales, the median indicates that half of the boundaries matched at least two distinct features at three distinct timescales, showing boundary perception to be a function of multiple features at multiple timescales. The non-boundary histogram is more heavily skewed to low values than the boundary histogram, and they are about equal when the number of novelty peaks matched is nine. Hence, if exactly nine novelty peaks match a particular point, then that point is about equally likely to be perceived as a boundary as not; the odds of the point being a boundary steadily increase as more novelty peaks match that point.

The light gray regions in Figure 4.7 indicate the subset of boundaries that are "symmetric," i.e., those where the labels of the sections before and after the boundary are the same (prime symbols attached to segment labels were disregarded here, so the labels A and A' were treated as equal). Symmetric boundaries are hypothesized to indicate less novelty than non-symmetric boundaries, and this is borne out modestly by the data. Of the boundaries that match no novelty function, 34.3% are symmetric, whereas only 26.7% of all boundaries were symmetric. The median number of matching novelty functions for non-symmetric boundaries is eleven; for symmetric boundaries, it is nine. A Wilcox rank-sum test showed that this was a significant effect (U = 11,406,306, p $< 10^{-15}$), with a small effect size (r = 0.10). The effect here is slight, but the measure of "symmetry" used is very rough, and does not take into account the annotated changes in lead instrumentation. In many of the jazz pieces, for instance, nearly every section is given the same label, and the most salient structural information lies with the changing soloists. Still, this result provides some support for the hypothesis that the perception of symmetric boundaries owes less to novelty and perhaps owes more to factors such as

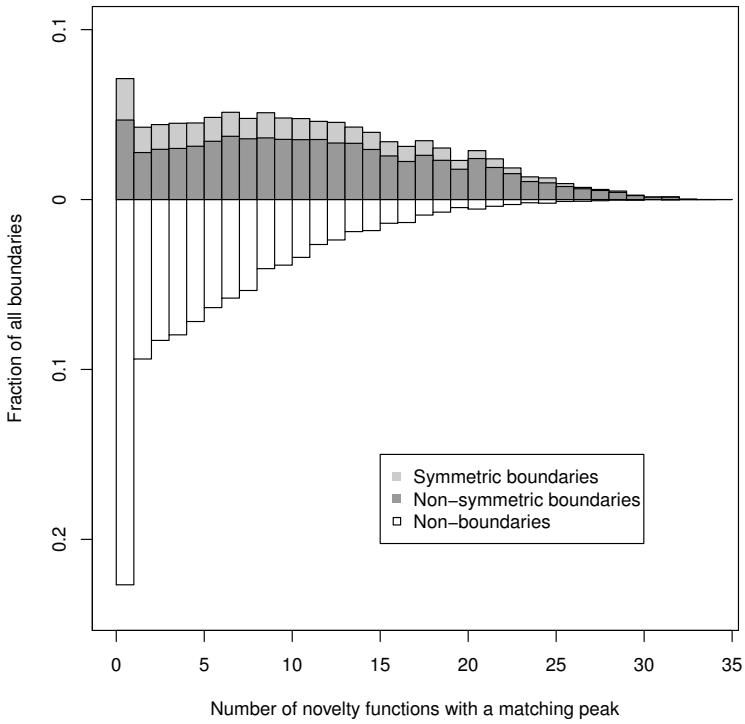


Figure 4.7: Comparison of histograms for boundaries (gray) and non-boundaries (white) according to number of novelty functions with a matching peak. Symmetric boundaries, i.e., those between sections with the same letter label, are distinguished from non-symmetric boundaries.

parallelism.

4.4.3 Can boundary salience be estimated by annotation concurrence?

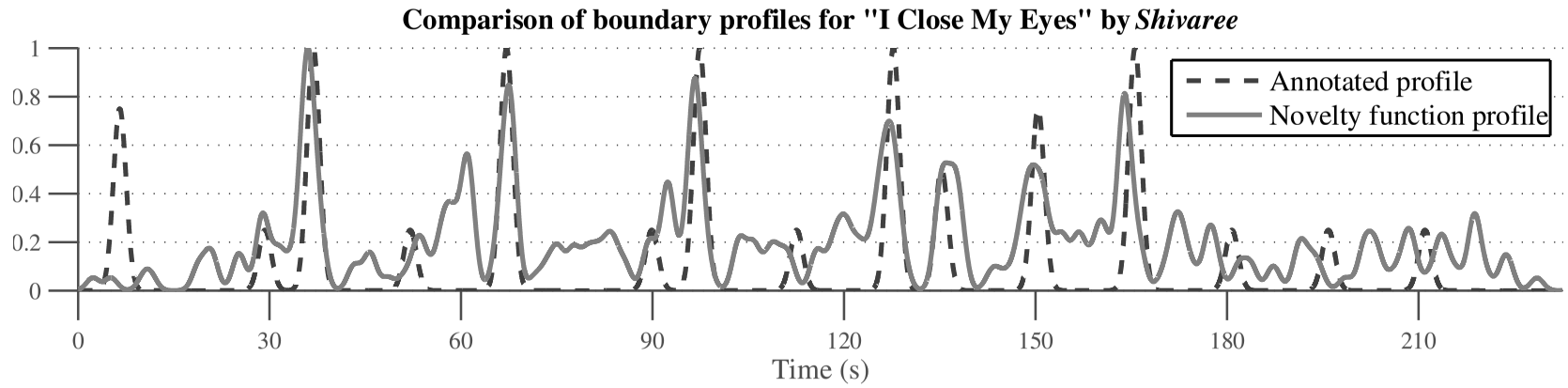
We have observed that boundaries vary in the number of novelty functions they match: nearly all boundaries match a few novel points, and a minority match several. This is curiously analogous to the finding in [BMK09] that, in each piece studied, a few boundaries stood out as salient to all listeners, while the majority of boundaries were indicated by only a handful of listeners. They further found that the perceptual salience of a boundary correlated strongly with the number of people who indicated that boundary. Bruderer et al. [BMK09] assembled the boundary indications of many listeners to pro-

duce a continuous boundary profile, indicating at each moment the potential salience of a boundary in that position. We conjecture that we could obtain a similar result by collecting information from a set of automated listeners (i.e., novelty functions), each indicating boundaries according to the parameter (i.e., a given musical feature at a given timescale) to which they are attending.

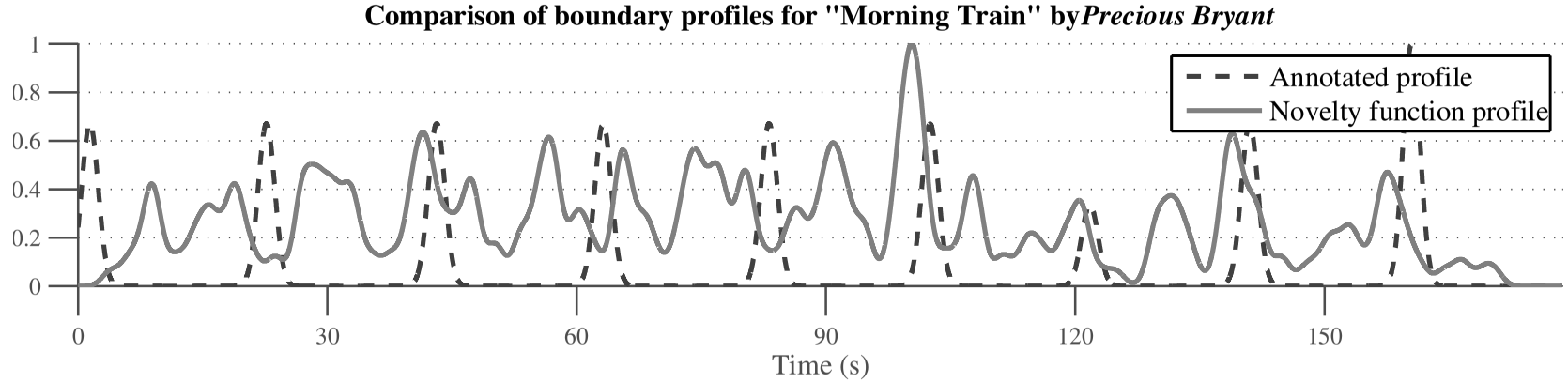
We do not have the boundary salience data to test this claim, but we may approximate salience by combining the annotations of two listeners and giving more weight to non-symmetric boundaries. We combined annotations with the technique proposed by [BMK09]: all the boundary indications were collected (non-symmetric ones counted twice), and the result was convolved with a Gaussian function (we used a full-width half-maximum of 1.5 seconds instead of 1.25 seconds given by [BMK09]).

Figure 4.8(a) shows the result of applying this procedure to the two annotations for the song "I Close My Eyes" by the band Shivaree. The dashed line gives the boundary function as estimated from the two annotations; the solid line gives the boundary function estimated from the 35 novelty functions. There is very close agreement with the largest peaks in the novelty functions, and less agreement among the less significant peaks. The Pearson correlation between the two time series is 0.60, a close overall fit. When we performed this procedure on all 492 pieces for which two annotations were available, we found the mean Pearson correlation to be 0.33 ($sd = 0.18$), suggesting a moderate relationship throughout the corpus. An example of a pair of boundaries that matched the novelty functions poorly is given in Figure 4.8(b). These are the annotation- and novelty-derived boundary functions for Precious Bryant's "Morning Train," and the Pearson correlation between them is -0.03. Even so, the fit is qualitatively good for the second half of the song.

This result shows that the simple measure of novelty defined in this chapter, versions of which are already used regularly in the MIR community, actually does seem to converge on the same information contained in the annotations. Moreover, this information, when collected from a variety of features at different timescales, can be combined into an overall



(a)



(b)

Figure 4.8: Comparison of boundary profiles estimated from annotations (solid line) and from novelty functions (dashed line) for (a) "I Close My Eyes" by Shivaree (SALAMI ID 4), and (b) "Morning Train" by Precious Bryant (SALAMI ID 36).

novelty function that seems to reflect the same patterns of salience that listeners display.

4.5 Conclusion and future work

We have investigated a large corpus of recordings and annotations to show that acoustic novelty, as estimated by features reflecting timbre, harmony, key, rhythm and tempo, relates strongly to the position of boundaries indicated by listeners. The strength of this relationship was shown to be moderately affected by the feature used to estimate the novelty—specifically, the novelty of tempo and key were found to be less informative than the others. However, the range of features used was small; if other features and other parameter settings for the features were used, the results could be different. The high performance of structural analysis algorithms at MIREX relative to ours suggests that there is plenty of room for more sophisticated features to reveal the greater relevance of novelty cues. By repeating these experiments with different features, future research could pin down the extent to which these findings are feature-dependent.

We also saw that the relevance of novelty for explaining boundaries depended on the timescale and genre. In contrast to the situation with musical features, the range of timescales used here seems wide enough to make a general claim that novelty at the 10-second timescale is most relevant to listeners. And given the breadth of the SALAMI corpus with respect to genre, the conclusion that the boundaries in classical music were less consistently novel than those in other genres seems robust as well. However, both conclusions again come with the caveat that the experiment used a limited set of features; other features may have interacted with timescale and genre differently.

Finally, we saw that a boundary profile derived from novelty functions correlated modestly with a boundary profile estimated from the annotations. Since [BMK09] found that the fraction of listeners who indicated a boundary correlated with the judged salience of that boundary, our findings extend this result to suggest that the salience of a boundary is correlated to its acoustic novelty.

All of these findings support the use of combinations of novelty functions as a major estimator of perceived boundaries. At the same time, our results show the limitations of predicting boundaries solely based on points of novelty: although nearly all boundaries corresponded to a peak in novelty, not all peaks in novelty indicated a boundary. (Indeed, according to Figure 4.7, peaks in more than nine novelty functions—or at least two features and two timescales—need to coincide before a point is more likely to be a boundary than not.) This indicates that as a predictor of boundaries, acoustic novelty has high recall but low precision. Thus, while novelty is important to listeners, it is not the final word: listeners reject many novel points as false positives. This is almost certainly due to the influence of top-down factors: metrical structure, parallelism, and other considerations lead listeners to perceive seemingly novel moments as moments of continuation. The success of state-of-the-art structural analysis algorithms which usually apply such constraints, whether by identifying long repetitions or by detecting a metrical grid, suggests this is indeed the case (see Section 4.4.1).

Bruderer and McKinney [BM08] demonstrated the perceptual validity of segmentation models that used score-based representations. The present study may help to ground comparable audio-based models that could be applied to any recorded music, whether or not a score exists—or whether the music even could be transcribed using Western music notation (as much electronic music cannot be).

An important caveat to our findings is that there is no proof of causality: boundaries do tend to occur at novel moments, but this novelty is not necessarily what motivates the listener to perceive a boundary. An alternative explanation would be that listeners identify repeated sequences and infer boundaries between them, and that the novelty of the boundaries arises from the fact that these sequences tend to differ acoustically. This alternative is perhaps supported by the observation that symmetric boundaries (those between repetitions) are less well explained by novelty than the other boundaries. While this experiment cannot settle the question of causation, the studies conducted in [CK90] and [BMK09] confirm that listeners often find the changes that occur at boundaries

their most salient aspect. Still, as illustrated in Figure 4.7, many boundaries remain unexplained by any kind of acoustic novelty. Further studies should test how well these boundaries are explained by parallelism, pauses, and changes in other musical parameters not tested here.

Further study could also focus on interactions between the factors considered here for a more complete picture of the importance of novelty. Although we did not investigate interaction effects systematically here, the results did show that: the usefulness of the tempo feature was higher at longer timescales; the timbre feature was less useful on the LMA database, perhaps because many of these recordings were noisier; and the best timescale on the classical music was 25 seconds (even though this was among the worst timescales for the other genres), perhaps indicating that boundaries in classical music tend to reflect long-term changes, or that the most significant short-term changes are often misleading with respect to finding boundaries in classical music.

Returning to the broader themes of this thesis, we recall that in Chapter 2, we argued that bottom-up approaches to structural analysis focusing on discontinuities fail to capture the variability in analyses found among listeners. The results of this chapter show why this may be the case: there is an important gap between acoustic novelty (the justification cited most often by listeners for perceiving points as boundaries) and the boundaries that were in fact indicated in the SALAMI annotations. In a piece of music, there may be many points that are acoustically novel but not perceived as boundaries, and many points perceived as boundaries that are not acoustically novel. If listeners perceive some boundaries where no changes occur, and perceive some big changes as points of continuation, this may be a result of top-down factors, and as noted in Chapter 2, top-down factors are more likely to be influenced by listener differences.

In Section 4.2.3, we pointed out that this experiment tests novelty because it is the standard justification for boundaries, and not the justification actually given by the SALAMI annotators—such justifications are not available for SALAMI nor for any other large collection of analyses. Thus, any experiments using annotations must make

assumptions, based on other studies, about what the annotators were likely to be thinking. Is it possible to estimate more rigourously what an analyst was thinking *post-hoc*? This is the project undertaken in the next chapter.

Chapter 5

Relating grouping structure to musical features

In Chapter 3 we noted, as in previous literature, that listeners attributed the perception of a boundary to stark changes in the music, and in Chapter 4 we tested the extent to which this explained the boundaries in a large corpus of annotations. In Chapter 3 we also noted that paying attention to a feature of the music was a common way to justify a given grouping decision. We would now like to test in a large corpus whether attention is a good explanation for grouping decisions. Unfortunately, we cannot conduct a study comparable to that in Chapter 4 for attention, since there is no way to compute or estimate the focus of the listener's attention from the audio the same way we could estimate novelty with common audio processing techniques.

However, MIR techniques for visualizing structure may offer a way to gain some insight into the minds of annotators. In this chapter we will attempt to solve the inverse of the problem usually posed in MIR: rather than process the audio to attempt to replicate a listener's analysis, we will use the analysis to discover those features in the audio that best support the listener's analysis. The goal is to have a tool to investigate the potential reasoning behind the grouping decisions of listeners.

We introduce a method that uses self-similarity matrices (SSMs) generated by musically-motivated audio features at various timescales. Since a listener's attention can shift among musical features throughout a piece, we further break down the SSMs into section-wise components and use quadratic programming (QP) to minimize the distance between a linear sum of these components and the annotated description. We posit that the optimal section-wise weights on the feature components may indicate the features to which a listener attended when annotating a piece, and thus may help us to understand why two listeners disagreed about a piece's structure. We discuss some examples that substantiate the claim that feature relevance varies throughout a piece, and use our method to study the differences between listeners' interpretations.

In the introduction that follows, we summarize the development and use of SSMs in MIR, and recap how the reasoning behind listeners' grouping decisions has been studied in the past.

5.1 Introduction

5.1.1 Previous methods in SSM calculation

One of the most important aspects of music is that it is repetitive: individual sounds, notes and chords, dynamic gestures, rhythmic patterns, instrumentations, and so forth are all elements liable to repeat, whether identically or with some variation. Discovering repetitions is important in MIR, since many MIR tasks can be performed or improved upon using information about music structure: for example, cover song detection [GSMA12] and chord transcription [MND09].

Recurrence plots, proposed by [EKR87] for analyzing the motion of dynamical systems, can reveal repetitions in sequential data. A self-similarity matrix (SSM), originally proposed by [Foo99], is a variation on such plots that moves beyond the usual feature-based methods of visualizing music, such as pitch rolls and other time-frequency

representations. Unlike these, SSMs do not represent the musical content itself, but only the pattern of repetitions and recurrences that it contains. The features themselves are of course still important in any given application: an SSM based on pitch content may be effective for tracking melodic repetitions but not repetitions of percussive sounds.

Whereas earlier experiments with extracting structure based on SSMs focused on one feature at a time (e.g., [Foo99] made SSMs derived only from Mel-frequency cepstral coefficients (MFCCs), and [BW01] used only chroma for chorus detection), it was soon realized that using multiple features could improve results. Eronen [Ero07] calculated SSMs from MFCCs and chroma and summed the result, while [Mar06] obtained three SSMs from chroma vectors, each calculated to reflect repetitions at different timescales, and took the element-wise product of the trio to reduce noise. Paulus and Klapuri's [PK06] optimization-based approach used information from separate SSMs reflecting timbral, harmonic and rhythmic similarity. In order to find transposed repetitions, [Got03] searched for maxima across multiple chroma-based SSMs. Rather than generate separate SSMs for separate features, [HKS12] concatenated the feature vectors for each frame and calculated a single SSM from the result.

5.1.2 Limitations of combining SSMs

However, simple combinations of SSMs rarely result in the exact structure that the experimenter hopes to extract. The features used to compute the SSMs may be too simple, failing to isolate the relevant musical parameters, and the patterns contained in the SSM may be hidden by noise, due to perceived repetitions actually being inexact or greatly varied. For example, the chroma feature may miss an important melodic repetition that has been obscured by a change in harmony. Successful methods of automatic music analysis based on SSMs invariably employ complex post-processing steps to mitigate these shortcomings: examples include low- and high-pass filtering [Got03], erosion/dilation operations [Ong07], dynamic programming [SJK06], non-negative matrix factorization (NMF) [KS10], re-emphasis of transitive relationships [Pee07], and so on.

We may conclude that a simple sum of SSMs does not reflect the similarity judgements that a listener may make across an entire piece.

What information could the sum of SSMs be missing? One straightforward suggestion is a weight for each feature or timescale. Perceptual evidence supports this strategy: for example, in investigating the relative importance of different laws in a Generative Theory of Tonal Music, [FC04] found that laws relating to some features are more important than others. The idea to tune feature weights before analyzing structure has appeared before in the MIR literature: to improve a structural segmentation algorithm, [PE04] chose feature weights to maximize the separability of vectors according to the Fisher criterion. In [KP12], the size of the window for calculating features was adapted to the estimated rate of change, improving the clarity of block patterns in SSMs. A hierarchical SSM proposed by [Jeh05] used different features and techniques at each timescale in a musicologically-informed manner.

Another aspect of listening that may be missed when SSMs are combined is that the focus of a listener's attention may shift at various points throughout a piece. For example, the self-similarity of a chorus of a given song may be very well accounted for by an SSM based on harmony, whereas the self-similarity of the guitar solo that follows may not be. Again, listener studies such as [BMK09] and [CK90], as well as our own findings in Chapter 3, demonstrate that listeners justify their section boundaries with various musical features throughout a piece. The timescale that is most pertinent to the listener may also vary: [JB89] hypothesized that listeners either focus their attention on short or long timescales, and that their focus may shift while listening, either willfully or as a result of changing attunement to the music.

In the study of vision, eye-tracking technology has enabled the direct study of visual attention for over 50 years, Unfortunately, no comparable research technique exists for studying musical attention. While a listener is certainly able to fixate on single aspects of music, this fixation is not expressed physiologically by the ear. Instead, researchers interested in what people are paying attention to in music must either perform the

laborious task of asking the listeners themselves (as in [CK90] and [Del87]), or manipulate the stimuli or experimental conditions. For example, in their study of change deafness, [AK08] manipulated the strength of listeners' attention by varying the metrical strength or the tonality of the tone that was changed in pairs of melodies. They found that changes in metrically weak tones, which are less likely to be focused on by listeners, were more likely to go unnoticed. By playing listeners the exact same pieces repeatedly, [Mar12] could deduce from their responses that with greater exposure, their attention had focused on longer timescales.

In this chapter, we present a method of combining SSMs that aims to model which features and which timescales a listener could have been paying attention to. The method does not manipulate the listening conditions or attention of the listener; rather, it is a *post hoc* approach that accounts for a given structural annotation produced by a listener. The method exploits the fact that listeners may focus on some features more strongly than others, and that their focus may change throughout a piece.

5.2 Proposed method

We first review how to calculate an SSM from acoustic data or from an annotation. In Section 5.2.2 we motivate our approach and define the algorithm using a simple example. In Section 5.2.3, we demonstrate its use on an audio recording of the song "Yellow Submarine" by The Beatles.

5.2.1 Self-similarity matrix calculation

A self-similarity matrix can be thought of as a real-valued recurrence plot where element e_{ij} indicates the similarity between frame i and frame j of a sequence of frames. It is typical to use Euclidean or cosine distance [PMK10] as a distance metric; here, we use cosine distance for its natural scaling between -1 and 1. Repeated sequences in

recurrence plots are revealed as diagonal lines. In SSMs based on music, it is also common to see off-diagonal blocks, revealing the repetition of sections that are homogenous with respect to a given feature.

5.2.1.1 SSMs from annotations

Binary SSMs are commonly generated from structural annotations as diagrams (e.g., [PMK10]) or to illustrate examples of song structure. Using cosine similarity, we set $e_{ij} = 1$ if frames i and j belong to sections with the same label, and -1 otherwise (see example in Figure 5.1, left). This section describes some variations on the usual approach that is relevant for our data.

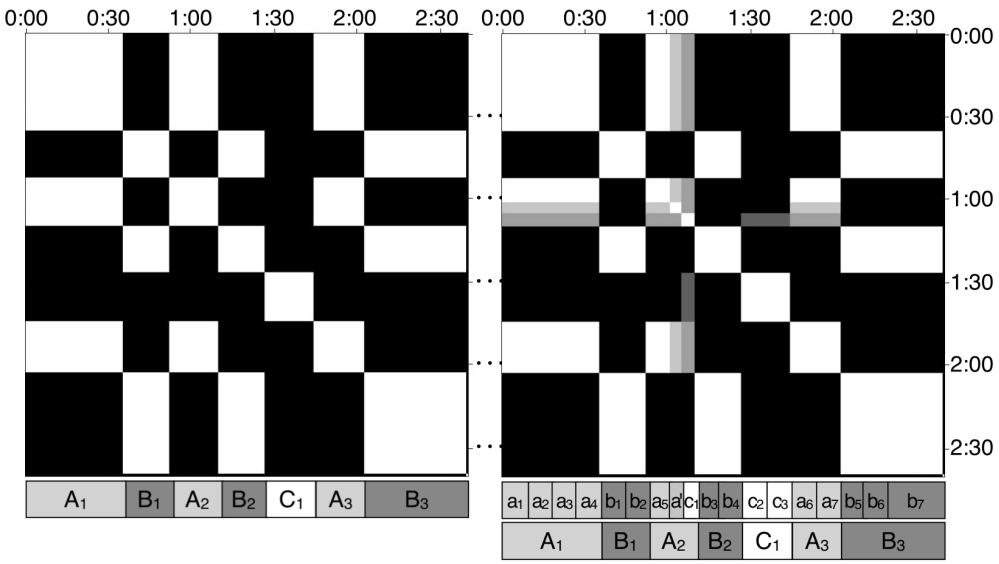


Figure 5.1: Left: SSM derived from annotation of "Yellow Submarine." Time progresses from left to right and from top to bottom. The large-scale annotation below it is from the SALAMI database (salami_id: 1634). Right: An alternative SSM derived using an additional layer of the annotation, where α, the relative weight of the large-scale labels, is set at 0.625 and β, the fractional similarity implied by primes, is 0.35.

The annotation in Figure 5.1, like all the examples in this chapter, are drawn from the Structural Analysis of Large Amounts of Music Information (SALAMI) dataset [SBF+11]. In the SALAMI annotation format, information about repeating sections

is given at large and small timescales, and sections may be distinguished with prime symbols (e.g., A vs. A'), which fuzzily indicates similarity with variation.

We include some of the richness of this description in the SSM by generating a separate SSM for each timescale and summing the results (see Figure 5.1, right). To emphasize one timescale over another, we can choose a weighting parameter $0 < \alpha < 1$, and multiply the large- and small-scale SSMs by α and $1 - \alpha$, respectively, before summing. A similar approach can act on prime symbols: when two frames have the same label but differ by a prime, instead of setting $e_{ij} = 1$, we can set it to some other value $-1 < \beta < 1$. Setting $\beta = 1$ would imply that A and A' are identical; $\beta = -1$ would imply they are completely distinct; and $\beta = 0$ would ignore the symbol.

5.2.1.2 SSMs from audio

The five different SSMs in Figure 5.2 were all calculated from a recording of "Yellow Submarine". Each one uses a different audio feature to represent a different musical parameter: MFCCs for timbre, chroma for pitch, fluctuation patterns (FPs) for rhythm, periodicity histograms for tempo, and RMS for loudness. These are the same audio features that were used in Chapter 4, with the addition of RMS and the subtraction of center of effect. A full description of the features is found in Section 4.3.1, but are summarized here.

MFCCs derive from the shape of a rescaled spectrum and can characterize a sound's timbre; chroma vectors estimate the power of each pitch class and characterize the harmonic content of the audio; fluctuation patterns estimate the strength of low-frequency periodicities within Bark-scale frequency bands over windows that are several seconds long and hence characterize the rhythmic content [PRM02]; periodicity histograms reflect the relative strength of different tempi by looking at sharp attacks in the audio and measuring the strength of periodicities in the tempo range of 40 to 240 beats per minute (0.6 to 4 Hz) [PDW03]; and finally, the root mean square (RMS) of the waveform and the

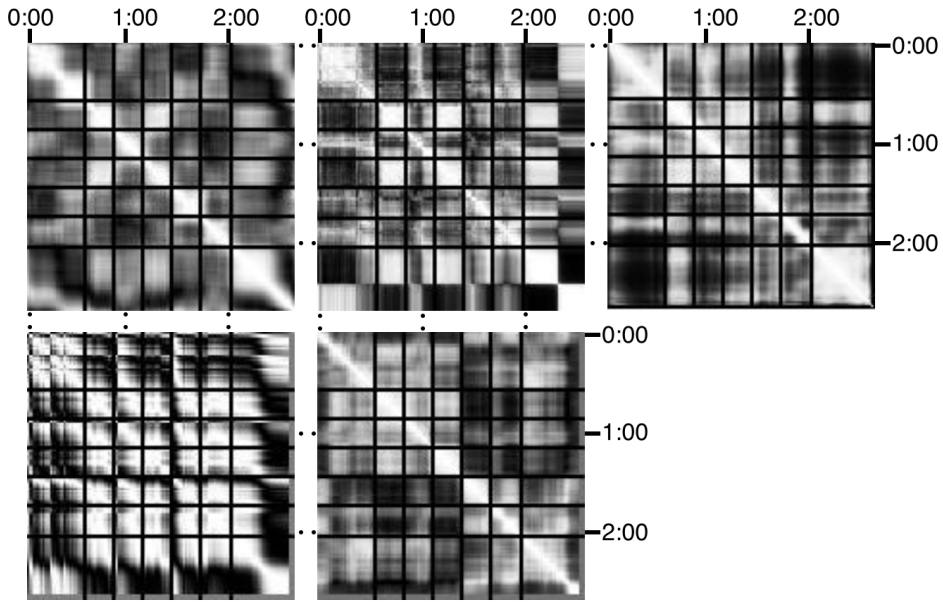


Figure 5.2: Five SSMs calculated from a recording of "Yellow Submarine." Highly similar frames produce white pixels, dissimilar frames produce black, and independent frames gray. From left to right, the SSMs represent (top row) MFCCs, chroma, FPs, and (bottom row) RMS, and periodicity histograms. The black lines indicate the boundaries of the structural annotation seen in Figure 5.1.

derivative of RMS estimate loudness and dynamic variations.

