Modelling Perception of Large-Scale Thematic Structure in Music

Submitted in partial fulfilment of the requirements of the
Degree of Doctor of Philosophy

by

Edward Thomas Ragdale Hall

School of Electronic Engineering and Computer Science
Queen Mary University of London

2023

Statement of Originality

I, *Edward Thomas Ragdale Hall*, 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.

Details of collaboration and publications

Work presented in Chapters 3 and 4 was published in the following paper:

Hall, E. T. R., & Pearce, M. T. (2021). A model of large-scale thematic structure. *Journal of New Music Research*, *50*(3), 220–241. https://doi.org/10.1080/09298215.2021.1930062

Work presented in Chapters 5 and 6 was presented at the following conference and is under revision for journal publication:

Hall, E. T. R., & Pearce, M. T. (2021). Perception of thematic structure in music over small and large scales. *16th International Conference on Music Perception and Cognition and 11th Triennial Conference of the European Society of the Cognitive Sciences of Music*

Hall, E. T. R., & Pearce, M. T. (*under revision*). Perception of thematic structure in music over short and long timescales. *Music Perception*.

Work presented in Chapter 8 was presented at the following conference:

Hall, E. T. R., & Pearce, M. T. (2023). Modelling the perception of large-scale order in music. *17th International Conference on Music Perception and Cognition and 7th Conference of the Asia-Pacific Society for the Cognitive Sciences of Music*

Abstract

Large-scale thematic structure—the organisation of material within a musical composition—holds an important position in the Western classical music tradition and has subsequently been incorporated into many influential models of music cognition. Whether, and if so, how, these structures may be perceived provides an interesting psychological problem, combining many aspects of memory, pattern recognition, and similarity judgement. However, strong experimental evidence supporting the perception of large-scale thematic structures remains limited, often arising from difficulties in measuring and disrupting their perception.

To provide a basis for experimental research, this thesis develops a probabilistic computational model that characterises the possible cognitive processes underlying the perception of thematic structure. This modelling is founded on the hypothesis that thematic structures are perceptible through the statistical regularities they form, arising from the repetition and learning of material. Through the formalisation of this hypothesis, features were generated characterising compositions' intra-opus predictability, stylistic predictability, and the amounts of repetition and variation of identified thematic material in both pitch and rhythmic domains.

A series of behavioural experiments examined the ability of these modelled features to predict participant responses to important indicators of thematic structure. Namely, similarity between thematic elements, identification of large-scale repetitions, perceived structural unity, sensitivity to thematic continuation, and large-scale ordering. Taken together, the results of these experiments provide converging evidence that the perception of large-scale thematic structures can be accounted for by the dynamic learning of statistical regularities within musical compositions.

Contents

List of Figures

Page

List of Tables

Acknowledgements

First and foremost, I wish to thank my supervisor, Marcus Pearce, for whose dedicated guidance and support throughout my studies I am immensely grateful. I would also like to thank the remaining members of my progression panel, Emmanouil Benetos, Simon Dixon and Elaine Chew, for the constructive feedback they have given during the critical stages of progression through this research.

I would like to give particular thanks to the members of the Music Cognition Lab at QMUL who have always offered valuable feedback, support and friendship, making my time as a Lab member hugely profitable and enjoyable.

Finally, the research presented in this thesis would not have been possible without the Centre for Doctoral Training in Media and Arts Technology and its members, which through my studies has provided the training and support needed, as well as providing the funding for this research (EPSRC and AHRC EP/L01632X/1).

Chapter 1

Introduction

Sensory input from the environment is rarely uniform but contains patterns that recur both exactly and approximately at a range of scales, allowing observers to learn and anticipate structural regularities. The same is true of human cultural domains such as language and music where large-scale temporal organisation has an impact on the meaning conveyed by an utterance or piece of music. In language, the emphasis is usually on communicating specific referential semantic content, which places strict requirements on word order and compositional hierarchy prescribed by syntactic considerations. In music, where the semantic content is usually much less specifically referential and stylistic syntax less prescriptive, more emphasis is often placed on the (often approximate) repetition of musical material throughout a piece of music, in general accordance with the schematic norms of a musical style. Examples of exact repetition are the literal reoccurrence of a chorus or re-entry of a principal theme, whereas approximate repetition often arises through progressive development and variation of a theme, reflecting an incremental process of change or developments upon successive reoccurrences.

Collectively, this general organisation of material within musical compositions are referred to as *thematic structure* (as opposed to other kinds of large-scale structures that can be present in music, such as tonal structures). From a philosophical perspective, Kivy (2017) argued that thematic structure arising from repetition of musical material constitutes the most distinctive characteristic of music, and an essential ingredient to its aesthetic experience. However, the perception of thematic structure in music is currently poorly understood to the extent that we cannot even be sure that such large-scale thematic structure can be perceived consistently by listeners.

Traditionally, composers and music theorists have argued in favour of the existence and importance of thematic structures in music, and have gone to great lengths to create, label and analyse them (Epstein, 1980; Galeazzi, 1796/2012; Meyer, 1989; Reti, 1978; R. Reynolds, 2002). Thematic struc-

ture is also—even when not explicitly stated—of importance to several highly-influential cognitive theories of music perception. This includes theories that involve structured representations of musical knowledge (for example, Lerdahl & Jackendoff, 1983; Temperley, 2007) and others that focus more on psychological processes of expectation (for example, Huron, 2006; Meyer, 1956; Narmour, 1990). However, despite this prevalent position in theoretical research on music perception, there is limited experimental research, and a resultant lack of empirical evidence, in support of the perception of large-scale thematic structure in music.

The research presented in this thesis constitutes an effort to rectify this situation. A probabilistic understanding of human cognition is used as the basis for the consideration of possible cognitive mechanisms most likely to allow large-scale thematic structure in music to be perceived. The research of this thesis has the underlying hypothesis that the perception of thematic structure is based on psychological mechanisms of statistical learning and probabilistic prediction of repeated structure in music—that such structures can be perceived through the statistical regularities that they form.

The approach taken by this thesis mixed computational modelling of cognition with empirical experimentation. The thesis hypothesis was used in the construction of a computational model that could provide a concrete specification of the cognitive processes involved in the perception of large-scale thematic structure. Through this formalisation of the hypothesis, features were generated characterising compositions based on their intra-opus predictability, stylistic predictability, and the amounts of repetition and variation of identified thematic material in both pitch and rhythmic domains.

A series of behavioural experiments examined the ability of these modelled features to predict listener responses to important indicators of thematic structure. Namely, similarity between thematic elements, identification of large-scale repetitions, perceived structural unity, sensitivity to thematic continuation, and large-scale ordering.

1.1 Research objectives and scope

Given this hypothesis and its motivation, this thesis has the following three general objectives:

1. To empirically investigate whether listeners can perceive large-scale thematic structures through the use of behavioural experiments.

2. To develop computational methods that can characterise the features of a given composition that are of hypothesised importance to the perception of thematic structure, based on the thesis hypothesis that the perception

of large-scale thematic structures can be accounted for by the dynamic learning of statistical regularities within musical compositions.

3. To examine whether these computational methods can predict listeners' behavioural responses to thematic structures, providing some insight into cognitive mechanisms underlying the perception of thematic structure.

In order to reduce the complexity of meeting these objectives, several practical constrains are introduced on the scope of this research. First, the study of thematic structure, in this research, is limited to the domain of western-classical tonal music. While the perception of thematic structure, and the proposed cognitive processes by which it may operate, are not considered to be culturally-constrained—indeed, many of the components contributing to the hypothesised processes, such as repetition, have evidence of being musically universal (Nettl, 2010)—the focus of the preceding literature and computational tools are overwhelmingly focused on this domain (largely based on their strong underpinnings in western music theory and its terminology). Second, music is considered at the symbolic level, in which compositions consist of a finite number of discrete features representing pitch or temporal properties of notes; this provides the assumption that the cognitive processing of thematic material and structure takes place at a higher level to the processing of auditory input. Third, the present research is limited to monophonic representations of melody, or polyphonic material containing a small number of independent voices under certain assumptions (see Chapter 3, Section 3.3).[1] Fourth, all pieces of music are considered from the perspective of a novel listener; it is outside the scope of this investigation to consider elements perception over repeated exposure to the same composition. Finally, the modelling of thematic structure presented in this thesis is aimed at the formalisation of plausible cognitive mechanisms leading to the perception of thematic structure; while this model does provide practical methods for processes such as repetition and theme detection, these are envisaged in their purely cognitive setting—the more general application of these processes is outside of the scope of this investigation.

1.2 Thesis outline

Background and related work

Chapter 2 contains a summary of the existing psychological and computational research relevant to examining the perception of large-scale thematic structure in music. The chapter, firstly, reviews the research from within the field of music

[1] Music in which voices can be considered as separate auditory streams.

perception that explicitly investigates the abilities of listeners to perceive large-scale structures, and that has produced such inconclusive findings, before summarising research into the perception of thematically-relevant properties and the computational techniques used to characterise them.

Statistical modelling of thematic structure

Chapter 3 presents the probabilistic computational model of the perception of thematic structure, based on the hypothesis that thematic structures can be perceived through the statistical regularities they form over the course of a composition. This chapter introduces IDyOM (Pearce, 2005), a computational model of auditory expectation, which provided a platform for the modelling of thematic structure, summarises the main components of the model itself, and presents four model-based measures of thematic structure to characterise properties of thematic structure for any given composition and to provide the foundation for empirical research.

Empirical experiments

Chapter 4 provides the first step in the testing of the thesis hypotheses through the analysis of the computational measures of thematic structure, when applied to a corpus of 623 full-length monophonic compositions. As there are no existing annotated corpora of large-scale structure, this analysis aimed to demonstrate that measures of hypothesised importance to the perception of thematic structure vary systematically when applied to the corpus, reflecting the inherent variation in structure present within the corpus itself.

Chapter 5 presents the first behavioural experiment testing the ability of computational measures derived from those of the model to predict listeners' perception of relationships on a small time-scale. This experiment aimed to understand how modelled features interacted and influenced responses is isolation as an important first step to understanding their effects when integrated into compositions over far longer timespans.

Chapter 6 presents a behavioural experiment testing listeners' perception of two highly-important indicators of large-scale thematic structure—identification of large-scale repetition and the perception of structural unity within compositions—over compositions of 2 minutes in duration. This experiment aimed to test how modelled statistical features could predict listeners' performance in these two tasks.

Chapter 7 presents an experiment that tested listeners' judgements of large-scale continuation. Through this task, the experiment aimed to investigate the relative importance of statistically learned features, informed by the model, on the perception of structure over entire compositions.

Chapter 8 presents a final behavioural experiment in which a puzzle-based task was used to test listeners' sensitivity to large-scale ordering of material. This experiment aimed to evaluate the performance measures of thematic and tonal structure as predictors of participants' orderings. The chapter additionally presents a computational equivalent experiment identifying patterns of ordering within existing compositions.

Summary and conclusions

Chapter 9 includes a summary of the research presented in this thesis, a discussion of the key findings across all of the experimental research, the limitations and scope of the research, and a discussion of promising directions for developing the contributions and addressing the limitations in future research.

1.3 Research contributions

Following the objectives presented in this introduction, and the interdisciplinary nature of this topic, the research reported in this thesis makes original contributions within the field of music perception and cognition and several related areas. The model presented in Chapter 3, and evaluated and expanded upon in the later empirical experiments, provides the key theoretical output of this thesis. The model brings together theories from traditional music theory and analysis, experimental psychology, and computer science in the development of cognitively appropriate methods for characterising thematic properties of compositions.

The behavioural experiments presented in Chapters 5, 6, 7 and 8 contain the key empirical contributions of this research. Through these experiments, this research builds on, and contributes to, a growing existing understanding of statistically learned elements of perception, extending it to psychological processing of large-scale thematic structure in music perception.

Chapter 2

Background and Related Work

2.1 Overview

In this chapter, a summary is given of psychological and computational research relevant to examining the perception of large-scale thematic structure in music. The chapter, firstly, reviews research from within the field of music perception that explicitly investigates the abilities of listeners to perceive large-scale structures. These studies are divided into two broad subcategories, those concerning thematic structures—the topic of this thesis—and those chiefly concerned with tonal structure, however, that still have some bearing on thematic elements. This limited body of behavioural experiments provides the closest background to the behavioural research presented in this thesis.

Secondly within this chapter, psychological research investigating the perception of musical features that relate to how large-scale thematic structures may be perceived is reviewed. Work is considered that tests the abilities of listeners to perceive similarity between excepts taken from within compositions, and examines listeners' abilities to perceive repetition.

Thirdly, computational methods are considered that seek to provide techniques for the modelling of similarity, categorisation, theme and repetition detection, and phrase boundary detection. These methods, while not directly providing cognitively-relevant or complete accounts of thematic structure, are useful tools for comparison and in forming the cognitive model of thematic structure in Chapter 3. Finally, in anticipation of this model, research that considers elements of structure in the context of statistical learning is summarised.

2.2 Structure in music

The term *structure*, when applied to music, is often used to describe a huge range of different features existing across a whole spectrum of varying magnitudes of scale. At its heart, it is the organisation of musical material that is taken as the

structure; this may be on the level of the arrangement of pitches in a chord, local harmonic features, voice leading, patterns in melodic phrases, metre, or grouping structures, and can span all the way up to global musical systems such as form, tonality and style.[1]

The concept of structure that is of principal concern to the research presented in this thesis is that most akin to musical form—a composition's large-scale organisation of material, made up from any of melodic, rhythmic or harmonic elements (Whittall, 2001). *Form*, however, is heavily imbued with the ideas of western-classical music, often implying a relationship between the construction of a work and certain tonal and stylistic conventions (for example, both of which are embodied in sonata form); indeed, for the purposes of the present research, it may be better to consider any individual piece as having its own unique form. It is *structure* (on a large-scale), therefore, that can be more generally applied, particularly with an emphasis on structural *unity*—the extent to which all elements of a piece can be considered to form a coherent unified whole. *Form* becomes the discrete sectioning of the large-scale structure, and the relations between such divisions (Salzer, 1962, p. 223).

Further distinction within large-scale structures can be made between *thematic* and *tonal* structures, both relating strongly to their own notions of closure. It is the properties of *thematic structure* that this research plans to explore—the perception that each component of a composition's thematic material relates convincingly to the others and to the whole, that every part 'belongs where it is' (Aschenbrenner, 1985, p. 159).

The primitive concepts of the structuring of music are ancient in their origin, being a fundamental part of any composition process; in producing a compositional work, certain levels of organisation and strategy must be put in place (Reti, 1978; R. Reynolds, 2002). Musical analysis has long sought to understand these structural elements and develop theories of music that follow accordingly. The feedback loop created between the analysis of monumental works and the pedagogy of composition results in marked conventions emerging. Some theories of analysis may prove to be cogent arguments for the workings of cognition—the creation of music and its analysis are tightly entwined—however, it should be used to act more as an informant of the areas on which to focus, rather than as creating a 'folk psychology' of music (Cross, 1998).

2.3 The perception of large-scale musical structures

While the perception of structures in music covers a huge span of music cognition research, those specific to the large-scale thematic structures that are the

[1] And even beyond, rising to the level of cyclic forms in which thematic material is repeated and developed across movements in symphonies or across entire operas.

target of this thesis are relatively scarce. The origins of the studies that do address this kind of structure often stemmed from a wish to test the empirical validity of long-held music-theoretical convictions: do examples following such highly-prized concepts perform better in behavioural tests of perception than those in violation?

At their simplest, early studies into the importance of large-scale structure in music sought to understand any perceived change in enjoyment or sense of overall unity if the structural elements of a work were presented in different orderings—some simply on the level of cyclic form between movements or discrete variations. This often resulted in binary conclusions being made on the ability of listeners to perceive large-scale structures, providing no intermediate detail or insight to the cognitive mechanisms at work.

Following a study in which movements of Beethoven piano sonatas and string quartets were rearranged with little effect on enjoyment (Konečni, 1984), Gotlieb and Konečni (1985) tested the abilities of listeners to perceive differences in arrangements of the variations in J. S. Bach's *Goldberg Variations*,[2] with three versions rated on 15 subjective feature scales (for example, pleasing/not-pleasing, surprising/not-surprising, emotional/not-emotional). Out of the different factors for all three versions, only one significant effect was found—the original ordering had a significant preference on the warm/cold scale. Overall, they concluded that changing the order of the variations in this piece had only a minimal effect on the composition's impact.