MFCCs and chroma were calculated using 0.19- and 0.10-second windows, respectively, with 50% overlap, using Queen Mary's Vamp Plugin set [LGC+11]. The twelve lowest MFCCs were kept, aside from the first, which correlates with loudness. FPs and periodicity histograms were calculated using 3-second windows and 0.37-second hops with the MA Toolbox [Pam04]. FP vectors have 1200 elements, measuring 60 modulation frequencies in 20 Bark-scale frequency bands, while periodicity histograms have 2000 elements, indicating whenever any of 40 tempo ranges is activated beyond 50 fixed thresholds. While it is common to use dimensional reduction techniques to reduce the large size of the feature vectors, the relative differences between the raw vectors are still well captured in the SSMs in Figure 5.2. Lastly, RMS was calculated using 0.1-second windows and 50% overlap.

Each of the above features gives, for every frame, a vector of some length. We transformed the values in two ways: first, each vector dimension was standardized over the length of the piece to have zero mean and unit variance. Since no frame-wise normalization was used for any feature, this ensures the variance in each dimension is weighed equally, ensuring that repetitions in low-magnitude signals are detected, albeit at the cost of some additional noise. The features were then smoothed in time; for the SSMs in Figure 5.2, a 10-second moving-average filter was used. Finally, the SSMs were calculated using cosine similarity.

These features and processing steps are not integral to the algorithm. The algorithm merely requires that some set of feature representations be chosen, but the choice is arbitrary. In this chapter, the choice was made based on convenience (software for computing the features was readily available) and a desire for a set of varied, easily-interpreted features.

5.2.2 Combining SSMs

Suppose we have an annotation for a song's structure, expressed as an SSM like in Figure 5.1, and want to find how best to explain it in terms of SSMs generated from the song's acoustic features, like those in Figure 5.2. We know that summing the feature matrices is a useful technique, and since we have the annotation we could try to calculate the optimal linear combination of feature matrices to reconstruct the annotation. This would provide a relative weight to each feature corresponding to its salience with respect to the entire song. However, knowing that the salience of different features can vary throughout the piece, we may wish to explain the annotation section by section.

Previous approaches to decomposing SSMs focused on discovering the structure of recordings, and hence used estimation techniques such as singular value decomposition [FC03] or NMF [KS10]. However, since our goal is to learn about the relationship between the known structure and the recording, we can use straightforward optimization

techniques. We propose to use a quadratic program (QP), a generic formulation of an optimization problem, to find the optimal combination of feature-derived SSMs to reconstruct the annotation-derived SSM in a piecewise fashion.

To illustrate the approach, we use a very simple example: suppose we have a piece with structure ABC, where the last section is twice as long as the previous sections (in general we may have s sections). The annotation matrix N could be:

$$N = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \tag{5.1}$$

Recall that -1 indicates a contrasting pair of frames and $+1$ indicates an identical pair. We would like to explain the reason behind each section using two features: a harmony-based feature and a timbre-based feature. (Again, in general we may have f features.) Suppose the pitch content of the song is identical for sections A and B, and the timbre is identical for sections B and C. For example, A, B and C could be the introduction, verse and chorus of a pop song, where the instrumentation changes after the introduction but stays constant thereafter, and where the pattern of chords only changes at the chorus. The two matrices F_1 and F_2, derived from the harmonic and timbral audio features, respectively, would be:

$$F_1 = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix} \qquad F_2 = \begin{bmatrix} 1 & -1 & -1 & -1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \\ -1 & 1 & 1 & 1 \end{bmatrix} \tag{5.2}$$

Neither matrix equals the true underlying structure N, but we would like to reconstruct the annotation matrix using subsets of the feature matrices that correspond to

the annotated sections. Since there are three sections and two feature matrices, there are six available components, shown below. We can generate these by applying three "masks," one for each section, to each feature. $M_{i,j}$, the element-wise product of the j^{th} mask with F_i, will show how the j^{th} section relates to the other sections with respect to the i^{th} feature.

$$M_{1,1} = \begin{bmatrix} 1 & .5 & -.5 & -.5 \\ .5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \end{bmatrix} \quad M_{2,1} = \begin{bmatrix} 1 & -.5 & -.5 & -.5 \\ -.5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \\ -.5 & 0 & 0 & 0 \end{bmatrix}$$

$$M_{1,2} = \begin{bmatrix} 0 & .5 & 0 & 0 \\ .5 & 1 & -.5 & -.5 \\ 0 & -.5 & 0 & 0 \\ 0 & -.5 & 0 & 0 \end{bmatrix} \quad M_{2,2} = \begin{bmatrix} 0 & -.5 & 0 & 0 \\ -.5 & 1 & .5 & .5 \\ 0 & .5 & 0 & 0 \\ 0 & .5 & 0 & 0 \end{bmatrix}$$

$$M_{1,3} = \begin{bmatrix} 0 & 0 & -.5 & -.5 \\ 0 & 0 & -.5 & -.5 \\ -.5 & -.5 & 1 & 1 \\ -.5 & -.5 & 1 & 1 \end{bmatrix} \quad M_{2,3} = \begin{bmatrix} 0 & 0 & -.5 & -.5 \\ 0 & 0 & .5 & .5 \\ -.5 & .5 & 1 & 1 \\ -.5 & .5 & 1 & 1 \end{bmatrix}$$

The $M_{1,i}$ matrices correspond to how the sections interrelate with respect to harmony, and the $M_{2,i}$ matrices show the same with respect to timbre. The masks halve all of the elements in the off-diagonal sections so that the feature matrices can be reconstructed by summing the components (i.e., $\sum_{j=1}^{s} M_{i,j} = F_i$). We would like to find a linear combination of the component matrices $M_{i,j}$ that will approximate the annotation N as closely as possible. That is, we want the vector of coefficients $x = \{x_{1,1}, x_{1,2}, ..., x_{f,s}\}$ that minimizes the following expression for the squared distance between the annotation and the reconstruction:

$$\left(\left(\sum_{j=1}^{s} \sum_{i=1}^{f} x_{i,j} M_{i,j} \right) - N \right)^2 \tag{5.3}$$

This problem is solvable as a quadratic program (QP) if we imagine each component matrix $M_{i,j}$ to be a single row in a larger array $\mathbf{M}$. If each $M_{i,j}$ is an $n \times n$ matrix, then letting $k = (i-1) \cdot f + j$ we can let $\mathbf{M}_k$, the k^{th} row of $\mathbf{M}$, be the horizontal concatenation of the n rows of $M_{i,j}$:

$$M = \begin{bmatrix} M_{1,1_{(1,1)}} & M_{1,1_{(2,1)}} & \cdots & M_{1,1_{(1,2)}} & \cdots & M_{1,1_{(n,n)}} \\ M_{1,2_{(1,1)}} & M_{1,2_{(2,1)}} & \cdots & M_{1,2_{(1,2)}} & \cdots & M_{1,2_{(n,n)}} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ M_{f,s_{(1,1)}} & M_{f,s_{(2,1)}} & \cdots & M_{f,s_{(1,2)}} & \cdots & M_{f,s_{(n,n)}} \end{bmatrix} \tag{5.4}$$

(In our example, $\mathbf{M}$ would have 6 rows and 16 columns, since there are 16 values in each of the 6 components $M_{i,j}$.)

If we similarly reshape N into a single row vector, and treat x as a column vector, then we may rewrite expression (1) as $(x^T\mathbf{M} - N)^2$. Expanding, we obtain the following expression for c, the reconstruction cost:

$$c(x) = x^T\mathbf{M}\mathbf{M}^T x - 2N\mathbf{M}^T x + NN^T \tag{5.5}$$

Here, $\mathbf{M}\mathbf{M}^T$ is a square matrix with f rows and columns, and NN^T is a constant term which can be ignored in the QP. Our goal is to minimize $c(x)$ subject to any constraints we may place on x. We set $x \geq 0$ and interpret each coefficient $x_{i,j}$ as the relevance of feature i in explaining the similarity of section j to the entire piece. The final QP formulation is:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & x^T\mathbf{M}^2x - 2N\mathbf{M}^Tx \\ \text{subject to} \quad & x_{i,j} \geq 0, \forall\, i = 1, \ldots, f, \forall j = 1, \ldots, s \end{aligned}$$

This is the standard form for QPs, and is quickly solvable on commercial software. All the QPs in this article were solved using the `quadprog` function in MATLAB's optimization package. The inequality is the only constraint placed on the solution; we do not enforce the common constraint $\sum x_{i,j} = 1$ as its interpretation is unclear and we never encountered any problems with degenerate solutions. (In linear and quadratic programming, programs with insufficient constraints may lead to degenerate solutions: ones where any point on a given line or surface satisfy the constraints.)

Solving this quadratic program for our example gives $x = \{0,\ 0.6875,\ 1.0625,\ 1.25,\ 0.3125,\ 0\}$. The reconstruction of the annotation using these coefficients is:

$$\mathbf{M}^Tx = \begin{bmatrix} 1.25 & -.44 & -1.16 & -1.16 \\ -.44 & 1.00 & -.72 & -.72 \\ -1.16 & -.72 & 1.06 & 1.06 \\ -1.16 & -.72 & 1.06 & 1.06 \end{bmatrix} \tag{5.6}$$

($\mathbf{M}^Tx$ is actually a column vector, but here we have reshaped the result into the reconstructed matrix it represents.) The largest components are $x_{1,3}$ and $x_{2,1}$; indeed, the most explanatory components are $M_{1,3}$, which perfectly shows how section C is distinguished from A and B on the basis of its harmony, and $M_{2,1}$, which shows how A differs from the others on the basis of its timbre. The coefficient $x_{1,1}$ is 0, which properly reflects that the harmony of section A is meaningless for distinguishing it from the rest of the piece, and vice versa for $x_{3,2}$. The intermediate values of $x_{1,2}$ and $x_{2,2}$ reflect that it is relatively difficult to explain the middle section with these features. Component $M_{1,2}$ distinguishes section C at the expense of conflating A and B, while $M_{2,2}$ distinguishes A and conflates B and C. Since there is a greater cost for mischaracterizing the longer

section, $x_{1,2}$ is larger than $x_{2,2}$.

The reconstruction cost $c(x)$ for the above solution is 1.125, compared to the maximum cost of 16.0 which is reached when every $x_{i,j} = 0$. If instead of this section-wise approach we had used a matrix-wise approach with two components, F_1 and F_2, we would have found the optimal coefficients 0.67 and 0.33, which gives a reconstruction cost of 5.33. Hence the section-wise approach gets over four times closer to the annotation than the matrix-wise approach in this artificial example. More importantly, the coefficients $x_{i,j}$ reveal when in the piece the different features are most relevant for determining its structure: in this case, harmony is an unimportant feature near the beginning of the piece, but becomes important later on, and vice versa for timbre.

The section-wise QP contains all solutions to the matrix-wise QP as a subset. The solution to the section-wise QP is thus guaranteed to be at least as good, and the reduction in reconstruction cost is no surprise. Although this prevents us from quantitatively evaluating the effectiveness of the section-wise QP, we may use the matrix-wise QP as a performance ceiling and evaluate the result qualitatively.

5.2.3 Reconstructing an annotation SSM from audio SSMs

Using SSM derived from features and the annotation for a song as described in Section 5.2.1, we can formulate a QP using the method in Section 5.2.2. We demonstrate this procedure for the song "Yellow Submarine." Figure 5.3 illustrates how the five feature-derived SSMs and seven section masks produce 35 components. Labeling the component matrices $M_{1,1}$ through $M_{5,7}$, our goal is to find the coefficients $x = \{x_{1,1}, \ldots, x_{5,7}\}$ that minimize $c(x)$. We solve the QP and illustrate the weights x in Figure 5.4.

The results suggest that the first verse, A_1, is best explained by a combination of its chroma and fluctuation pattern vectors, that the first chorus, B_1, is explained almost wholly by its chroma vectors, and so forth. The prominence of FPs in the first and last sections reflects the fact that the FPs were very robust to the changes that occur

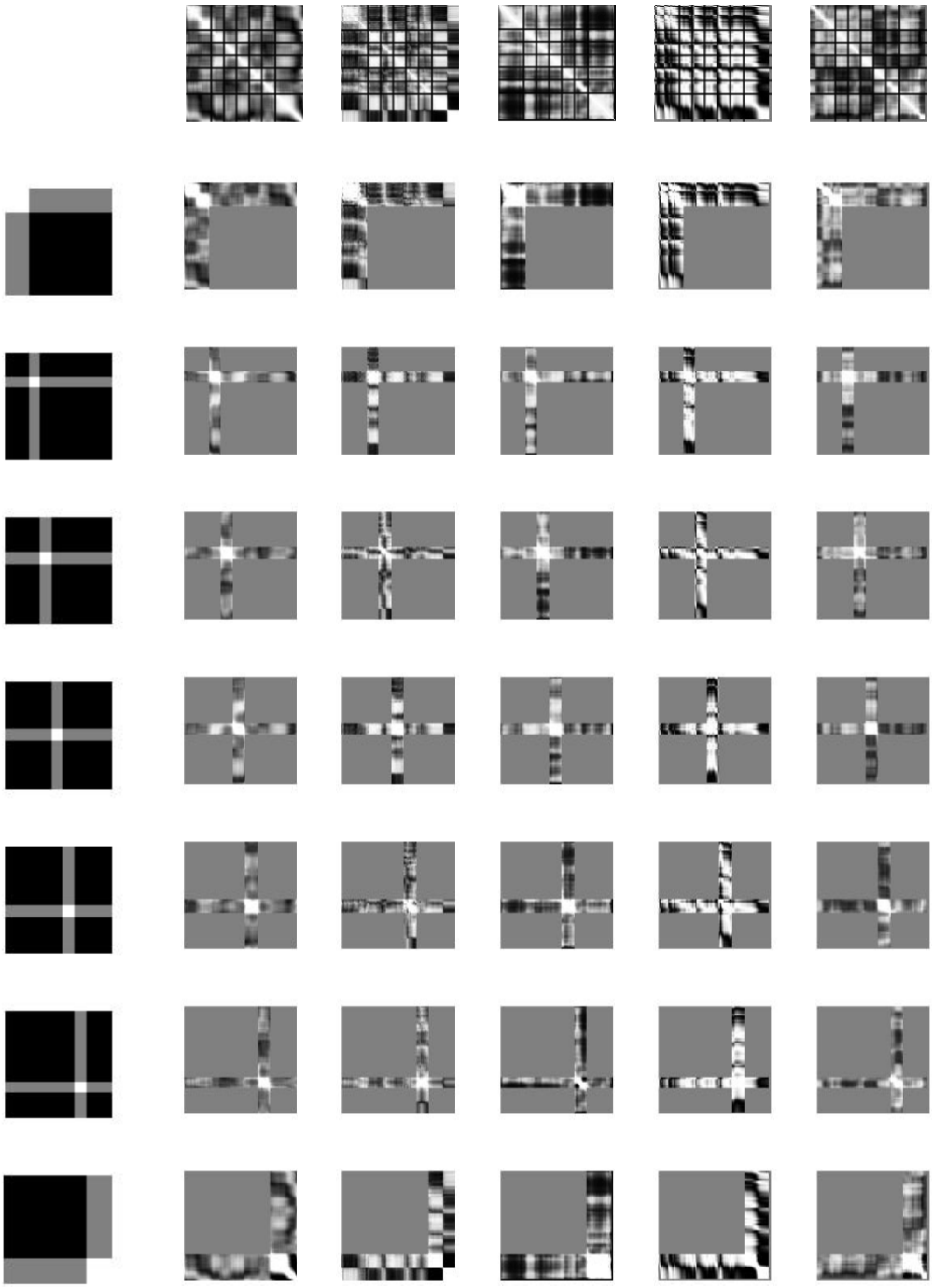


Figure 5.3: Illustration of matrix components. The feature matrices are given in the top row (left to right: MFCCs, chroma, FPs, RMS, and periodicity histograms); the masks in the left column represent each section in the annotation. Each component matrix is the element-wise product of a feature matrix and a mask. The feature matrices and the products are scaled from −1 (black) to +1 (white), while the masks are scaled from 0 (black) to +1 (white).

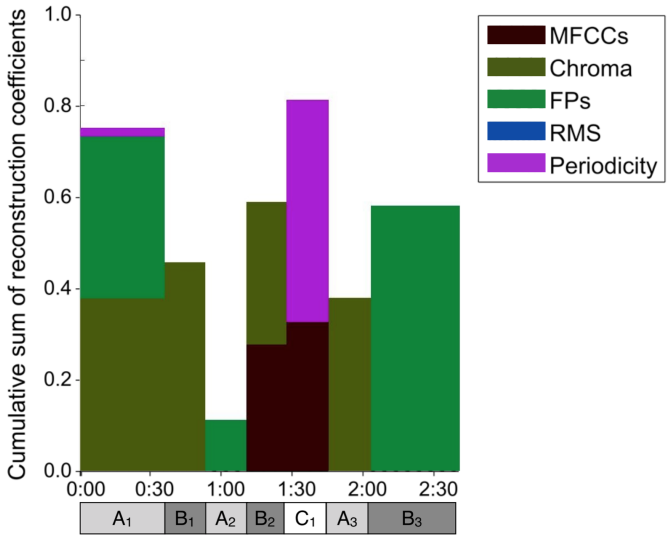


Figure 5.4: The optimal reconstruction coefficients for five different features for all sections of the song Yellow Submarine. The height of each block is the value of the reconstruction coefficient for the section indicated on the x-axis. The annotation is given below the graph.

partway through each section: midway through A_1, a number of nautical sound effects intrude and affect the MFCCs and chroma, and the fadeout in B_3 affects the similarity of other features. Referring to the top and bottom rows of components in Figure 5.3, it is clear that FPs best represent the homogeneity of the first and last sections.

Section C_1 stands out from the piece as being best explained with a combination of timbre and tempo features. Indeed the most distinguishing feature of this section is its timbre, since it is an instrumental portion that contains many unusual sound effects like splashes and bells. Perhaps the arrhythmic nature of these sounds led the periodicity histograms to detect a strong dissimilarity with the rest of the piece.

Our method estimates connections between an analysis and the features, and our analysis of this example suggests that the connection plausibly relates to the listener's experience. However, whether the feature weights obtained by the QP actually correlate to the listener's justifications for their analysis remains a matter of conjecture. Settling this question would require paired data—annotations coupled with listener's self-reported justifications—that is not presently available, though we do plan to collect such data in

the future.

5.2.3.1 Reconstruction cost

Subjectively, the x-values found by the QP are reasonable, but we would like to obtain some quantitative estimate of how well this method works compared to others. One measure of the quality of the output is the reconstruction cost c, which is the average squared deviation between the reconstructed matrix and the target annotation. (This is also the value of the objective function (2) at the solution found by the QP.) The maximum allowable reconstruction cost cannot exceed N^2, since this can be obtained trivially by setting $x = 0$. We can thus express the fractional reconstruction cost c/c_0, where c_0 is the cost at $x = 0$.

With this metric, we can compare the quality of different quadratic programs. To fairly estimate how much analyzing the song section by section instead of all at once improves the reconstruction, we need to run a second quadratic program: this one simply finds the coefficients $x = \{x_1, x_2, \ldots, x_f\}$ that makes the sum of the feature matrices as close to the annotation as possible. This method gives a fractional reconstruction cost of 0.81, whereas the section-wise method garnered a fractional cost of 0.68. This result is expected, since (as noted in Section 2.2) the coarse matrix-wise solution can never be better than the finer-grained solution. Still, by examining how this improvement tapers off with finer-grained formulations of the QP, as done in the next section, we can assess the limit of this method's effectiveness.

The matrices reconstructed using these two methods are pictured in Figure 5.5, along with a plot of the mean squared deviation from the annotation, which highlights those regions that are poorly reconstructed. The latter plots show that both approaches have trouble reconstructing the edges of the SSM (the parts that describe how the very beginning and end of the song relate to the rest). The matrix-wise approach also has particular trouble reconstructing the third verse (section A_3)

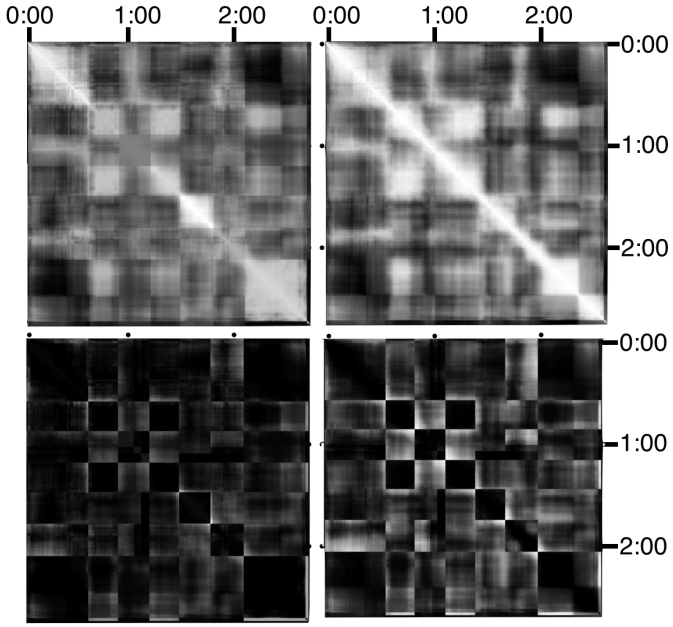


Figure 5.5: Optimal reconstruction of the annotation (see Figure 5.1) using the section-wise approach (top left) and the matrix-wise approach (top right). The mean squared deviation between each reconstructed matrix and the annotation is shown below the reconstruction. (Here, black pixels indicate where the reconstruction is most accurate.)

5.2.3.2 Reconstruction using smaller sections

The section-wise approach was mathematically guaranteed to result in at least as good an approximate of the annotation as a linear combination of full matrices. We may expect even better approximations if we divide the matrix into smaller sections. However, if further segmentation is structurally irrelevant, the reductions in reconstruction cost will taper off. We repeated the previous QP using the finer segmentation of the small-scale sections as well as a "finest-scale" segmentation with segments every 2.5 seconds—shorter than the longest feature windows.

Looking at finer timescales reveals new insights: for example, it is very noticeable that section c_1 is poorly explained by all the features (Figure 5.6(a)). Indeed, this small section contains a novel tune played on a brass instrument and sounds nothing like the rest of the piece. It doesn't even sound like the later sections c_2 and c_3, although

their relatedness could be argued by their both containing sound effects and by their being some kind of variation of the usual A section. Also, whereas in the previous analysis (Figure 5.4) we saw that section C_1 was explained both by its distinct timbre and potentially confusing tempo, we can see now that each half of the section is better explained by one of these features. The first half, c_2, has less percussion than the second half and is arguably the more confusing.

Drilling down further to 2.5-second segments (Figure 5.6(b)), the result is more detailed but not necessarily more informative: the sum of the coefficients leaves a similar "skyline", indicating that approximately the same amount of information is explained in each. However, a few parts are better explained at this smaller scale: the best explanation for a_5 and a' switches from FPs to chroma, bringing A_2 in line with the other A sections. Also, the coefficients in b_7 are higher at this scale.

As stated earlier, finer-grained segmentations are guaranteed to lead to better reconstructions. In the above example, the fractional cost for the small-scale segmentation was 0.67 (compared to 0.68 for the large-scale segmentation), and for the short-window analysis it was 0.61. An analysis of reconstruction costs over a larger corpus shows that, as expected, improvement tapers off at finer timescales. The QP algorithm was executed on annotations for 704 recordings in the SALAMI dataset at the four levels of granularity: matrix-wise, large-scale, small-scale, and finest-scale. The reconstruction costs were computed for each and are plotted in Figure 5.7. We see that while large reductions in reconstruction cost are typical when moving from matrix-wise to large-scale QP formulations, there is less improvement moving from small-scale to short window, indicating diminishing returns when the segmentation proceeds beyond what the annotation contains.

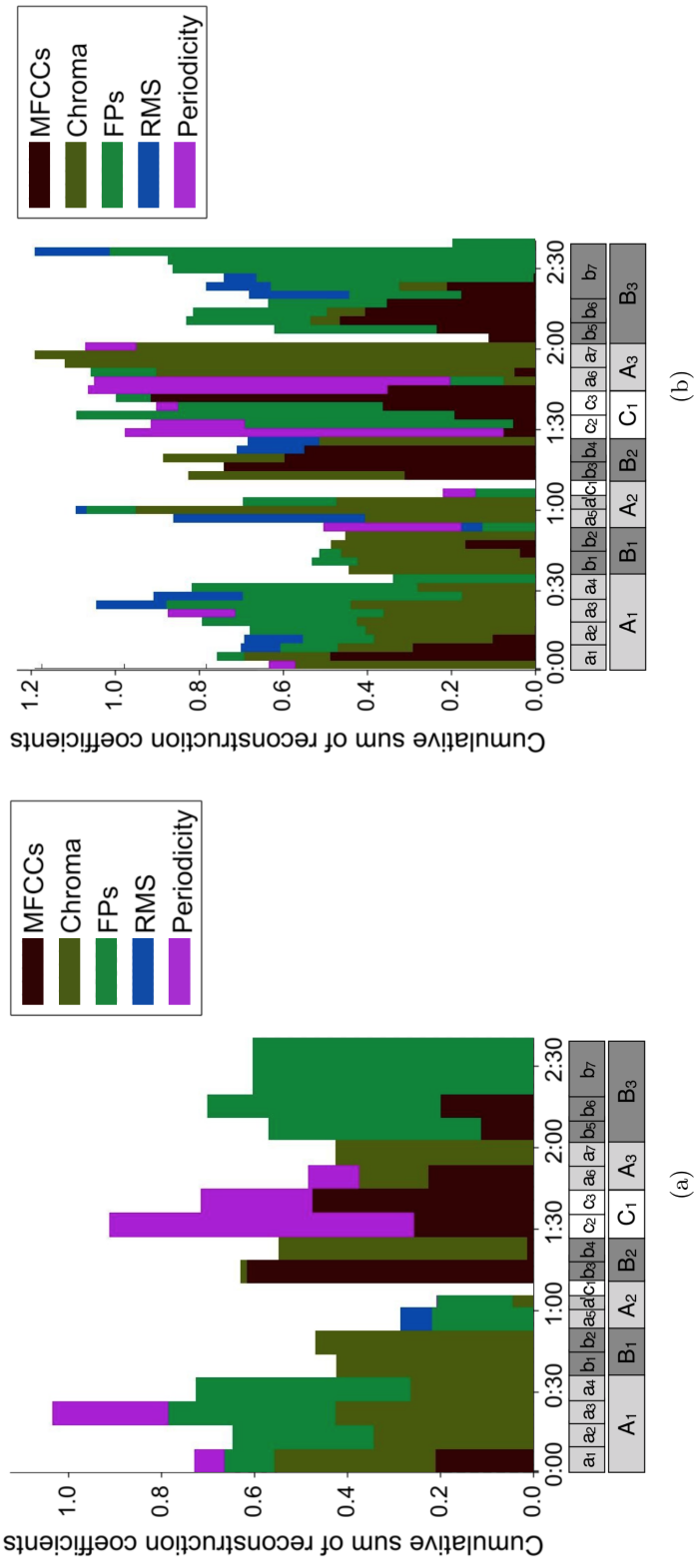


Figure 5.6: The optimal reconstruction coefficients (a) for each small-scale section, and (b) for each 2.5-second frame in the piece.

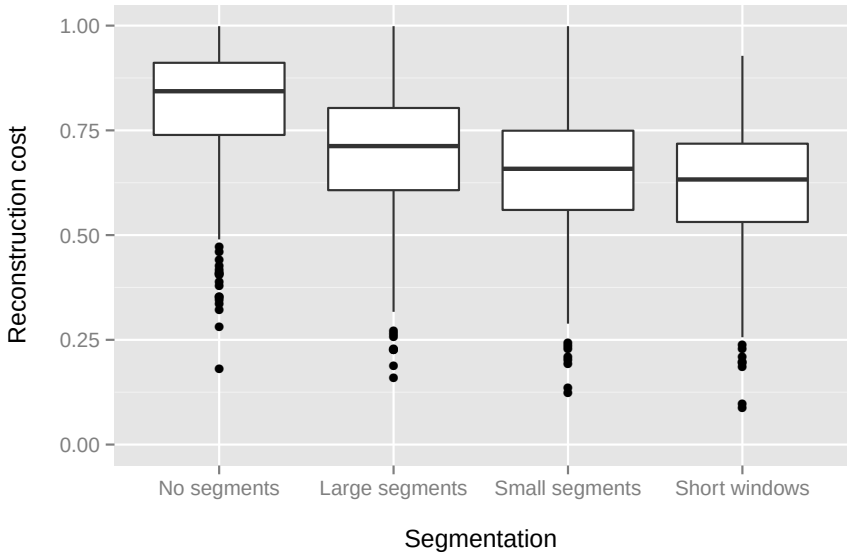


Figure 5.7: Box plot of QP reconstruction costs obtained for QPs using no segmentation (matrix-wise sum) and the piece-wise decomposition at three timescales. The data cover 704 recordings from the SALAMI corpus.

5.3 Visualizing structural differences

The previous examples show that the relationship between the structure of a piece of music and its feature-derived SSMs can vary over time: repetitions in a feature that are irrelevant at one point in a piece may be foregrounded at another. Just as [KP12] and [PE04] argued that dynamic feature weighting could improve structural analysis, our examples show that dynamic feature interpretation could aid in applications based on structural information. Here, we focus on its use as a visualization tool.

The data obtained by our approach may provide interesting visualizations for projects like SALAMI, which plans to execute several algorithms to annotate the musical structure for a large library [SBF+11]. To facilitate browsing in it they have developed a system to visualize each structural description with a diagram, like the annotation in the lower part of Figure 5.6(b) [EBD+11]. Providing the section-wise estimates of which features are estimated to be most salient could enrich these diagrams.

Our method of decomposing annotations is also suited to comparing annotations prepared by different listeners. We illustrate this with two examples. The first shows how a single large difference between two annotations is reflected in the reconstruction coefficients, and the second demonstrates the power of the approach to reconstruct greatly divergent analyses from the same set of components.

5.3.1 Investigating a single difference

As noted in the introduction, it is common for two listeners to analyze the same piece of music differently, and this may be because they are paying attention to different acoustic features. Our analysis method allows one to investigate the differences between two analyses, showing how each may have arisen by emphasizing certain features over others at certain times.

We illustrate this potential by reconstructing two different annotations of the piece "Garrotin", a solo flamenco guitar piece recorded by Chago Rodrigo. Two SALAMI annotators gave analyses that were similar overall (Figure 5.8): the piece begins and ends with many repetitions of the same main melodic gesture, and the middle of the song (0:15 to 1:25) consists of a number of different melodic episodes separated by short reprises of the main theme.

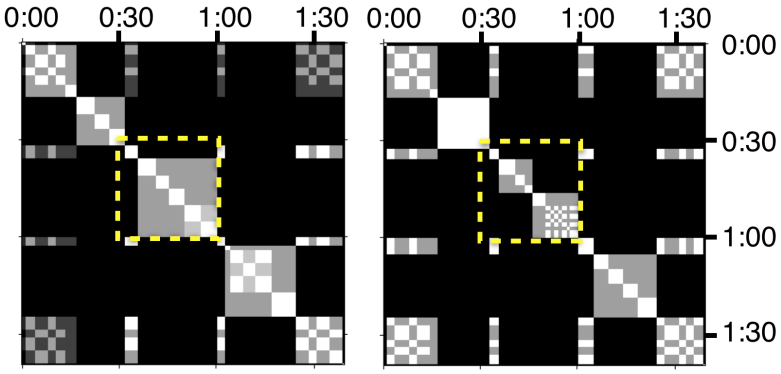


Figure 5.8: Annotation-derived SSMs by two listeners for "Garrotin" (salami_id: 842).

The analyses differ mainly in their treatment of the middle episode (roughly 0:40

to 1:00): the first listener interprets it as a single episode, while the second listener analyzes it as two distinct episodes. (The region from 0:30 to 1:00 is outlined in a yellow dotted line in Figure 5.8.) The feature SSMs (Figure 5.9) show that this portion of the piece is quite self-similar with respect to MFCCs and chroma, but an internal contrast to the section is revealed in the FP-derived SSM at the shorter timescale (5 seconds). This difference is reflected in the solutions to the QP (Figure 5.10): the middle section of the piece is reconstructed best by chroma features when kept as one section, but reconstructed better by FPs when divided in two. And, crucially, this contrast is borne out by the music: both halves of the section mainly consist of an alternation between tonic and dominant chords, leading to similar overall pitch content; but while the meter of the first half is expressed relatively evenly, a $3+3+2$ rhythm is strongly emphasized in the second half.

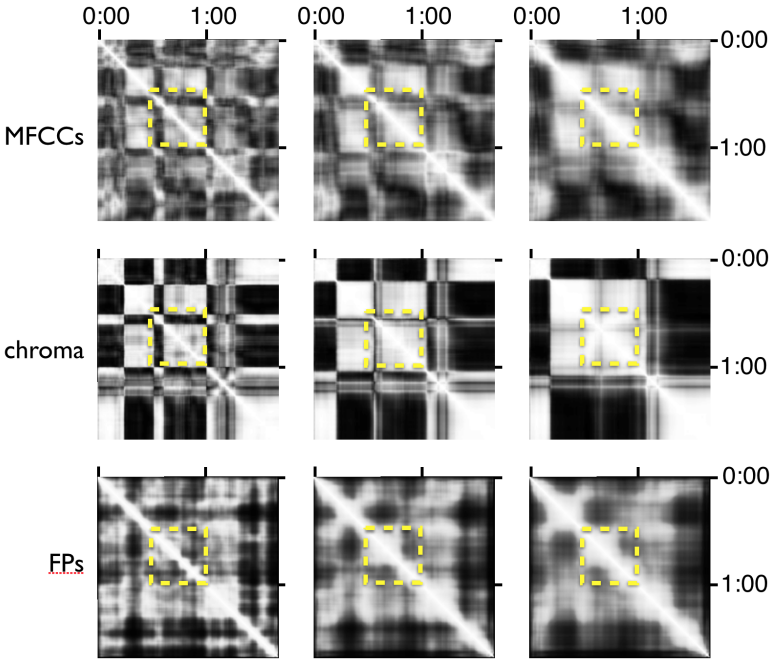


Figure 5.9: Feature matrices for the reconstruction of "Garrotin." In this example, three features are used: MFCCs (top row), chroma (middle row), and FPs (bottom row), as well as three smoothing window sizes: 5, 10 and 15 seconds (left, middle and right columns, respectively). The yellow dotted boxes outline the region between 0:30 and 1:00 to enable comparison with Figure 5.10.

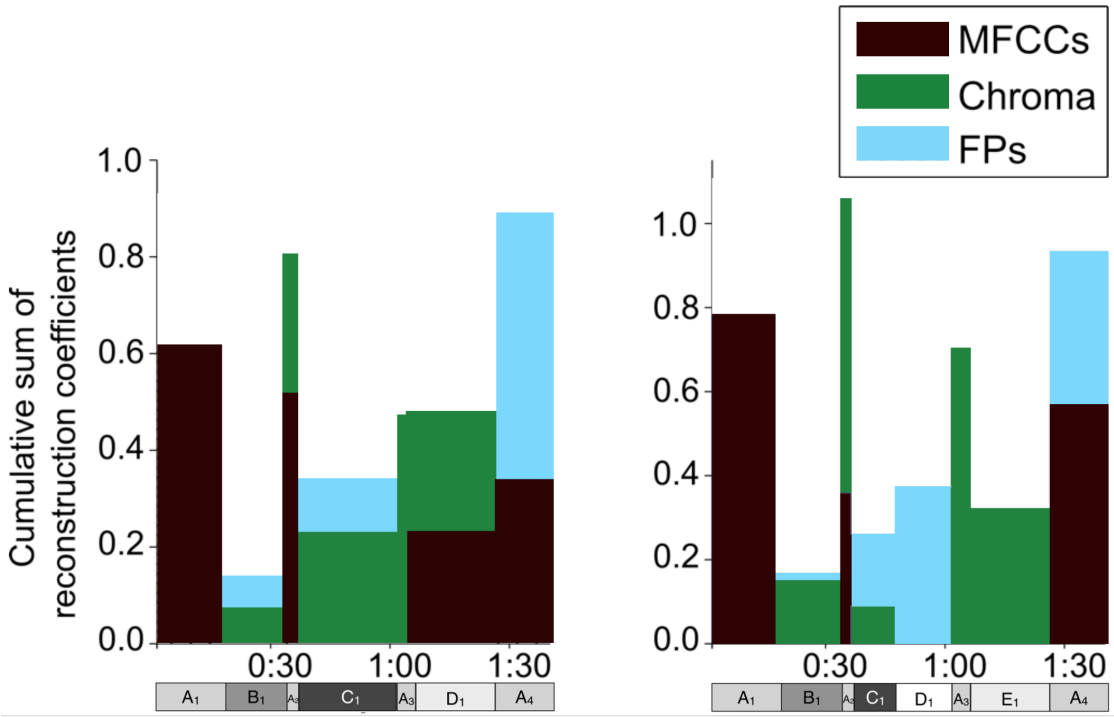


Figure 5.10: The optimal reconstruction coefficients for two different annotations of the song "Garrotin." Three features and large-scale decomposition were used. The coefficients shown here are the average of the coefficients for the three different timescales. The yellow dotted boxes outline the region between 0:30 and 1:00 to enable comparison with Figure 5.9.

5.3.2 Reconstructing dissimilar analyses

In the previous example, a small disagreement was investigated. What if listeners have vastly different interpretations—is it still possible to find QP solutions that justify each interpretation equally well?

In the song "As the Bell Rings the Maypole Spins" by the World music band Dead Can Dance, a singer and bagpiper play a series of reels, and the pattern of reels repeats a few times before a long repetitive coda section ends the piece. The stark difference between the two annotations is apparent from the SSMs (Figure 5.11, top row). The first listener has identified a sequence of three reels as a self-contained repeating group, leading to large off-diagonal blocks in the middle of the SSM. The second listener has not

identified these larger groupings, but does indicate that many of the reels are identical or similar to the coda section (from 3:40 onward), resulting in a series of thin bars in the SSM.

Despite the differences in the annotations, a QP using five features and three timescales has reconstructed both annotations qualitatively well (Figure 5.11, middle row). The fractional reconstruction costs for the first and second annotations, when using the large-scale segmentation, are .63 and .57, compared to .74 and .76 when using no segmentation.

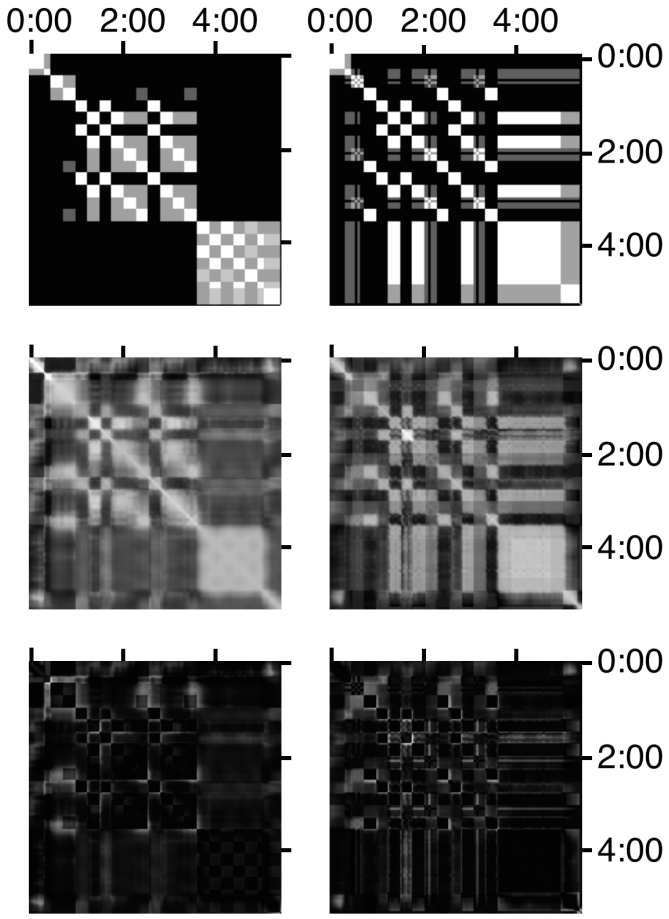


Figure 5.11: SSMs, reconstructions, and errors for "As the Bell Rings the Maypole Spins," by Dead Can Dance (salami_id: 860). Top row: Annotation-derived SSMs from two annotations. Middle row: reconstructed matrices for each annotation Bottom row: mean squared reconstruction error.

Examining the reconstruction coefficients (Figure 5.12), we can observe that the two

solutions depend on fairly distinct sets of features. The first annotation (Figure 5.12(a)) is explained best by FPs in the first section (up to 1:00), and thereafter mainly by MFCCs. By contrast, much of the second annotation is explained best by RMS. Both solutions involve a mixture of features between 1:00 and 2:30, but the mixtures are distinct.

Listening to the song, the different solutions match the different interpretations of the piece. In the second analysis, the coda is given the same label as several earlier sections. What these sections have in common, musically, is that the two main voices, the singer and the bagpipes, play together. Since they play separately in most of the other sections, this gives these sections a unique timbre and makes them among the loudest sections. Hence, RMS and to a lesser extent MFCCs are a good explanation for these sections.

In contrast, the first analysis indicates no relationship between the coda and earlier sections, and identifies larger groupings of reels as single sections. Hence, these sections vary internally with respect to loudness, and RMS is not an important part of the reconstruction. On the other hand, the larger groupings exclude the sections that do not contain bagpipes, so MFCCs remain as the main part of the explanation. In addition, FPs and periodicity histograms, which reflect rhythmic patterns, are important to both the earlier sections and the coda. A subtle change in meter can be noted between the coda and the earlier sections: the piece, which is in 6/8 time, is mostly played as $3 + 3$, but in the coda, the percussion presents a hemiolic $2 + 2 + 2$ rhythm.

5.4 Conclusion and future work

We have introduced the problem of estimating the relevance of different acoustic features to different sections of a structural analysis, and proposed a method of solving the problem based on quadratic programming. The approach is founded on the intuition that while acoustically-derived SSMs may not always reflect the perceived structure of a piece, components of SSMs for specific features may explain the perception of structure

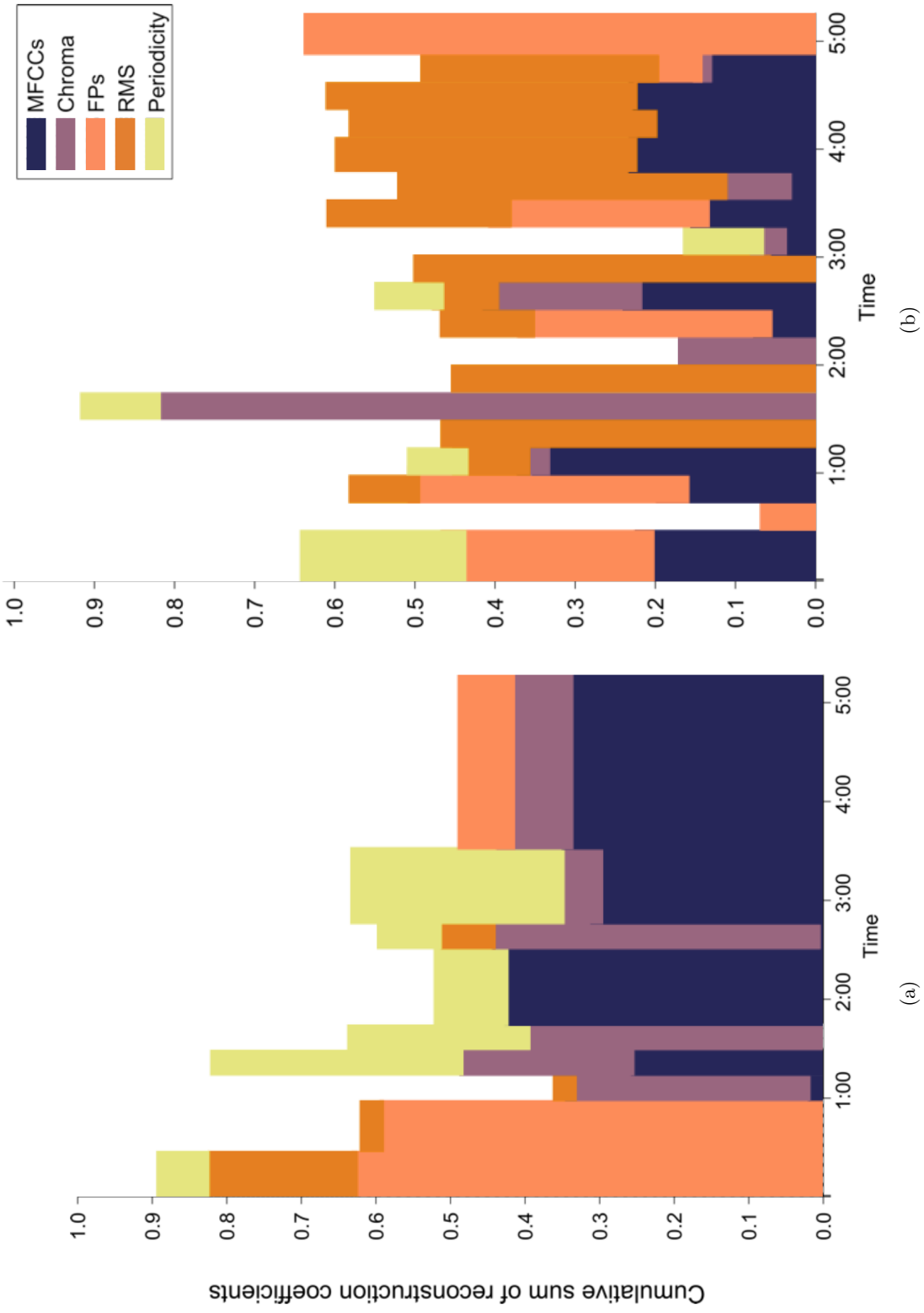


Figure 5.12: The optimal reconstruction coefficients for five features for large sections of the song "As the Bell Rings the Maypole Spins." The coefficients shown here are the average of the coefficients for the three different timescales.

in a piece-wise fashion.

The method presented is a general one and several aspects could be changed. Many more (and more sophisticated) features could be used than the ones tested here, leading to more detailed reconstructions at the cost of a larger QP. One could also use many of the same SSM transformations used in the literature (see Section 5.1), filtering the SSMs to emphasize the desired types of patterns.

Separately from the choice of features, the parameters of the QP, such as the mask shape, could be modified to suit different problems. For example, if the masks (as in Figure 5.4) were altered to only include the upper left portion of the SSM, a future-agnostic analysis would result. This might be useful for modeling how a listener's first hearing of a piece compares with a later one in which the listener is more familiar with it. The annotations considered here were all produced after the music had been heard and briefly studied, so such an approach did not make sense here, but response data that reflected one's real-time perception of structure would perhaps best be analyzed with such a future-agnostic framework.

Similarly, a QP using masks that emphasized or de-emphasized the main diagonal would alter the importance of long-term memory in listening. A reconstruction that emphasized the main diagonal would provide a solution that explained only local similarities, modeling how a listener might account for their small-scale decision about structure. This version would be useful if, for instance, one considered the acoustic similarity between the very beginning and end of a piece to be unimportant. (The more narrowly one focuses on the diagonal axis, the more this method resembles the novelty function calculation proposed by [Foo00]—except, instead of correlating the diagonal axis with a checkerboard kernel, we would be correlating it with the ground truth annotation.) On the other hand, emphasizing the off-diagonal would relax the need to explain the precise segmentation given in the annotation, and focus on finding a justification for grouping distant parts of the piece.

One important caveat with our approach is that it is crucial that the annotation used properly reflect the information that one seeks to explain. The algorithm, as presented, will work only so far as the "states" hypothesis applies to the annotation rather than the "sequences" hypothesis. That is, our method assumes that a section with a given label is homogenous, and uniformly distinct from any differently-labeled section. It would be useful to extend this work to account for "sequences" interpretation of annotations, in which a section B is presumed to be a heterogeneous sequence of events that recurs exactly whenever B repeats. The use of structural information at multiple timescales, described in Section 5.2.1.1, is intended to mitigate this shortcoming, since in practice the short-timescale annotation often charts a "sequence"-like path through the blocks of the large-scale annotation.

A simple adaptation of our method to account for sequences would be to alter the block-based masks so that diagonal stripes appeared in those locations where a repetition was annotated. However, this is not as straightforward as it sounds, since two annotated sections with the same label may have different lengths, and thus the stripe locations cannot be predicted. The method would have to be redesigned to incorporate an automatic alignment algorithm such as [MA08].

The annotations investigated here are the products of listeners, and it is possible that our method reveals insights into how listeners analyzed the pieces: what features they found most salient, and what groupings they paid the most attention to. However, the findings of this chapter are strongly limited by the fact that we have relied on close examination of a few songs as anecdotes. To establish a general correlation between the SSMs and the listeners' salience and grouping judgements would require new experimental data. The next chapter will describe an experiment, whose data could be used for such purposes in the future.

Chapter 6

The effect of attention on grouping decisions

Chapters 3–5 have built up and supported the hypothesis that listeners may pay attention to different musical features, and that this can cause them to differ in their interpretations of grouping structure in music. However, in each chapter, it was not clear whether there was a causal link between the attention of the listener and their analysis. In this chapter, we test this hypothesis in an experiment with listeners.

We study the influence of attention to musical features (including harmony, melody, rhythm and timbre) on grouping decisions. The experiments use composed musical stimuli exhibiting changes in particular features by design; some stimuli exhibit a single change, while others exhibit changes in different features at different times, leading to ambiguous segment boundaries and groupings.

The parts of the experiment address four questions: first, are listeners able to attend to different features within a piece of music? Second, does the salience of a change in music increase when one is focusing on the feature that changes, rather than listening normally? Third, does focusing on a feature make a listener more likely to group sections

in accordance with how that feature changes? And finally, are listeners able to correctly extend an analysis based on an arbitrary feature? Our findings suggest that all of these questions may be answered in the affirmative.

6.1 Introduction

6.1.1 The role of attention

Jones and Boltz [JB89] proposed that listeners can adopt one of two listening strategies: future-oriented attending and analytic attending. In future-oriented attending, listeners use the structure of the music to anticipate what will happen; in analytic attending, listeners are focused on a shorter timescale and are only tracking events, not anticipating them. The authors develop a formalism to describe how hierarchical levels are related and how listeners attune to them. Most music involves deviations from an ideal hierarchy, which can lead to more complex structures. This model is comparable to the proposal by Hanninen that listeners may adopt either a sonic or associate orientation when analyzing music [Han12]. The sonic orientation searches for local discontinuities and is thus aimed at the shortest timescale, whereas the associative orientation builds on this and projects backwards and forwards across the music in order to establish similarity relationships. In a chapter about the relationship between repetition and attention, Margulis pointed out that with repetition can come either ritualization or routinization [Mar14]. In ritualization, the focus narrows to consider the subtleties of individual gestures; in routinization, the focus broadens to grasp the scope of the narrative.

In each of these accounts of how listeners' attention can shift, it is uncertain how much control the listener has over how their attention shifts, and how much their attention is guided by the music. In Margulis' case, repetition enables both routinization and ritualization, but how or whether one's focus actually changes is up to the goals or whims of the listener. In Jones and Boltz's model, a listener chooses a mode of attending,

but attuning to events at larger timescales may become too difficult if the information content of the music is too high, and listeners fail to attune to the correct pulse (i.e., fail to generate proper expectations about the music). And although sonic and associative orientations are distinct in Hanninen's view, she explains that as modes of attending, they are interdependent: detecting boundaries requires identifying coherent groupings, and vice versa.

As an example, consider Steve Reich's "Clapping Music." Its overall structure is very simple to articulate: two performers clap the same rhythm repeatedly, with one performer skipping ahead in the pattern every eighth measure. But this simple description results in a complex musical surface, and a listener's attention may be divided: one is partly aware of the piece's overall structure and one's place in it (i.e., how many different-sounding sections have been traversed), and yet the extreme repetition and occasional small or stark change can draw one's attention to the smallest timescale.

In the works above, Jones and Boltz and Margulis mainly discuss the various timescales at which a piece of music may be regarded. But unlike "Clapping Music," most music is multi-dimensional, in the sense that several parameters—timing, loudness, melody, harmony—are usually changing at once or independently. A listener's attention may be divided among all of these dimensions: one can pay attention primarily to the melody of a piece, or to the changes in key centre. The musical feature being attended to, like the timescale, can shift throughout a piece. This might happen as a matter of course: if one is focused on the singer of a rock song but the singing stops, attention shifts to what remains. But a shift in attention may be inspired by subtler musical changes; if the melody ceases to evolve and begins to repeat a single gesture, attention may change to a different timescale (as described by Margulis) or shift to the other parts of the music, which may be continuing the evolution. Finally, attention can shift as a result of the listener's goals: a listener may lose interest in the lyrics and choose to focus on the other parts.

In Chapter 3, we argued that disagreements in grouping structure arise when listeners

pay attention to different features of a piece of music, and sought to follow the chain of causation backwards from there. However, we could not be certain that attention was what caused the listeners to prefer the groupings they did, or if this attention was merely a *post-hoc* attribution. This is much like the question of whether novelty causes listeners to perceive boundaries or if this is merely decided *post-hoc*, which was addressed in Chapter 4. The main goal of this chapter is to determine whether paying attention to a particular feature in music actually does *cause* listeners to perceive groupings according to that feature.

6.1.2 Proposed experiments

Does paying attention to a feature lead one to prefer a grouping analysis that matches that feature? The first and second experiments described in this chapter build further support for this hypothesis, which is tested directly in Experiment no. 3. Experiment no. 4 addresses a question that arose earlier in this thesis. In Chapter 3, we hypothesized that after listeners have interpreted the beginning of a piece of music, this interpretation colours the rest of their listening. Although this is a complex hypothesis and is not tested completely in this thesis, we begin to answer it by assessing how plausible it is that listeners internalize an analysis from initial excerpt and apply that understanding to a longer excerpt.

We start with a very basic claim: that all listeners, musicians and non-musicians alike, are capable of multi-dimensional listening. That is, listeners can discern changes in musical patterns that are only expressed by a single feature (such as melody or timbre), and furthermore are able to identify the feature that expressed the change. While we may wish to take this claim for granted, it is important to test here for two reasons: first, to establish that the patterns in the artificial stimuli we designed for all of the experiments are in fact discernible by our participants; and second, to establish what the importance of musical training is in developing this listening skill.

This hypothesis is tested in **Experiment no. 1: Change identification**, which has a very straightforward design. Participants listen to short excerpts which are static with respect to all musical parameters save one. Listeners are tasked with identifying the parameter that changed.

Assuming that participants can identify which feature expressed a change in the music, it follows that they can pay attention to this feature and track it over time. This leads to our second hypothesis: when listeners are paying attention to a feature, changes expressed by that feature become more salient than they normally are.

The test for this is also straightforward. In **Experiment no. 2: Salience judgements**, participants are asked to focus on a single feature in a short excerpt that, like the stimuli in Experiment no. 1, exhibits a change in only one feature. The feature they pay attention to may match the changing feature or not. Listeners rate the salience of the change, and are expected to find the changes more salient when they have paid attention to it.

Bruderer et al. showed that the salience of change points correlates to the probability that listeners will perceive them as boundaries [BMK09]. Therefore, if focusing on a feature increases the salience of changes in that feature, it should follow that listeners who pay attention to a feature are more likely to segment a piece according to that feature. Our third hypothesis extends this claim, suggesting that the influence of attention goes beyond boundary detection and also affects grouping analysis. We propose that paying attention to a feature leads one to perceive a grouping structure that accords with that feature.

In **Experiment no. 3: Pattern detection and Grouping preference**, we present listeners with an ambiguous three-part excerpt, which has structure AAB according to one feature and ABB according to another, and ask them which grouping they prefer. For each trial, the listener's attention is directed toward a single feature, that may match either of the two structures. Instead of overtly asking listeners to pay atten-

tion to a given feature, we aim here to manipulate their attention covertly. To do so, we ask listeners to detect whether a given pattern occurs in the excerpt; in this way, their attention is directed to the feature that expressed the pattern.

If confirmed, this hypothesis supports the notion that attention is a mechanism that can explain listener disagreements among otherwise similar listeners. In our examination of listener disagreements in Chapter 3, after arguing that attention was the proximate cause of listener disagreements, we argued that one cause of differences in attention is how the opening moments of a piece are perceived. When a piece begins, a listener's ears are a blank slate and they can pay attention to whatever they happen to find most salient. If the beginning of a piece is especially ambiguous, then how a listener perceives these moments may have an impact on how the rest is heard.

In **Experiment no. 4: Analysis continuation**, we test the hypothesis that listeners are able to extend an analysis to the remainder of a piece. Participants are played two very brief samples with sparse texture, identified as A and B. We then present them with a longer, full-textured excerpt whose structure is very ambiguous—it has form $AABB$, $ABAB$ and $ABBA$ with respect to three different features. If participants are able to continue the analysis correctly, this will demonstrate that it is possible for listeners to stick with an analysis conceived in the opening moments of a piece.

Section 6.2 describes the participants, the musical materials developed for the experiments, and details of the procedure. The results are presented in Section 6.3, and discussed in Section 6.4.

1. I have never been complimented for my talents as a musical performer.
2. I can't read a musical score.
3. I would not consider myself a musician.
Answers for questions 1–3: (1) Completely disagree, (2) Strongly disagree, (3) Disagree, (4) Neither agree nor disagree, (5) Agree, (6) Strongly agree, (7) Completely agree.
4. I engage in regular, daily practice of a musical instrument for _____ hours.
5. At the peak of my interest, I practiced _____ hours per day on my primary instrument.
6. I have played or sung in a group, band, choir, or orchestra for _____ years.
7. I have had formal training in music theory for _____ years.
8. I have had _____ years of formal training on a musical instrument.
9. I can play _____ musical instruments.
Answers for questions 4–9: 0, 1, 2, 3, 4–5, 6–9, 10 or more.

Table 6-A: Musical training survey questions

6.2 Method

6.2.1 Participants

Participants were recruited via emails to academic and social lists at universities in the UK and in Canada, and on international academic lists. Aside from a stipulation that participants be at least 18 years old, no participant was refused.

A total of 87 participants completed all four parts of the experiment, including 50 men and 35 women (2 did not report their gender). Participants ranged in age from 20 to 71 years with a median of 30 ($M = 34.26, SD = 12.55$).