However, this result may possibly have been influenced by the small sample size used in the experiment; only 14 participants of mixed musical abilities (however, mostly non-musicians) were tested. The nature of the work chosen could also provide some grounds to question whether any strong preference for the original should be hypothesised on the basis of music theory; the individual movements of the *Goldberg Variations* can function individually as complete works in their own right[3]—there is no line of thematic developments or continuation between movements.

Some criticism of these conclusions was drawn from the musicological side of the aisle—contesting that altering the structure of a highly well-known work, such as the first movement of Mozart's Symphony No. 40 (so testing intra-opus, rather than cyclical structure), for musically-trained listeners, would *surely* decrease enjoyment of it (Batt, 1987). Karno and Konečni (1992) conducted a more rigorous experiment in response, making structural interventions to the symphony movement as suggested. Five versions (including the original) were created following rearrangements of the sections of its sonata form. Three rat-

[2] Bach's *Goldberg Variations* differ significantly to the later ideas of classical or romantic 'theme and variations' form. The concept of a theme reoccurring and being constantly developed in each subsequent variation does not exist here.

[3] And are often performed as such.

ings of pleasure, the wish to own a recording, and interest were recorded for each, as well as an explicit question as to which version they perceived had the best structure.

Again, only a small sample size is used—11 musician participants—for whom no significant effect of version on the scales measured was found. However, for the untrained listeners, with a much larger sample size of 42, some minimal effects were reported. For these participants, when the original version was played first it was rated higher on all scales. However, overall the original was not the most preferred, it was only preferred over three of the five versions. It is interesting to note that, for all three stimuli order groups, the first version heard scored highest in all factors, indicating primacy effect biasing listeners towards the first version heard.

In a similar experimental paradigm, Tillmann and Bigand (1996) opted for the division of compositions into chunks of around 6 seconds in length, using piano works by Bach, Mozart and Schoenberg. These chunks were either arranged in a forwards order (*i.e.*, the original) or in a retrograde version, with the intention that that small-scale structures could remain intact and still make musical sense while those on the large-scale would be disrupted completely. Forty participants with minimal musical experience were used—half listened to the original three compositions, rating their expressiveness on 27 scales—the other half completed the same task for the retrograde stimuli. The findings showed little difference between perceived expressiveness in the original and retrograde versions. Only in the Schoenberg was some significant effect observed, with the modified version being deemed less expressive. Tillmann and Bigand (1996) therefore concluded that the perception of expressiveness was strongest within the short, local chunks but became much weaker beyond that.[4]

An interesting addition to this field of study is that of Tan and Spackman (2005), in which a particular focus is made specifically on structural unity. The paradigm of the previous studies was modified to include more comprehensive disruptions of musical works and a large range of compositions. Fifteen short piano solos were used: five intact; five with one section from a composition repeated thrice; and five made by the combination of three sections from three different works.[5] A rating of each piece's 'unity' was given by 20 participants who were additionally asked qualitative questions as to the features that they thought created unity, and those that disrupted it. The ratings did show some sensitivity (albeit within the small sample size) to the varying degrees of repetition and variety in the stimuli. Musically-trained participants (half of the

[4]Further review of the perception of musical structure—comparing both large and small—to this point is given in Tillmann and Bigand (2004).

[5]Works were used by Bach, Blitzstein, Bowles, Chopin, Copland, Godowsky, Hofmann, Liebermann, Liszt, Ravel, Ravina, Rogers, Schumann, Shostakovitch, Tausig and Tchaikovsky. Participants were not made aware that any of the works had been altered.

participants) had a focus on repetition, themes and motifs when answering the descriptive questions showing some influence of music theory—at least in the language used. A later experiment of Tan et al. (2006) replicated this effect for original versus patchwork versions of compositions, but found that the effect diminished over repeated hearings

Relatedly, expanding the styles of music studied outside that of western-classical tonal music, Lalitte and Bigand (2006) found significantly higher coherence ratings for original over reorganised versions of contemporary and popular music and McAdams et al. (2004) found an influence of large-scale organisation on listeners' continuous ratings of intra-opus familiarity between two differently-structured versions of a contemporary composition.

Looking more specifically at 'inner form', Eitan and Granot (2008) used stimuli of the intact opening movements of two Mozart piano sonatas and a hybrid that mixed corresponding sonata form sections of the two movements. Participants rated examples on aesthetic scales, mostly relating to the works' perceived interest, whether 'the version is a masterpiece', but also one of coherence. One hundred and sixteen participants were used—54 musically trained. No significant preference for the original versions was found.

The lack of any statistically significant evidence for the ability for listeners to perceive structures on this scale may indicate that the kind of experimental paradigm used up to this point is unable (at least for the participant sample sizes used) to properly account for effects of large-scale thematic structures and their ability to create a sense of unity. There is also the additional problematic assumption in many of these studies that the original ordering of the piece is indeed the absolute best possible in terms of unity and structure; that the composer is infallible and lack of preference for original version equates to no preference for large-scale structure at all. When combined with the fact that relatively small groups of participants were used in many of the studies, these issues may account for the difficulty in finding a difference in the perception of modified and original structures—differences in structure that may be too subtle to give a large enough effect. Similar problems can also be found in the application of this paradigm to the perception of song form in popular music (Rolison & Edworthy, 2012).

Two linked studies by Granot and Jacoby (2011) and Granot and Jacoby (2012) employed a different paradigm, partly in response to some of the problems arising in the prior research, particularly removing the focus on the composer's original. The task took the form of a musical puzzle—ten disordered sections from a Mozart piano sonata first movement (2011) and eight from a Haydn sonata first movement (2012) were presented to 87 and 82 participants, respectively, who were tasked with creating an order that best made a coherent whole. The analysis of the participants' orderings was not focused on their

relation to the original, instead seeking patterns between participants. The approach of Granot and Jacoby (2011) yielded encouraging implications for listeners' abilities to perceive large-scale structure, and the importance of thematic considerations; results indicated sensitivity to (form-like) structure, grouping and placement of developmental material, and placement of opening and closing gestures. Distance score measures of Levenshtein (1966) edit distance and 'arrow of time distance' were applied between participant orders and revealed some sensitivity to 'directionality'; there was agreement as to the relative positioning of sections, even if not in the exact order.

Further to these behavioural findings, Farbood et al. (2015) found evidence of differences in neural processing in musicians listening to an excerpt of a Brahms piano concerto in its original form and versions scrambled at measure, phrase, and section levels. Functional magnetic resonance imaging responses showed a hierarchy of auditory processing within the brain, with only the original version giving reliable responses at the top-most level. The work gives evidence as to the ability of listeners to make structural connections over larger musical sections.

2.3.1 The perception of tonal structures

Although it is large-scale thematic structure that this research is primarily concerned with, it has been customary in the study of the perception of large musical structures to combine both elements of thematic and tonal structure. In many cases, the respective differences and properties of the two are never satisfactorily disentangled. For many, their emergence from western-classical musical theory and analysis—in particular, those with a desire to test long-held theoretical concepts by empirical behavioural testing—creates an emphasis on notions of form. In form, the conventions dictating the ordering of thematic material and its development go hand-in-hand with an ordered progression of related tonal areas; sonata form, being one of the most prescriptive in terms of material content and key, and revered in traditional music theory, is habitually used as the large-scale structure in question. However, work in this area can still inform our current research; information on non-tonal large structure can be revealed in this body of literature, and the effects of tonal structure should be considered as requiring experimental control for experiments investigating thematic structures over the same timescales.

Cook (1987) was one of the first such studies looking at large-scale intra-opus structure, and one that became quite influential, giving rise to much of the work in a similar vein. In this experiment, the modification of the tonal structure of a work was reported to have little significant impact on listeners' preferences. Cook particularly led with the idea that there could be a maximum time

scale in which the structure could have an influence on a listener's perception—particularly for tonal closure. Criticism of both conception and of execution were levelled at Cook (1987), particularly by Gjerdingen (1999). Among such issues are some largely applicable to much of the similar literature—the insubstantial participant sample size (here 19), and the importance placed on the subsequent failure to reject the null hypothesis (that structure is not perceptible at this timescale) through a lack of significance.

Subsequent studies that primarily intended to examine tonal closure, but were implemented through the manipulation of formal musical units, produced similar findings. The second experiment in Marvin and Brinkman (1999) applied structural variations to six keyboard works by Handel, disrupting both thematic structure and key progressions. Thirty-three participants performed no better than chance when identifying beginning and ending key as the same. This study has much in common with the previously discussed early work on non-tonal structure and faced many similar problems.

More recently, research into global tonal and harmonic structure—and that also involves non-tonal structures—have reported significant evidence in support of perception of organisation on this scale. Koelsch et al. (2013) harmonically impeded the hierarchical structure of Bach chorales while leaving local structure intact. EEG brain responses differed between regular hierarchical and disrupted stimuli, indicating cognitive processes are present that can deal with dependencies over the large time spans needed. This finding is supported by the probe-cadence study of Woolhouse et al. (2016) over temporally non-adjacent keys; listeners could hold non-adjacent keys in memory 'globally' (for over ten seconds). However, similar work by Farbood (2016) and Spyra et al. (2021) found little evidence that such effects of tonal memory persisted after around 20 seconds.

2.3.2 Summary

In summary, existing empirical research into large-scale structures in music has produced conflicting evidence in support of their perception. The work of McAdams et al. (2004), Tan and Spackman (2005), Tan et al. (2006), Lalitte and Bigand (2006), Granot and Jacoby (2011), Granot and Jacoby (2012), and Farbood et al. (2015), all provided results that suggest that listeners could perceive differences between musical stimuli in terms of large-scale structure. The work of Karno and Konečni (1992), Tillmann and Bigand (1996), Eitan and Granot (2008), and Rolison and Edworthy (2012) all found no evidence to suggest such structures can be perceived. Although the collection of studies that found evidence in support of the perception of specifically thematic structure is smaller still, several of the more recent approaches are among them. Even so, the un-

derlying mechanisms that may facilitate the perception of thematic structure are still little researched.

2.4 Structural similarity and repetition

The lack of much substantial research into the properties of large-scale thematic structure and its ability to create a sense of unity is the target of the research presented in this thesis. Due to the difficulties that studies encountered and that have hampered the field's development, reviewed above, this research aimed to provide a new approach that could provide the apparent nuance needed for the potential effects of these structures to be understood. Instead of attempting to disrupt thematic structures through the scrambling of example pieces, this research aimed to investigate the possible psychological processes that could facilitate a perception of thematic structure.

Aside from composition-wide measures, such as those used in many of the behavioural experiments reported in the previous section, the effects of structure can be examined within the composition itself, in the form of the internal relationships that underlie the perception of thematic structure. Two such relationships have received attention: perception of similarity and direct repetition of musical material. Both repetition and similarity are widely thought to underlie perception of thematic structure in music and contribute to properties such as unity in both hierarchical and linear theories of musical structure (Deliège, 2007; Kivy, 1993; Schoenberg, 1967; Smyth, 1993).

2.4.1 Similarity and categorisation

Similarity is considered a fundamental process in psychology and cognitive science as it allows for the classifying of perceptual phenomena into useful categories (Goldstone & Son, 2005; Shepard, 1987). Similarity and categorisation in the music perception literature encompass a vast field, and one that has indispensable elements when considering repetition and structure. Similarity can help to explain how thematic repetitions can be subjected to variation yet still retain their connection to an original, integrating into such theories of structure. Studies of similarity are wide-ranging and, in some cases, quite disparate. This may be due partly to the highly context-specific nature of similarity in music, particularly when judging between material arising from the same work (Ahlbäck, 2007; Cambouropoulos, 2009). However, several offerings of similarity research provide noteworthy perspectives on themes, repetition and structure.

To investigate similarity as an indicator of thematic structure, research has attempted to uncover the importance of different types of musical feature when making similarity judgements, using the following experimental logic. If

the evidence suggests that similarity judgements are made purely on the basis of surface properties of the music—such as its general dynamics or texture—compared with deeper comparisons of its thematic material, the case for listeners' perception of large-scale thematic structure would be weakened. The results of experiments following this approach are ambiguous and require further research.

One particular study that is of relevance to the ability of listeners to perceive connections between themes and their repetitions is that of Welker (1982). Using a set of transformations, based on five rules, a set of variants of a melody was created. Without being presented the original, participants were tasked with recreating the melodic contour of the central tendency of the variant set. Participants were found to be able to recreate the contour of the original melody successfully.

Lamont and Dibben (2001) conducted an investigation into the ways in which similarity relationships are perceived in different kinds of music, and how these relate to motivic structure. Participants (of both musically trained and untrained backgrounds) were asked to rate the similarity of pairs of extracts taken from two piano works—by Beethoven and Schoenberg—and provided descriptive adjective ratings for each. Both trained and untrained listeners' similarity ratings were roughly equal but were context specific. Similarity ratings were primarily found to be based on surface-level features—such as dynamics, articulation, texture and contour. For Lamont and Dibben (2001), however, motivic relations belong to the 'deeper' features and so were not among those perceived by listeners. While this may be true for some of the views of motifs based in more traditional musical analysis and theory, if motifs are seen more as salient cues that are extracted then they operate more on the musical surface (Deliège et al., 1996). Lamont and Dibben (2001) do suggest that a lack of significant evidence to support judgments based on thematic and motivic similarity may have been due to the short-term nature of the stimuli used; with greater exposure to repetition, affording a stronger thematic structure and development, valuable contexts for judging similarity may be learned. The context-specific nature of intra-opus similarity is highlighted here by the results from pieces of highly contrasting styles—in many cases, the two works create their own similarity criteria.

Similar conclusions of the use of musical surface features in the judgment of similarity were made by Mélen and Wachsmann (2001) and Koniari et al. (2001), testing the abilities of infants from 6 to 10 months and children aged 10 to 11, respectively, to categorise musical motifs in compositions by Schubert and Diabelli. They concluded that categorisations primarily used musical surface features, such as melodic line, register or dynamics, however, counter to Lamont and Dibben (2001), also used elements related to the underlying har-

monic structure, such as harmonic properties.

Counter to the findings of Lamont and Dibben (2001), Ziv and Eitan (2007) likewise conducted an experiment of similarity perception using the same stimuli. Participants, for each composition individually, were tasked with categorising and rating extracts as belonging to one of the given compositions two principal themes. By comparing listeners' responses to those collected by Lamont and Dibben (2001) and to published musicological thematic analyses by experts, significant agreement was found for excerpts from the Beethoven composition, but not for the Schoenberg. Characteristic musical features for Beethoven's two themes—such as phrase structure, melodic intervals, voice-leading, melodic schemas, and rhythmic features, and Schoenberg's—tone-row structures—were gathered. Correlations between ranked combinations of these features and those provided by participants' thematic categorisation ratings were tested for, with features of texture, rhythm, melodic contour and dynamics corresponding best to listeners' results. These results corroborate other findings within research of the perception of musical similarity, that established the relative importance of mode, contour, and rhythmic elements (in descending order) in the judgment of melodic similarity (Bartlett & Dowling, 1980, 1988; Halpern et al., 1998). These findings provide an alternative explanation for the findings of Lamont and Dibben (2001); surface features were the most important in the similarity and categorisation judgements tested, however, it is possible that these surface features could also facilitate the necessary thematic categorisation.

Volk and van Kranenburg (2012) investigated the musical features used by experts in their categorisation of Dutch folk melodies (the *Meertens* tune collection) into tune families. As the precise historical contexts of the melodies are lost, experts formed tune families based on their expert intuition of similarity between melodies. In an annotation study, experts provided annotations of feature categories used and ratings of the strength of these judgements. Analysis of annotations found that, in general, short characteristic motifs were the most relevant to experts' judgements—more so than features such as global melodic contour, often used in the description of melodies. These findings provide evidence that motif-like repeated patterns are highly important for similarity judgements in music.