The level of musical training of the participants was assessed at the end of the experiment using a set of nine questions from the Goldsmiths Musical Sophistication Index [MGMS14], shown in Table 6-A. Each answer was scored on a scale of 0 to 6 and the scores were summed to produce an overall ability score. The scores fell mostly between 20 to 50 with a median of 36 ($M = 34.28, SD = 7.11$).

Since the balance of these characteristics was not controlled, we tested for spurious

correlations. Gender was well decorrelated from age ($d.f. = 84, r = .03, p = .78$) and ability ($d.f. = 82, r = .06, p = .57$). There was a slight negative correlation between age and ability, although this trend only approached significance ($d.f. = 82, r = -.18, p = .11$) and is probably due to a few outlying participants with higher age and lower reported ability. These outliers were not removed, leading to some spurious age-related effects among the results, as will be seen later.

We also tested for whether any of these characteristics correlated with the main blocking variables in the experiment: these were the musical environment and the stimulus set used in each part (described below). Without applying any correction for the multiple comparisons required (which would have increased all p-values beyond 0.05), we did discover some slight anomalies: those who received the second (out of four) sets of stimuli skewed male, as did those who heard the "HT-MR" music (explained in the following section) on Experiment no. 4. But overall, a fair cross-section of participants completed each version of the experiment.

6.2.2 Material

Musical environments were created in which four musical attributes could be systematically manipulated: chord progression, melody, rhythm, and timbre. A great many other attributes could have been chosen (e.g., loudness, tempo, dynamics, register), and subtler aspects of these attributes could have been manipulated systematically (e.g., degree of syncopation in rhythm, level of dissonance in a melody). However, the more attributes involved, the greater the experiment size. We chose the four attributes based on their ability to be manipulated independently, their importance (among other features) in communicating form (as reported, e.g., in [CK90]), and their use in previous work [BMK09].

For each musical environment, two parts were composed that expressed these four attributes. The environments differed in how the attributes were varied in the voices;

in other words, which voices were "convolved" (in the sense of "entwined," not the mathematical sense). In the "HR-MT" environment, harmony is convolved with rhythm, and melody with timbre: that is, one voice plays two different chord progressions with two different rhythms (a total of four possible chord parts), while the other voice plays two different melodies with different timbres (four possible melody parts). In the "HM-RT" environment, harmony is convolved with melody, and rhythm with timbre; in the "HT-MR" environment, harmony is convolved with timbre, melody with rhythm. Given an environment, a measure of music can be created by choosing one of the four possible parts for one voice, and one for the other, for a total of 16 "stems." The eight parts composed for the "HT-MR" environment are shown in Figure 6.1.

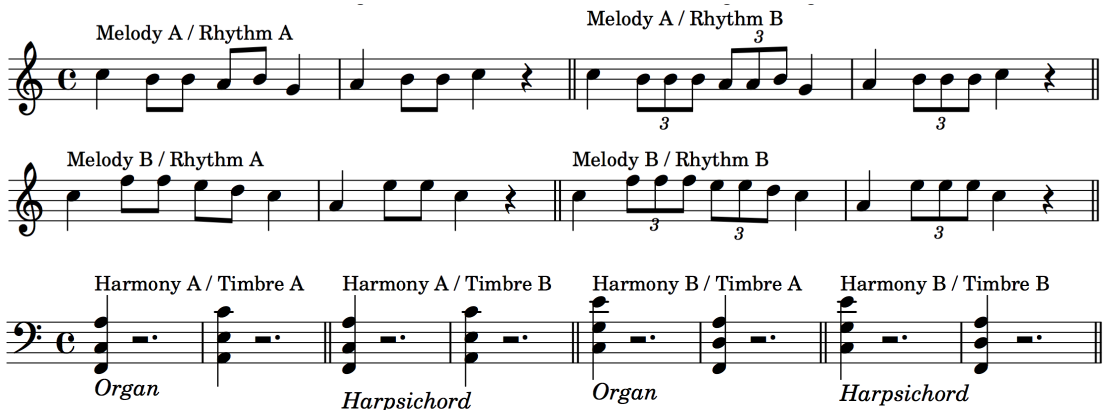


Figure 6.1: Voice parts for the HT/MR environment. Above: four combinations of melody and rhythm. Below: four combinations of harmony and timbre.

Using these two-voice stems, we can compose brief passages of music that express whatever form, with respect to whatever musical attributes, we like. The most basic passages, used in experiments 1 and 2, present listeners with examples that have an AB structure by having a single feature change its pattern in the middle. An example is shown in Figure 6.2 in which the harmony, melody and timbre are constant throughout, but the rhythm changes halfway through. Experiment 3 included three-part stimuli in which two features varied, each expressing pattern AAB or ABB (Figure 6.3), and in Experiment 4, each stimulus had form $AABB$, $ABAB$ and $ABBA$ with respect to three different features (Figure 6.4). (To keep the length of the four-part stimuli manageable,

the stems were not repeated.)

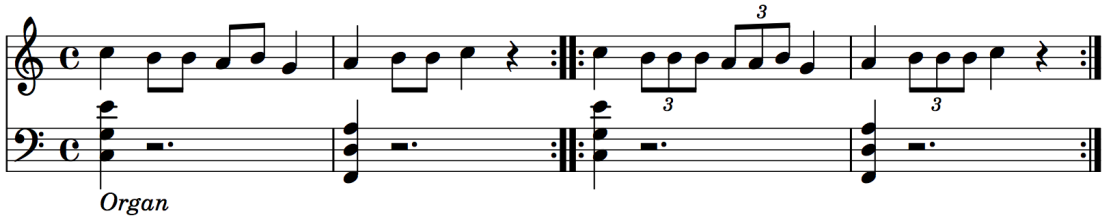


Figure 6.2: Example two-part stimulus with pattern *AB* with respect to rhythm. This particular pattern has codename *HB-MA-RAB-TB*.

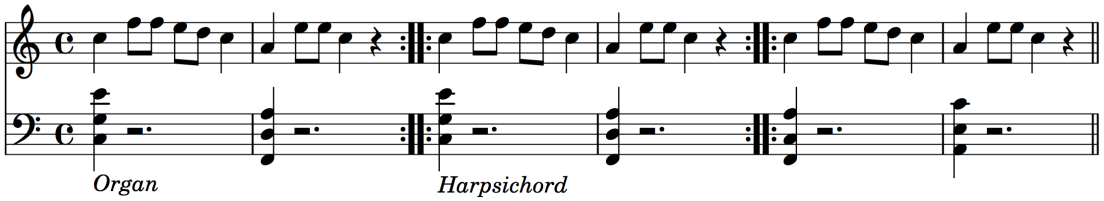


Figure 6.3: Example three-part stimulus with pattern *AAB* with respect to chord progression and *ABB* to melody. This particular pattern has codename *HBBA-MB-RA-TABB*.

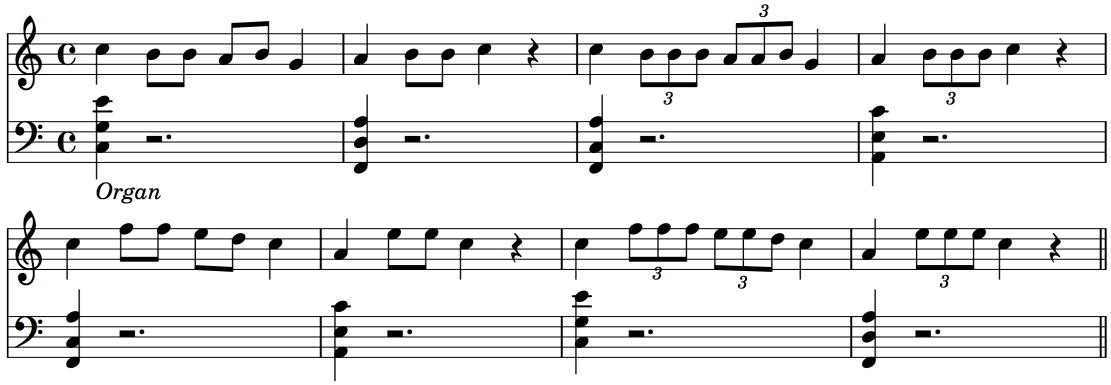


Figure 6.4: Example four-part stimulus with pattern *ABBA* with respect to chord progression, *AABB* to melody, and *ABAB* to rhythm. This particular pattern has codename *HBAAB-MAABB-RABAB-TB*.

The music was composed using Digital Audio Workstation software with standard available instrument sounds. All features other than the four being manipulated were held approximately constant across the sets of stimuli and within each set. Specifically, all the clips had the same exact tempo (140 beats per minute), and no dynamic patterns (e.g., crescendos or diminuendos) were used. Although the software was not capable of precisely equalizing loudnesses, each voice and each set of music were set to have roughly

equal loudness. Finally, when a voice was not meant to express a feature, that feature was held constant across all versions of the voice. For example, in the set where rhythm was convolved with harmony, the melody voice still had a rhythm, but this rhythm was held constant for this set of stimuli.

6.2.3 Procedure

The sequence of the experiment was as follows. Screen captures of each portion of the experiment are reproduced in Appendix A.

Explanation of music analysis. The experiment began with a very broad definition of music analysis as "a compact description of the patterns in the music", and explained how strings such as $ABAB$ can be used to record an analysis, in a manner similar to rhyme schemes in poetry.

Experiment no. 4: Analysis continuation. This experiment contained 12 trials, preceded by an introductory set of 4 trials.

Definition of musical terms. This page defined the terms melody, chord, rhythm and timbre. The page included a rendition of "Twinkle Twinkle, Little Star", and following each definition, participants could hear an example taken from this tune. Participants then had to say how confident they were that they understood the definition on a scale from 1 ("not at all confident") to 5 ("totally confident").

Experiment no. 3: Pattern detection & grouping preference. This experiment contained 12 trials, preceded by an introductory set of 3 trials.

Experiment no. 2: Salience judgements. This experiment contained 12 trials.

Experiment no. 1: Change identification. This experiment contained 12 trials.

Additional information. This section contained the musical training survey questions described in Table 6-A, and a text box for participants to leave any comments they

wished.

The sequence of the experiments is summarized in Table 6-B. Note that participants completed the parts in the order opposite to their presentation here: participants began with Experiment no. 4, the least directed task, and progressed to Experiment no. 1, the most directed task. This scheme was used so that the more specific instructions of experiments no. 3 and no. 4 were not used by listeners to guide their listening in no. 1 and no. 2.

Table 6-B also indicates how the musical environments varied across trials: although randomized across participants, every participant heard the same musical environment for the first and final parts. The table also summarizes the form the stimuli took in each part, and the anticipated time to complete each part told to each participant. We gave participants this time estimate to prevent participants from overthinking; during testing we found that participants could become indecisive and spend too long on each trial. The prompt had the desired effect: the mean time taken to solve the four main parts of the experiment was at most 10, 15, 5 and 5 minutes, compared to a projected 8, 12, 6 and 6, respectively.

In every experiment below, two independent variables changed across participants: (1) the musical environment assigned to each task, and (2) the stimuli sequence. Four unique stimuli sequences were constructed for each task, although not every sequence contained 12 unique sounds (details below).

6.2.3.1 Experiment no. 1: Change identification

Hypothesis: Listeners (musicians and non-musicians alike) are able to consciously segment a piece of music according to changes in a specified feature when asked. That is, their understanding of these stimuli is multi-dimensional. To test this, ask listeners to identify what changed in a stimulus and expect them to answer correctly.

Stage	Number of questions	Approximate time to complete (minutes)	Musical environment	Stimulus form
Experiment no. 4 Intro	4	2	Music 1	AABB-ABAB-ABBA
Experiment no. 4: Analysis continuation	12	8	Music 1	AABB-ABAB-ABBA
Explanatory stage	3	1	Twinkle Twinkle, Little Star	
Experiment no. 3 Intro	4	2	Music 2	AAB-ABB
Experiment no. 3: Pattern detection & grouping preference	12	12	Music 2	AAB-ABB
Experiment no. 2: Salience judgements	12	6	Music 3	AB
Experiment no. 1: Change identification	12	6	Music 1	AB

Table 6-B: Summary of experiment sequence

Task: Participants were presented a stimulus with a single change in the middle (structure AB) and asked to listen. The question posed was: "Please indicate the musical feature that changed during this excerpt." The possible answers were "Chord progression", "Melody", "Rhythm", "Timbre", or "No change."

Variables: Each participant's 12 stimuli included 8 different sound samples, with each of the four features changing in two different examples. The last four stimuli were repetitions of 4 earlier stimuli, one for each feature. Thus, two independent variables were varied within participant: the feature that changed, and whether the stimulus had been heard before.

The response variable was whether the change was correctly identified.

6.2.3.2 Experiment no. 2: Salience judgements

Hypothesis: Paying attention to a feature causes a change in that feature to be more salient than a change in another feature. Since change salience correlates with boundary placement [BMK06], our hypothesis implies that focusing on a feature leads to the perception of boundaries when that feature changes. To test this, we ask listeners to concentrate on a given feature and expect the rated salience to be greater when they are concentrating on the feature that changed.

Task: Participants were presented a stimulus with a single change in the middle (structure AB) and told to pay attention to a specific feature (either chords, melody, rhythm or timbre). The question posed was: "How strong is the change at the midpoint of the excerpt?", with answer ranging from "1. Not strong at all" to "5. Extremely strong."

Variables: The independent variables varied within each participant were the focal feature (the feature they were told to pay attention to) and the changing feature. The combination of focal and changing feature was varied, with three possible outcomes:

"match" (the focal and changing features were the same), "convolved" (the focal feature was expressed by the same voice as the changing feature), and "wrong" (the focal and changing features were carried by different voices). Each participant's 12 stimuli included only 4 unique sound stimuli—each containing a different changing feature—and the focal feature was varied across the three presentations per stimulus. Participants never heard the same stimulus twice in a row.

The response variable was the salience of the boundary, rescaled from -2 to $+2$.

6.2.3.3 Experiment no. 3: Pattern detection & grouping preference

Hypothesis: Paying attention to a feature (even when this attention is not completely conscious) makes one more likely to analyze a piece according to that feature. To test this, we direct participants' attention to a feature in an ambiguous stimulus and expect them to prefer the analysis that matches this feature.

Task: Each trial had two subtasks. First, participants were shown a target musical pattern (labelled as either a melody, chord progression, rhythmic pattern or timbre) and had to answer whether the pattern occurred in a longer three-part stimulus; the possible answers were "yes", "yes, but only a variation", "no" and "I don't know." Second, participants were asked to re-listen to the three-part stimulus and indicate the analysis that they thought fit best: AAB or ABB. They then indicated their confidence on a 5-point Likert scale from "not at all certain" to "totally certain."

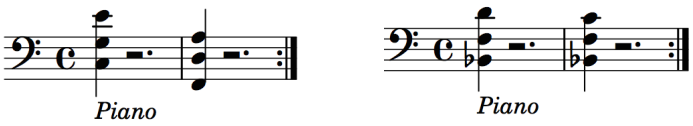
We further hypothesized that confidence in one's answer would be greater when the target pattern was present and when the target feature was relevant to the analysis.

A short introduction to this experiment was used to establish a baseline of preference for each stimulus. Participants heard three different three-part stimuli and were asked only to indicate which analysis they preferred, AAB or ABB. No target pattern was given.

Variables: Each participant heard 12 unique stimuli, each of which was defined by two variables: the feature expressing AAB and the feature expressing ABB. All six combinations of two changing features appeared twice. The choice of target pattern defined three more variables: the focal feature (the type of pattern being focused on), the presence of the target (in half of the trials, the target pattern was absent from the stimulus; in half, it was present), and the relevance of the target (in half of the trials, the focal feature matched one of the two changing features of the stimulus; in half, it did not match either).

The response variables for each trial were whether the target's presence was determined correctly or not, whether the suggested analysis matched the implied analysis, and the participant's confidence in their choice of analysis.

Note that the target pattern was always somewhat abstracted from the musical environment. The harmony and melody parts were always played on piano; the rhythmic patterns were played with a single drum sound; and a single note was played for the timbre examples. Also, "fake" targets that were comparable to the composed patterns were written for those cases when a target was required to be absent. For example, a target chord progression was created that matched the spacing and rhythm of the actual chord progressions, but which did not appear; see Figure 6.5.



(a) (b)

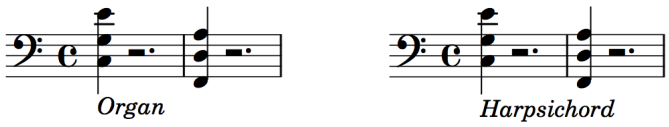
Figure 6.5: Example targets for Experiment no. 3. (a) A target chord progression that is present in the stimulus in Figure 6.3 (implying analysis AAB). (b) A "fake" target chord progression that is not present but which matches other properties of the voice.

6.2.3.4 Experiment no. 4: Analysis continuation

Hypothesis: Listeners who interpret the beginning of a piece in one way can use this to guide their continued analysis. To test this, we present listeners the beginning of an analysis, ask them to continue it, and expect them to do so correctly. (Note: we are here only testing whether listeners *can* do this, not whether they *do* do this in general listening.)

Task: Participants were asked to "imagine someone has listened to a piece and analyzed its structure", and in doing so labelled two short clips as A and B. These prompts only contained one voice of the musical environment; for example, in the "HR-MT" environment, only the harmony-rhythm voice might be used (see Figure 6.6). After being presented with the prompts, participants heard a four-part clip and were asked to guess how the hypothetical listener would have analyzed it: as $AABB$, $ABAB$ or $ABBA$. They then indicated their confidence on a 5-point Likert scale from "not at all certain" to "totally certain."

A short introduction to this experiment was used to establish a baseline of preference for each stimulus. Participants heard four different four-part stimuli and were asked only to indicate which analysis they preferred, $AABB$, $ABAB$ or $ABBA$. No prior analysis was given.



(a) (b)

Figure 6.6: Example prompt for Experiment no. 4. The patterns labelled "A" and "B". This would be followed by the presentation of a stimulus like that in Figure 6.4.

Variables: Each stimulus was defined by three variables: the features expressing forms $AABB$, $ABAB$ and $ABBA$. Participants heard four unique stimuli presented three times each, though the same stimulus was never heard twice in a row. The focal

feature was the feature that varied across the two prompts. The three presentations of each stimulus used a different focal feature each time, and the focal feature was always one of the three changing features.

The two response variables were whether the participant guessed the correct form or not, and the confidence of their answer.

6.3 Results

6.3.1 Experiment no. 1: Change identification

Participants were shown 12 stimuli and asked which feature changed in the middle of each one. The likelihood of guessing the correct answer by chance is 25%, since there were four features. (There was a fifth answer option, "No change," which was given 6% of the time, but something did change each time.) The variables in the experiment are summarized in Table 6-C.

How varied among participants	Feature	Codename	levels
Within	Changing feature	change_feat	*harmony, melody, rhythm, timbre*
Within	Whether stimulus was heard before	heard_before	*no, yes*
Between	Musical environment	music	*HR-MT, HT-MR, HM-RT*
Between	Stimulus sequence	stimset	1, 2, 3, 4
Uncontrolled	Musical training score	ability	15–48 points
Uncontrolled	Age	age	20–71 years
Uncontrolled	Gender	gender	female, male
Dependent	Answer value	correct	1, 0

Table 6-C: Summary of variables in Experiment no. 1

Among 87 participants, there were 886 correct answers out of 1044, an accuracy of 85.87%. A binomial test, whose null hypothesis is that the chance of success on each trial was 0.25, confirmed the obvious: success was significantly better than chance ($p < 10^{-16}$), and we may conclude that overall, listeners were capable of multi-dimensional listening, or of abstracting individual musical features from the whole. Still, there were a number of people who fared poorly. For an individual's 12 trials, 7 successes were necessary for the binomial test to show that the individual performed better than chance. Out of 87 participants, 6 did not surpass this threshold.

This gives an overall picture, but the factorial experiment design allows us to look deeper. Since the output variable, *correct*, was binary, we used binomial logistic regression to model participants' response as a function of the independent variables: *change_feat, heard_before, music, ability, age* and *gender*. We treated each individual and each stimulus sequence as a random block, since each of these were drawn from a larger population to which we wish to extrapolate the results. All the other factors were fixed effects and we looked for first- and second-order interactions between them. The model was computed in *R* (like all the models in this chapter) and the full table of fitted coefficients (like all the tables of effects in this chapter) may be seen in Appendix B. For brevity, we show only the significant effects in Table 6-D.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.67	0.96	3.84	0.00
ability	1.25	0.57	2.20	0.03
ability:musicHTMR	-1.22	0.58	-2.08	0.04
change_featTimbre:musicHMRT	-2.25	0.81	-2.80	0.01
change_featMelody:age	-0.63	0.27	-2.34	0.02
musicHMRT:heardbefore	1.29	0.58	2.21	0.03

Table 6-D: Significant effects in linear model for Experiment no. 1

The single significant main effect was the self-reported musical training of the participants ($p = 0.028$). A scatter plot of participant's average scores against their training reveals a strong positive correlation (see Figure 6.7).

The strongest interaction effect was a negative one between "Timbre" and "HM-RT"

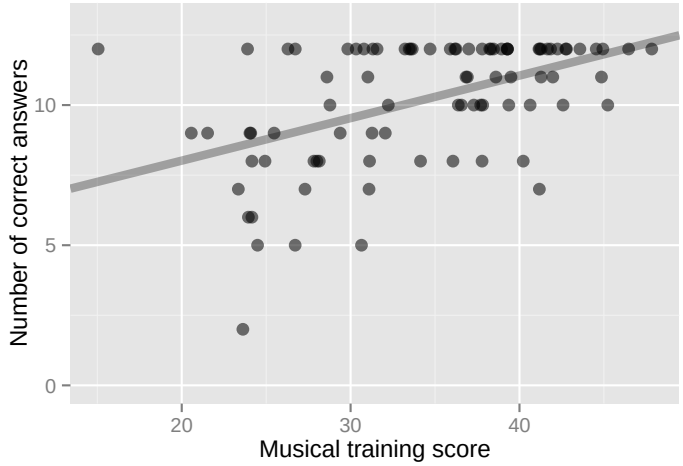


Figure 6.7: Experiment no. 1: Scatter plot showing main effect of musical training on identification accuracy, with line of best fit. Jitter has been applied to the training scores to help distinguish points.

($p = 0.005$), indicating that many participants failed to correctly identify timbre changes in this environment (see Figure 6.8).

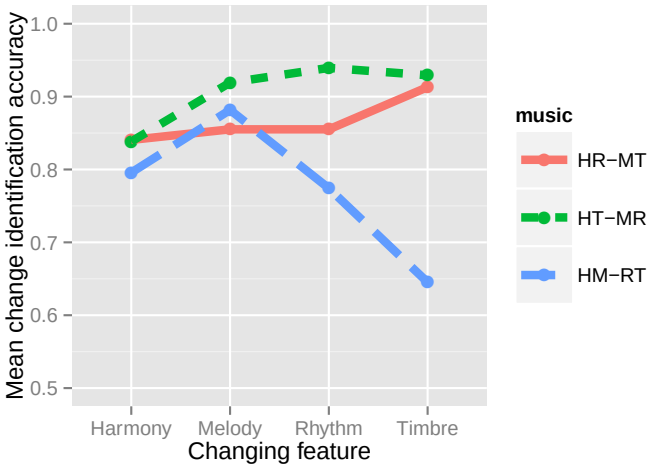


Figure 6.8: Experiment no. 1: Interaction plot of changing feature and musical environment on identification accuracy.

Whether the stimulus had been heard before or not did not have a main effect, but there was a significant interaction between it and the environment. Listeners in the "HR-MT" environment more often changed their answer for the worse (see Figure 6.9), while in the "HM-RT" environment, they often corrected their mistakes, and the difference between these changes was significant.

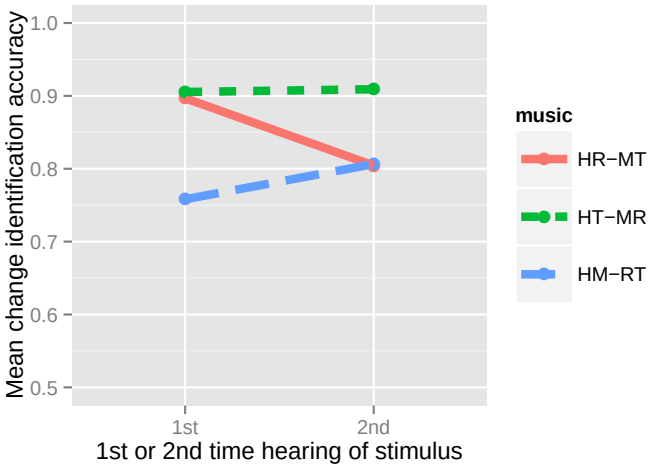


Figure 6.9: Experiment no. 1: Interaction plot of exposure (whether the stimulus had been heard once before or not) and musical environment on identification accuracy.

A negative interaction between ability and the "HT-MR" environment was also found, reflecting the fact that participants with less musical training still did well in this set of stimuli. And finally, the interaction effect between age and the melody feature suggests that younger participants identified melodic changes more accurately than other features, and older participants worse than others. However, this may be due to undersampling among older participants.

6.3.1.1 Convolved features

So far, we have only treated the changing feature as the correct answer. However, this feature was convolved with another feature by being expressed by the same voice, and providing the "convolved feature" as the answer is arguably also correct.

First, we observe that of the 158 incorrect answers, 63 of them were "No change" and the other 95 were genuine misattribution errors. Of these 95 misattributions, a large number were to the convolved feature: 61.1%, rather than the chance level of 33%. A binomial test confirmed the difference is significant ($p < 10^{-7}$).

Re-running the model analysis from the previous section, we can obtain a new set

of significant effects (Table 6-E). Most noticeably, the main effect of *ability* has been reduced to marginal significance ($p = 0.090$); evidently, many of the mistakes made by less-trained participants were convolved-feature errors.

In addition, the *change_featTimbre:musicHMRT* interaction has disappeared ($p = 0.29$), since most of the errors in the "HM-RT" environment had been confusion about what had changed in the percussion voice, which expressed either changes in rhythm or in timbre. The new interaction effects between *age* and other factors, as before, seem spurious and attributable to the undersampling of older participants, as noted in Section 6.2.1.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.03	1.06	3.79	0.00
ability:age	0.57	0.26	2.17	0.03
change_featMelody:musicHMRT	3.10	1.34	2.32	0.02
change_featMelody:age	-0.83	0.36	-2.30	0.02
change_featRhythm:age	-1.10	0.36	-3.05	0.00
musicHMRT:heardbefore	1.61	0.69	2.34	0.02

Table 6-E: Significant effects in updated linear model for Experiment no. 1, treating convolved-feature errors as correct.

Summary

Overall, participants were skilled at identifying the feature that changed. As expected, errors tended to be misattributions of the changing feature to the other feature expressed by the same voice; for example, if the rhythm of the melody changed, it was common to respond that the melody had changed. Whether such responses were even errors is debatable. If treated as errors, there was a significant effect of musical training on answer correctness; if treated as correct, then the effect of musical training was marginal.

How varied among participants	Feature	Codename	levels
Within	Changing feature	change_feat	*harmony, melody, rhythm, timbre*
Within	Match between attention and change stimulus	match	*match, convolved, wrong*
Between	Musical environment	music	*HR-MT, HT-MR, HM-RT*
Between	Stimulus sequence	stimset	1, 2, 3, 4
Uncontrolled	Musical training score	ability	15–48 points
Uncontrolled	Age	age	20–71 years
Uncontrolled	Gender	gender	*female, male*
Dependent	Salience	ans_sals	$-2, -1, 0, 1, 2$

Table 6-F: Summary of variables in Experiment no. 2

6.3.2 Experiment no. 2: Salience judgements

The variables in the experiment are summarized in Table 6-F. For each trial, participants were asked to pay attention to one feature (the focal feature), and then to rate how salient the change was. In the *match* condition, the focal feature matched the changing feature; in the *convolved* condition, the focal feature and the changing feature were expressed in the same voice, and in the *wrong* condition there was a complete mismatch.

As before, we first take a broad look at the data, and then analyze the experiment with a linear model. We first test the null hypothesis that there is no difference in salience among the match conditions; a Kruskal-Wallis test easily rejects that ($H = 362.7, df = 2, p < 10^{-15}$). We then perform a Mann-Whitney U test for each pair of conditions (with Bonferroni correction applied). The test indicates significant differences between the *match* condition and each of the other conditions ($W > 98,851, p < 10^{-15}$), as well as a much smaller but still significant difference between the *convolved* and *wrong* conditions ($W = 67,773, p = 0.0024$). Participants in the *wrong* condition found the

changes slightly less salient than in the *convolved* condition and much less than in the *match* condition.

We analyzed the experiment with a general linear mixed effects model, which characterized salience as a function of the independent variables and their secondary interactions: *change_feat, match, music, ability, age* and *gender*. As in Experiment no. 1, the individuals and stimulus sequences were treated as random blocking effects. The significant effects appear in Table 6-G.[1]

	Estimate	Std..Error	t.value	p.z
(Intercept)	-1.33	0.19	-6.89	0.00
change_featRhythm	-0.49	0.21	-2.33	0.02
matchMatch	2.43	0.20	12.24	0.00
ability:change_featMelody	-0.21	0.09	-2.22	0.03
ability:matchMatch	0.29	0.08	3.58	0.00
change_featRhythm:matchMatch	-0.77	0.22	-3.50	0.00
change_featTimbre:matchMatch	-0.60	0.22	-2.69	0.01
change_featRhythm:matchWrong	-0.44	0.22	-1.98	0.05
change_featTimbre:matchWrong	-0.52	0.22	-2.34	0.02
change_featRhythm:musicHT-MR	0.99	0.23	4.34	0.00
change_featMelody:musicHM-RT	0.74	0.22	3.41	0.00
change_featRhythm:musicHM-RT	0.48	0.22	2.21	0.03
change_featTimbre:musicHM-RT	0.43	0.22	1.96	0.05
matchMatch:musicHT-MR	-0.48	0.20	-2.42	0.02
matchMatch:musicHM-RT	-0.95	0.19	-5.06	0.00
matchWrong:musicHM-RT	-0.47	0.19	-2.48	0.01

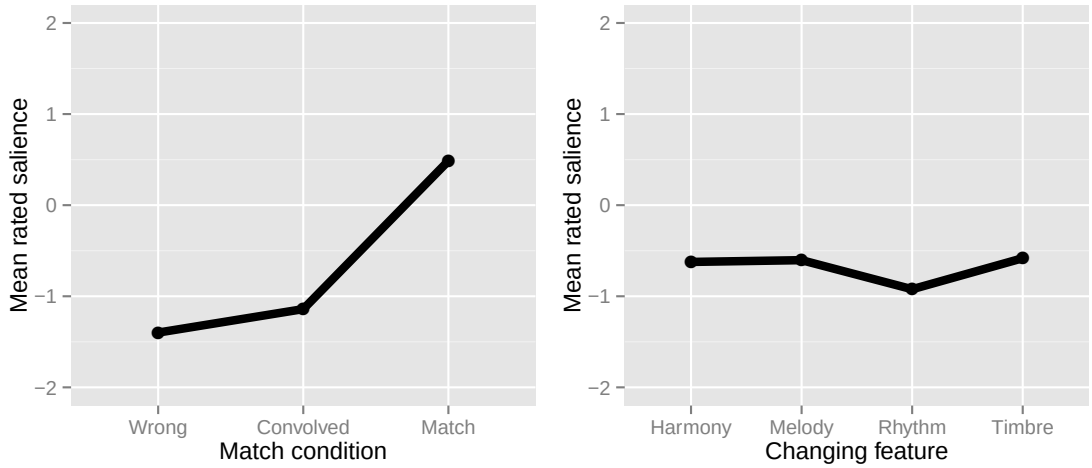
Table 6-G: Significant effects in linear model for Experiment no. 2.

There were two main effects, with variation observed among match conditions and features. In addition, there were several significant interactions between these two factors and the musical environment.

The match condition was the most important factor ($p < 10^{-15}$). The main effect plot in Figure 6.10(a) illustrates what the Mann-Whitney U tests showed earlier: matching one's focus to the change in the stimulus led to a great increase in salience, and changes in the *convolved* condition were slightly more salient than in the *wrong* condition.

[1] Significant effects were those whose 95% confidence interval did not span 0. The p-values presented were estimated from the t-statistic, since the t-distribution converges to the normal distribution when the degrees of freedom are large, and our second-order model had over 800 unused degrees of freedom.

The main effect of changing feature shows that changes in rhythm were less salient than other changes (see Figure 6.10(b)) . This may be a result of how the musical changes were expressed: while rhythmic changes always required some time to recognize (at least the time until the next onset), the other changes did not: new timbres were sounded immediately, new chord progressions always began on a new chord, and two of the three melody changes involved a change in the first note. Hence, it may be the lack of suddenness in rhythmic changes that made them less salient.



(a) (b)

Figure 6.10: (a) Main effect of match condition on rated salience of change. (b) Main effect of changing feature on rated salience of change.

Musical training interacted with both match condition ($p = 0.0003$) and the melody changes ($p = 0.026$). As seen in Figure 6.11(a), with increasing musical training came and increasing contrast in the salience of the *match* and *wrong* conditions. Meanwhile, the interaction with melody amounts to the trend that those with less musical training found melodic changes much more salient than average (see Figure 6.11(b)). Perhaps there is a tendency to focus on changes in melody, and with greater musical training comes the ability to ignore these changes.

The changing feature, the musical environment and the match condition all had several significant interactions between them, shown in Figure 6.12. In the "HM-RT" environment, harmony changes were less salient than the others, and in the "HT-MR"

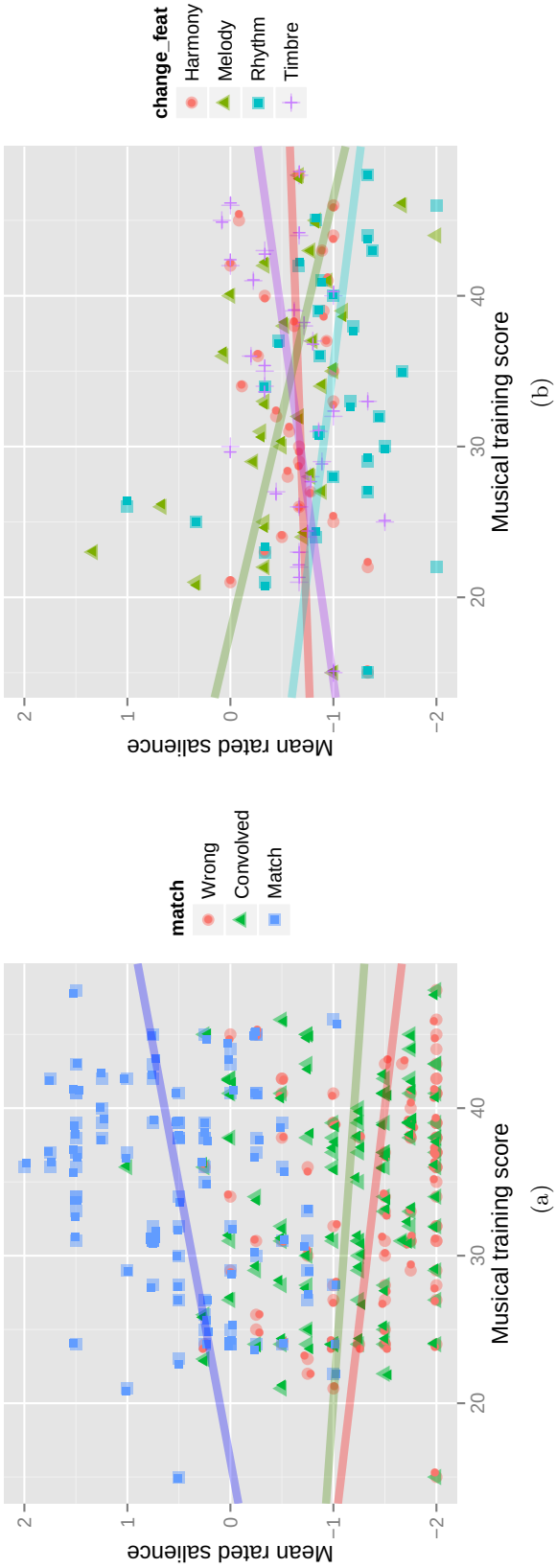


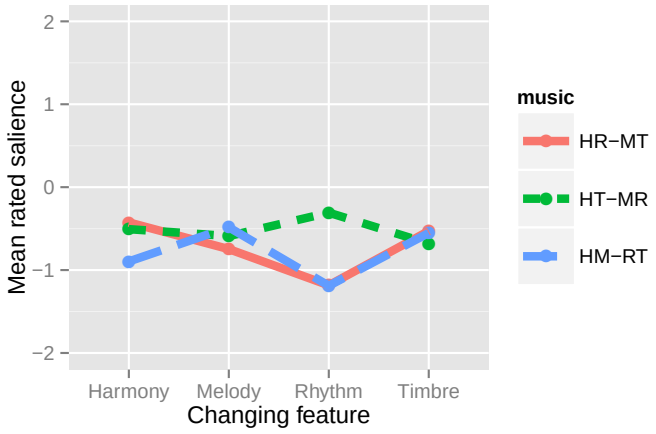
Figure 6.11: (a) Scatter plot showing the interaction effect of musical training and match condition on rated salience of change. The line of best fit for the salience as a function of musical training is plotted for each match condition.
(b) Scatter plot showing the interaction effect of musical training and changing feature on salience. The line of best fit for salience as a function of musical training is plotted for each feature.

environment, rhythm changes were more salient than usual (Figure 6.12(a)). In the latter case, it seems that changes in rhythm, usually deemed less salient than average, became more salient thanks to being associated with the melody. However, the opposite was the case for changes in harmony: they were least salient when convolved with melody in the "HM-RT" environment. This may be due to the specific nature of this convolution: although the melody and chord progression in "HM-RT" are played using the same instrument, they are still functionally independent. Therefore, the convolution in this environment may be less strong.

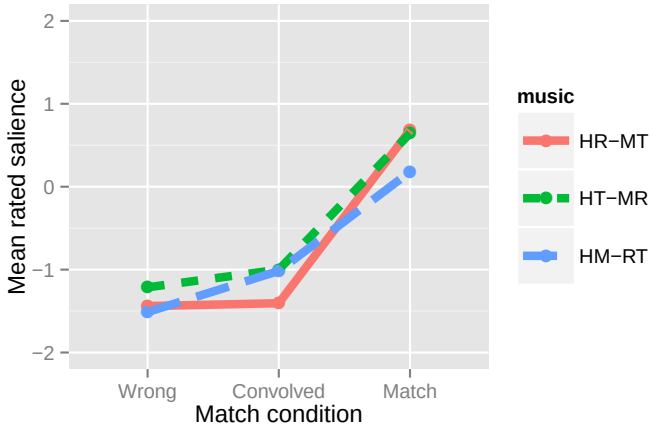
The interaction between match condition and musical environment (Figure 6.12(b)) can be summarized as: in the "HR-MT" environment, salience was especially high in the *match* condition, while the "HM-RT" environment, salience was higher than average in the *convolved* condition. Finally, in Figure 6.12(c), we can see that rhythm and timbre changes were especially salient in the *convolved* condition—that is, when they were expressed by another feature, especially the melody: rhythm and timbre changes were each strongest in environments in which they were convolved with melody (see Figure 6.12(a)).

Summary

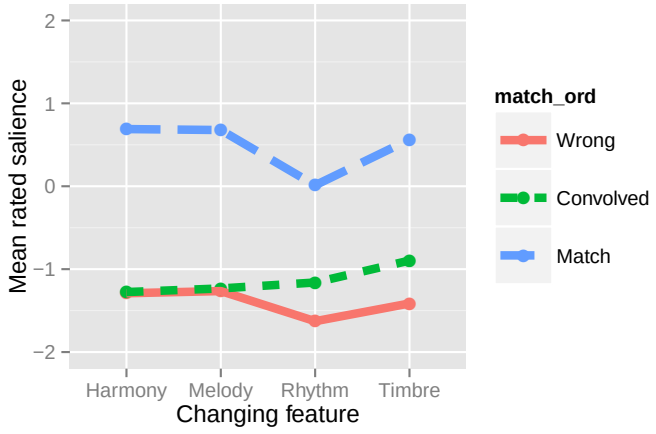
The judged salience of changes was greatest when listeners were asked to pay attention to the feature that expressed the change. There was also a slight but significant increase in salience when attention was directed not to the change but to another aspect of the voice that did change (this was the *convolved* condition). The contrast between these conditions increased with greater musical training. Overall, changes in rhythm were judged as less salient; rhythm and timbre changes were judged to be especially salient when convolved with melody. However, the salience of melodic changes lessened with greater musical training.



(a)



(b)



(c)

Figure 6.12: Interaction effect on rated salience of change for (a) changing feature and musical environment, (b) match condition and musical environment, and (c) changing feature and match condition.

How varied among participants	Feature	Codename	levels
Within	Focal feature	focalfeat	*harmony, melody, rhythm, timbre*
Within	Target presence	presence	*absent, present*
Within	Target relevance	relevance	*irrelevant, relevant*
Between	Musical environment	music	*HR-MT, HT-MR, HM-RT*
Between	Stimulus sequence	stimset	1, 2, 3, 4
Uncontrolled	Musical training score	ability	15–48 points
Uncontrolled	Age	age	20–71 years
Uncontrolled	Gender	gender	*female, male*
Dependent	Pattern detection	correct_pres	0, 1
Dependent	Form agreement	correct_form	0, 1
Dependent	Analysis confidence	ans_cons	$-2, -1, 0, 1, 2$

Table 6-H: Summary of variables in Experiment no. 3

6.3.3 Experiment no. 3: Pattern detection & grouping preference

The variables in the experiment are summarized in Table 6-H. Each of the 12 trials had two parts: first, participants were shown a target pattern of a particular feature type (the *focal feature*), and asked whether the pattern occurred in a longer excerpt (it was either *absent* or *present*). This was essentially a distractor task to get participants to focus on a particular feature. Second, participants indicated the analysis of the excerpt they preferred (AAB or ABB) and their confidence in their answer. The focal feature was either *relevant* or *irrelevant* to this grouping; in cases where it was relevant, we recorded the agreement between their answer and the form implied by the focal feature.

When participants were asked if the target pattern appeared, they had four options: "yes", "no", "I don't know", and "yes, but only a variation." The last option was intended to allow participants who detected the pattern, but were reluctant to say so

because it appeared inexactly, to give a qualified "yes." Our original intention was thus to group "yes" and "variation" responses together. However, this answer was given about as often when the pattern was present and absent (see Table 6-I). Treating "variation" as "yes," as intended, the pattern identification rate was 75.3%, which a binomial test confirms is above the chance level of 50% ($p < 10^{-15}$). If we disregard "variation" answers and only consider trials where a clear yes/no/IDK answer was given, success rises to 81.8%.

	Yes	No	IDK	Variation
Present	**406**	26	15	***73***
Absent	91	**307**	27	*95*

Table 6-I: Answers provided in the pattern-detection task. Bold answers were treated as correct. Italicized answers could arguably be discarded, but were retained in the analyses.

Among trials where the target pattern's feature was relevant to the form of the excerpt, we are interested in how many chose the form associated with the feature. Out of 522 trials, the implied form was chosen 341 times, a rate of 65.3%, significantly above the 50% chance rate ($p < 10^{-11}$).

Our next hypothesis was that confidence in one's answer would increase when the pattern was present, and also when the pattern was relevant. This was indeed the case: a Mann-Whitney U test found a difference in confidence between the relevant and irrelevant conditions ($W = 125,407, p = 0.044$) and between the present and absent conditions ($W = 123,813, p = 0.018$).

Grouping decision agreement

With the same approach as in Experiment no. 1, we used binomial logistic regression to model *correct_form*—whether the listener's grouping decision matched the focal feature or not—as a function of the independent variables. There were several significant main effects (see Table 6-J): focal feature, environment, and age.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.41	0.55	2.58	0.01
focalfeatRhythm	-2.04	0.63	-3.27	0.00
focalfeatTimbre	-1.68	0.63	-2.66	0.01
musicHT-MR	-1.70	0.70	-2.43	0.01
age	-0.78	0.32	-2.45	0.01
focalfeatMelody:musicHT-MR	2.55	1.10	2.32	0.02
focalfeatRhythm:musicHT-MR	4.28	0.82	5.24	0.00
focalfeatTimbre:musicHT-MR	2.10	0.74	2.85	0.00
focalfeatTimbre:musicHM-RT	1.63	0.72	2.25	0.02
focalfeatRhythm:age	0.63	0.32	1.96	0.05
musicHT-MR:age	0.85	0.33	2.61	0.01

Table 6-J: Significant effects in linear model for Experiment no. 3: Grouping preference

The focal feature main effects indicate that listeners were less likely to be swayed by the influence of the target when a rhythmic or timbral target was given. This trend is evident in Figure 6.13. However, although participants always agreed with target feature above 50%, it is not necessarily true that the chance level was 50% for each feature. We saw in Experiment no. 2 that melodic changes were more salient than rhythmic ones, for example, and it stands to reason that listeners are more likely to analyze a piece according to melody than rhythm anyway.

To account for this, Experiment no. 3 was preceded by a pre-test. Participants were asked to provide their preferred analysis for a few stimuli in the absence of any direction. The average rate at which participants' preferred analysis matched each feature is plotted as the baseline in Figure 6.13. Finally, this figure also plots the average rate at which listeners agreed with each feature despite the distractor task directing their attention towards a different feature.

Comparing the "direct" condition to the "baseline" condition, we see that although agreement seemed lower in rhythm, there was a similar boost in salience in rhythm compared to harmony and melody. The timbre feature is the odd one out: despite being a more salient feature than harmony and rhythm when attention was free to wander, the act of focusing on timbre was less influential. (However, binomial testing on the

difference in agreement between the "direct" and "baseline" conditions still turned up significant differences: $p = 0.006$ for timbre, $p < 10^{-7}$ for the other features.)

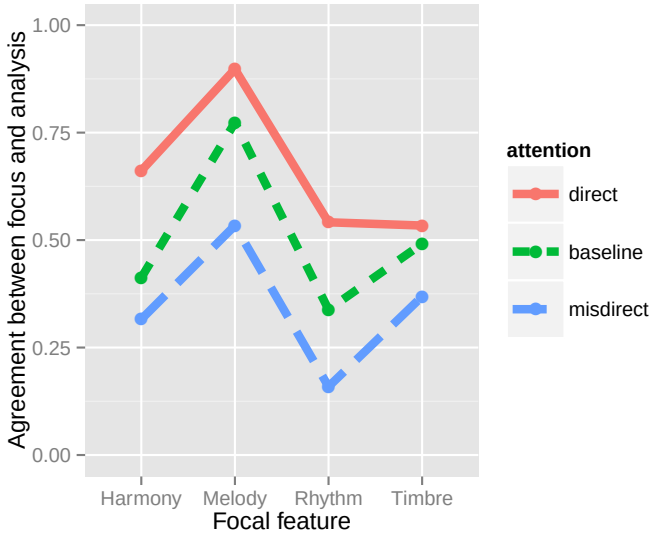
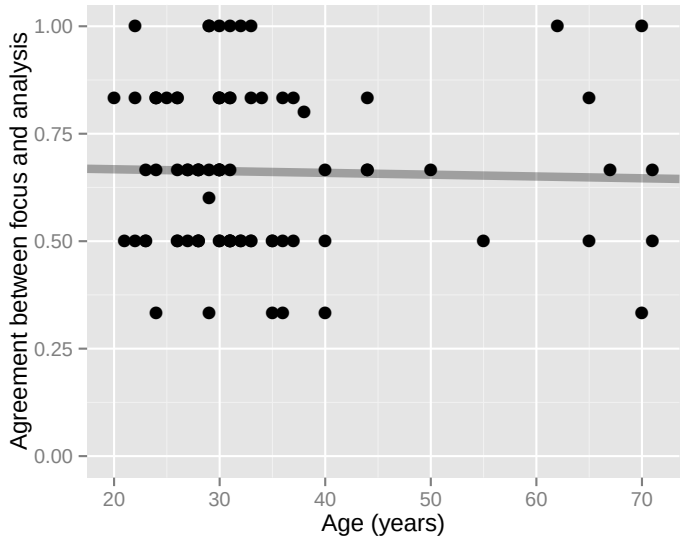


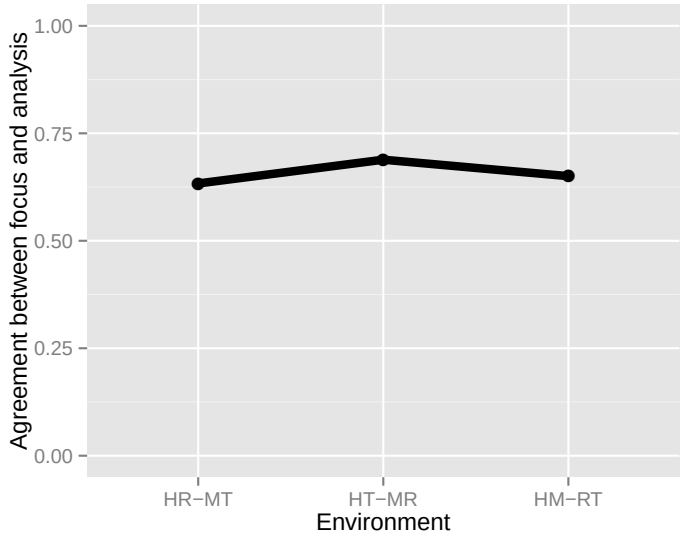
Figure 6.13: Main effect of focal feature on grouping agreement. The "direct" line gives the rate at which participants' analyses matched each feature when attention was directed towards it. The "misdirect" line gives the same rate for when participants' attention was directed away from it. The "baseline" gives the same rate when participants' attention was no directed either way.

The main effect plots for musical environment and age are unimpressive (see Figure 6.14) and suggest that these factors are not in fact significant on their own. However, environment interacted strongly with focal feature. The interaction between feature and music is plotted in Figure 6.15. (We have for each environment a separate baseline of grouping preferences.) The influence of the target task was less for harmony in the "HT-MR" environment, and the influence of timbre particularly bad in the "HR-MT" environment—in fact, this is the only case where the attention task had a negative effect on agreement.

The remaining interaction effects, between age and both rhythm and "HT-MR", suggest that older participants were more swayed by the influence of the distractor task in these cases. Both effects are very weak and perhaps a result of outliers.



(a)



(b)

Figure 6.14: (a) Scatter plot showing main effect of age on grouping agreement, with line of best fit.
(b) Main effect of musical environment on grouping agreement.

Grouping decision confidence

We next examine how the experimental factors affected the confidence with which people provided the grouping decisions analyzed above. However, we expand our view to include those trials where the target pattern was not relevant to the grouping. We computed a general linear mixed effects model to characterize *confidence* as a function of the

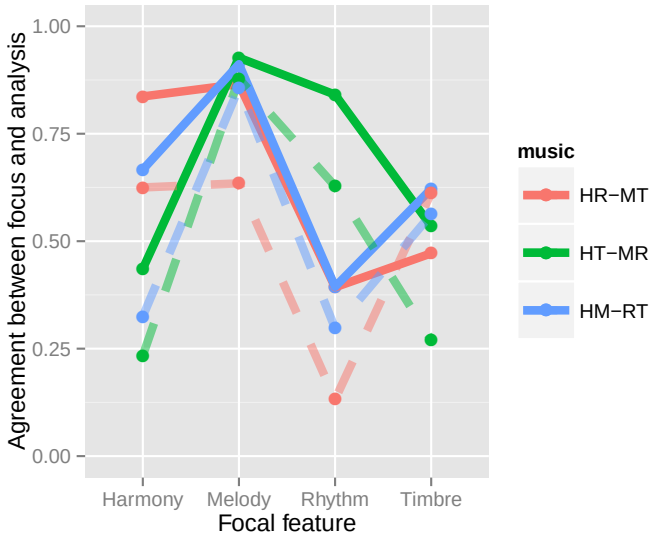


Figure 6.15: Interaction between feature and musical environment on grouping agreement. The baseline for each environment is plotted as a fainter dotted line (compare with the average baseline in Figure 6.13).

independent variables. However, the second-order model turned out to be a very poor fit to the data, discovering only one significant main effect (of musical environment). So, we present instead a much simpler first-order model. The significant factors are presented in Table 6-K, and include several main effects—although, curiously, not musical environment.

	Estimate	Std..Error	t.value	p.z
(Intercept)	0.54	0.18	3.03	0.00
presencePresent	0.12	0.05	2.64	0.01
focalfeatMelody	0.18	0.06	2.89	0.00
ability	0.17	0.08	2.08	0.04
age	-0.23	0.08	-2.87	0.00
relevanceRelevant	0.13	0.04	2.98	0.00

Table 6-K: Significant effects in linear model for Experiment no. 3: Answer confidence (all trials)

We had hypothesized that when the pattern was present, compared to absent, and when the pattern was relevant, compared to irrelevant, that listeners would have greater confidence in their grouping decision. The main effect plots in Figure 6.16 show this was indeed the case, with both factors having about an equal, if slight effect. An interaction

plot in the same figure demonstrates their additive effect.

The other main effect plots (Figure 6.17) show that confidence correlated positively with musical training, and negatively with age, and that participants were more confident than usual when the target pattern was a melody.

Pattern detection performance

Although it is of secondary interest, we also constructed a linear model using logistic regression to analyze the effect of each factor on whether the participant correctly determined the presence of the target pattern. The significant effects are reported in Table 6-L (see Table 2-E for all factor estimates).

	Estimate	Std. Error	z value	Pr(>|z|)
ability	0.81	0.30	2.69	0.01
focalfeatRhythm	1.63	0.58	2.83	0.00
musicHM-RT	2.09	0.62	3.36	0.00
presencePresent	2.85	0.54	5.24	0.00
relevanceRelevant	1.07	0.45	2.34	0.02
focalfeatRhythm:musicHM-RT	-2.52	0.69	-3.67	0.00
focalfeatMelody:genderf	1.58	0.54	2.91	0.00
focalfeatRhythm:genderf	1.50	0.56	2.65	0.01
musicHM-RT:age	-0.64	0.29	-2.22	0.03
musicHM-RT:relevanceRelevant	-1.12	0.50	-2.25	0.02
presencePresent:relevanceRelevant	-1.26	0.46	-2.77	0.01

Table 6-L: Significant effects in linear model for Experiment no. 3: Pattern recognition.

There are several main effects, plotted in Figures 6.18 and 6.19. As in Experiment no. 1, success at the task increased with greater musical training. Participants were also best at identifying the melodic and rhythmic patterns, and worst at identifying the chord patterns. The chord patterns, admittedly, were often difficult to discern, since the chord qualities were very simple (only major or minor). Pattern recognition was best in the "HM-RT" musical environment.

Whether the pattern was present or not evidently had a large impact (Figure 6.19(b));

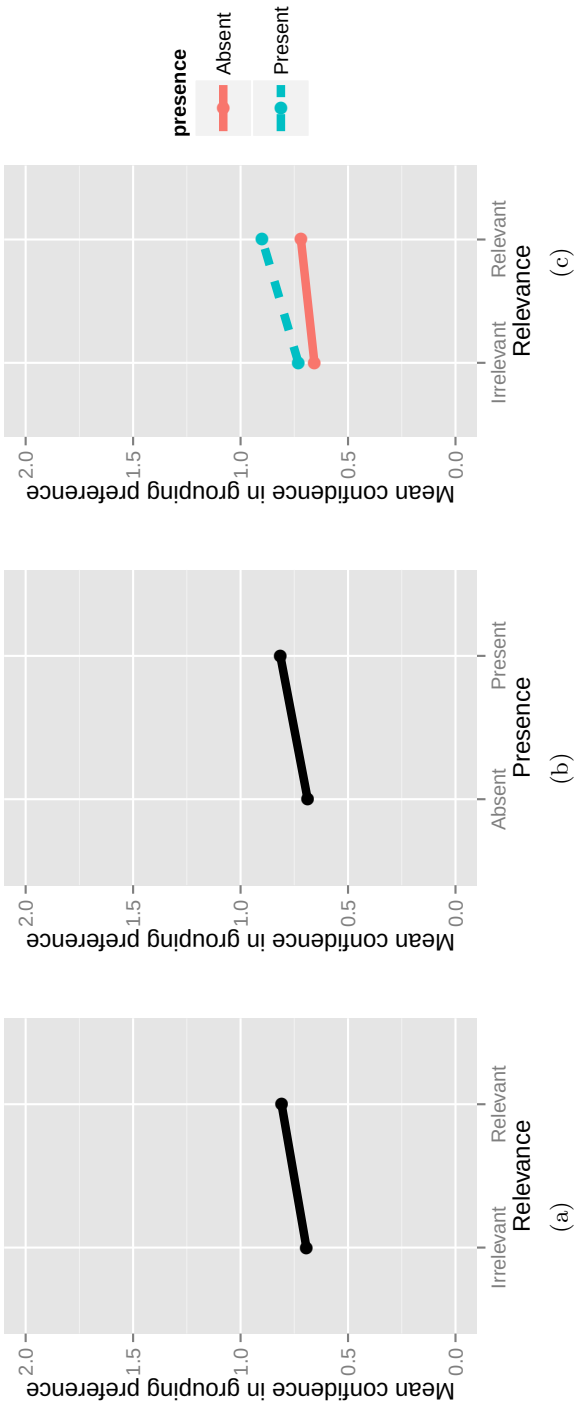


Figure 6.16: (a) Main effect of relevance on grouping confidence.
(b) Main effect of presence on grouping confidence.
(c) Interaction plot of relevance and presence on grouping confidence, demonstrating their additive effect.