2.4.2 Repetition

The idea that the structure of a work is built up by a listener throughout a composition by the repetition and reaffirmation of certain salient fragments has been suggested several times in music perception in a few different manifestations. Here, these salient features or landmarks will be described as *thematic*

material, or *motifs*.[6] While the basis of this theory—that it is the repetition of material that builds up the structure—has been suggested constantly, the theoretic reasoning as to the precise nature of how variation and development actually can exist often is disputed.

Adam Ockelford's extensively developed *Zygonic*[7] theory of music (Ockelford, 1991, 2004, 2005, 2006, 2009, 2010) proposed that: 'the cognition of structure stems from a sense of derivation arising from the presence of repetition in certain contexts' (Ockelford, 2006, p. 81). The repeated instances of material are related by similarity judgements of the listener and can be classed as one imitating, or deriving from, the other. 'Zygonic relationships' can be used to connect similarities in pitch and rhythm, as well as larger distributions of pitches and auditory qualities such as loudness. The most useful and convincing examples of strong similarity relationships are those given by material that may be considered motivic landmarks—a perspective informed, and supported, by work in cue abstraction by Deliège et al. (1996).

Although this theory sets out how motifs can generate a larger structure, the nature of, and contexts needed for similarity tend to be purely a descriptive list of operations taken to translate from one motif to another—noting the removal or addition of notes and their pitches, transposition, and manipulations such as retrograde. The supposed similarity is primarily identified by the analyst, often conforming to previously held notions of music theory; while some attempt is made to consider the cognitive processing of structure, Zygonic Theory only really succeeds as a method of formal analysis.

Huron (2013) explores hypothetical strategies of composition, based on balances between repetition leading to processing fluency (and so positive pleasure) and habituation (resulting in a reduction in responsiveness). Three such strategies were derived from these principles: (1) a 'trance' strategy involving high levels of repetition that can fully exploit pleasure induced from processing fluency, with the habituation causing an inward focus of a trance-like state (the effectiveness of this strategy may be highly-dependant on the listener and their cultural context); (2) a 'variation' strategy in which material is repeated with constant slight modification; and (3) a 'rondo' strategy involving sequences of (more exact) repetition that shorten throughout a composition, with new material introduced to limit effects of habituation. Due to the heavy reliance on the disposition of the listener for the trance strategy, the latter two were considered more likely to produce reliable psychological effects (with combinations of strategies possible). Huron (2013) identified that a characteristic feature of this final strategy would be that compositions would favour 'early repetition'—

[6] Borrowing the terms from music theory and analysis where they are often used to describe the structuring of a work in the composition process (Drabkin, 2001a, 2001b; Reti, 1978).

[7] A term derived from the Ancient Greek for 'yoke'—a union between two similar things (Ockelford, 1991).

repeating thematic material to a large extent in the beginning portions of the composition while processing fluency still outweighs habituation. An analytical study of a cross-cultural sample of music found evidence supporting the prevalence of early repetition.

In addition to these predominantly theoretical discussions, behavioural studies have also explored the repetition of thematic material and its contribution to structure. Pollard-Gott (1983) used an approach centred on the change in listeners' conceptions of a composition—how they are altered through repeated presentation of short passages taken from one single work. Experiments were devised to observe any variation in a listener's appreciation of harmonic, melodic, rhythmic and contrapuntal thematic categories. Musically-trained and non-trained participants were played passages and were asked to rate their similarity. Repeated listenings were conducted over three sessions. After the similarity test was completed, participants were informed of two themes in the extracts. A mixture of old and new stimuli were then played and listeners were to identify to which theme they belonged. A final task asked the participants to rate the excerpts on 15 bipolar adjective scales. The repeated testing found that listeners perceived the relationships among passages corresponding with more global thematic structures after repeated exposure. However, this failed to manifest with any significance after only a single exposure to the music.

In a study by Margulis (2012) (with further discussion in Margulis, 2014) that investigated the repetition of short musical stimuli, listeners were asked to detect—by the press of a button—exact repetitions of material within excerpts of between one and two-minutes long, taken from four classical recordings. Responses indicated that participants found it difficult to actively identify repetitions on a first listening, with the likelihood that participants would correctly identify a given repetition only marginally above chance for the best-performing excerpt. However, not all types of repetition were equally salient; shorter repeating units were more easily identified, with an optimal duration of units of approximately 6 seconds. Additionally, within-phrase repetition was identified more often than repetition of material across phrases, while complete phrases were more easily identified than fragments. It was also found that extra exposures facilitated better repetition detection for longer units; conversely, detection for shorter units became impaired.[8] These additional exposures had the effect of shifting attention towards larger time spans, possibly indicating how a motif could be established by frequent repetition over short time periods. After the initial exposure, the motif would no longer need to be repeated on such a small scale.

As a study into the basic features of repetition in music, Margulis (2012)

[8] While this research focuses particularly on intra-opus repetition, Margulis (2014) continues the with further in-depth discussion and review of repeated playing of individual pieces.

uncovers several interesting aspects that could be useful for the study of how repeated material could contribute the perception of thematic structures. The results for an optimal repetitive unit size could help to show the types of repetition that would be most successful in creating a sense of structural unity (although motifs are usually much shorter than the 6 seconds found to be optimal in this study). As Margulis (2012) identified, there were potential issues with this study, such as the limited repertoire. It was also discussed that similarity has a large part to play in the study of musical repetition and motifs. The task presented to participants— that of identifying when a repeat in material was heard—does not necessarily require the repetition to be exact; it may, therefore, be much more likely that some judgement of similarity needed to be used on the part of the participant.

However, despite the apparent difficulty of explicitly identifying repetitions in real time, there is evidence that repetition can implicitly influence the ways in which listeners perceive music. Margulis (2013) demonstrated a significant implicit effect of repetition on listeners' aesthetic responses to contemporary music—increasing enjoyment, interest, and judgements of artistic ability. Though not addressed directly in that study, repetition may also influence perception of thematic structure in a similarly implicit manner. Indeed, it is possible that the enhanced aesthetic experience of the music with increasing repetition actually depended on perception of increased thematic unity or coherence.

Outside of the purely musical context, research into the learning of repetitions in auditory stimuli provides strong evidence that listeners are both highly sensitive to auditory repetition and that repetitions are implicitly learned (Agus & Pressnitzer, 2013; Bianco et al., 2020). In a series of experiments, Bianco et al. (2020) tested the speeds with which participants could identify emerging repetitions in sequences of otherwise random tones. Reaction times significantly decreased after each reoccurrence of a repeated pattern. Of particular importance to the perception of music are the timescales over which these effects were observed. Repetitions were sparsely spaced within stimuli, occurring at intervals of approximately three minutes, with faster reaction times to repeated stimuli persisting for weeks after the initial trial. In the context of music, this spacing indicates that even a little repetition should be able to influence perception of structures at the level of a musical composition. Similar to the findings of Margulis (2013), Bianco et al. (2020) found these effects of repetition to be implicit, with participants being largely unaware of having any memory for the repeated stimuli.

2.4.3 Summary

In this section, empirical research into the perception of musical similarity and repetition in music was reviewed. Both similarity and repetition play important roles in the thematic structuring of music; if thematic structures are to be perceived, listeners must be sensitive to patterns of repetitions and internal relationships within musical compositions. This body of work provides evidence supporting the perception of similarity and repetition in music (Margulis, 2012; Ziv & Eitan, 2007), and gives some indications as to the musical properties that can best facilitate them. In particular, for repetition, the presence of early repetition (Huron, 2013), and repetition of whole phrase units (Margulis, 2012).

2.5 Computational modelling of musical properties

Due to the usefulness of some of the musical properties, discussed so far in this chapter, to a wide range of applications, computational methods attempting to replicate these properties have been developed across a number of different fields.

2.5.1 Similarity

The computational modelling of musical similarity, due to its importance to many sub-disciplines of the study of music, its usefulness in allowing for the classification of material, and its highly complex nature, has produced a wide range of different approaches. Similarity modelling has been the topic of specific journal issues (Toiviainen, 2007; Volk et al., 2016) and specialised workshops (Benetos, 2015).

It has generated a particularly large output of methods in the field of Music Information Retrieval (MIR), thanks, in part, to the 'symbolic melodic similarity' category of the Music Information Retrieval Evaluation Exchange (MIREX) competition (Downie, 2008; Downie et al., 2010), which provided the necessary 'ground truths' against which models could be evaluated. Broadly, these approaches can be categorised (based on the model taxonomy of Velardo et al., 2016) into: those based on cognitive constraints (de Carvalho Jr & Batista, 2012; Roig et al., 2013; Vempala & Russo, 2015); those based on music-theoretical principles—often using elements of Lerdahl and Jackendoff (1983), Narmour (1992), or Schenkerian analyses (Grachten et al., 2004; Orio & Rodà, 2009; Yazawa et al., 2013); those based on mathematical concepts (that were more abstract from the computational mathematics used by other approaches) (Aloupis et al., 2006; Bohak & Marolt, 2009; Frieler, 2006; Laitinen & Lemström, 2010; Lemström, 2010; Urbano, 2013; Wolkowicz & Kešelj, 2011); and those employing hybrid approaches, combining several different techniques (Frieler & Mül-

lensiefen, 2005; Müllensiefen & Frieler, 2004; Rizo & Inesta, 2010; Suyoto & Uitdenbogerd, 2010).

Of particular relevance to the approach taken to modelling thematic structure in this thesis are those, firstly, that can function as parallels to cognitive processes and, secondly, those based on a probabilistic understanding of music—such as those based on information theory-derived models (Laney et al., 2015; Pearce & Müllensiefen, 2017).

Pearce and Müllensiefen (2017) used an implementation of a 'compression distance' between two sequences—a string similarity metric based on the shortest program needed to compute one string, given another (Li et al., 2003).[9] The statistical model of IDyOM (Pearce, 2005) was employed to estimate the compressed length of musical sequences through the use of the Prediction by Partial Matching algorithm (Bell et al., 1990; Bunton, 1997; Cleary & Teahan, 1997; Cleary & Witten, 1984; Moffat, 1990). Pearce and Müllensiefen (2017) found that the compression distance metric provided a good fit to listeners' similarity judgments over three experiments. The metric also showed a comparable performance to the best-performing algorithms from the MIREX competition (Grachten et al., 2004; Orio & Rodà, 2009; Suyoto & Uitdenbogerd, 2010), with one configuration of the model achieving a higher average dynamic recall than these algorithms (Typke et al., 2005).

2.5.2 Theme and repetition identification

The ability to computationally identify patterns within music (more specifically, within symbolic representations of music) is an important task, with applications in computational musical analysis and MIR. The classes of patterns that the present research is chiefly concerned with are the detection of the principal themes within a composition and the identification of repeated material (or motifs).

The use of *theme*, where it is used in the research reviewed in this chapter (Pollard-Gott, 1983; Ziv & Eitan, 2007)—and in music theory in general—overlaps in many properties with that of motif, but the two may be distinguished in several ways. While motifs, in the context of this research, can be any small piece of salient material, theme can be something considerably lengthier; themes are customarily complete in their own right and, classically at least (and certainly in the previous research reviewed above), the original memorable material from which later repetitions, or motifs, are derived. Theme is quite often taken for granted in much of the research it is used—particularly in music perception. For example, in both Pollard-Gott (1983) and Ziv and Eitan (2007), principal themes were given, to which the connection of further material was

[9] This process is given in greater detail in Chapter 3.

to be judged. Themes were identified by traditional analyses or by ear; the actual cognitive mechanisms behind the process of theme detection are little studied.

Computational identification of themes in music is a task of some difficulty. Existing theme detection methods are themselves derived as specific cases of string-matching repetition detection; different selection criteria are used to identify themes from repetitions, often using the longest matching substring or the most frequent substantial repetition. Explicit methods for theme detection using the exact matching of substrings are given by Hsu et al. (2001), Meek and Birmingham (2001) (with some deviations in rhythm allowed), Wang et al. (2006), and Karydis et al. (2007).

Approximate matching methods also exist that find themes based on repeated similar material. Uitdenbogerd and Zobel (1999) used an approach to theme detection that produced a melodic string, against which theme queries could be compared. With the acknowledgement that melody queries may be inexact, several similarity measures were employed. Variations on edit distance (Levenshtein, 1966) were used: 'longest common substring'—the length of the longest contiguous string identified was used to rank the pieces—and using n-gram measures to count the matches of a certain length; 'longest common subsequence'—the queries were matched with no penalty for gaps of any size; and 'local alignment'—'dynamic programming' determined matching of strings on a local basis. Uitdenbogerd and Zobel (1999) found some of these techniques largely unsuccessful—in particular the longest common subsequence similarity function, and with problems caused by rests—when applied to a large MIDI database. For theme extraction based on repetition and similarity, local alignment performed best, followed by n-gram counting.

Methods for repetition detection in general for symbolic music have been produced for an array of purposes within MIR, and the MIREX category 'discovery of repeated sections and themes' (2013–17) provided an outlet for the testing of many approaches, given 'ground-truths' for five pieces (with summaries given of submissions and related approaches given in Janssen et al., 2014 and Ren et al., 2017 as well as the more recent work of Melkonian et al., 2019 and Laaksonen and Lemström, 2019, 2021). The task required models to output a list of patterns repeated within musical pieces—any subsequence of pitch and time that appeared at least twice—and accounting for patterns shifted in time or transposed. Approaches used in the symbolic version of the task covered geometric approaches, that treated melodies as a set of points in a multidimensional space (Chen & Su, 2017; Collins et al., 2013; Meredith, 2015); approaches using self-similarity matrices (Nieto & Farbood, 2014); those using underlying sequential representations (Lartillot, 2014); wavelets (Velarde et al., 2016); and machine learning (Pesek et al., 2017). Based on the evaluation metrics, the algorithms performed well in precision, recall, and F1-scores, however, none could fully

reproduce human-annotated patterns. Lartillot (2014) is notable among these approaches as being the only algorithm to operate sequentially along pieces.

Using the annotations of Volk and van Kranenburg (2012), Boot et al. (2016) employed a MIR approach to compare six pattern discovery methods (Collins et al., 2013; Conklin, 2010; Lartillot, 2014; Meredith et al., 2002; Nieto & Farbood, 2014) for identifying the repeated motifs that were the most relevant features used by experts. Melodies were then compressed so as to contain only the repeated identified material. Similarity metrics were then used to compare between compressed melodies and categories them into tune families, comparing their classification accuracy to the experts' annotations. Using their proposed framework for information compression based on pattern categorisation, expert annotations allowed for a compression of up to 60% with minimal loss in accuracy; this outstripped that of any of the computational algorithms, suggesting such methods still are not capable of competing with expert human classifications.

For the purposes of the research of this thesis, while these methods of computational pattern detection provide a useful background, they were not intended as simulations of possible cognitive mechanisms behind the perceptual task of theme and repeated-pattern identification. In particular, for those identifying themes, they operate offline, first identifying repetitions throughout a composition, then selecting the most likely theme. By contrast, in real-world listening to music, this process occurs online and dynamically during a single hearing. For a cognitive model simulating perception of thematic material on first listening, detection of themes must precede detection of thematic repetitions, and both must be achieved dynamically on a single passing of the composition. As it is beyond their original scope, some MIR theme detection methods have additional limitations for use in a cognitive model—many only consider finding a single primary theme for each composition in which many may exist, and there is also a significant difficulty in determining a theme's length.

2.5.3 Segmentation and phrase detection

A consistent problem in theme detection is identifying a musically meaningful theme length. In the case of repetition and similarity matching methods—particularly those searching for theme fragments—the start of a theme is found but its duration and endpoint are often unclear. To try to rectify this problem, these theme detection methods have been combined with some form of musical phrase-based boundary detection (Lu & Zhang, 2003; Takasu et al., 1999).

Melodic segmentation has been targeted in music cognition and MIR as a modelling task, subdividing melodies into small musically meaningful sections. While it is often the case that segmentation has no one exact correct answer—

listeners may disagree on the location of these boundaries—a formal rule-based criteria were presented by Lerdahl and Jackendoff (1983) in *A Generative Theory of Tonal Music*, deriving a set of rules from ideas of proximity and similarity in Gestalt psychology ('Grouping Preference Rules' or GPR). GPR 2 considers temporal proximity, with large rests of inter-onset intervals, relative to surrounding events, indicators of boundaries. GPR 3 considers similar large relative changes in register, dynamics, duration and articulation. These rules have formed the basis for some of the best-performing phrase detection models created.

The *Local Boundary Detection Model* created by Cambouropoulos (2001) uses local changes in IOI, rests and pitch, quantified to follow the GPRs. The Change Rule places a boundary strength that is proportional to the degree of change between any two consecutive non-identical intervals—with respect to IOI, rests, and pitch. A Proximity Rule gives a greater boundary strength for the larger interval between any two consecutive non-identical intervals. Peak detection, or a threshold, needs then to be applied to produce segmentations.

The *Grouper* algorithm given by Temperley (2001) segments a melody according to three Phrase Structure Preference Rules (PSPRs): the gap rule, PSPR 1, tries to locate boundaries at large IOIs or large offset–onset intervals; the phrase length rule, PSPR 2, favours phrases of a predetermined length—usually 10 notes; and the metrical parallelism rule, PSPR 3, favours boundaries occurring on the same position within a bar. In Cenkerová et al. (2018), *Grouper* and *LBDM* were compared to each other and a combined model of the two. While the compound model provided the most accurate detections, the accuracy of *Grouper* was not substantially lower.

These two phrase detection models, although no-longer recent, have been extensively tested and compared, both to each other and to newer competing models. More recent approaches to automatic segmentation have sought to produce models that are not based on meeting rules defined by *Gestalt* principles. Of particular interest to this research project are those taking information-theoretic approaches. Juhász (2004) uses a memory-based maximum entropy model to segment compositions, inferring a boundary either before an unexpected melodic event or after an event for which the continuation is hard to predict. Likelihood estimates generated by IDyOM have been demonstrated to be effective at melodic segmentation, marking boundaries at peaks in information content (Pearce, Müllensiefen & Wiggins, 2010). However, a comparison of IDyOM, Grouper and LBDM favoured the older rule-based approaches, *Grouper* in particular. Similarly, the model of Lattner et al. (2015) also takes an information-theoretic approach, using a restricted Boltzman machine to model the probability of melodic events, identifying boundaries at peaks. This model provided some small improvement over the IDyOM based approach.

Other approaches to segmentation have employed repetition identification

in compositions, based on the assumption that the start and end points of repetitions form section boundaries. Several of these approaches based on the identification of repetitions in musical sequences (Cambouropoulos, 2006; Rafael et al., 2009; Rafael & Oertl, 2010) but segmentation based on repetition has played an important role in segmentation and musical structure analysis using self-similarity matrices. A self-similarity matrix represents the similarity or dissimilarity between different segments or frames of a music piece, with respect to itself. Each element of the square matrix corresponds to a comparison between two time frames in the piece. In the audio domain, matrices can be computed based on the comparison of chroma and timbral features, and more successfully with compacted spectral representations such as constant-Q transforms and log-mel spectrograms (Müller, 2015; Nieto & Bello, 2016). Symbolic approaches have applied similarity measures between pairs of fragments exhaustively across a composition (Rodríguez-López and Volk, 2015, which used cosine similarity between vectors of pitch interval and duration ratios). Using these matrices, segments can be determined based on the combination of different properties (Nieto et al., 2020): (1) homogeneous regions found in blocks of high similarity (Ullrich et al., 2014); (2) repetition found in high-similarity diagonal paths (Rodríguez-López & Volk, 2015); and (3) the regularity of structures identified.

However, by the nature of their construction—the pairwise comparison of all time frames in a composition—these methods can not be considered to be analogous to the cognitive processes involved in the detection of repetition. Identification of structure and repetition from these matrices takes place after their construction, rather than dynamically throughout a piece.

2.5.4 Summary

This section provides a brief overview of computational research in a wide range of fields related to thematic structure. These include methods for characterising musical similarity, identifying repeated patterns, and detecting structural boundaries. While each of these topics has many different approaches—with all three being active areas of research in the field of music information retrieval—for the purposes of the research in this thesis, particular importance needs to be placed on those methods that can provide a convincing simulation of human cognition. For this to be the case, methods must be applicable to a first listening to a piece of music; they must be able to be applied linearly as a work progresses. Chapter 3 makes use of some of the methods introduced here—the similarity metric of Pearce and Müllensiefen (2017) and the phrase detection model of Temperley (2001).

2.6 Statistical learning of structural elements

The problems experienced in much of the literature exploring large-scale structure and its related properties of repetition and similarity mean that no convincing cognitive model of thematic structure has been produced. Much of this gap can be addressed by a probabilistic model; motivic salience and repetition can be understood in terms of predictions made by a listener as a piece progresses, their development can be viewed as variation according to intra-opus thematic rules or extra-opus stylistic congruity. The repetition, and so strengthening, of salient material in a statistical model can be considered as a mechanism by which large-scale structure and coherence is achieved. A model aiming to fulfil these processes is described in detail in Chapter 3.

As was found in the similarity literature, many of the features important for the perception of large-scale structure are highly context dependant. These dependencies could be categorised in two ways: the first as stylistic conventions—some variation is made in ways that are stylistically predictable, such as cadential embellishments in classical-era music, and so may still be judged as highly similar in the context of style; the second involves piece-specific rules—particularly how the work's motifs produce a sense of thematic development.

The application of statistical modelling to music has proved itself to be successful in creating predictions for pitch likelihoods based on style (Pearce, 2005, 2018). While the accurate modelling of statistical properties of style is not needed to achieve the aims of this research—all that is required for stylistic elements to be accounted for, rather than distinguishing between disparate styles—the contextual information it can provide is valuable. This can prove highly useful for determining the stylistic variation of motifs, allowing us to find the extent to which motivic embellishments can be predicted by a piece's style. The idea of style in this way is proposed by (Meyer, 1989) as the cultural constraints dictating compositional decisions—giving rise to notions of harmony and tonality, as well as motivic features in melody. The transformation of these constraints into grammatical probabilities through exposure to a music generates the syntax we consider to be the style (Pearce & Rohrmeier, 2018; Rohrmeier & Pearce, 2018).

Aside from stylistically-congruent embellishment of thematic material, material can also vary in terms of the individual thematic development of a composition. This aspect provides the key to examining structure in this research, with little previous work tackling this concept. Temperley (2014) looks at thematic development in music, characterising its variation in terms of 'information flow'. A corpus analysis of folk song themes with intra-opus repetition shows an expansion in complexity of material upon repeat—often with more chromaticism and expanded interval sizes. Temperley (2014) reasons this development is in order to maintain an optimal information density (negative log of prob-

ability), maintaining interest at a suitable level; he does not, however, take into account properties of stylistic variation.

2.7 Summary

This chapter presented a review of the elements of psychological and computational research most relevant to the topic of this thesis. The relatively small number of studies that specifically investigate the perception of large-scale structures in music prove to be rather inconclusive in their outcomes; as discussed, many face problems in sample size and assumptions in methodology, thus reporting no significant indication for listeners' understanding of structure in this way. The work of McAdams et al. (2004), Tan and Spackman (2005), Tan et al. (2006), Lalitte and Bigand (2006), Granot and Jacoby (2011), Granot and Jacoby (2012), and Farbood et al. (2015), all provided results that suggest that listeners could perceive differences between musical stimuli in terms of large-scale structure. The work of Karno and Konečni (1992), Tillmann and Bigand (1996), Eitan and Granot (2008), and Rolison and Edworthy (2012) all found no evidence to suggest such structures can be perceived. Although a small collection, smaller still when considering a focus on thematic structure, more recent approaches to this field have produced encouraging evidence to the contrary. Even so, the underlying mechanisms that may facilitate the perception of thematic structure are still little researched.

Research into the perception of musical similarity and repetition in music, when combined, provided some insight into the ways in which thematic structures may be perceived, and also uncovered additional concerns for this topic; the nature of similarity is highly context dependant. By creating a probabilistic model informed by statistical learning, large-scale thematic structure in music can be examined though repetition of salient material, provision of similarity contexts through stylistic congruity and thematic development, and emergence of large-scale hierarchical structure. Such a model can provide a foundation against which hypotheses of the cognitive mechanisms underlying the perception of thematic structures can be tested.

The following Chapter 3 aims to develop such a model, making use of some of the computational models summarised here in characterising similarity and identifying phrase boundaries. The behavioural experiments of the later part of this thesis aimed to test the model with reference to the behavioural studies of structure perception discussed in this chapter.

Chapter 3

Modelling Thematic Structure

3.1 Overview

This chapter sought to provide a concrete specification of the cognitive processes involved in the perception of large-scale thematic structure. In this chapter, it is proposed that statistical learning provides a plausible underlying mechanism for the perception of thematic structure. According to this proposal, large-scale thematic structures are perceived through implicit recognition of statistical regularities learned through both exact and inexact repetition and variation of material. Based on this proposal, a probabilistic computational model of the perception of thematic structure was implemented.

Within this chapter, first, the case for an understanding of thematic structure based on the learning of intra-opus statistical regularities is given, providing the motivation for the modelling presented here. Second, the chapter introduces IDyOM (Pearce, 2005), a computational model of auditory expectation which provides a platform on which the modelling of thematic structure is built, and discusses its relevant features and the ways in which this base is extended through the research of this thesis. Third, the main components of the model of thematic structure itself are summarised, including methods for the detection of thematic candidates and repeated thematic material, based on their intra-opus statistical regularities. Finally, the chapter presents a series of four measures of thematic structure, based on this model, that provide a foundation to underpin the empirical research of this thesis.

3.2 Statistical learning of thematic structure

Repetition and structure are intrinsically linked in music; the idea that, through the repetition and variation of material, large-scale structures can be created has a long history in music theory (Epstein, 1980; Meyer, 1989). The basic building blocks of this repeated material are small salient landmarks—or motifs—the

combination and variation of which create thematic development. The motivic structuring of music in this way is certainly no recent idea; the concept of the motif as the building block of a musical work can be found in many manuals of composition (even as early as Galeazzi, 1796/2012). Through such inclusion of repeated motifs, sections of a composition are explicitly linked, and a coherent sense of large-scale unity can be achieved across the work. It is also consistent with the implications of cognitively informed models of music, such as the hierarchical structures of Lerdahl and Jackendoff's (1983) *Generative Theory of Tonal Music*.[1]

Large-scale structure—the global organisation of a work's material—encompasses several different concepts. First, a distinction can be made between *thematic* and *tonal* structures, the first concerning the structuring of repeated musical material, the second the hierarchical organisation of harmonies relating to key. The effects of large-scale structures over a composition can be summarised by the term *unity* (or *coherence*)—the extent to which all elements of a piece can be considered to form a unified coherent whole—the perception of which, on the part of a listener, requires that a work's material is sufficiently closely related to be experienced as belonging to the same entity.[2] For both thematic and tonal structures, repetition of material plausibly leads to an increase in its perceived salience and a greater sense of unity. It is large-scale thematic structure with which the current research is concerned.

Repetition seems very likely to play an important role in the perception of large-scale thematic structure but has received relatively little attention in empirical research on music perception, producing limited convincing evidence (as reviewed in the previous chapter). In part, this may reflect the lack of a formalised model characterising the cognitive processes involved in perceiving thematic structure. The present chapter presents an outline of such a formalised model. The model was based on statistical learning (Saffran et al., 1999) and probabilistic prediction, since repetition plays to the strengths of such accounts of music perception. The model is based on the premise that motivic salience and repetition can be understood in terms of the positive effect they have on predictions made by a listener as a composition progresses, while motivic development could be viewed as variation according to learned regularities, based either on intra-opus (within a composition) thematic or extra-opus (between compositions) stylistic models.

The goal of the present research was to understand the perception of them-

[1] The *Generative Theory of Tonal Music*'s Time-Span Reduction Preference Rule of parallelism (TSRPR 4)—to assign parallel heads in the hierarchy to time spans if their material is similar—when applied on the scale of a complete composition, gives some account of large-scale structure created through repetition.

[2] Form, from music theory, is a specific manifestation of large-scale structure, often implying a relationship between the construction of a work and certain stylistic thematic and tonal conventions—such as in sonata form (Whittall, 2001).

atic structure by breaking-down the overall process into its component parts; to logically sequence how this process may function in cognition—and, equally importantly, where it fails, leading to the weak results of past behavioural studies (as previously reviewed). For this to be fulfilled, a probabilistic model of the cognitive mechanisms hypothesised to underlie perception of thematic structure in music was developed based on the *Information Dynamics of Music* (IDyOM) framework (Pearce, 2005, 2018). Through this model, a set of fundamental components were developed, with which thematic structures may be measured.

3.3 Modelling statistical learning

Cognitive modelling of thematic structure in real-world music provides advantages over the stimulus manipulations used in existing experimental work. We can avoid artificially manipulating a composition's structure—decisions that have to be motivated by some prior knowledge or expectation about the functioning of form—and compositions can be used in their original entity, greatly aiding ecological validity. In modelling large-scale thematic structure, it is possible to propose explicitly and test multiple hypothesised cognitive mechanisms, rather than trying to interpret the implications of one or more experimental manipulations for underlying cognitive mechanisms.

A probabilistic interpretation of music, founded in statistical learning, lends itself particularly well to this task. Using such a conception, we can construct a model of large-scale thematic structure from small base-units of repetition. Repetition can function as it does in practice, allowing for the inclusion of variation and embellishment. Employing statistical learning in this way also provides a fruitful way of operationalising perceived coherence or unity—high intra-opus predictability would indicate greater structural unity.

3.3.1 IDyOM

The present implementation of a model of thematic structure uses as its basis the *Information Dynamics of Music*, or IDyOM, framework to represent the statistical learning of musical material (Pearce, 2005, 2018). As a computational model, IDyOM aims to simulate the cognitive processes of statistical learning for symbolic representations of auditory sequences, generating conditional probabilities for sequential events.

Specifically, given a sequence e_1^k of symbolic events with total length k, IDyOM can be used to obtain the conditional probability of each event, given the preceding context and the prior experience of the model m. This likelihood can be estimated using a finite context of the n preceding events.

$$p_m(e_i|e_1^{i-1}) \approx p_m(e_i|e_{(i-n)+1}^{i-1}), \forall i \in \{1, \ldots, k\}$$

These distributions are computed using the Prediction by Partial Matching (PPM*) algorithm, a variable-order Markov model that tallies the occurrences of subsequences (or n-grams) of varying length within a training sequence, smoothing between predictions of different orders (Bunton, 1997; Cleary & Teahan, 1997). Due to this smoothing, the contexts used for prediction need not have a fixed maximum length.

Given the conditional distributions of estimated likelihoods returned for each note-event, a value of *information content*, h (measured in *bits*), provides a measure of unpredictability, or surprisal, of the note that actually occurs (where *p* is the estimated probability for the occurring event in the distribution), given the context and the prior experience of the model.

$$h_m(e_i|e_1^{i-1}) = -\log_2 p_m(e_i|e_1^{i-1})$$

Low information content indicates a predictable note-event—one where much of the information provided is redundant—whereas high information content indicates an unexpected note-event.

IDyOM may be configured as a *short-term model* (STM) which learns incrementally from an initially empty state within a given musical sequence—such as the pitches of notes in a melody—representing a listener's short-term acquisition of statistical knowledge about repeated structure within an individual piece of music (*i.e.*, $m_x = e_1^{i-1}$). It may also be configured as a *long-term model* (LTM), in which case it is trained on a separate set of musical sequences, representing long-term learning of the statistical structure of a musical style, before being applied to predicting the notes of a musical composition.

To date, IDyOM has been used to model perceptual expectation and uncertainty (Egermann et al., 2013; Hansen & Pearce, 2014; Hansen et al., 2016; Omigie et al., 2012; Omigie et al., 2013; Pearce, 2005; Pearce, Ruiz et al., 2010; Sauvé, 2018), boundary perception (Pearce, Müllensiefen & Wiggins, 2010), metre induction (van der Weij et al., 2017), similarity (Pearce & Müllensiefen, 2017), memory (Agres et al., 2018), emotional response (Egermann et al., 2013; Gingras et al., 2016) and aesthetic experience (Cheung et al., 2019; Gold et al., 2019).