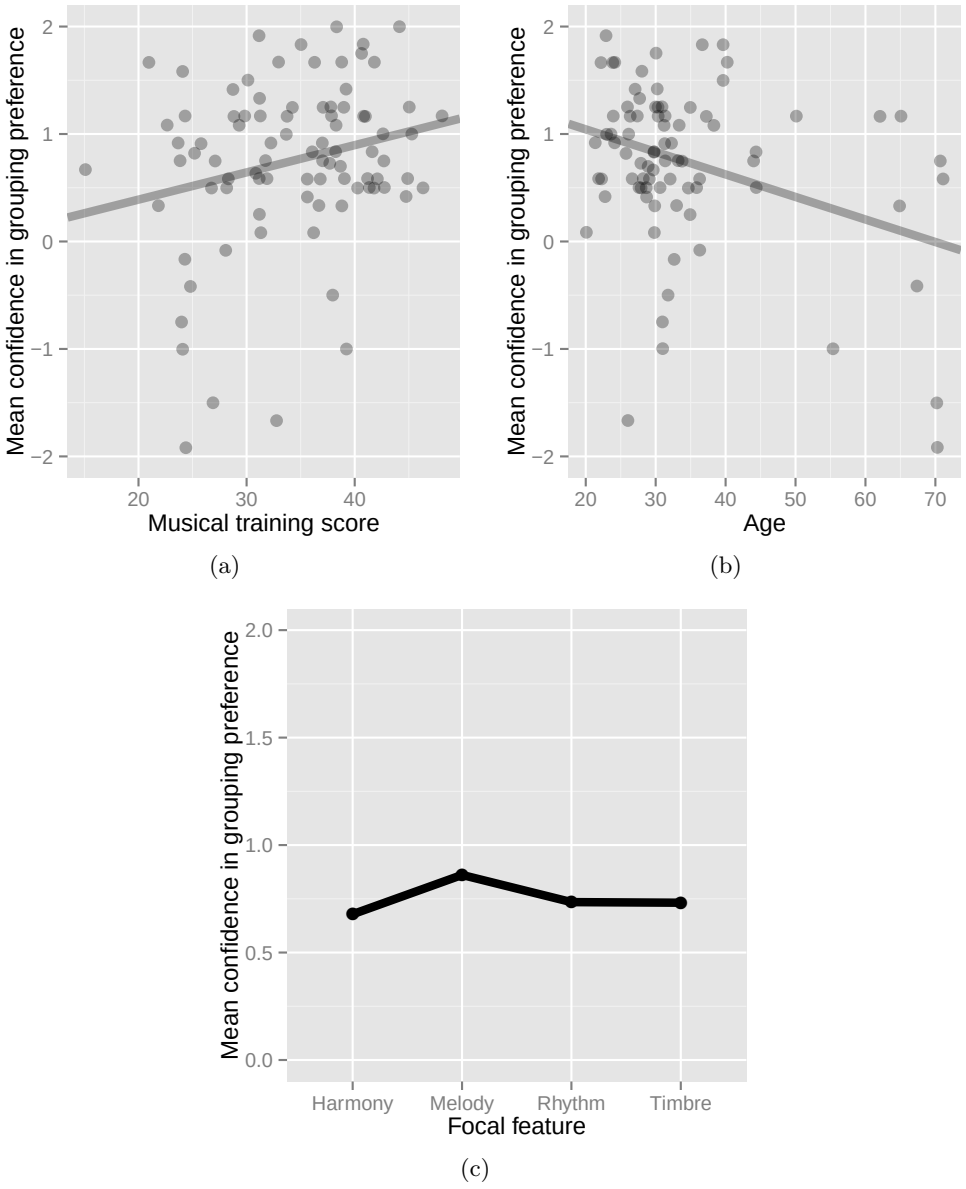
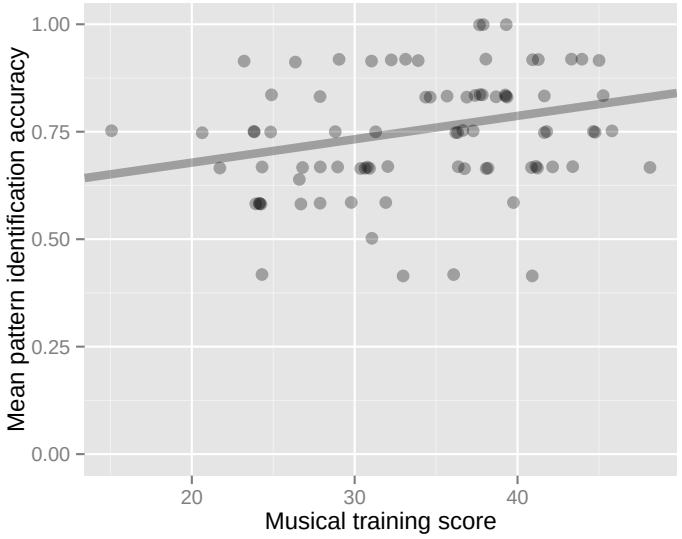


Figure 6.17: (a) Scatter plot showing main effect of musical training on grouping confidence.
(b) Scatter plot showing main effect of age on grouping confidence.
(c) Main effect of focal feature on grouping confidence.

the plot suggests that participants were more likely to make Type I errors, claiming the pattern occurred when it was in fact absent, than to say it did not occur when it was present (this can be confirmed in Table 6-I).



(a)

Figure 6.18: Scatter plot showing main effect of musical training on pattern detection accuracy.

The final main effect plot (Figure 6.19(d)) is rather unimpressive: there is a very small and yet significant increase in accuracy when the pattern was relevant—that is, when the pattern was present for only part of the time, but changed partway through. However, turning to the interaction plots in Figure 6.20, we see an interaction between relevance and presence (Figure 6.20(a)). When the target pattern was present, relevance decreased accuracy; when the pattern was absent, relevance increased accuracy. In the *absent* condition, the change from irrelevance to relevance meant that listeners, who were paying attention to the focal feature, heard that feature change; this change may have sharpened their focus and allowed them to realize that the pattern did not occur. In the *present* condition, relevance meant that the target pattern occurred for less of the stimulus; this may have made listeners less certain it had occurred.

Relevance also interacted with the musical environment. While there was, overall, a clear difference between the identifiability of patterns in the three environments, with "HM-RT" patterns easier to spot, relevant patterns in the "HR-MT" environment were especially well-identified (Figure 6.20(b)). The reason seems to lie with the harmonic-rhythmic pattern. The chord pattern was in fact a single chord, either major or minor,

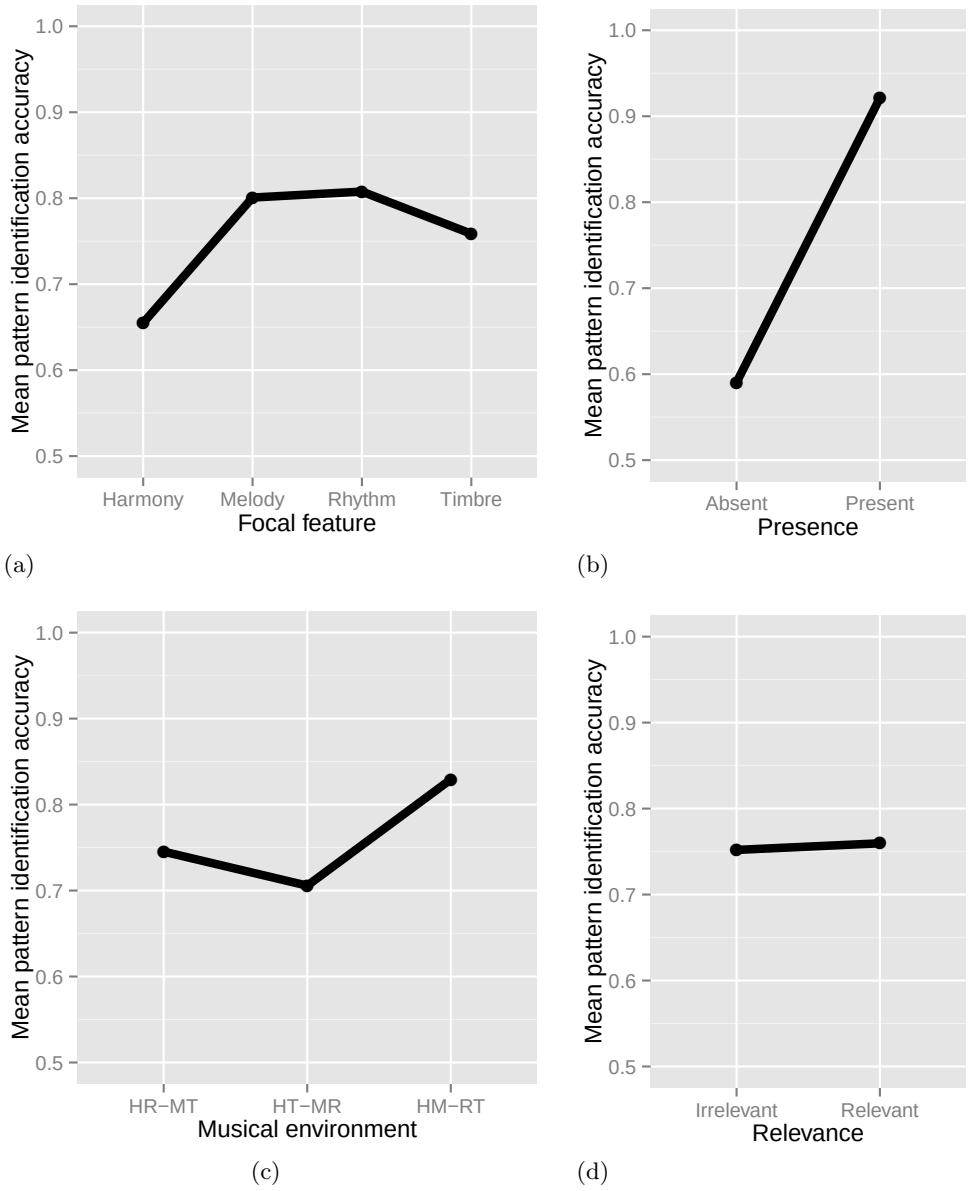


Figure 6.19: Main effect on pattern recognition accuracy of (a) focal feature, (b) presence of pattern, (c) musical environment, and (d) relevance of pattern.

repeated. A target chord quality may have been easier to discern when the chords in the excerpt changed, explaining the improved performance in the *relevant* condition.

An interaction effect between focal feature and musical environment, seen in Figure 6.21(a), reveals that the patterns composed for some environments were harder to detect than others. For example, although detection was poorest for harmonic patterns overall

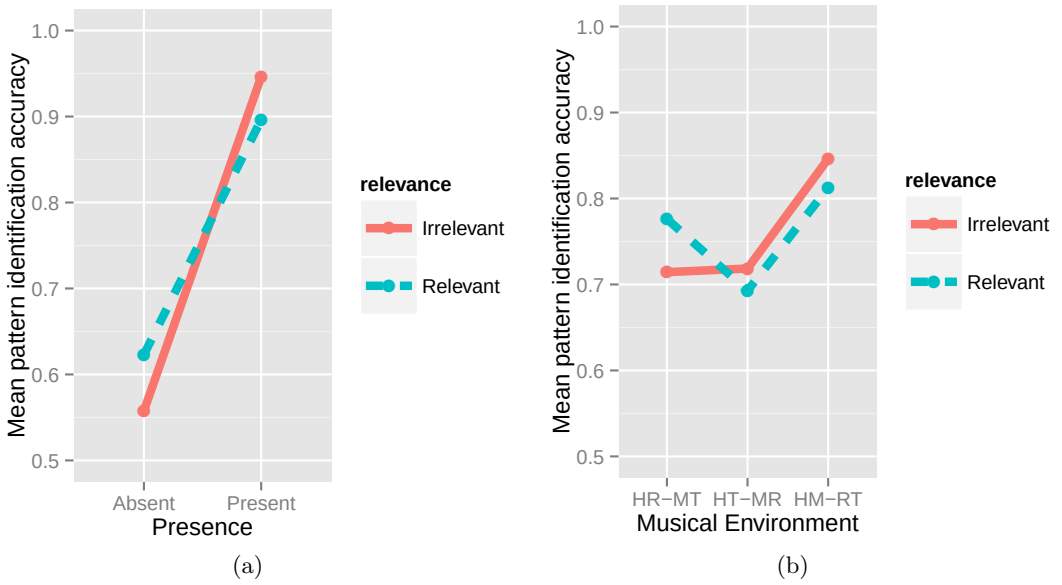


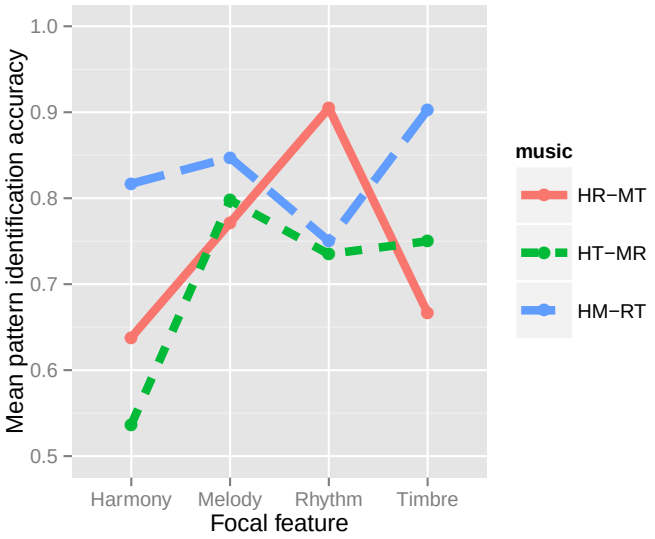
Figure 6.20: Interaction effect on pattern recognition accuracy (a) between presence and relevance, and (b) between musical environment and relevance.

(see Figure 6.19(a)), the harmonic patterns in the "HM-RT" environment were detected much more successfully.

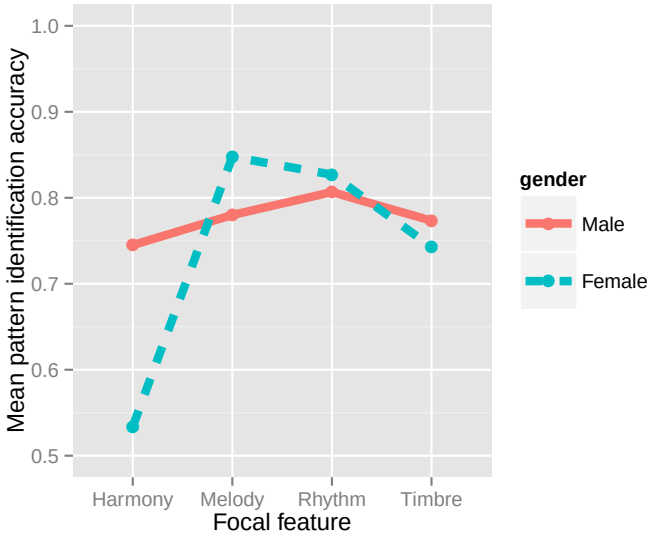
The final interaction is between gender and feature (Figure 6.21(b)). Men's detection accuracy was similar across each feature, while women's accuracy varied substantially, with poorer performance in recognizing chord progressions and better performance recognizing melodic progressions. As was the case with age, undersampling may have played a role here, since there were many more men than women in the study (50 vs. 35); the greater variance in detection accuracy observed here among women is consistent with this explanation.

Summary

Listeners were asked to pay attention to a target pattern while listening to an ambiguous musical excerpt, which had form AAB or ABB. When the target pattern was expressed by a feature that also implied one of these two forms, listeners were more likely to prefer that



(a)



(b)

Figure 6.21: Interaction effect on pattern recognition accuracy (a) between focal feature and musical environment, and (b) between focal feature and gender.

form, compared to both the chance level of 50% and compared to the baseline likelihood of choosing each analysis in the control condition where attention was not manipulated. This effect did not depend on musical training. When the target pattern was a timbre, the influence of the target was very weak, and in the "HR-MT" environment, had the opposite impact than every other feature. There was a strong baseline tendency to prefer

an analysis according to melody, although tracking a target melody still increased this preference significantly.

Confidence in one's analysis was greater both when the target pattern was present and when it implied an analysis (i.e., when the focal feature varied). Participants with more musical training and younger participants claimed greater confidence in their grouping decisions.

Finally, we note that musical training also improved participants' accuracy in detecting the pattern. Participants were more prone to Type I errors than Type II (when erring, they were more likely to claim incorrectly that the pattern was present). Accuracy in detecting the pattern also varied across the musical environments and features, with complex interactions between them.

6.3.4 Experiment no. 4: Analysis continuation

In each trial, participants heard two excerpts and were told that another listener had labelled them as A and B; the feature that defined the difference between these was the *focal feature*. Participants were asked to give the analysis they thought that listener would have given to a longer excerpt. The main independent variables of interest are the focal feature and the *static feature*, the sole feature that did not vary over the longer excerpt. In a given trial, the other two features were the *distractor features*. The variables in the experiment are summarized in Table 6-M.

Out of 1026 trials, the correct analysis was chosen 869 times, an accuracy of 84.7%, above the chance level of 33.3% (binomial test: $p < 10^{-15}$). This is a similar achievement on the part of participants as in Experiment no. 1, where changes were identified with 85.9% accuracy, although that was compared to a chance level of 25%. For an individual to answer beyond the chance level, 8 out of 12 successes were required, and 8 participants scored lower than this. It seems that even in this highly abstract task, which did not explain to participants the nature of the musical manipulations to expect, participants

How varied among participants	Feature	Codename	levels
Within	Focal feature	focalfeat	*harmony, melody, rhythm, timbre*
Within	Static feature	static_feat	*harmony, melody, rhythm, timbre*
Between	Musical environment	music	*HR-MT, HT-MR, HM-RT*
Between	Stimulus sequence	stimset	1, 2, 3, 4
Uncontrolled	Musical training score	ability	15–48 points
Uncontrolled	Age	age	20–71 years
Uncontrolled	Gender	gender	female, male
Dependent	Answer value	correct	1, 0

Table 6-M: Summary of variables in Experiment no. 4

were overall highly qualified.

We once again used binomial logistic regression to model participants' accuracy as a function of the independent variables. Due to the poor fit of the second-order model, we present here the simpler model with only main effects; the significant ones are listed in Table 6-N.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.86	0.44	4.26	0.00
ability	0.30	0.12	2.42	0.02
focalfeatMelody	2.23	0.37	5.95	0.00
focalfeatTimbre	1.16	0.29	4.07	0.00
music2	-1.01	0.36	-2.77	0.01
music3	-1.51	0.37	-4.04	0.00
age	-0.67	0.12	-5.68	0.00
static_featTimbre	0.82	0.31	2.66	0.01

Table 6-N: Significant effects in linear model for Experiment no. 4: Grouping preference

The main effects are shown in Figures 6.22 and 6.23. Again, repeating the findings from Experiment no. 1, we find that answer accuracy improves with musical training

($p = 0.015$). Probably due to the influence of outliers, we also observed a decrease in performance with age ($p < 10^{-7}$). Answer accuracy was best in the "HR-MT" environment ($p = 0.005$), and when the timbre was constant throughout ($p = 0.008$). Paradoxically, accuracy was also better when timbre was the focal feature, along with melody. That is, when A and B were defined by different timbres or melodies, the analysis was more accurately continued. This suggests that timbre was especially effective as a distractor: when it was neither the focus nor static, accuracy decreased.

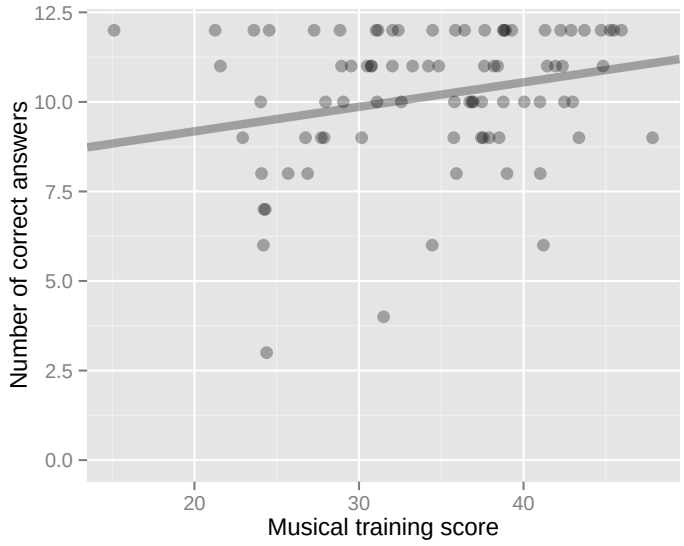
The final effect plot (Figure 6.23(c)) shows how success in continuing the analysis varied across different focal features. Included in the plot is the baseline preference for analyzing the excerpts according to different features. As in Experiment no. 3, these were established in a pre-test in which similar excerpts were analyzed but with no guidance given to the listener. The shape of the baseline is similar to that in Figure 6.13, but with an even more pronounced preference to use melody as the defining grouping principle.

Summary

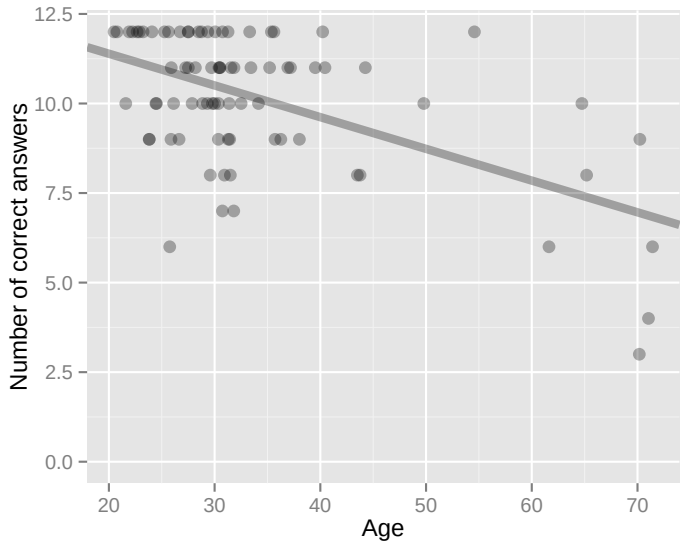
Participants were able to continue the analysis begun by a hypothetical listener with high accuracy. Doing so required discerning which feature changed between two short stimuli, and then tracking this feature in a longer excerpt. Performance in this task increased with musical training, decreased with age, and varied across the musical environments and feature combinations. As in Experiment no. 2, we found that listeners were predisposed to analyze excerpts according to melody, but were still able to complete the task, focusing on the correct feature.

6.4 Discussion

Each of the experiments supported the hypothesis it was designed to test. We first confirmed that listening among our participants was multi-dimensional; second, that bound-



(a)



(b)

Figure 6.22: Scatter plot showing main effect on analysis continuation accuracy of (a) musical training and (b) age.

ary salience increased when listeners focused on the feature that defined the boundary; third, that listeners preferred the grouping that matched the feature they paid attention to; and fourth, that participants were able to interpret the structure of longer pieces based on a prior decision about grouping.

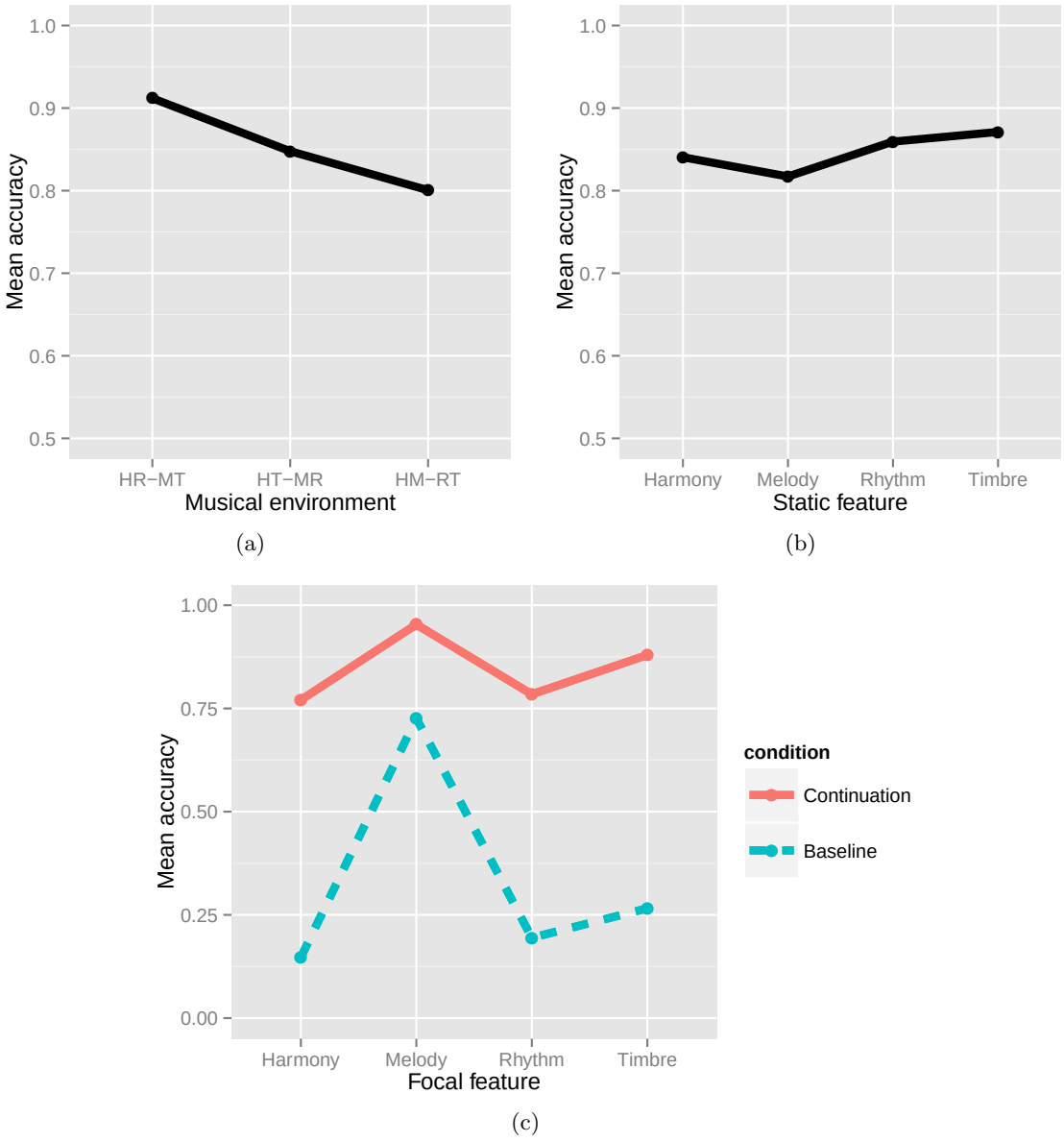


Figure 6.23: Main effect on analysis continuation accuracy of (a) musical environment, (b) static feature and (c) focal feature. The extra "baseline" in plot (c) indicates the analysis preferences established in the pre-test.

Methodology

However, the outcome of Experiment no. 2 (Salience judgements) may be disputed. Unlike in Experiment nos. 3 and 4, we did not establish a baseline for Experiment no. 2: that is, we did not collect salience judgements from listeners who were not told to pay

attention to any feature at all. As a result, there may be skepticism about the contrast between the *match* and *wrong* conditions. Participants were told to pay attention to a given feature, then asked: "How strong is the change at the midpoint of the excerpt?" This instruction was arguably ambiguous: was it interpreted (as intended) to mean, "How strong is the change *in the music globally* at the midpoint", or instead as, "How strong is the change *in that feature* at the midpoint"? If the latter was the interpretation, then the observed difference in salience is uninteresting.

Although the interpretation of the participants cannot be confirmed now, one observation may convince us that the question was interpreted as intended: the fact that the "least salient" and "most salient" ends of the scale were used in both the *match* and *wrong* conditions. That is, there were many occasions where a listener rated a change as "Not strong at all" even though that change was exactly in the feature being focused to; conversely, sometimes a change was rated as "Extremely strong" even though that change was completely unrelated to the musical attribute being focused on. If a participant were focusing on timbre, and the timbre did not change but the melody did, and the change was rated as "extremely strong," then it seems likely that this participant was assessing the global salience of the changes. This behaviour was not uncommon; in fact, 30 out of 87 participants gave one of these answers at some point. This behaviour suggests participants really were rating the salience of the change globally (a rating that was influenced by attention), and not the salience only of the feature being focused on.

Still, we acknowledge that the overt instruction of Experiment no. 2—"Please pay attention to *feature X* while listening"—is awkward. Attention can be difficult to manage consciously, as anyone told not to imagine an elephant can attest. This is why attention was manipulated covertly in Experiment no. 3. This is also why Experiment no. 3 preceded no. 2: we did not want to cue participants to the fact that their attention was being manipulated too soon. Still, one may wonder whether participants, after being told to look for different kinds of patterns in the music and then told to analyze the same passages, began to sense from the experiment that there was a "correct" way to

interpret each passage based on the target task that accompanied it. To settle this doubt, we may look at whether the impact of the focal feature was less in the first few trials of the experiment than in the last. It turns out that exactly the opposite is the case! Whereas the likelihood of agreeing with the focal feature was reported as 65.33% overall, the rate of agreement among the first four trials was 80.60%, and in the first trial, 90.63%. Although the number of samples per trial number is low, there does not seem to be a basis to believe that it was participants' awareness of the purpose of the experiment that led to the positive results.