While the hypotheses underlying the present approach are sufficiently generalisable that any statistical predictive framework for symbolic music could be applied, IDyOM (and, more broadly PPM) provides certain features that make it particularly advantageous. Firstly, the ability to configure separately an LTM, trained on the entire corpus, and STM, constructed only for an individual composition, allow a distinction to be made between predictions deriv-

ing from inter-opus stylistic (including tonal) knowledge, and predictions based on thematic structure within a work (reflecting its *internal unpredictability*). Secondly, the dynamic nature of IDyOM STMs provides an online model of music listening—all predictions are made sequentially as a work progresses—that can be used to simulate online continuous perception of music; a feature that precludes many techniques from Music Information Retrieval (see Chapter 2). Finally, models generated using different representations of the musical surface (detailed below) can be directly compared, providing insight into the specific representations that are most relevant to the perception of thematic structure.

An outline as to IDyOM can form the basis of a model of large-scale thematic structure is as follows. A composition with a large amount of repeated material will have a low average *internal unpredictability* (mean information content for the STM), indicating that it has high thematic unity. Thus, IDyOM usefully embodies the hypothesised links between repetition, prediction and thematic structure. Furthermore, inexact repetitions still have some degree of increased predictability by virtue of the variable-order smoothed Markov modelling of PPM*, and the multiple levels of abstraction provided by the use of different viewpoint representations. Embellishment of repeated material in accordance with stylistic conventions[3] can be accounted for by searching for material possessing a low information content for the LTM, making it stylistically coherent (with respect to the corpus), but relatively high information content in the STM, making it thematically less coherent.

3.3.2 Representations of the musical surface

IDyOM takes as its input musical sequences—as initially used here, monophonic melodies. Through its implementation of a multiple-viewpoint framework (Conklin & Witten, 1995), IDyOM has the ability to generate probabilities based on different representations of the musical surface. As the present research aimed to tease apart the musical features that have the most pronounced effect on perception of thematic structure, an initial selection of six pitch-domain and four rhythm-domain representations was made to best cover important features in these domains.[4]

In the pitch domain, the selection covered several different levels of abstraction:

- Pitch, the exact MIDI chromatic pitch number—unable to account for transpositional invariance

[3] An example is given by material becoming increasingly scalic when approaching a cadence in a work from the Classical era. Variation not due to this stylistic embellishment could be considered thematic development, taking place over the course of the piece.

[4] A complete list of the representations available to IDyOM is given at https://github.com/mtpearce/idyom/wiki/List-of-viewpoints.

- Pitch interval, the number of (directional) steps between pitch values—that is transposition invariant (*i.e.*, does not distinguish between the same material at different transpositions)
- Pitch contour, whether the pitch ascends, descends, or remains the same
- Scale degree', representing pitch relative to a tonal centre

The final two pitch representations used were linked representations that assume alphabets corresponding to the Cartesian product of the alphabets of their two respective components:

- Pitch linked with pitch interval
- Pitch interval linked with scale degree

Likewise, the representations of rhythm consisted of different levels of abstraction:

- Inter-onset interval (IOI), the time interval between event onsets
- IOI ratio, the relative size of an IOI compared to the previous IOI
- IOI contour, whether the IOI lengthens, lessens, or remains constant
- Event position in bar linked with bar length, the placement of events relative to the start of the current bar distinguishing metres with different bar lengths

This set of representations allowed for models to be used in competition with each other, comparing the ability of each to predict elements of thematic structure using a series of metrics generated within the model. All of these representations were used in examining the modelling of thematic structure when applied to a large corpus of western-classical melodies (Chapter 4), with single representations for pitch and rhythmic domains selected (pitch interval and inter-onset interval, respectively) for use in the modelling of behavioural data (Chapters 5, 6, 7 and 8).

While representations in both pitch and rhythmic domains were used here, it should be noted that rhythmic representations of music behave substantially differently to those of pitch. Specifically, it is perfectly viable in the rhythmic domain to have only a single note duration (or perhaps a single rhythmic pattern) for the vast majority of a composition, with the useful structure lying wholly in the pitch domain in these instances.[5] The effects of these isochronous or isorhythmic compositions have the ability to mask other effects of repetition and

[5]Such an effect is particularly common in the music of J. S. Bach, for example, the first prelude of *The Well-Tempered Clavier* or as can be found in his Cello Suites.

structure when using a rhythmic representation. More generally, for the styles of western-classical music used in the analyses of this thesis, it seems likely that pitch structure offers stronger influence on thematic structure than rhythmic structure.

3.3.3 Modelling polyphony

The PPM algorithm models sequences of symbolic events, providing event-by-event estimations of unpredictability. Likewise, when modelling music, IDyOM makes its estimations based on sequences formed from discrete symbolic representations of the musical surface; primarily, these representations consist of melodic pitch or rhythmic features (for example, modelling patterns in the pitch of note-events). As PPM models a single sequence at one time, IDyOM, when modelling melodic content, is applied to monophonic material.[6,7] While the syntactic structures of musical melody that IDyOM is concerned with, and the thematic structures hypothesised in the current research, exist in the musical surface (as the large amount of research using the framework has demonstrated), they are not necessarily limited to it. Monophonic music only forms a small portion of the music in the Western-classical canon, with polyphonic works being widespread—often incorporating multiple simultaneous melodically-relevant lines. For behavioural applications of such models, as the length of the musical sequences of interest increases (reaching the large-scale structures that are the focus of this thesis), and so retaining listener engagement and attention becomes a greater consideration, the ability to model polyphonic material would provide an important step towards even greater ecological validity.

The modelling of polyphonic music presents many complex challenges; the individual note-events that make up polyphonic material can be organised in multiple combinations in different dimensions. The music could be considered in its 'vertical' dimension—taking slices of all note-events occurring at a given point in time (which provides a harmonic representation, to varying degrees of abstraction)—or in a 'horizontal' dimension of multiple perceptually independent voices (Cambouropoulos, 2008). For the task of modelling thematic features in music, it is this horizontal aspect of music that is of chief interest.

For an example of the importance of the horizontal aspect in perceiving thematic elements, we can consider the processes needed in a composition con-

[6] Under the multiple-viewpoint framework used by IDyOM (Conklin & Witten, 1995; Pearce, 2005) multiple representations of the surface may be used to estimate likelihoods for events at any one time; however, crucially, while each event is represented multiple times with different representations, only a single sequence of events occurs.

[7] Polyphonic material can still be modelled by its harmonic content if all parts are reduced to a single sequence of symbols (*i.e.*, chords) representing the harmonic attributes at a given point (Harrison et al., 2020).

sisting of two perceptually-distinct voices. For a given repetition of material occurring in either of the voices, we are interested in how predictable the repetition is in the context of the composition up to that point—the context provided by both voices. This predictability, therefore, should not be dependent on the voice in which the original material occurred, but rather, each voice should be able to contribute equally to predictability of material in either voice.

This concept can be implemented by expanding the use of PPM modelling. Given N multiple streams, each of which may contain a different number of events, $(s_1^N)_1^{k_j}$, events can be modelled based on their preceding context in the same sequence.

$$p_m\big(s_{j,i}|(s_j)_1^{i-1}\big) \approx p_m\big(s_{j,i}|(s_j)_{(i-n)+1}^{i-1}\big), \forall i \in \{1,\ldots,k_j\}, \forall j \in \{1,\ldots,N\}$$

In applications in which no dynamic learning is needed—such as in the example given above, should we only be concerned with modelling the predictability of a single point in the composition—all of the distinct streams in the training material can be used as independent sequences in the training of a single model that is then used to predict the note-events of the target material, given its context (the preceding n-grams) in its specific voice (for example, this could be achieved using an IDyOM LTM, trained on sequences from within the composition, rather than from a stylistic corpus). To dynamically train and predict events in multiple parallel sequences or streams (as IDyOM's STM does using single-voice sequences), a variant of PPM can be implemented in which a model learns incrementally from all sequence streams, adding n-gram subsequences of different orders in individual sequences to a shared representation, then predicts events in each sequence given the context provided in that stream. As the model proceeds through the composition, events are added to their respective n-grams as they occur.[8] As illustrated in Figure 3.1), for the event $s_{j,i}$ (event i in voice j), the experience of the STM is given by:

$$m_x = \big[(s_1)_1^{\max\{l \mid t_{1,l} < t_{j,i}\}}, ..., (s_N)_1^{\max\{l \mid t_{N,l} < t_{j,i}\}}\big]$$

This technique relies on being able to access the perceptually independent streams within a polyphonic composition. The processes by which listeners are able to perceive multiple parallel streams, and how these streams may be computationally identified, are the focus of the large body of voice separation and auditory stream analysis research (Bregman, 1990; Cambouropoulos, 2008; Sauvé, 2018; Temperley, 2009). While the modelling of this process is beyond

[8] Estimations of likelihood are made for all events at a given time-point, before updating the training model. This prevents voice ordering from influencing predictions, as may otherwise be the case when events in multiple voices occur at the same time-point.

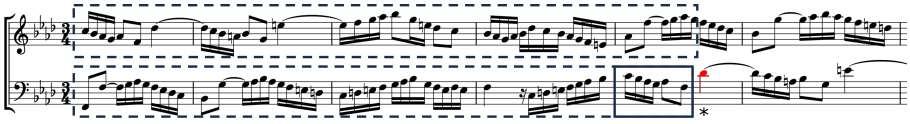


Figure 3.1: Illustrative example of a polyphonic PPM STM, J. S. Bach, Invention in F minor, No. 9, BWV 780. For a given note-event for which a likelihood is to be estimated (bar 5, lower voice, marked with asterisk), the context for prediction is bounded by a solid box (*i.e.*, in this case with an order of six), and the model's experience bounded by dashed boxes. As can be seen, material useful for predicting the target note occurs in both voices, particularly in the opening of the upper voice

the scope of this thesis, horizontal thematic structures can still be modelled in a limited selection of real-world polyphonic compositions, under specific assumptions. Compositions with multiple voices need to contain:

1. Only a small number of distinct voices to reduce the cognitive load that may lead to voices being combined into a smaller number of streams
2. Limited unison, octave, or other parallel movement between voices (*principle of tonal fusion* and *pitch co-modulation principle*, Huron, 2001)
3. Limited homorhythmic passages (*onset synchrony principle* and *synchronous note principle*, Cambouropoulos, 2008; Huron, 2001)
4. Limited possibility of material possessing 'implied polyphony', when multiple perceptual streams can outlined by a single monophonic line (Davis, 2006)
5. Substantially melodic material in each voice (*i.e.*, a voice should not only contain accompanying material).

Due to the exploratory nature of the horizontal probabilistic modelling described here, the implementation and use of these techniques were limited to the later behavioural experiments presented in this thesis (Chapters 7 and 8). Compositions of two voices (meeting the other assumptions above) were used as stimuli, alongside the monophonic stimuli used throughout.

3.4 Modelling thematic structure

The purpose of the computational model outlined in this chapter is to implement an integrated collection of hypothesised cognitive processes that produce a set of quantitative measures of thematic structure, such that a given composition can be described in a multidimensional space. Within this space, the composition can be defined as a point, representing the extent to which it possesses various features of hypothesised importance to the perception of thematic structure. To achieve this, the model also needs to be able to extract potential themes

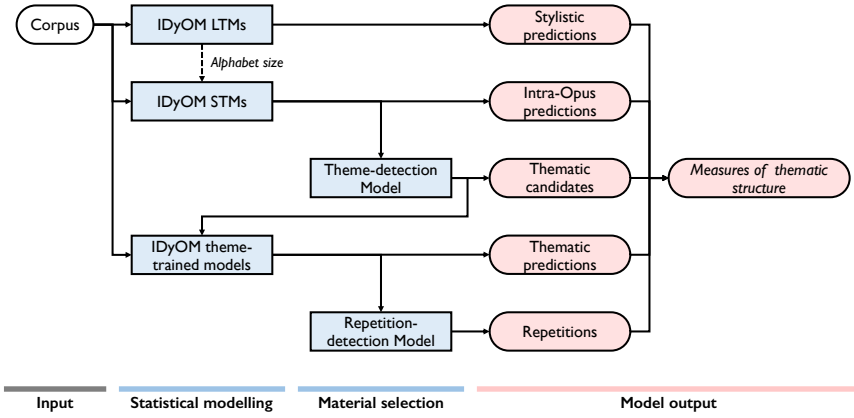


Figure 3.2: Outline of the statistical model of large-scale thematic structure using the IDyOM framework.

and repeated thematic material from the works of the corpus. For a given composition, multiple IDyOM models were applied, based on different training material, representations of its musical surface and LTM/STM configurations. These statistical models were combined and used to simulate listeners' perceptions of repetition and variation as it varies dynamically throughout listening to the composition.

An overview of the modelling process for a given composition from a larger corpus is as follows (and illustrated schematically in Figure 3.2). The symbolic music data of the large corpus of compositions was first used to train an LTM; this corpus-trained LTM was used to calculate information content values for each note-event in the target composition. Training on the entirety of the corpus made these LTMs models of stylistic structure, generating predictions based on learned stylistic conventions (thus yielding measures of *stylistic unpredictability*).

The training of the LTMs over the whole corpus also provided the full alphabet covered in that representation—for example, the complete list of absolute pitch values or the entire collection of note durations used in the corpus. Since PPM* produces non-zero probabilities over the entire alphabet defined for the model, information content is sensitive to alphabet size; by maintaining the use of these full alphabets in the subsequent creation of models that are trained only on subsets of the corpus, these information content values are directly comparable between all works for a given representation.

A short-term model for the composition was then implemented. Once again, an information content value was generated for each note-event, based on the online accumulation of context data within that composition. These information contents provided a measure of the *internal unpredictability* of each note-event in a composition.

In addition to the extra-opus LTMs and intra-opus STMs, a third IDyOM model type was implemented and used in this analysis. Using a theme-detection model based on the patterns present in the STM (see Figure 3.2; described fully below), thematic candidates were identified within the compositions. For each composition, these thematic candidates were then used as the training material for new models, one for each candidate. These models produce information content for each note-event giving the predictability of that event relative to the chosen thematic parent.

3.4.1 Large-scale structure of internal unpredictability

Based on the principles of statistical learning, repetitions of material become more predictable as they occur successively. In the information content values of *internal unpredictability*, estimated by the IDyOM STM, improved prediction of these later repetitions results in a reduction of information content. The dynamic online processing of the STM means that all repeated material in a composition is subject to this effect. If material is repeated a greater number of times, its information content will continue to be lowered. Figure 3.3 displays the *internal unpredictability* (information content) for each note-event modelled using the pitch interval representation in the upper-most voice of the first movement of Mozart's Piano Sonata No. 12, K. 332 (serving as an example throughout this chapter). Visually, the prediction set provided by the STM for this composition contains several prominent areas of densely populated low-information-content (highly predictable) events, with a complete absence of more unpredictable material. These correspond to the exact repetitions of sections in the movement. The first, between 280 and 558 quarter-notes from the start of the piece, is identifiable as the repeat of the sonata form exposition section—compressing the same patterns as the first exposition into a lower range of information content. Across the piece, repeated material with high internal predictability (low information content) is interspersed with higher-information-content, more unexpected, pitch intervals; however, repeated thematic material can still be distinguished. The visibly different section at 559 quarter-notes corresponds to the start of the development section. Here patterns of thematic statements—not necessarily in their exact form—intersperse regions of more distantly-related material.

The clear identification of repetition patterns in the STM is undoubtedly obscured by other factors contributing to these information contents; even over the duration of a single composition, the model is beginning to learn other statistical regularities present—forming a basis for tonality and style that is independent of thematic repetition. Were our perception of thematically-salient material to function solely in this manner, it would take many repetitions of a

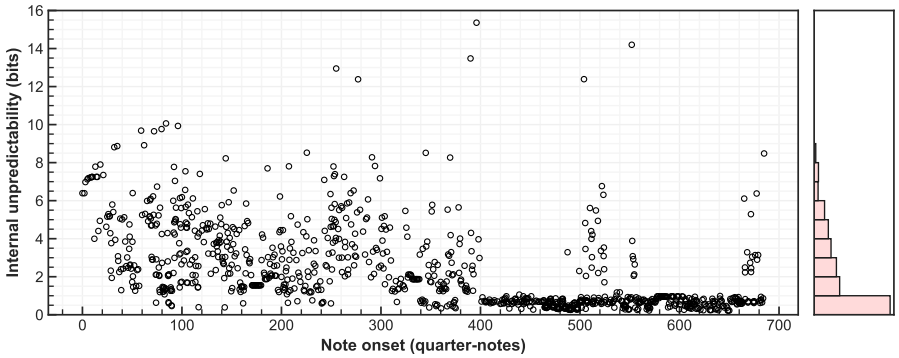


Figure 3.3: Information contents generated by a pitch interval interval short-term model predicting each note in Mozart's Piano Sonata No. 12, K. 332, first movement. The histogram in the right subplot shows the overall distribution of information content.

theme for reliable identification—while possible, this is not a wholly convincing representation of our cognitive processes. More likely, this statistical salience is reinforced by other mechanisms and prior knowledge of stylistic convention—phrasing, positioning, local melodic structure, and others. We assume, as a first approximation, that these conventions manifest themselves together under the concept of *theme*, in a way more imminent and meaningful than mere thematic repetition; our perceived salience through repetition may be bolstered through the identification of one or more themes, patterns that play a particularly important role in perception of thematic structure. In the rather noisy STM, the identification of possible thematic candidates allows repetitions of derived material to be prominently predictable.

3.4.2 Theme-detection model

The computational identification of themes in music is an extremely challenging task. As summarised in Chapter 2, Methods for general repetition detection in symbolic music have been produced for a wide range of purposes (a summary of this body of research is given in Janssen et al., 2014). However, all of these methods operate in a manner contrary to that needed for our model of large-scale thematic structure—they first identify repetitions throughout a composition, then select the most likely theme. For a cognitive model simulating perception of thematic material on first listening, detection of themes must precede detection of thematic repetitions, and both must be achieved dynamically on a single passing of the composition. For many computational methods, there is also a significant difficulty in determining a theme's length.

This chapter presents a method by which values of *internal unpredictability* generated by the IDyOM STM can be employed to find potential themes—or

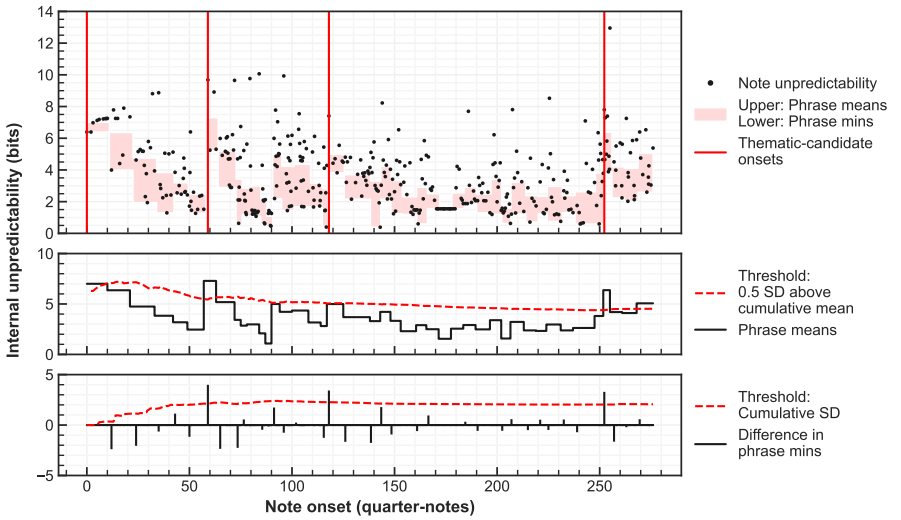


Figure 3.4: Theme detection for Mozart K. 332, first movement, exposition. Four onsets were identified at 0, 60, 121 and 253 quarter-notes, based on pitch interval internal unpredictability. Thematic candidates were identified by phrase mean information contents being greater than a threshold of a 0.5 SD above the cumulative mean (subplot 2), and having a difference in phrase minimums greater than 1 SD (subplot 3).

thematic candidates—in an incremental online manner, avoiding the need for exact repetition matching. While the level of definition in the STM information contents is not, in practice, great enough to allow for direct identification of thematically-derived material (as exemplified above), it can clearly identify the locations of multiple new thematic candidates by the presence of sudden increases in information content. If a composition were to contain a single thematic candidate—presented at the beginning—from which all subsequent material derived in some way, the information contents of the STM would show a decrease in overall information content as the composition progressed. The introduction of a subsequent new thematic candidate in the composition would present unrelated, and so unpredictable, material—causing an increase in information content (as illustrated in Figure 3.4). A similar approach in MIR has been found to be successful in the identification of segment boundaries in audio self-similarity matrices based on novelty and repetition (Nieto et al., 2020; Serra et al., 2012).

In order to give some estimation as to the length of a thematic candidate, it is useful to take into account the sequential grouping of material in music perception. Here, grouping boundaries between phrases were identified using the *Grouper* algorithm of Temperley (2001).[9] As summarised in the previous chapter, *Grouper* segments a melody according to three Phrase Structure Preference Rules (PSPRs) (Temperley, 2001, pp. 68–71): the 'gap rule', PSPR 1, tries to

[9] *Grouper* was used via the implementation available in *The Melisma Music Analyzer* (Sleator & Temperley, 2003) accessible at https://www.link.cs.cmu.edu/music-analysis/.

locate boundaries at large IOIs or large offset–onset intervals; the 'phrase length rule', PSPR 2, favours phrases that are close to a predetermined length—here, ten notes;[10] and the 'metrical parallelism rule', PSPR 3, favours boundaries occurring on the same position within a bar. Although it has been demonstrated that IDyOM can be extended to provide a probabilistic method of boundary segmentation (Pearce, Müllensiefen & Wiggins, 2010), *Grouper* continues to be one of the best performing and most robust methods available (Cenkerová et al., 2018). Its relative simplicity allows for it to be adapted and implemented incrementally note-by-note—rather than offline, as originally intended—producing similar boundary locations.

The theme-detection model defined thematic candidates based on sequential comparisons of *Grouper*-segmented phrases to the cumulative body of all preceding material (*i.e.*, all of the composition before a given phrase). The first event in the composition was considered an implicit beginning of a thematic candidate. Otherwise, a candidate was declared at the start of a phrase based on two conditions.

1. The mean *internal unpredictability* of the phrase was greater than the cumulative mean of the preceding material by at least one-half the standard deviation. For a sequence of n segmented phrases, s_1^n, with l_1^n denoting the number of note-events in each phrase, this condition for phrase s_i was met when:

$$\frac{1}{l_i}\sum_{j=1}^{l_i} s_{i,j} < \frac{1}{\sum_{k=1}^{i-1} l_k}[s_{1,1} + ... + s_{i-1,l_{i-1}}] + 0.5\sigma[s_{1,1}, ..., s_{i-1,l_{i-1}}]$$

2. There was a complete absence of low *internal unpredictability* material in the STM, indicated by the phrase minimum information content (*i.e.*, the note with the lowest information content in the phrase) rising by more than one standard deviation over the minimum of the preceding phrase.[11] This condition for phrase s_i was met when:

$$\min\{(s_i)_1^{l_i}\} > \min\{(s_{i-1})_1^{l_{i-1}}\} + \sigma[s_{1,1}, ..., s_{i-1,l_{i-1}}]$$

The values of these thresholds were chosen through manual optimisation with the intention of providing a robust identification of significantly novel material within all compositions. Figure 3.4 illustrates this process in the detection

[10] This value was arrived at by Temperley through an optimisation of the algorithm to an annotated subset of the Essen Folk Song Collection (Temperley, 2001, p. 74).

[11] This process is the same when modelling polyphonic compositions as it is described here. For polyphonic compositions, an initial thematic candidate was selected at the beginning of each voice and a dynamically-trained STM was used to identify novel material in all voices based on information content.

of thematic candidates in the Mozart K. 332 movement's exposition section, where, in this instance, all the detected candidates lie. Four thematic candidates were identified with onsets at 0, 60, 121 and 253 quarter-notes.

The precise length of thematic candidates, once a start point is detected, is still unknown. Using the phrase boundary segmentation, this length can be considered a free parameter, defined in terms of a given number of phrases. A length of two phrases, for example, functions well to account for the antecedent/consequent phrase pattern in much music of the Classical period and is used in the present analysis. To limit the number of models, theme detection was run using a single representation for each domain: the pitch interval (transposition insensitive) representation for the pitch domain, the inter-onset interval representation for the rhythmic domain. Thematic candidates were returned as symbolically notated fragments, used in the training of a new set of statistical models for the composition.

It should be stressed again that the thematic candidates extracted by the model do not necessarily possess all the properties traditionally associated with the concept of theme in music analysis. True themes possess additional perceptual salience. In many cases this salience may occur through the repetition of the theme's material, but also through other influences—such as form, where there is often a strong tonal element not covered at all here. This process is not, therefore, intended as a tool by which a new analysis can be performed. Thematic candidates here should be considered simply as regions within the piece at which novel material is introduced. The extent to which thematic candidates actually contribute to the model's output—and are likely to be considered as actual themes—is determined through the identification of repetition of their material.

The four, two-phrase thematic candidates extracted from the example K. 332 movement using the pitch interval representation are shown in Figure 3.5. As with many of the opening movements of Mozart piano sonatas, copious analyses for this work exist belonging to numerous different schools (Allanbrook, 1992; Beach, 1994; Beghin, 2014; Caplin, 2001; Galand, 2014; Hatten, 2014; Hepokoski & Darcy, 2006; Irving, 2010; Kinderman, 2006; Rumph, 2014; Schenker, 1994). It is perhaps unusual in that it contains more unique thematic material than may otherwise be expected in a sonata form exposition, and, although a direct alignment with any of these is not intended, the comparison below of the extracted thematic candidates to those identified by music theorists provides a concrete illustration of how the theme-detection model performs.

A brief summary of a formal analysis of this sonata form exposition section, based on those of Beach (1994) and Kinderman (2006), is as follows: there is an opening theme that (slightly unusually) is three phrases long (bars 1–12); this is followed by a new thematic idea (bars 13–22) in the second half of this

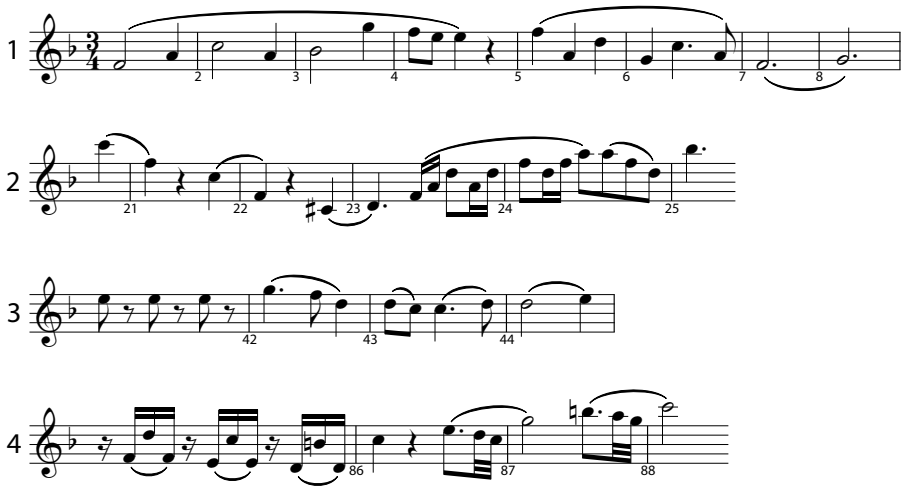


Figure 3.5: Extracted thematic candidates from Mozart K. 332, first movement.

first thematic group (Beach, 1994); a dramatically different transition passage in bars 23–40 occurs in the relative minor (Kinderman, 2006, p. 52); then, the second thematic group is presented in the dominant (bars 41–48)—variations of this material alternate with darker syncopated passages (bars 56–66) until the closing codetta of the exposition in bars 86–93 (Kinderman, 2006, p. 52). The components of this analysis are shown in the annotated score in Appendix A, alongside the four detected thematic candidates and modelling values.

In the model's detection of thematic candidates, we can see, in Figure 3, that many of these inner thematic ideas within the wider groups were not found to be distinct-enough for separate classification (at least in part due to many of their distinguishing features being lost in the removal of texture, harmony and rhythm). Instead, what was identified (in Figure 3.5) was (1) the opening theme; (2) the start of the transition section, prepended by a small amount of material leading into it;[12] (3) the second subject; and (4) material leading into, and the beginning of, the codetta. After the initial theme, the thematic candidate of the transition section was found to be highly prominent, even more so than the second subject which only just qualified above the threshold for mean information content. The codetta material presented the possibility of a spurious classification—the material was novel in the STM, but it might be considered to contain purely stylistic content that not directly relevant to thematic organisation.

When compared to the traditional analyses, these thematic candidates did not completely cover all the themes of the original. In particular, the second theme in the first thematic group (bars 13–22) was not identified; the pitch con-

[12]This premature thematic-candidate identification could likely be attributed to a small difference in phrase boundary placement by *Grouper*, when compared with those made in the analyses described above. The inclusion of the C-sharp (bar 22, beat 3) in the preceding phrase, rather than the following, altered the phrase in which novel pitch content occurs.

tent of these phrases was not sufficiently novel for a candidate to be detected. However, using the counterpart rhythmic theme detection (using inter-onset interval) a candidate was located at this position in the music.

The detection using inter-onset interval identified a different set of thematic candidates. Aside from the opening candidate, thematic candidates were identified beginning at bar 12, beat 3 and at bar 86, beat 3. The latter two candidates were notably different from those identified using the pitch interval representation. The placement of the second candidate matched that of the second theme of the first thematic group in the traditional musical analyses of the work. This candidate was not identified using pitch interval as the majority of its novelty only exists in the rhythmic domain. The third inter-onset interval thematic candidate overlapped with the fourth identified by pitch interval, the latter occurring in the preceding phrase. Both versions identify material at the start of the codetta, however, that of inter-onset interval more accurately identified its starting point. The identification of these two thematic candidates indicates that—for non-isorhythmic compositions—addition of the rhythmic domain can better inform how thematic material is identified.

3.4.3 Repetition-detection model

By pre-training an IDyOM model on the data for a single thematic candidate (creating theme-trained models—TTMs), models were created that could provide a predictive probability for each note-event in a composition, based on the thematic candidate; using these models, the noise in the STM *internal unpredictability* was largely reduced as the models are not incrementally updated with every new note-event in the composition. When multiple thematic candidates were identified in a composition, TTMs were constructed for each one, creating separate models for each candidate beginning at that candidate's onset. Figure 3.6 shows the note-event information contents generated for the first movement of Mozart K. 332 when a model was trained on each of the four thematic candidates extracted using the pitch interval representation (shown in Figure 4). The resulting note information contents can then be classified as to whether or not they constitute thematically-derived repetitions (or 'motivic' material). Once again, this process needed to function incrementally within the work's progression.

Models generated in this way have the property of producing clearly stratified patterns of information content, reflected in a degree of bimodality in their distribution. The information content of each note-event can be considered as either belonging to a low-information-content thematic distribution, or to the high-information distribution of the remainder.

To perform this classification, clustering with Gaussian Mixture Models

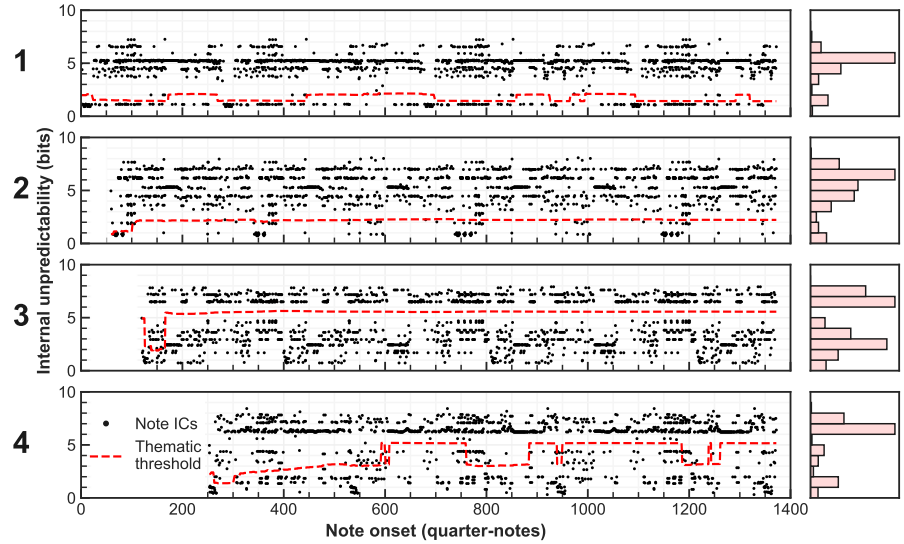


Figure 3.6: Prediction sets and their distributions for the four theme-trained models of Mozart K. 332, first movement. A thematic threshold was generated by incrementally updated GMMs at the phrase level—material below this boundary was classified as thematic.