Feature differences

One trend that was apparent throughout the experiments was that melody was the most salient feature to listeners, while rhythm was among the least salient. In the pretests for Experiment nos. 3 and 4, listeners overwhelmingly preferred to group melodic patterns rather than other patterns, while rhythm ranked near the bottom (see Figures 6.13 and 6.23(c)). Rhythmic changes were the least salient ones in Experiment no. 2 (Figure 6.10(b)), and received a significant boost in salience when they were expressed by the melody (Figure 6.12(a)). Melodic changes were among the easier to identify in Experiment no. 1 (Figure 6.8).

Next to melody and rhythm, harmony and timbre mostly occupied middle ground. However, timbre did stand out in Experiment no. 3 as the feature which had the least influence via attention on grouping preference. In the pretest, we found that participants were unlikely to prefer a harmony- or rhythm-based grouping (see Figure 6.13). With their attention focused by the pattern-detection task, the preference for these groupings increased substantially. However, only a very small rise was observed when participants had to detect a timbre. The weakness of the influence of timbre may be due to how the target task was framed. For the other features, the target was a pattern: a melodic, harmonic, or rhythmic sequence that spanned the length of a measure. For timbre, the target was just a sound: a single note or chord. Alternatively, the difference may be

due to timbre's inherent salience; as Margulis wrote, "timbres and chords aren't 'catchy' in the way a tune can be," since it is the repetition of something that makes it catchy [Mar14] (66).

In light of this, we may want to attribute the influence of the pattern-detection task not to attention *per se*, but to the crystallization in the listener's mind of that pattern as a unit. When the unit was recognized in the excerpt, the rest of the excerpt was understood in relation to that unit, resulting in a preference for AAB or ABB depending on the occurrence of the unit. If this were true, then we should expect that when listeners paid attention to a feature that was relevant, but to a pattern that was absent, the effect would be diminished. In fact, *presence* was not found to be a significant factor in the model, and average agreement with the focal feature was roughly the same in both conditions (*present*: 65.90%; *absent*: 64.75%).

Participant differences

Finally, we note that in every task, the level of musical training of the participants affected the outcome. Those with greater musical training were able to identify changing patterns with better accuracy (Figure 6.7), recognize patterns with better accuracy (Figure 6.18), and provide the continuation of an analysis with better accuracy (Figure 6.22(a)). In Experiment no. 2, the contrast in salience between when the correct feature was attended to or not increased with musical training, perhaps reflecting an increased ability to focus (Figure 6.11(a)).

However, in Experiment no. 3, although confidence in one's grouping decision increased with musical training (Figure 6.17(a)), there was no effect of training on the influence of the target task. In fact, as can be seen in Figure 6.24, the influence of the target task appears to be totally independent of training. This suggests that, while musical training affects the salience of individual musical features, the role of attention in directing one's interpretation of a piece does not depend on training. This bolsters the view that atten-

tion is a fundamental mechanism that guides how listeners interpret grouping structure in music. While a listener can learn to pay more or less attention to different features, this training does not seem to affect *how* attention affects their perception of structure.

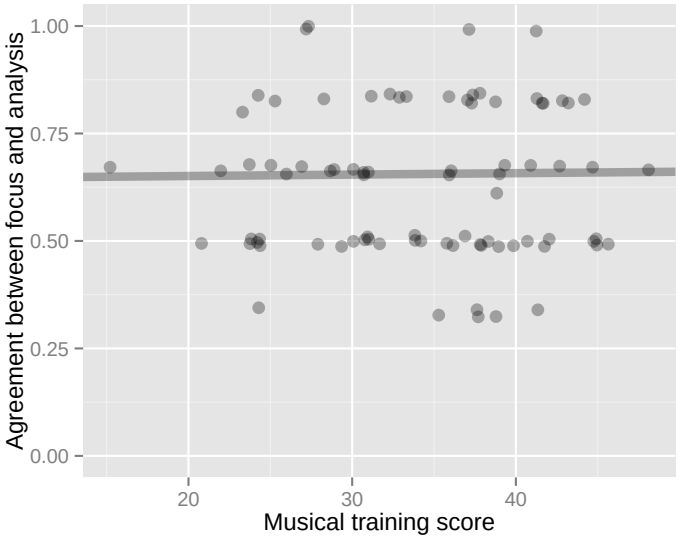


Figure 6.24: Scatter plot showing the main effect of musical training on grouping agreement.

6.5 Conclusion and future work

The results of these experiments support the view that attention plays a key role in determining how listeners arrive at the grouping decisions they do. The sequence of experiments offers a possible model for how grouping decisions are made: attention is directed toward a given feature, discontinuities in that feature become more salient to the listener, and this leads the listener to prefer groupings according to that feature. This description does not account for what determines the initial focus of the listener, but the fourth experiment suggests that whatever this initial focus is, it is possible for listeners to continue in the same vein.

This view also suggests explanations for how listener disagreements arise. When confronted with an ambiguous passage, two listeners may choose to focus on different aspects of the music, leading them to two different interpretations. Or, two listeners may

have different *a priori* habits of attention, such as a tendency to focus on specific musical features, perhaps finding metrical changes or nuances in melody especially salient.

As discussed in Chapter 2, when listening to a piece of music, one's attention is partly controlled by oneself and partly guided by the music. While the preceding experiments should convince us that attention is an important factor in determining perception, it is still difficult to judge, for a given piece of music, the balance between the top-down influence of attention and the bottom-up influence of the many discontinuities and associations perceivable in the music.

Chapter 7

Conclusion

The main question asked by this thesis is: why is it that two listeners will often disagree about the structure of a piece of music? The past four chapters have built up a case for the importance of a listener's attention in determining what changes they find most salient and what groupings they prefer. In this conclusion, we recount how each of the chapters has contributed to this case. We also remind the reader of the shortcomings of each chapter, and what could be done to make our case more strongly. Lastly, we suggest some directions for future research that would build on the work accomplished here.

7.1 Summary

In Chapter 2, we described how most models of the perception of grouping structure have worked in a ground-up fashion, predicting groupings based on Gestalt principles or statistical models of musical patterns—that is, based on musical content. We also discussed how top-down influences and endogenous factors (those arising from within the listener rather than the music) affect perception. We argued that understanding these endogenous factors will be crucial to understanding why similar listeners can disagree.

In Chapter 3, we examined the justifications given by two listeners (myself and Isaac Schankler) and traced the disagreements back to their original causes. We concluded that the attention of the listener was the best proximate explanation for the disagreements. On the basis of our study, we conjectured that one's initial understanding of a piece, one's prior information about a piece and one's expectations were key factors in explaining how one's attention was focused, and thus in accounting for listener disagreements.

In Chapter 4, we determined that acoustic novelty was a necessary but not sufficient condition for a point to be perceived as a boundary. That is, nearly all boundaries can be explained by pointing to a significant change, but the fact that many other such changes are not perceived as boundaries implies that top-down processes help to guide boundary perception. Nevertheless, the degree of novelty was a good predictor of whether a point would be considered a boundary.

In Chapter 5, we presented an algorithm that connects a listener's analysis of a piece to similarities in audio features, automatically producing a plausible "explanation" of the piece. We used the algorithm to demonstrate with some examples how a disagreement between two listeners could result from their having paid attention to different features.

Finally, in Chapter 6, we tested the hypothesis that one's perception of grouping in music depends on what feature in the music one is focusing on. This finding was significant because it supported the view that differences in attention *cause* listener disagreements, and do not merely accompany them. The experiments also demonstrated that musical changes are more salient when one is focusing on the musical parameter that changes.

Together, these findings support the view that what a listener chooses or is influenced to focus on is a crucial top-down factor affecting how the listener perceives boundaries and groupings in the music. Differences in attention were the best explanation for the listener disagreements in Chapter 3; they may be the factor that best explains the differences noted among the annotations studied in Chapters 4 and 5; and they were

certainly the best explanation for the outcome of the experiments in Chapter 6.

7.2 Limitations

On the other hand, the limitations of the present studies leave room for skepticism. The small size of the case study in Chapter 3 means that its findings should be treated as a set of new hypotheses to test, rather than firm conclusions. (At least one of these hypotheses, of course, was supported by the work in Chapter 6.) In addition, the later chapters all required us to choose a limited set of features to study. The differences among the features were carefully noted in each case (in Chapter 4, we commented on the significant variation in f-measure contrast among features, and in Chapter 6, the combination of feature and musical environment was a significant source of noise), but it remains an open question whether the subset of features we have studied are representative of all musical attributes, or merely outliers.

The results of Experiment no. 3 seem unequivocal at first: listeners were more likely to prefer the grouping (AAB or ABB) implied by the feature they were paying attention to, which implies that attention guides analysis. But how generalizable is this result? The participants were in a highly contrived listening situation, so it is essential to replicate the finding with even more ecological stimuli. Another part of the experiment's design that deserves to be varied is how the attention of the listeners was controlled. As noted in the chapter, the lack of an "ear-tracking" device means we must control attention as an independent variable. We did so in two ways: directly asking listeners to focus on a particular feature, and asking listeners to detect a target pattern. Another possible approach is to prime participants before the trial by having them do a particular attention-narrowing task, like ranking a set of chords from least dissonant to most dissonant, to focus their attention on chords. Or, participants' attention could be guided by having them detect not measure-length patterns, but isolated fragments (for example, asking whether a particular word occurs to direct attention to whichever part

is sung). This variation may be of particular interest since in our experiment the timbre probe, which was a fragment and not a measure-long excerpt, had the least influence on listeners.

The main unanswered question of Chapter 5 was whether the algorithm actually did produce reconstructions of SSMs matching what a listener was focusing on. We could only surmise that it seemed to, since the SALAMI annotations we studied did not have any information about the listener's focus. However, the stimuli we created for Chapter 6, along with the data about what listeners were guided to pay attention to, is exactly the material we would need to construct a validation experiment of the reconstruction algorithm. This study is planned for immediate future work.

If that study validates the algorithm from Chapter 5 (or some more advanced version of the algorithm—several ideas for improvements were mentioned at the end of that chapter), then we could treat it as a tool to address several questions. Among them: based on their annotations, are there differences among listeners in what they attend to? Are people more likely to focus on some features and at some timescales in certain genres? In short, we could conduct a data-mining study similar to that in Chapter 4, but using the validated tool to probe a deeper interpretation of the annotations being studied.

7.3 Future work

Taking a broader view of the work, our conclusion—that attention can influence how structure is perceived—begs another question: what generates and controls attentional orientations to begin with? In Chapter 6, we directed participants' attention in an experimental setting. In natural listening, it is unknown how attention shifts and drifts, although we mentioned some ideas in Chapter 6: Margulis explained that repetition enables one to attend to either shorter or longer timescales, through the processes of ritualization and routinization [Mar14]. Depending on which features characterize the

music at which timescale, these processes could direct one's focus toward or away from particular features. On the other hand, it seems that novelty (in a way, the opposite of repetition) plays a role in directing attention—after all, "novel" is a synonym for "attention-getting." We found in Chapter 4 a positive but imperfect relationship between novelty and boundary placement, and novel events may alert one's attention to particular features; on the other hand, the results from Chapter 6 suggest that once one's attention has been directed toward one feature, it makes changes in other features less salient. All of these conflicting points of view show that without further study, it is unclear how to interpret the role of attention in natural listening settings. Research in this area is also needed to investigate another possibility: that although attention can influence perception in constrained experimental situations, it does not actually rank as an important factor in natural settings.

In any case, the fact that attention, a factor that listeners can wilfully control, may have an impact on how listeners perceive musical events, is problematic for those pursuing bottom-up, content-based models of structural analysis. Previous research has already shown that top-down factors such as repeated listenings and musical training have an influence on analysis, but compared to these, attention is a much more idiosyncratic influence. Attention can be guided by the musical events themselves, as considered above, but can also be controlled, to some extent, by the listener's own whims and desires. The influence of attention also seems to not be restricted to the largest timescale of analysis: attention can reach deep and affect how low-level structure is perceived (for example, by changing the salience of local boundaries). Thus the border between bottom-up perception and top-down interpretation is not simply hard to find, as described in Chapter 2, but porous. To account for this, newer musicological and perceptual models should seek to integrate more flexibility into how groupings are predicted based on lower-level musical events.

One of the central problems for structural analysis work in MIR is the lack of a precise problem definition, which is related to the ambiguity of the analysis task. The discov-

ery that disagreements among listeners could be attributable to differences in attention (among other factors) may help those who are developing new annotation procedures that are repeatable and well-defined. These procedures should either direct annotators to focus on particular things (a suggestion raised by Befus, Sanden and Zhang, who had listeners provide boundary indications while focusing on a given feature [BSZ10]), or record in some what what the annotator found salient about the music (in the manner of the case study in Chapter 3, but in some more codified way). Peeters and Deruty argued that analyses were multi-dimensional, distinguishing layers of similarity, function, and instrumentation [PD09]. We argue that it is important to recognize that the similarity layer is itself multi-dimensional with respect to feature. All of these suggestions aim to either reduce disagreements between listeners or make them more transparent.

This research may set an example for MIR by testing the assumptions of common models on a ground truth collection. Mining evaluation data for insights and using these to improve one's models may seem like cheating, but it is more akin to demonstrating a proof of concept. For example, before developing a new genre prediction algorithm which uses instrument identification, based on the hypothesis that genres are distinguished by their use of different instrumental mixtures, one should first ask, and test, whether this hypothesis is in fact correct. Doing so can ensure that one's new algorithm is based on sound principles, but is also an opportunity to ensure that the test collection is properly understood.

There is one final lesson to draw from this research: noise can be more interesting, and more explicable, than it seems. The subject of this thesis was listener disagreements, and in previous research, such disagreements have often been treated as noise. This is at least the case in the fields of music psychology and MIR, which are interested in the behaviour of general listeners more than the idiosyncracies of individuals. So we began in the field of music theory, in which it is more usual to question how individuals come to the interpretations they do. We examined disagreements which may at first have seemed superficial—certainly there were many ways in which the pairs of analyses in Chapter

3 were congruent—but which we argued had deep causes. We studied these effects in a more focused way in the fields of music psychology and MIR, using a larger set of examples and more listeners, and in demonstrating their importance, we cast new light on our understanding of the original phenomenon that was deemed noisy: the perception of musical structure.

7.4 Summary of key contributions

In developing the case for the role of attention in explaining disagreements among listeners, each chapter of this thesis produced new pieces of evidence for this claim, and developed or introduced novel methods and materials. The chapters also span several disciplines. These contributions are summarized below.

New evidence that attention affects grouping

- A case study indicated that attention was the best proximate explanation for listener disagreements (Chapter 3)
- A corpus study indicated that on a broad scope of music, acoustic novelty correlated with boundaries, and hence possibly with the salience of changes (Chapter 4);
- Distinct annotations provided by listeners were found to be well explained by distinct sets of features (Chapter 5)
- An experiment presented direct evidence that attention to a particular feature led listeners to prefer a grouping structure aligned with that feature (Chapter 6)

Novel methods and materials

- The case study was novel in its depth of analysis into the justifications for

grouping decisions, and in the material used (Chapter 3)

- We mined a corpus, normally used for evaluation, to derive insights into human perception (Chapter 4)
- Introduction of a novel tool for linking musical analyses to the musical features that support them (Chapter 5)
- Development of a novel set of stimuli in which grouping structure varies systematically across different features (Chapter 6)

Interdisciplinarity Contributions included:

- A submission to a music theory journal (Chapter 3) [SSC14]
- Submissions to journals and conferences on multimedia (Chapters 4 and 5) [SCC14, SC13b]
- A submission (in preparation) to a journal of music perception and cognition (Chapter 6)

Bibliography

[Aga94] Kofi Agawu. Ambiguity in tonal music: A preliminary study. In Anthony Pople, editor, *Theory, Analysis and Meaning in Music*, pages 90–107. Cambridge University Press, 1994.

[AK08] Katherine Agres and Carol Krumhansl. Musical change deafness: The inability to detect change in a non-speech auditory domain. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 969–974, 2008.

[AP09] Samer Abdallah and Mark Plumbley. Information dynamics: Patterns of expectation and surprise in the perception of music. *Connection Science*, 21(2–3):89–117, 2009.

[APS05] J.-J. Aucouturier, François Pachet, and Mark Sandler. "The way it sounds": Timbre models for analysis and retrieval of music signals. *IEEE Transactions on Multimedia*, 7(6):1028–1035, 2005.

[ARC+13] Daniel A. Abrams, Srikanth Ryali, Tianwen Chen, Parag Chordia, Amirah Khouzam, Daniel J. Levitin, and Vinod Menon. Inter-subject synchronization of brain responses during natural music listening. *European Journal of Neuroscience*, 37:1458–1469, 2013.

[AS01] J.-J. Aucouturier and Mark Sandler. Segmentation of musical signals using

hidden Markov models. In *Proceedings of the Audio Engineering Society Convention (AES)*, Amsterdam, The Netherlands, 2001.

[ASRC06] Samer Abdallah, Mark Sandler, Christophe Rhodes, and Michael Casey. Using duration models to reduce fragmentation in audio segmentation. *Machine Learning*, 65(2-3):485–515, 2006.

[ATT12] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338:1225–1229, 2012.

[Bam06] Jeanne Bamberger. What develops in musical development? A view of development as learning. In Gary McPherson, editor, *The Child as Musician: Musical Development from Conception to Adolescence*, pages 69–92. Oxford: Oxford University Press, 2006.

[BDSV12a] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. Complementary report to the article "semiotic structure labeling of music pieces : concepts, methods and annotation conventions". Technical Report PI 1996, Institut national de recherche en informatique et en automatique (INRIA), June 2012.

[BDSV12b] Frédéric Bimbot, Emmanuel Deruty, Gabriel Sargent, and Emmanuel Vincent. Semiotic structure labeling of music pieces: Concepts, methods and annotation conventions. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 235–240, Porto, Portugal, 2012.

[BLBSV10] Frédéric Bimbot, Olivier Le Blouch, Gabriel Sargent, and Emmanuel Vincent. Decomposition into autonomous and comparable blocks: A structural description of music pieces. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 189–194, 2010.

[BM08] Michael Bruderer and Martin McKinney. Perceptual evaluation of models for music segmentation. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*, Thessaloniki, Greece, 2008.

[BMK06] Michael Bruderer, Martin McKinney, and Armin Kohlrausch. Structural boundary perception in popular music. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 198–201, Victoria, Canada, 2006.

[BMK09] Michael Bruderer, Martin McKinney, and Armin Kohlrausch. The perception of structural boundaries in melody lines of Western popular music. *MusicæScientæ*, 13(2):273–313, 2009.

[BSZ10] Chad Befus, Chris Sanden, and John Zhang. Psychoacoustic feature based perceptual segmentation. In *Proceedings of the International Computer Music Conference (ICMC)*, 2010.

[BW01] Mark Bartsch and Gregory Wakefield. To catch a chorus: using chroma-based representations for audio thumbnailing. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 15–8, 2001.

[CAFW13] Tom Collins, Andreas Arzt, Sebastian Flossman, and Gerhard Widmer. SIARCT-CFP: improving precision and the discovery of inexact musical patterns in point-set representations. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 549–554, Curitiba, Brazil, 2013.

[Cam01] Emilios Cambouropoulos. The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC)*, Havana, Cuba, 2001.

[Cam06] Emilios Cambouropoulos. Musical parallelism and melodic segmentation: A

computational approach. *Music Perception*, 23(3):249–267, 2006.

[Cap98] William Caplin. *Classical Form: A Theory of Formal Functions for the Music of Haydn, Mozart, and Beethoven.* Oxford University Press, 1998.

[CC07] Ching-Hua Chuan and Elaine Chew. Audio key finding: Considerations in system design and case studies on Chopin's 24 Preludes. *EURASIP Journal on Advances in Signal Processing*, 2007.

[Che00] Elaine Chew. *Towards a mathematical model of tonality.* PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 2000.

[Che05] Elaine Chew. Regards on two regards by Messiaen: Post-tonal music segmentation using pitch context distances in the Spiral Array. *Journal of New Music Research*, 34(4):341–354, 2005.

[CK90] Eric F. Clarke and Carol L. Krumhansl. Perceiving musical time. *Music Perception*, 7(3):213–51, 1990.

[Cla89] Eric F. Clarke. Mind the gap: Formal structures and psychological processes in music. *Contemporary Music Review*, 3(1):1–13, 1989.

[Del87] Iréne Deliège. Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's Grouping Preference Rules. *Music Perception*, 4(4):325–359, 1987.

[Deu99] Diana Deutsch. *The Psychology of Music*, chapter Grouping Mechanisms in Music, pages 299–348. Academic Press, 1999.

[EBD+11] Andreas F. Ehmann, Mert Bay, J. Stephen Downie, Ichiro Fujinaga, and David De Roure. Exploiting music structures for digital libraries. In *Proceeding of the International ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 479–480, Ottawa, Ontario, Canada, 2011.

[EKR87] J.-P. Eckmann, S. Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 4(9):973–7, November 1987.

[Ero07] Antti Eronen. Chorus detection with combined use of mfcc and chroma features and image processing filters. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 229–236, Bordeaux, France, 2007.

[Far12] Morwaread Farbood. A parametric, temporal model of musical tension. *Music Perception*, 29(4):387–428, 2012.

[FC03] Jonathan Foote and Matthew Cooper. Media segmentation using self-similarity decomposition. In Minerva Yeung, Rainer Lienhart, and Chung-Sheng Li, editors, *Proceedings of the SPIE: Storage and Retrieval for Media Databases*, volume 5021, pages 167–75, Santa Clara, CA, USA, 2003. SPIE.

[FC04] Bradley Frankland and Annabel Cohen. 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, 2004.

[FCT07] Alexandre R. J. François, Elaine Chew, and Dennis Thurmond. Visual feedback in performer-machine interaction for musical improvisation. In *Proceedings of the Conference on New interfaces for Musical Expression (NIME)*, pages 277–280, 2007.

[Foo99] Jonathan Foote. Visualizing music and audio using self-similarity. In *Proceedings of the ACM International Conference on Multimedia*, pages 77–80, New York, NY, USA, 1999. ACM.

[Foo00] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME)*, pages 452–455, 2000.

[Fra58] Robert Francès. *The Perception of Music*. Psychology Press, 1987 translation by w. jay dowling edition, 1958.

[Fra09] Alexandre R. J. François. Time and perception in music and computation. In Gérard Assayag and Andrew Gerzso, editors, *New Computational Paradigms for Computer Music*, pages 125–46. Editions Delatour France / IRCAM, 2009.

[GCJM13] Harald Groghanz, Michael Clausen, Nanzhu Jiang, and Meinard Müller. Converting path structures into block structures using eigenvalue decompositions of self-similarity matrices. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2013.

[GJN10] Rolf Inge Godøy, Alexander Refsum Jensenius, and Kristian Nymoen. Chunking in music by coarticulation. *Acta Acoustica united with Acoustica*, 96(4):690–700, 2010.

[Got03] Masataka Goto. A chorus-section detecting method for musical audio signals. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 5, pages 437–440, 2003.

[Got06] Masataka Goto. A chorus section detection method for musical audio signals and its application to a music listening station. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1783–1794, 2006.

[GSMA12] Peter Grosche, Joan Serrà, Meinard Müller, and Josep Ll. Arcos. Structure-based audio fingerprinting for music retrieval. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 55–60, Porto, Portugal, 2012.

[Han12] Dora A. Hanninen. *A Theory of Music Analysis: On Segmentation and Associative Organization*. University of Rochester Press, Rochester, NY, 2012.

[HHT04] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. Automatic generation of grouping structure based on the GTTM. In *Proceedings of the International Computer Music Conference (ICMC)*, 2004.

[HHT06] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. Implementing a *Generative Theory of Tonal Music*. *Journal of New Music Research*, 35(4):249—277, 2006.

[HKS12] Steven Hargreaves, Anssi Klapuri, and Mark Sandler. Structural segmention of multitrack audio. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(10):2637–2647, 2012.

[HSN+14] Eric J. Humphrey, Justin Salamon, Oriol Nieto, Jon Forsyth, Rachel M. Bittner, and Juan Pablo Bello. JAMS: A JSON Annotated Music Specification for reproducible MIR research. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2014.

[HTH07] Keiji Hirata, Satoshi Tojo, and Masatoshi Hamanaka. Techniques for implementing the *Generative Theory of Tonal Music* (tutorial session). In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Vienna, Austria, 2007.

[HTS02] Karin Höthker, Belinda Thom, and Christian Spevak. Melodic segmentation: evaluating the performance of algorithms and musical experts. In *Proceedings of the International Computer Music Conference (ICMC)*, 2002.

[HVP13] Niels Chr. Hansen, Peter Vuust, and Marcus Pearce. Predictive processing of musical structure: effects of genre-specific expertise. Abstract at the Annual Meeting of the Cognitive Science Society, 2013.

[JB89] Mari Riess Jones and Marilyn Boltz. Dynamic attending and responses to time. *Psychological Review*, 96(3):459–491, 1989.

[Jeh05] Tristan Jehan. Hierarchical multi-class self similarities. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 311–314, New Paltz, NY, United States, 2005.

[Kar12] Gary S. Karpinski. Ambiguity: Another listen. *Music Theory Online*, 18(3), 2012.

[KP12] Florian Kaiser and Geoffroy Peeters. Adaptive temporal modeling of audio features in the context of music structure segmentation. In *International Workshop on Adaptive Multimedia Retrieval*, Copenhagen, Denmark, 2012.

[Kru96] Carol L. Krumhansl. A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas. *Music Perception*, 13(3):401–432, 1996.

[KS10] Florian Kaiser and Thomas Sikora. Music structure discovery in popular music using non-negative matrix factorization. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Utrecht, The Netherlands, 2010.

[LC00] B. Logan and S. Chu. Music summarization using key phrases. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 749–752, Washington D.C., USA, 2000. IEEE Computer Society.

[Lew09] George Lewis. The condition for improvisation. Keynote address at the International Society for Improvised Music, Santa Cruz, CA, 2009.

[LGC+11] Christian Landone, Martin Gasser, Chris Cannam, Chris Harte, Matthew Davies, Katy Noland, Thomas Wilmering, Wen Xue, and Ruohua Zhou. QM Vamp Plugins. Audio feature extraction plugins from Queen Mary, University of London, 2011.

[LJ83] Fred Lerdahl and Ray S. Jackendoff. *A Generative Theory of Tonal Music*. MIT Press, 1983.

[Luk08] Hanna Lukashevich. Towards quantitative measures of evaluating song segmentation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 375–380, Philadelphia, PA, USA, 2008.

[MA08] Meinard Müller and Daniel Appelt. Path-constrained partial music synchronization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 65–68, Las Vegas, NV, USA, 2008.

[Mar06] Matija Marolt. A mid-level melody-based representation for calculating audio similarity. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 280–285, Victoria, Canada, 2006.

[Mar12] Elizabeth Margulis. Musical repetition detection across multiple exposures. *Music Perception*, 29(4):377–385, 2012.

[Mar14] Elizabeth Margulis. *On Repeat*. Oxford University Press, Oxford, UK, 2014.

[MGJ11] Meinard Müller, Peter Grosche, and Nanzhu Jiang. A segment-based fitness measure for capturing repetitive structures of music recordings. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 615–620, 2011.

[MGMS14] Daniel Müllensiefen, Bruno Gingras, Jason Jiří Musil, and Lauren Stewart. The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLOS One*, 9(2), 2014.

[MK07] Meinard Müller and Frank Kurth. Towards structural analysis of audio recordings in the presence of musical variations. *EURASIP J. Appl. Signal Process.*, 2007(1), 2007.

[MND09] Matthias Mauch, Katy Noland, and Simon Dixon. Using musical structure to enhance automatic chord transcription. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 231–236, Kobe, Japan, 2009.

[MOG00] Massimo Melucci, Nicola Orio, and Marco Gambalunga. An evaluation study on music perception for music content-based information retrieval. In *Proceedings of the International Computer Music Conference (ICMC)*, 2000.

[Nar90] Eugene Narmour. *The Analysis and Cognition of Basic Melodic Structures: The Implication-Realization Model.* University of Chicago Press, Chicago, IL, USA, 1990.

[NFJB14] Oriol Nieto, Morwaread Farbood, Tristan Jehan, and Juan Pablo Bello. Perceptual analysis of the f-measure to evaluate section boundaries in music. In *Proceedings of the International Society for Music Information Retrieval Conference*, Taipei, Taiwan, 2014.

[Ong07] Bee Ong. *Structural Analysis and Segmentation of Music Signals.* PhD thesis, University Pompeu Fabra, Barcelona, Spain, February 2007.

[Pam04] Elias Pampalk. A Matlab toolbox to compute similarity from audio. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 254–257, Barcelona, Spain, 2004.

[PD09] Geoffroy Peeters and Emmanuel Deruty. Is music structure annotation multi-dimensional? A proposal for robust local music annotation. In *Proceedings of the International Workshop on Learning the Semantics of Audio Signals (LSAS)*, pages 75–90, Graz, Austria, 2009.

[PDW03] Elias Pampalk, Simon Dixon, and Gerhard Widmer. Exploring music collections by browsing different views. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2003.

[PE04] R. Mitchell Parry and Irfan Essa. Feature weighting for segmentation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[Pea05] Marcus Thomas Pearce. *The construction and evaluation of statistical models of melodic structure in music perception and composition.* PhD thesis, City University, London, London, United Kingdom, 2005.