(GMM) was applied to the distributions (D. Reynolds, 2009). Two Gaussians were fitted to the information content data using expectation–maximisation (Dempster et al., 1977). The starting parameters of the GMM were calculated so that the intended distributions were identified—for example, so that the two Gaussians avoid favouring some other multimodal components in the remainder distribution, or at other local minima—and to ensure consistent results. To provide the best chance of finding the correct thematic distribution and the remainder, starting means of the lowest value and the median (respectively) were used. Gaussians were initially equally weighted and the standard deviation of the remainder was specified as double that of the thematic. After the initial occurrence of a thematic candidate, for which the question of thematic association is not needed, this GMM could categorise each note as thematic or remainder (*i.e.*, non-thematic) based on the cumulative previous distribution, updated at the phrase level. This is shown in Figure 3.7 for a TTM trained on the first thematic-candidate and applied to remainder Mozart K. 332, first movement—*i.e.*, the right-most classification made for theme 1 in Figure 3.6.

The categorisation of thematic material on an individual note basis, as described above, provides rather a harsh and exacting process; the local context of each note is not taken into account and the resulting material extracted is highly fragmented. Performing the same operation with smoothing on a phrase-based level allowed larger—still low-information-content and salient—sections to be extracted, following perceived sequential groupings of material, providing a larger-scale output more akin to motifs. For this purpose, Temperley's *Grouper*

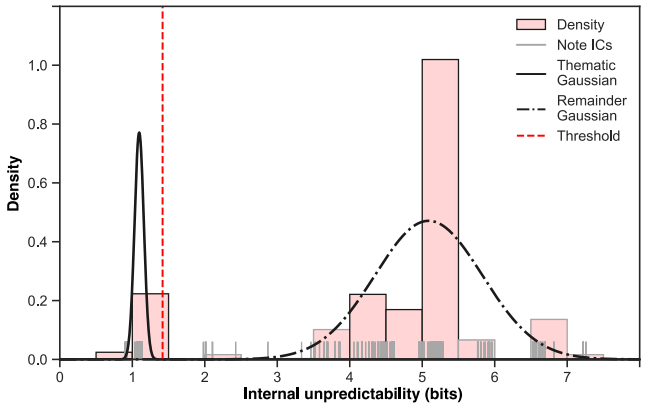


Figure 3.7: Note *internal unpredictability* and Gaussian Mixture Model Clustering for a TTM trained on thematic-candidate 1 and applied to Mozart K. 332, first movement. A lower cluster is identified as the thematic material and the upper the remainder. The vertical dashed line indicates the threshold identified.

was once more applied. The threshold information content was still computed on the note-event level; the mean information content for each phrase was then categorised based on this threshold.

3.4.4 Compression Distance

The repetitions of thematic candidates identified by the repetition-detection model can be used to simulate the degree of intra-opus variation of thematic material. A computational measure of similarity between each phrase categorised as being thematic and its parent thematic candidate (*i.e.*,the candidate used in the identification of the phrase) can describe how much variation thematic material undergoes. As with all the steps of this model so far described, the similarity metric should be well-motivated in terms of representing actual cognitive processes. The information-theoretic measure of compression distance was used, having previously shown promise in simulating perceived melodic similarity (Pearce & Müllensiefen, 2017).

A similarity metric based on normalised compression distance was introduced by Li et al. (2003). The *dissimilarity* between two sequences, x and y was taken as a function of the predictability (or compressed length) of one sequence, given a model trained on the other. Pearce and Müllensiefen (2017) employed IDyOM to estimate dissimilarity between two musical sequences, $D(x|y)$, as the normalised summed information content of the latter sequence (of length k), given a PPM model trained on the former.

$$D(x|y) = \frac{1}{k}\sum_{i=1}^{k} h_{m_y}(x_i)$$

In the same way, the dissimilarity between an identified thematic fragment

and its parent thematic candidate could be modelled as the sum total of the information contents for each note in the thematic fragment, given the IDyOM model trained on the parent thematic candidate. The sum of the information contents was then normalised with respect to the longest sequence—the candidate. This asymmetric measure of compression distance was appropriate for the current model of thematic structure, given that it is only concerned with the amount of variation moving forward through the composition.

3.5 Measures of thematic structure

The process of modelling thematic structure described above extracts thematic candidates and thematic repetitions for a given composition. In the research presented in this thesis, we are particularly interested in the application of these methods to produce quantitative measures that can characterise features of hypothesised importance to the perception of thematic structure. By applying the measures to a large corpus, we can describe quantitatively how real-world musical compositions differ in a number of key dimensions relating to large-scale thematic structure. Of these measures, we are primarily interested in those characterising intra-opus features of compositions, or directly arising from repetition, as the main parameters hypothesised to influence the perception of structural structure. The primary four measures are summarised below.

Internal unpredictability. The STM (using IDyOM or other PPM equivalent) information content of a composition. For note-events, this measure provides the *internal unpredictability* of each event, in its representation, given a model training and predicting dynamically throughout the composition. This measure is also used to refer to the *internal unpredictability* of an entire composition, in which case the mean is taken for individual note-event values. This measure uses the information-theoretic unit of *bits*. For a composition, e_1^k of k events long, where h_{stm} is the STM information content:

$$U(e_1^k) = \frac{1}{k} \sum_{i=1}^{k} h_{\text{stm}}(e_i | e_1^{i-1})$$

Thematic repetition. Using the theme and repetition-detection processes described above, repetitions of potential thematic material were identified in a given composition. The measure of *thematic repetition* quantifies the proportion of note-events out of the whole composition that were identified as being thematically repeated material—*i.e.*, the proportion of notes that were below a threshold in w_1^N, found when using GMM clustering on a model trained on the corresponding theme in T_1^N, where N is the number of themes.

$$R(e_1^k) = \frac{1}{k}\text{count}_{i \in k}\Big\{\text{any}\big\{(h_{T_1}(e_i|e_1^{i-1}) < w_1), ..., (h_{T_N}(e_i|e_1^{i-1}) < w_N)\big\}\Big\}$$

Thematic variation. Using the dissimilarity measure of compression distance, the amount of thematic variation within a composition was characterised by the mean dissimilarity between all thematic phrases (material identified using repetition detection, smoothed over detected phrases) and their closest parent theme. As this measure was implemented using *dis*similarity, low *thematic variation* indicated that identified repetitions did not vary far from their original instances. Specifically, for a list of N themes, T_1^N, each with a list of phrases identified as belonging to that theme, $(s_T)_1^m$:

$$V(T_1^N) = \frac{1}{Nm}\sum_{i=1}^{N} D(T_i|x) \forall x \in s_{T_i}$$

Stylistic unpredictability. The repeated thematic material identified (at the phrase level) was modelled using an LTM trained on a large corpus of western-classical tonal music (presented in the following chapter). The mean LTM information content for thematic note-events (their individual *stylistic unpredictability*) was used to provide a measure of the *stylistic unpredictability* of thematic material. For a composition, e_1^k of k events long, where, h_{ltm} is the LTM information content:

$$S(e_1^k) = \frac{1}{k}\sum_{i=1}^{k} h_{\text{ltm}}(e_i|e_1^{i-1})$$

when

$$\text{any}\big\{(h_{T_1}(e_i|e_1^{i-1}) < w_1), ..., (h_{T_N}(e_i|e_1^{i-1}) < w_N)\big\}$$

These model-based measures, as presented here, are designed to apply to complete pieces of music. In their use and evaluation throughout this thesis, these four primary measures were divided into corresponding experimental measures that reflect the precise paradigms to which they were applied, within the context of their respective experiments. The mappings of these measures on to their experiment-specific variants are laid-out in Table 3.1 (entries in parentheses indicate experiment measures that are not directly related to model measures, but still partly fulfil their purpose).

Table 3.1: Summary of Experiment Measures Used in This Thesis, Based on the Model Measures of Thematic Structure

Measure	Experiment 1 (Chapter 5)	Experiment 2 (Chapter 6)	Experiment 3 (Chapter 7)	Experiment 4 (Chapter 8)[a]
Internal unpredictability	*(Dissimilaity)*	*Internal unpredictability* *Internal unpredictability of moment* The unpredictability of a given phrase, when trained on the composition preceding it *Internal unpredictability at moment* The mean *internal unpredictability* of a composition before a given moment	*Composition unpredictability* The unpredictability of continuation material given a preceding composition *Late-composition unpredictability* The unpredictability of continuation material given the second half of a preceding composition *(Continuation unpredictability)* The *internal unpredictability* of continuation material	*Internal unpredictability (perfect)* Perfect-memory *internal unpredictability* (*i.e.*, the original measure) *Internal unpredictability (buffer)* Buffer-limited *internal unpredictability* *Internal unpredictability flow* Mean change in *internal unpredictability* between segments of a composition
Thematic repetition	–[b]	*Thematic repetition* *Thematic repetition at moment* The average *thematic repetition* in a melody before a given moment	*Thematic unpredictability* The unpredictability of continuation material given the themes identified in a composition	–[b]
Thematic variation	*Dissimilarity* The normalised compression distance between a parent theme and a subsequent phrase identified as a repetition	*Thematic variation* *Thematic variation at moment* The average *thematic variation* in a melody before a given moment	*(Thematic unpredictability)*	*(Variation)* The extent to which material in ordered segments of a composition develops linearly
Stylistic unpredictability	*Stylistic difference* The absolute difference in *stylistic unpredictability* for two given sequences *Mean stylistic unpredictability* The (grand) mean of *stylistic unpredictability* for two given sequences	*Stylistic unpredictability* *Stylistic unpredictability at moment* The average *stylistic unpredictability* of a melody before a given moment	*Stylistic unpredictability* The unpredictability of continuation material given a stylistic corpus	*Stylistic unpredictability flow* Mean change in *stylistic unpredictability* between segments of a composition

[a] Experiment used additional measures not listed in this table
[b] Theme-detection modelling not applicable to the paradigms used in these experiments

3.6 Summary

This chapter presented an overview of how large-scale thematic structure in music can be computationally modelled, based on the underlying theory that the perception of such structures is facilitated by the statistical regularities caused by repetition, and variation, of material within a composition. While structure is considered important in music theory and in theoretical models of music cognition, past psychological work in search of evidence for the ability of listeners to perceive large-scale thematic structure has often proved inconclusive (see Chapter 2). The statistical model is presented as the beginnings of a concrete specification of the cognitive processes involved in the perception of such structures. The model expresses explicitly a plausible computational account of how large-scale thematic structure might be perceived and allows us to derive a set of formal, quantitative measures of thematic structure.

The model used the IDyOM framework (Pearce, 2005) to create extra-opus models of style—in which a large corpus of compositions can be used to calculate the stylistic unexpectedness of the notes in each, given the context—intra-opus models of composition interrelatedness—where the training context is provided dynamically as a music progresses—and theme models trained on the extracted thematic candidates of a composition. Not only does this model provide novel techniques for extracting thematic candidates and thematic repetitions (or motifs) from a composition, it also produces a multidimensional set of measures, with which the variation of structurally important elements present in music can be captured.

The model presented in this chapter presents some of the key theoretical contributions of this thesis. The subsequent chapters of this thesis are concerned with the empirical evaluation of this model, and the overarching hypothesis on which it is based.

Chapter 4

A Corpus Analysis

4.1 Overview

In this chapter, the first steps in testing the theory behind our modelling of thematic structure are taken. This chapter presents a corpus analysis that explores the output of the computational measures when applied to an assembled corpus of full-length monophonic compositions. As there is a lack of music corpora annotated with thematic structure (either perceived or analysed) against which the performance of measures can be tested, this analysis takes a exploratory approach that aimed to demonstrate that measures of hypothesised importance to the perception of thematic structure vary systematically when applied to the corpus, reflecting the inherent variation in structure present within the corpus itself.

The chapter, firstly, gives a summary of the computational measures used in the analysis, including the four primary model measures introduced at the end of the previous chapter, and several additional measures characterising general composition properties and elements of the theme detection process. Secondly, a description is given of a corpus of 623 complete monophonic compositions, assembled for use in this analysis—and also providing ecologically-valid stimuli for stimulus selection in subsequent behavioural experiments. The analysis of the model output when applied to the corpus is then explored in two stages, tackling separately the output when applied using pitch and rhythmic representations, respectively. For both representations, the analysis first considers the relative performance of different representations of the musical surface within its respective domain, then examines the ability of measures to explain variance inherent in the corpus, and finally, explores the function of measures through example compositions selected from the extremes of the model output.

4.2 Modelling

The model of thematic structure, introduced in the previous chapter, was founded on the idea that repetition is the principal enabler for the perception of thematic structure in music. More specifically, according to the probabilistic conception of music espoused, through learning, this repetition increases the intra-opus predictability of a given composition, giving rise to perceived structural unity. Repetition of thematic material strengthens its salience—the more material is repeated, the greater its perceived prominence. By following stylistic or work-specific structural regularities, such repetitions can undergo embellishment and variation and still reinforce future predictions. The purpose of the analysis presented in this chapter was to examine the ways in which four measures that were the model's principal output contributed to systematic variation of thematic structure within a corpus of western-classical melodies.

4.2.1 Measures

The four primary model measures of thematic structure described at the end of the previous chapter were used to quantify compositions in a corpus of western-classical melodies. Alongside these measures hypothesised to be predictors of thematic structure, additional measures characterising more general properties of compositions and the theme-detection process were also included in the analysis presented in this chapter. The measures of this analysis belonged to four categories: (1) those concerned with general properties of the compositions; (2) those describing features of a given composition's thematic candidates (as detected by the model); (3) those characterising the properties of repeated thematic material (again, as detected by the model); and (4) *internal unpredictability*.

General composition properties

Composition year. The year in which the work was composed.

Composition length. The number of note-events in a composition.

Properties of thematic candidates

Number of themes. The number of thematic candidates that were identified in a composition, using the theme-detection method described in Chapter 3.

Length of themes. The mean number of note-events for the identified themes within a composition.

Strength of theme prominence. The difference in mean *internal unpredictability* of note-events between the first phrase of a detected theme and the mean *internal unpredictability* of the portion of the composition that preceded it—averaged across all themes identified in a single composition. For N themes in T_1^N, where $T_{1,\text{start}}$ and $T_{1,\text{end}}$ denote the indices of events within a composition where the first phrase of T_1 starts and ends:

$$G(T_1^N) = \frac{1}{N}\sum_{j=1}^{N}\Big(\frac{1}{T_{j,\text{length}}}\sum_{i=T_{j,\text{end}}}^{T_{j,\text{end}}} h_{\text{stm}}(e_i|e_1^{i-1}) - \frac{1}{T_{j,\text{start}}-1}\sum_{i=1}^{T_{j,\text{start}}-1} h_{\text{stm}}(e_i|e_1^{i-1})\Big)$$

Stylistic unpredictability of themes. The mean *stylistic unpredictability* of thematic candidates identified in a composition.

$$S(T_1^N) = \frac{1}{N}\sum_{j=1}^{N}\Big(\frac{1}{T_{j,\text{length}}}\sum_{i=1}^{T_{j,\text{length}}} h_{\text{ltm}}(T_{j,i}|(T_j)_1^{i-1})\Big)$$

Properties of thematic material

The three model measures of *thematic repetition*, *thematic variation*, and *stylistic unpredictability* (of thematic material) fall into this category.