[Pee04] Geoffroy Peeters. Deriving musical structures from signal analysis for music audio summary generation: "sequence" and "state" approach. In Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen, editors, *Computer Music Modeling and Retrieval*, volume 2771, pages 169–185. Springer Berlin / Heidelberg, 2004.

[Pee07] Geoffroy Peeters. Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 35–40, Vienna, Austria, 2007.

[Pei07] Ewald Peiszer. Automatic audio segmentation: Segment boundary and structure detection in popular music. Master's thesis, Technische Universität Wien, 2007.

[PK87] Caroline Palmer and Carol L. Krumhansl. Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, 41(6):505–518, 1987.

[PK06] Jouni Paulus and Anssi Klapuri. Music structure analysis by finding repeated parts. In *Proceedings of the ACM Workshop on Audio and Music Computing Multimedia (AMCMM)*, pages 59–68, New York, NY, USA, 2006. ACM.

[PK08a] Jouni Paulus and Anssi Klapuri. Acoustic features for music piece structure

analysis. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 309–312, Espoo, Finland, 2008.

[PK08b] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and an integrated musicological model. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 369–374, Philadelphia, PA, USA, 2008.

[PK09] Jouni Paulus and Anssi Klapuri. Music structure analysis using a probabilistic fitness measure and a greedy search algorithm. *IEEE Transactions on Audio, Speech & Language Processing*, 17(6):1159–1170, 2009.

[PMK10] Jouni Paulus, Meinard Müller, and Anssi Klapuri. Audio-based music structure analysis. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 625–636, Utrecht, The Netherlands, 2010.

[PMW08] Marcus Pearce, Daniel Müllensiefen, and Geraint Wiggins. A comparison of statistical and rule-based models of melodic segmentation. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 89–94, Philadelphia, PA, USA, 2008.

[PMW10a] Marcus T. Pearce, Daniel Müllensiefen, and Geraint A. Wiggins. Melodic grouping in music information retrieval: New methods and applications. In Zbigniew W. Raś and Alicja A. Wieczorkowska, editors, *Advances in Music Information Retrieval*, volume 274 of *Studies in Computational Intelligence*, pages 364–388. Springer Berlin Heidelberg, 2010.

[PMW10b] Marcus T. Pearce, Daniel Müllensiefen, and Geraint A. Wiggins. The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, 39:1367–1391, 2010.

[PRM02] Elias Pampalk, Andreas Rauber, and Dieter Merkl. Content-based organi-

zation and visualization of music archives. In *Proceedings of the ACM Multimedia*, pages 570–579, Juan les Pins, France, December 1-6 2002. ACM.

[PW06] Marcus T. Pearce and Geraint A. Wiggins. Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–405, 2006.

[Roy95] Matthew S. Royal. Review of *The Analysis and Cognition of Basic Melodic Structures* and *The Analysis and Cognition of Melodic Complexity* by Eugene Narmour. *Music Theory Online*, 1(6), 1995.

[SAPM02] E. Glenn Schellenberg, Mayumi Adachi, Kelly T. Purdy, and Margaret C. McKinnon. Expectation in melody: Tests of children and adults. *Journal of Experimental Psychology*, 131(4):511–537, 2002.

[SBF+11] Jordan Bennett Louis Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen Downie. Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 555–560, Miami, FL, United States, 2011.

[SBZ12] Chris Sanden, Chad R. Befus, and John Z. Zhang. A perceptual study on music segmentation and genre classification. *Journal of New Music Research*, 41(3):277–293, 2012.

[SC13a] Jordan B. L. Smith and Elaine Chew. A meta-analysis of the MIREX Structure Segmentation task. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 251–256, Curitiba, Brazil, 2013.

[SC13b] Jordan B. L. Smith and Elaine Chew. Using quadratic programming to estimate feature relevance in structural analyses of music. In *Proceedings of the ACM International Conference on Multimedia*, pages 113–122, Barcelona, Spain, 2013.

[SCC14] Jordan B. L. Smith, Ching-Hua Chuan, and Elaine Chew. Audio properties of perceived boundaries in music. *IEEE Transactions on Multimedia*, 16(5):1219–1228, August 2014.

[SCF14] Isaac Schankler, Elaine Chew, and Alexandre R. J. François. Improvising with digital auto-scaffolding: how Mimi changes and enhances the creative process. In Newton Lee, editor, *Digital Da Vinci: Computers in Music*, pages 99–125. Springer, 2014.

[SJK06] Yu Shiu, Hong Jeong, and C.-C. Jay Kuo. Similarity matrix processing for music structure analysis. In *Proceedings of the ACM Workshop on Audio and Music Computing Multimedia (AMCMM)*, pages 69–76, New York, NY, USA, 2006. ACM.

[Smi10] Jordan B. L. Smith. A comparison and evaluation of approaches to the automatic formal analysis of musical audio. Master's thesis, McGill University, Montreal, QC, Canada, 2010.

[SP94] Dale Stammen and Bruce Pennycook. Real-time segmentation of music using an adaptation of Lerdahl and Jackendoff's grouping principles. In *Proceedings of the International Conference on Music Perception and Cognition (ICMPC)*, pages 269–270, Liege, Belgium, 1994.

[SSC14] Jordan B. L. Smith, Isaac Schankler, and Elaine Chew. Listening as a creative act: Meaningful differences in structural annotations of improvised performances. *Music Theory Online*, 20(3), 2014.

[SSFC11] Isaac Schankler, Jordan Bennett Louis Smith, Alexandre R. J. François, and Elaine Chew. Emergent formal structures of factor oracle-driven musical improvisations. In Carlos Agon, Moreno Andreatta, Gérard Assayag, Emmanuel Amiot, Jean Bresson, and John Mandereau, editors, *Mathematics and Computation in Music*, volume 6726 of *Lecture Notes in Computer*

Science. Springer Berlin / Heidelberg, 2011.

[Tem01] David Temperley. *The Cognition of Basic Musical Structure*. MIT Press, Cambridge, MA, 2001.

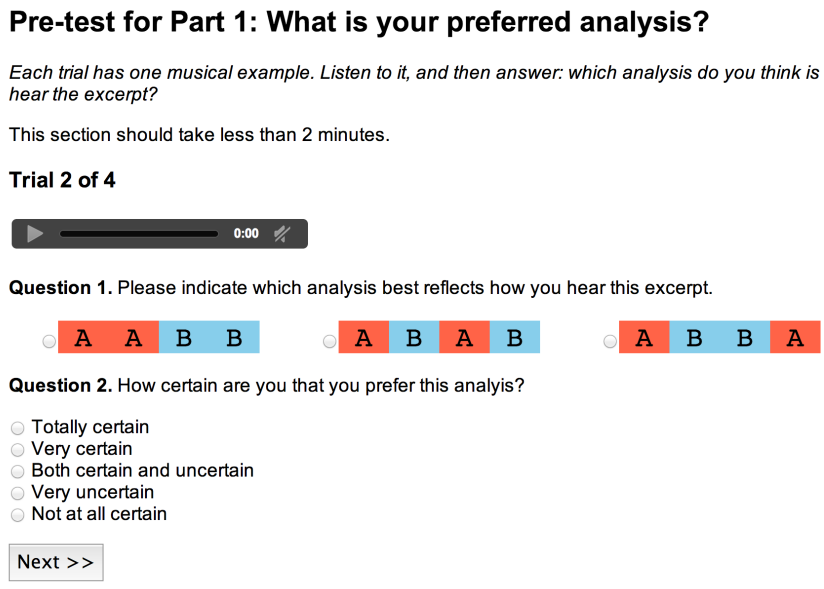
[TLPG07] Douglas Turnbull, Gert Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pages 51–54, Vienna, Austria, 2007.

[TP80] James Tenney and Larry Polansky. Temporal gestalt perception in music. *Journal of Music Theory*, 24(2):205–241, 1980.

[Tym11] Dmitri Tymoczko. *A Geometry of Music: Harmony and Counterpoint in the Extended Common Practice*. Oxford University Press, 2011.

Appendix A

Gallery of Web Pages from Experiments in Chapter 6



Pre-test for Part 1: What is your preferred analysis?

Each trial has one musical example. Listen to it, and then answer: which analysis do you think is the best description of how you hear the excerpt?

This section should take less than 2 minutes.

Trial 2 of 4

0:00

Question 1. Please indicate which analysis best reflects how you hear this excerpt.

A A B B A B A B A B B A

Question 2. How certain are you that you prefer this analyis?

- Totally certain
- Very certain
- Both certain and uncertain
- Very uncertain
- Not at all certain

Next >>

Part 1 of 4: Complete the analysis.

In this set of questions, you will be played two examples A and B, followed by a longer excerpt for you to analyze. Your goal is to guess how a listener who labelled A and B as such would label the excerpt.

You might not find any of the analyses to be perfect, but you must still choose one.

This section should take less than 8 minutes.

Trial 6 of 12

Imagine someone has listened to a piece and analyzed its structure. They have labelled the following clips as **A** and **B**:

Based on these labels, what is the best analysis of the following four-part clip?

Question 1. Please indicate the structure that the person who provided the previous analysis would provide.

Question 2. How certain are you that this would be the person's analysis?

- Totally certain
- Very certain
- Both certain and uncertain
- Very uncertain
- Not at all certain

Next >>

Pre-test for Part 2: What is your preferred analysis?

Each trial has one musical example. Listen to it, and then answer: which analysis do you think is the best description of how you hear the excerpt?

This section should take less than 2 minutes.

Trial 3 of 3

Question 1. Please indicate which analysis best reflects how you hear this excerpt.

Question 2. How certain are you that you prefer this analyis?

- Totally certain
- Very certain
- Both certain and uncertain
- Very uncertain
- Not at all certain

Next >>

Part 2 of 4: Does the pattern occur?

In this set of questions, a musical pattern of some kind will be shown to you. Your goal is to judge whether this pattern occurs in the longer musical excerpt that follows. We then ask you to re-listen to the excerpt, and state whether you prefer form AAB or ABB.

This section should take less than 12 minutes.

Trial 5 of 12

Please listen to the following **chord progression**

Please listen to the following excerpt of music and indicate whether that **chord progression** appears in it.

Question 1. Did the **chord progression** appear in the excerpt?

- Yes
- Yes, but only a variation
- No
- I do not know

Now, please listen to the excerpt again. (The following clip is identical to the previous clip.)

Question 2. Which of the following analyses do you think best fits the excerpt?

Question 3. How certain are you about your choice of analysis?

- Totally certain
- Very certain
- Both certain and uncertain
- Very uncertain
- Not at all certain

Next >>

Part 3 of 4: Salience of change

Every excerpt in this part has a single pattern repeated 4 times, with a change in some feature between the 2nd and 3rd instances; i.e., it has form AABB. We ask you to focus on a particular aspect of the music while listening, and tell us: how significant was the change at the half-way point?

This section should take less than 6 minutes.

Trial 4 of 12

Please pay attention to the **chords** of the following excerpt.

Question 1. How strong is the change at the midpoint of the excerpt?

- 5. Extremely strong
- 4.
- 3.
- 2.
- 1. Not strong at all

Next >>

Part 4 of 4: What changed?

For each question in this part, you will listen to one short excerpt. Like in the previous part, every excerpt has pattern AABB, with a change in some aspect of the music in the middle. We ask you to identify the aspect of the music that changes.

This section should take less than 6 minutes.

Trial 3 of 12

Please listen to the following excerpt.

Question 1. Please indicate the musical feature that changed during this excerpt.

- Chord Progression
- Melody
- Rhythm
- Timbre
- No change

Next >>

Appendix B

Computed Models for Experiments in Chapter 6

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.67	0.96	3.84	0.00***
ability	1.25	0.57	2.20	0.03*
change_featMelody	0.34	0.82	0.41	0.68
change_featRhythm	0.19	0.77	0.25	0.80
change_featTimbre	1.31	0.85	1.54	0.12
musicHT-MR	-1.74	0.99	-1.75	0.08
musicHM-RT	-1.34	0.99	-1.35	0.18
age	0.42	0.42	1.02	0.31
genderf	-1.52	0.96	-1.58	0.11
heardbefore	-0.37	0.68	-0.54	0.59
ability:change_featMelody	0.13	0.36	0.36	0.72
ability:change_featRhythm	-0.26	0.32	-0.82	0.41
ability:change_featTimbre	0.37	0.33	1.12	0.26
ability:musicHT-MR	-1.22	0.58	-2.08	0.04*
ability:musicHM-RT	-0.34	0.53	-0.63	0.53
ability:age	0.22	0.18	1.18	0.24
ability:genderf	0.10	0.40	0.25	0.80
ability:heardbefore	0.48	0.27	1.80	0.07
change_featMelody:musicHT-MR	0.85	0.90	0.94	0.35
change_featRhythm:musicHT-MR	1.30	0.88	1.49	0.14
change_featTimbre:musicHT-MR	0.09	0.90	0.09	0.92
change_featMelody:musicHM-RT	0.63	0.81	0.77	0.44
change_featRhythm:musicHM-RT	-0.41	0.73	-0.56	0.58
change_featTimbre:musicHM-RT	-2.25	0.81	-2.80	0.01**
change_featMelody:age	-0.63	0.27	-2.34	0.02*
change_featRhythm:age	-0.29	0.26	-1.13	0.26
change_featTimbre:age	0.09	0.27	0.34	0.74
change_featMelody:genderf	0.20	0.71	0.29	0.77
change_featRhythm:genderf	-0.01	0.63	-0.01	0.99
change_featTimbre:genderf	0.42	0.64	0.66	0.51
change_featMelody:heardbefore	0.08	0.70	0.12	0.91
change_featRhythm:heardbefore	-0.38	0.62	-0.62	0.54
change_featTimbre:heardbefore	-0.96	0.61	-1.57	0.12
musicHT-MR:age	-0.50	0.40	-1.26	0.21
musicHM-RT:age	-0.14	0.40	-0.35	0.72
musicHT-MR:genderf	1.67	1.11	1.51	0.13
musicHM-RT:genderf	0.25	0.99	0.25	0.80
musicHT-MR:heardbefore	0.65	0.67	0.97	0.33
musicHM-RT:heardbefore	1.29	0.58	2.21	0.03*
age:genderf	-0.58	0.36	-1.62	0.11
age:heardbefore	0.29	0.21	1.42	0.16
genderf:heardbefore	0.11	0.50	0.23	0.82

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-A: Full model results from Experiment no. 1. These are the parameters for modeling answer correctness, treating convolved-feature errors as errors.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4.03	1.06	3.79	0.00***
ability	1.11	0.65	1.70	0.09
change_featMelody	-0.24	1.02	-0.24	0.81
change_featRhythm	0.76	0.97	0.78	0.43
change_featTimbre	1.90	1.12	1.70	0.09
musicHT-MR	-1.12	1.16	-0.97	0.33
musicHM-RT	-1.50	1.08	-1.38	0.17
age	0.94	0.54	1.73	0.08
genderf	-1.71	1.05	-1.63	0.10
heardbefore	-0.56	0.78	-0.72	0.47
ability:change_featMelody	-0.13	0.53	-0.25	0.80
ability:change_featRhythm	-0.48	0.42	-1.14	0.25
ability:change_featTimbre	0.97	0.56	1.74	0.08
ability:musicHT-MR	-1.35	0.71	-1.90	0.06
ability:musicHM-RT	-0.11	0.59	-0.18	0.85
ability:age	0.57	0.26	2.17	0.03*
ability:genderf	0.16	0.48	0.33	0.74
ability:heardbefore	0.64	0.35	1.82	0.07
change_featMelody:musicHT-MR	1.80	1.13	1.60	0.11
change_featRhythm:musicHT-MR	18.70	1311.42	0.01	0.99
change_featTimbre:musicHT-MR	1.90	1.53	1.24	0.21
change_featMelody:musicHM-RT	3.10	1.34	2.32	0.02*
change_featRhythm:musicHM-RT	-1.08	0.82	-1.31	0.19
change_featTimbre:musicHM-RT	-0.94	0.90	-1.05	0.29
change_featMelody:age	-0.83	0.36	-2.30	0.02*
change_featRhythm:age	-1.10	0.36	-3.05	0.00**
change_featTimbre:age	0.03	0.37	0.07	0.94
change_featMelody:genderf	0.79	0.94	0.84	0.40
change_featRhythm:genderf	-0.34	0.76	-0.44	0.66
change_featTimbre:genderf	0.32	0.81	0.39	0.70
change_featMelody:heardbefore	-0.32	0.88	-0.36	0.72
change_featRhythm:heardbefore	-0.97	0.79	-1.23	0.22
change_featTimbre:heardbefore	-0.71	0.83	-0.86	0.39
musicHT-MR:age	-0.79	0.50	-1.59	0.11
musicHM-RT:age	-0.37	0.45	-0.83	0.41
musicHT-MR:genderf	2.05	1.38	1.49	0.14
musicHM-RT:genderf	0.77	1.07	0.72	0.47
musicHT-MR:heardbefore	-0.23	0.90	-0.26	0.80
musicHM-RT:heardbefore	1.61	0.69	2.34	0.02*
age:genderf	-0.49	0.42	-1.18	0.24
age:heardbefore	0.29	0.26	1.09	0.27
genderf:heardbefore	0.40	0.60	0.67	0.50

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-B: Full model results from Experiment no. 1. These are the parameters for modeling answer correctness, treating convolved-feature errors as correct.

	Estimate	Std. Error	t value	Estimated Pr(>|z|)
(Intercept)	-1.33	0.19	-6.89	0.00***
ability	-0.02	0.12	-0.18	0.85
change_featMelody	-0.27	0.21	-1.26	0.21
change_featRhythm	-0.49	0.21	-2.33	0.02*
change_featTimbre	0.29	0.21	1.37	0.17
matchMatch	2.43	0.20	12.24	0.00***
matchWrong	0.24	0.20	1.18	0.24
musicHT-MR	0.24	0.25	0.99	0.32
musicHM-RT	0.03	0.24	0.11	0.91
age	-0.03	0.14	-0.23	0.81
genderf	0.03	0.24	0.11	0.92
ability:change_featMelody	-0.21	0.09	-2.22	0.03*
ability:change_featRhythm	-0.14	0.09	-1.45	0.15
ability:change_featTimbre	0.15	0.09	1.65	0.10
ability:matchMatch	0.29	0.08	3.58	0.00*
ability:matchWrong	-0.03	0.08	-0.35	0.72
ability:musicHT-MR	-0.14	0.13	-1.09	0.28
ability:musicHM-RT	0.10	0.13	0.80	0.43
ability:age	-0.02	0.05	-0.36	0.72
ability:genderf	-0.03	0.11	-0.26	0.80
change_featMelody:matchMatch	-0.08	0.22	-0.38	0.71
change_featRhythm:matchMatch	-0.77	0.22	-3.50	0.00***
change_featTimbre:matchMatch	-0.60	0.22	-2.69	0.01**
change_featMelody:matchWrong	0.00	0.22	0.02	0.99
change_featRhythm:matchWrong	-0.44	0.22	-1.98	0.05*
change_featTimbre:matchWrong	-0.52	0.22	-2.34	0.02*
change_featMelody:musicHT-MR	0.29	0.23	1.29	0.20
change_featRhythm:musicHT-MR	0.99	0.23	4.34	0.00***
change_featTimbre:musicHT-MR	-0.05	0.23	-0.22	0.83
change_featMelody:musicHM-RT	0.74	0.22	3.41	0.00***
change_featRhythm:musicHM-RT	0.48	0.22	2.21	0.03*
change_featTimbre:musicHM-RT	0.43	0.22	1.96	0.05*
change_featMelody:age	-0.12	0.09	-1.32	0.19
change_featRhythm:age	-0.10	0.09	-1.08	0.28
change_featTimbre:age	-0.03	0.09	-0.30	0.77
change_featMelody:genderf	-0.17	0.19	-0.92	0.36
change_featRhythm:genderf	0.34	0.19	1.80	0.07
change_featTimbre:genderf	-0.04	0.19	-0.19	0.85
matchMatch:musicHT-MR	-0.48	0.20	-2.42	0.02*
matchWrong:musicHT-MR	-0.16	0.20	-0.82	0.41
matchMatch:musicHM-RT	-0.95	0.19	-5.06	0.00***
matchWrong:musicHM-RT	-0.47	0.19	-2.48	0.01*
matchMatch:age	-0.02	0.08	-0.31	0.76
matchWrong:age	0.03	0.08	0.42	0.67
matchMatch:genderf	0.11	0.16	0.70	0.48
matchWrong:genderf	-0.09	0.16	-0.53	0.59
musicHT-MR:age	0.00	0.14	0.01	0.99
musicHM-RT:age	-0.15	0.14	-1.14	0.26
musicHT-MR:genderf	-0.33	0.27	-1.22	0.22
musicHM-RT:genderf	0.01	0.27	0.06	0.95
age:genderf	0.10	0.11	0.89	0.37

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-C: Full model results from Experiment no. 2. These are the parameters for modeling boundary salience.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.41	0.55	2.58	0.01**
ability	-0.11	0.36	-0.32	0.75
focalfeatMelody	0.93	0.86	1.08	0.28
focalfeatRhythm	-2.04	0.63	-3.27	0.00**
focalfeatTimbre	-1.68	0.63	-2.66	0.01**
musicHT-MR	-1.70	0.70	-2.43	0.01*
musicHM-RT	-0.70	0.64	-1.10	0.27
age	-0.78	0.32	-2.45	0.01*
genderf	0.02	0.60	0.04	0.97
presencePresent	0.68	0.60	1.14	0.26
ability:focalfeatMelody	0.81	0.47	1.75	0.08
ability:focalfeatRhythm	-0.09	0.32	-0.29	0.77
ability:focalfeatTimbre	-0.17	0.30	-0.56	0.58
ability:musicHT-MR	0.23	0.31	0.74	0.46
ability:musicHM-RT	-0.26	0.33	-0.77	0.44
ability:age	0.07	0.12	0.61	0.54
ability:genderf	-0.17	0.28	-0.61	0.54
ability:presencePresent	0.33	0.23	1.41	0.16
focalfeatMelody:musicHT-MR	2.55	1.10	2.32	0.02*
focalfeatRhythm:musicHT-MR	4.28	0.82	5.24	0.00***
focalfeatTimbre:musicHT-MR	2.10	0.74	2.85	0.00**
focalfeatMelody:musicHM-RT	1.41	1.07	1.31	0.19
focalfeatRhythm:musicHM-RT	0.96	0.74	1.31	0.19
focalfeatTimbre:musicHM-RT	1.63	0.72	2.25	0.02*
focalfeatMelody:age	-0.09	0.36	-0.26	0.79
focalfeatRhythm:age	0.63	0.32	1.96	0.05*
focalfeatTimbre:age	0.34	0.31	1.10	0.27
focalfeatMelody:genderf	0.15	0.88	0.17	0.86
focalfeatRhythm:genderf	0.40	0.64	0.62	0.54
focalfeatTimbre:genderf	0.28	0.62	0.46	0.65
focalfeatMelody:presencePresent	-0.62	0.84	-0.73	0.46
focalfeatRhythm:presencePresent	-0.57	0.61	-0.93	0.35
focalfeatTimbre:presencePresent	-0.49	0.58	-0.85	0.40
musicHT-MR:age	0.85	0.33	2.61	0.01**
musicHM-RT:age	0.43	0.28	1.52	0.13
musicHT-MR:genderf	0.04	0.58	0.06	0.95
musicHM-RT:genderf	0.23	0.67	0.34	0.74
musicHT-MR:presencePresent	-0.15	0.57	-0.25	0.80
musicHM-RT:presencePresent	-0.54	0.55	-0.98	0.33
age:genderf	0.24	0.25	0.99	0.32
age:presencePresent	0.10	0.23	0.45	0.65
genderf:presencePresent	-0.21	0.47	-0.44	0.66

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-D: Full model results from Experiment no. 3. These are the parameters for modeling agreement between the given analysis and the focal feature's implied analysis, considering only relevant trials.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.65	0.42	-1.54	0.12
ability	0.81	0.30	2.69	0.01**
focalfeatMelody	0.49	0.52	0.96	0.34
focalfeatRhythm	1.63	0.58	2.83	0.00**
focalfeatTimbre	0.18	0.50	0.36	0.72
musicHT-MR	0.32	0.57	0.55	0.58
musicHM-RT	2.09	0.62	3.36	0.00***
age	0.16	0.31	0.53	0.60
genderf	-1.01	0.52	-1.96	0.05
presencePresent	2.85	0.54	5.24	0.00***
relevanceRelevant	1.07	0.45	2.34	0.02*
ability:focalfeatMelody	-0.08	0.27	-0.29	0.77
ability:focalfeatRhythm	-0.11	0.27	-0.42	0.68
ability:focalfeatTimbre	-0.20	0.26	-0.78	0.44
ability:musicHT-MR	-0.39	0.27	-1.43	0.15
ability:musicHM-RT	-0.24	0.33	-0.73	0.46
ability:age	0.21	0.12	1.79	0.07
ability:genderf	-0.18	0.26	-0.72	0.47
ability:presencePresent	0.00	0.23	0.01	0.99
ability:relevanceRelevant	-0.11	0.19	-0.59	0.56
focalfeatMelody:musicHT-MR	0.40	0.60	0.66	0.51
focalfeatRhythm:musicHT-MR	-1.05	0.64	-1.64	0.10
focalfeatTimbre:musicHT-MR	0.79	0.56	1.41	0.16
focalfeatMelody:musicHM-RT	-0.61	0.68	-0.91	0.37
focalfeatRhythm:musicHM-RT	-2.52	0.69	-3.67	0.00***
focalfeatTimbre:musicHM-RT	0.83	0.73	1.13	0.26
focalfeatMelody:age	0.31	0.28	1.08	0.28
focalfeatRhythm:age	0.33	0.28	1.18	0.24
focalfeatTimbre:age	-0.20	0.26	-0.77	0.44
focalfeatMelody:genderf	1.58	0.54	2.91	0.00**
focalfeatRhythm:genderf	1.50	0.56	2.65	0.01**
focalfeatTimbre:genderf	0.70	0.51	1.37	0.17
focalfeatMelody:presencePresent	-0.22	0.59	-0.36	0.72
focalfeatRhythm:presencePresent	1.01	0.87	1.15	0.25
focalfeatTimbre:presencePresent	-0.43	0.55	-0.79	0.43
focalfeatMelody:relevanceRelevant	-0.39	0.51	-0.76	0.45
focalfeatRhythm:relevanceRelevant	-0.30	0.53	-0.57	0.57
focalfeatTimbre:relevanceRelevant	-0.50	0.50	-1.01	0.31
musicHT-MR:age	-0.38	0.27	-1.39	0.16
musicHM-RT:age	-0.64	0.29	-2.22	0.03*
musicHT-MR:genderf	-0.21	0.52	-0.41	0.68
musicHM-RT:genderf	-0.69	0.69	-1.01	0.31
musicHT-MR:presencePresent	-0.00	0.49	-0.01	1.00
musicHM-RT:presencePresent	1.04	0.76	1.38	0.17
musicHT-MR:relevanceRelevant	-0.66	0.43	-1.52	0.13
musicHM-RT:relevanceRelevant	-1.12	0.50	-2.25	0.02*
age:genderf	-0.30	0.25	-1.17	0.24
age:presencePresent	0.31	0.24	1.28	0.20
age:relevanceRelevant	0.11	0.19	0.58	0.56
genderf:presencePresent	0.47	0.47	1.00	0.32
genderf:relevanceRelevant	0.19	0.39	0.50	0.62
presencePresent:relevanceRelevant	-1.26	0.46	-2.77	0.01**

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-E: Full model results from Experiment no. 3. These are the parameters for modeling pattern detection accuracy, considering all trials, and considering "variation" as a "yes" answer.

	Estimate	Std. Error	t value	Estimated Pr(>|z|)
(Intercept)	0.54	0.18	3.03	0.00**
presencePresent	0.12	0.05	2.64	0.01**
focalfeatMelody	0.18	0.06	2.89	0.00**
focalfeatRhythm	0.04	0.06	0.61	0.54
focalfeatTimbre	0.04	0.06	0.57	0.57
ability	0.17	0.08	2.08	0.04*
musicHT-MR	-0.15	0.19	-0.81	0.42
musicHM-RT	0.20	0.19	1.03	0.30
age	-0.23	0.08	-2.87	0.00**
genderf	0.08	0.17	0.48	0.63
relevanceRelevant	0.13	0.04	2.98	0.00*

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-F: Full model results from Experiment no. 3. These are the parameters for modeling answer confidence, considering all trials.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.86	0.44	4.26	0.00***
ability	0.30	0.12	2.42	0.02*
focalfeatMelody	2.23	0.37	5.95	0.00***
focalfeatRhythm	0.22	0.26	0.83	0.41
focalfeatTimbre	1.16	0.29	4.07	0.00***
musicHT-MR	-1.01	0.36	-2.77	0.01**
musicHM-RT	-1.51	0.37	-4.04	0.00***
age	-0.67	0.12	-5.68	0.00***
genderf	0.10	0.28	0.36	0.72
static_featMelody	0.39	0.28	1.38	0.17
static_featRhythm	0.26	0.31	0.84	0.40
static_featTimbre	0.82	0.31	2.66	0.01**

p-values: *** < 0.001, ** < 0.01, * < 0.05

Table 2-G: Full model results from Experiment no. 4: grouping preference