Thematic/stylistic balance. The proportion of note-events identified as thematic repetitions that were more internally than stylistically predictable.

$$B(e_1^k) = \text{count}_{i\in k}\big\{h_{\text{stm}}(e_i|e_1^{i-1}) < h_{\text{ltm}}(e_i|e_1^{i-1})\big\}$$

when

$$\text{any}\big\{(h_{T_1}(e_i|e_1^{i-1}) < w_1), ..., (h_{T_N}(e_i|e_1^{i-1}) < w_N)\big\}$$

4.3 The corpus

A corpus was constructed to provide an application domain for the computational model of thematic structure presented in the previous chapter.[1] The spread of this corpus was intended to provide a broad representative sample of western-classical tonal music. Corpus items were included or excluded subject to certain constraints: (1) compositions were completely monophonic—the focus on monophonic melodic structures in this initial analysis was chosen to

[1] The corpus, along with information on the individual works, can be found at https://osf.io/dg7ms/.

avoid the numerous obstacles posed by tracking interrelating thematic material through multiple polyphonic layers; (2) compositions were used in their entirety (or the entirety of a movement)—the model was intended to simulate perception of large-scale structure, so the full structures were required; and (3) compositions were of sufficient length that such structures could be considered unambiguously present. Therefore, melodies were only included if they contained in excess of one hundred notes.

The nature of corpus needed for this analysis falls in a somewhat limited area of focus for current digital symbolic music databases; datasets in music perception research tend towards smaller, segmented stimuli, while digitised collections of full compositions are largely polyphonic (with a particular bias towards compositions for the piano). The monophonic constraint, in particular, provided a significantly limiting factor in gathering the corpus; compositions were needed either to be originally composed as such, or—as was the case for the majority—manipulated to produce a monophonic line. Therefore, in order to create a corpus of sufficient size and breadth, manipulations to the original sources were made to extract melodic lines. These monophonic extractions, while certainly not the composers' original compositions, still provided melodic lines that could arguably present comprehensible works of music in their own right and, importantly, retained a large proportion of their thematic structures.

Original compositions were gathered from online databases of existing digitised symbolic scores, primarily from KernScores[2] MuseScore[3], and the Classical Archives MIDI[4] collections. Composition-selection aimed at providing the broadest-possible spread of composition date, balanced with the considerations of selecting types of work for which the process of extracting a melodic line would be appropriate. The corpus consisted of three categories of composition: (1) those (a relatively small number) originally composed for a single-stave instrument and so already monophonic in nature; (2) piano works to which a skyline algorithm (Uitdenbogerd & Zobel, 1998) was applied to extract the uppermost line; and (3) works for solo instrument with accompaniment—from which the solo part was used and any large gaps were filled by a skyline of the upper accompanying line.[5] While the skyline algorithm could not be guaranteed to always find the optimal melody line—melody by no means appears univer-

[2] https://kern.ccarh.org

[3] https://musescore.com

[4] https://classicalarchives.com/midi.html

[5] Certain genres of vocal music, such as songs with a single voice and piano accompaniment, could also be considered to fit in this category, however, they were not included in this corpus. Firstly, many songs that fit into this category (for example, the *lieder* of composers such as Schumann) exist as part of a larger song cycle and are often too short to meet the length requirements. Secondly, the stylistic period of these songs is one well covered by other instrumental works with accompaniment. Thirdly, the presence of text within songs allows for variation to take place outside of the domains studied this thesis. For example, many songs are written in a strophic form with no variation between verses other than in the lyrics.

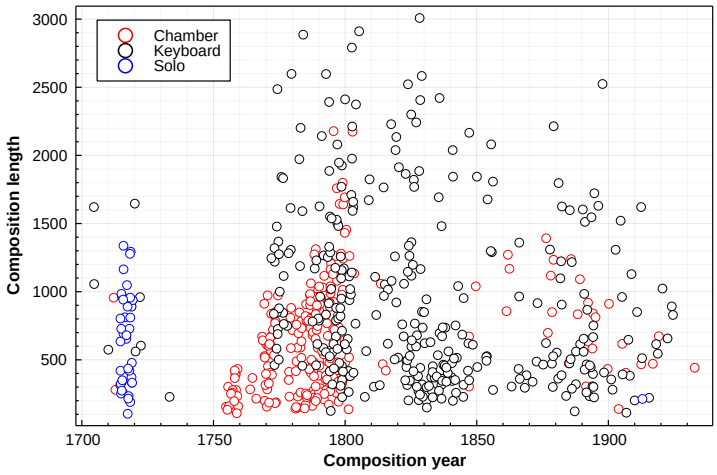


Figure 4.1: Distribution of corpus items by composition year, number of note-events and instrumentation type.

sally in the uppermost voice—it provided a robust technique for extracting the vast majority of melodic material. The chance of these errors occurring was, in part, mitigated by the use of compositions for monophonic instruments, with or without accompaniment, so giving an indication of where the primary melodic material is to be found.

It should be noted that, due to the nature of the constraints on the curation of the corpus, the database contains some inherent biases. A bias existed towards particular instrumentations, and so influences towards certain genres. As a result of the selection process, all works included were originally composed for either one or two instruments, no pieces for larger ensembles were present; the corpus, therefore, was confined to chamber music genres.

The resulting corpus contained 623 works—or self-contained work movements—with composition dates spanning from 1703 to 1934, encompassing western styles from Baroque to Early Twentieth-Century.[6] The works were distributed with a mean of 124.60 pieces for each of the five half-century divisions. The distribution of corpus items by composition year, length and instrumentation is shown in Figure 4.1, alongside the types of instrumentation requiring different techniques for extracting a monophonic line (further descriptive statistics for the corpus are given in the analysis in Table 4.1).

[6] All composition description labels and genre/style classifications were taken from the International Music Score Library Project (IMSLP, https://imslp.org).

4.4 The analysis

4.4.1 The pitch domain

In addition to the comparison of measures, the analysis contained the added dimension of representation of the musical surface used (applicable to all measures but those describing general composition properties). The IDyOM long-term, short-term and theme-trained models were all produced for each of the six chosen representations of musical pitch (described in the previous chapter). While it is acknowledged that, as listeners, we likely use a combination of multiple different representations of music at any one time, the precise way in which these representations are weighted and combined is unknown. However, to explore the output of the model, it is preferable to do so with only a single representation. Given that there is a great deal of overlap between representations, and for conciseness when exploring the multidimensional set of measures, only the notionally best-performing representation in each domain was selected for further analysis.

There is no globally-optimal method by which the best-performing representations within the model can be chosen; there is a lack of music corpora annotated with thematic structure (either perceived or analysed) against which representations could be tested. However, a comparison between the measures produced by the model allows for an exploration of the importance of different representations, relative to each other. Therefore, in the absence of a more explicit criteria, a comparison of the amounts of *thematic repetition* captured when using each representation was used—being a measure central to the hypothesis that thematic structures can be statistically learned, and one reliant on all theme and repetition-detection processes of the model. This comparison revealed that the pitch interval representation captured the highest proportion of thematic material, averaged across the corpus (see Figure 4.2).

For the selected representation of pitch interval, summary statistics of the measures of thematic structure and their respective pairwise correlations are given in Table 4.1. Measures derived directly from information content values—namely *internal unpredictability*, *stylistic unpredictability of themes*, and thematic material's *stylistic unpredictability*—have a scale that is lower for greater predictability/expectedness. For *strength of theme prominence*, higher values indicate more prominent thematic candidates in the STM. The compression distance-based measure of *thematic variation* gives a lower score when material is more closely related.

Due to the similarities between some measures, not all measures were found to be independent of each other. Those correlations with a coefficient $r > 0.5$ are displayed in bold in Table 4.1. *Composition length* and *number of themes* were found to be highly correlated—if a composition was longer, thematic-candidate

Table 4.1: Summary Statistics and Correlations for Measures of Thematic Structure Using Pitch Interval. Pearson's *r*, *n* = 623

	Measure	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1	*Composition year*	1809.50	49.03	–									
2	*Composition length*	801.94	562.04	.05	–								
3	*Number of themes*	2.48	1.98	.13	**.70**	–							
4	*Length of themes*	17.01	3.51	−.11	−.19	−.29	–						
5	*Strength of theme prominence*	1.85	0.86	−.07	−.09	−.29	−.13	–					
6	*Stylistic unpredictability of themes*	3.64	0.64	.24	.03	.14	−.20	.26	–				
7	*Thematic repetition*	0.42	0.18	.08	.19	.39	−.05	−.26	−.14	–			
8	*Thematic variation*	0.64	0.28	.17	.44	.48	−.24	−.29	.07	**.64**	–		
9	*Stylistic unpredictability*	2.91	0.54	.27	−.11	.03	.02	.10	**.67**	−.18	.00	–	
10	*Thematic/stylistic balance*	0.65	0.12	.04	.22	.01	.03	.40	.13	−.10	−.17	.26	–
11	*Internal unpredictability*	2.77	0.63	.08	**−.50**	−.20	.07	−.34	.18	−.15	−.11	.23	**−.67**

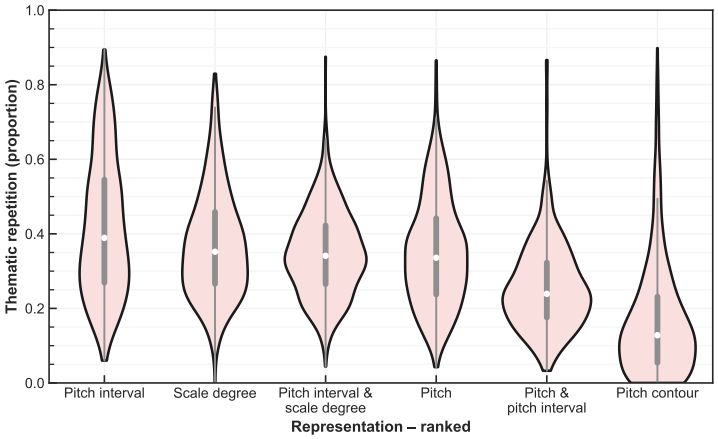


Figure 4.2: Distributions for the measure of *thematic repetition* across the corpus for all pitch representations, ranked in order of median value.

extraction was more likely to identify a greater number of candidates. *Stylistic unpredictability of themes* and thematic material's *stylistic unpredictability* were both derived from the same LTM, with the former limited just to the parent thematic candidate(s) and the latter calculated for thematic material identified by the repetition-detection model as being related to the parent candidates. *Thematic repetition* and *thematic variation* were highly correlated—for compositions in which a greater amounts of *thematic repetition* was found, there was an increased opportunity for repetitions to be embellished further from their original.[7] Finally, the correlation between *internal unpredictability* and *composition length* indicated that longer compositions, through their greater opportunity for repetition, were more predictable.

While examining the correlations between measures gives some insight into their relationships, the relative importance of these measures in actually explaining the variance of structure observed in the corpus still needs to be established. This variation was explored in greater depth by performing Principal Component Analysis and Independent Component Analysis on the set of measures applied to the corpus, providing dimensionality-reduction and importance exploration techniques.

Principal Component Analysis

Principal Component Analysis (PCA) allowed for the geometric reduction of the original measures into a smaller set of orthogonal components (Abdi & Williams, 2010). These new components consisted of linear combinations of the original measures, attempting to account for the maximum amount of variance

[7] No corresponding correlation was found between *composition length* and *thematic repetition*, as there was with thematic candidate measure. This is likely due to the latter being computed as a proportion of composition length.

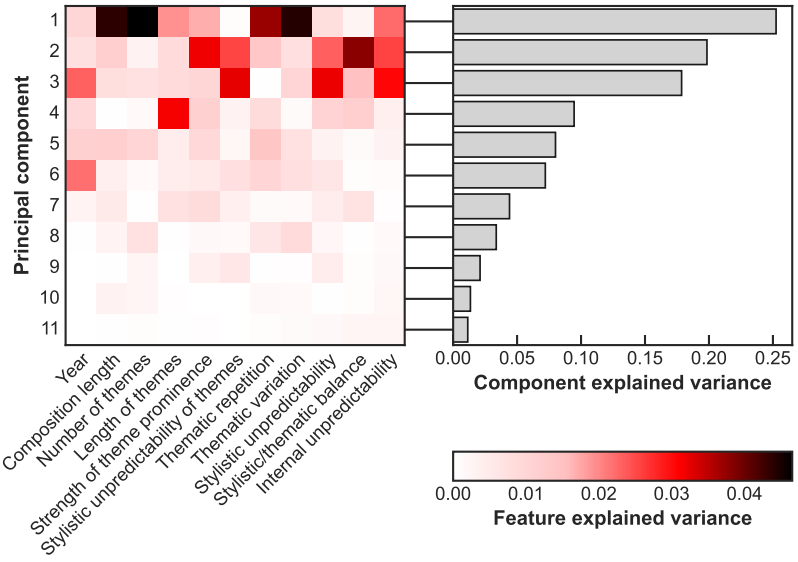


Figure 4.3: Overall explained variances for each measure using pitch interval and total explained variance for each output component in PCA.

in the structure of the data by the smallest number of components. While the output components do not correspond directly to the input measures, performing a PCA with equal numbers of each will produce components that account for all variance, and—importantly for its use here—the proportion of explained variance for which each component was responsible. The ordering of the principal components in a PCA is such that the first accounts for the largest proportion of variation in the set, the next attempts to explain additional variance while remaining orthogonal to the first, repeated for the number of components.

The purpose of using PCA in the present analysis was to assess the importance of each of the measures in terms of their ability to account for inherent variance within the corpus. Figure 4.3 displays the output of a PCA conducted with the 11 input measures, using the pitch interval representation (when appropriate). The first six components alone account for 88% of the total variance. The inner weightings of measures for each component are shown.

The first component was dominated by four measures—*composition length*, *number of themes*, *thematic repetition* and *thematic variation*, with a lesser influence of *internal unpredictability*. Based on this first component, and the significant correlations discussed for Table 4.1, these findings suggest that one of the biggest differences in thematic structure between compositions in the corpus can be attributed to the number of thematic candidates (which is constrained by the length of the piece), amount of thematically repeated material (normalised against the effects of length) and how much that material is varied within the composition.

The second and third principal components were dominated by measures in which stylistic congruence had the greatest influence—*stylistic unpredictability of themes*, thematic material's *stylistic unpredictability* and *thematic/stylistic balance* (with the balance favouring the LTM). Noticeably, though not a measure of style, *internal unpredictability* was also strongly weighted in these components; the presence of this measure in the second component was likely due to the strong (but opposite) relationship between it and *thematic/stylistic balance*, and in the third component, due to an element of stylistic learning occurring over these substantially-sized compositions. Component four was dominated by *length of themes*, and, while the fifth had no clear interpretation, the sixth accounted for variation in *composition year*.

Overall, the PCA provided insight into the relative importance of the model-based measures in their ability to explain the inherent variation in thematic structure within the corpus. However, we still know little about the relations between them. An Independent Component Analysis was employed to examine the independence of measures by isolating 'noise' present within each measure from similarity with others.

Independent Component Analysis

Independent Component Analysis (ICA) can be considered an extension of the previous PCA in its application to the data generated for these measures—instead of optimising components according to first- and second-order statistics in the covariance matrix, for an ICA, features with the ability to expose non-gaussian structures are optimised (Tharwat, 2018). While PCA aims to find orthogonal and uncorrelated components that could account for the maximum amount of variance in the data, ICA aims to find statistically-independent components, but that are not necessarily orthogonal. When the number of components matches the number of input features, ICA will effectively try to extract 'original' sources from the multivariate features—attempting to minimise the mutual information between components. ICA first performs 'whitening' of the data, so that it is centred and uncorrelated. Whitened data is then rotated so as to minimise Gaussianity in all dimensions, resulting in statistically independent components.[8]

When applied to the measures generated across the corpus, an ICA had the effect of isolating a component for each measure, in which any of the 'noise' related to other measures was effectively removed. In this analysis, the ordering of components does not carry any significant meaning, and there is no ordering the relative weight of each component in accounting for the data. The ICA mixings for each component were used to indicate how much each input measure

[8] Due to the Central Limit Theorem, the sum of independent random variables will fit the normal distribution more closely than the parent distributions.

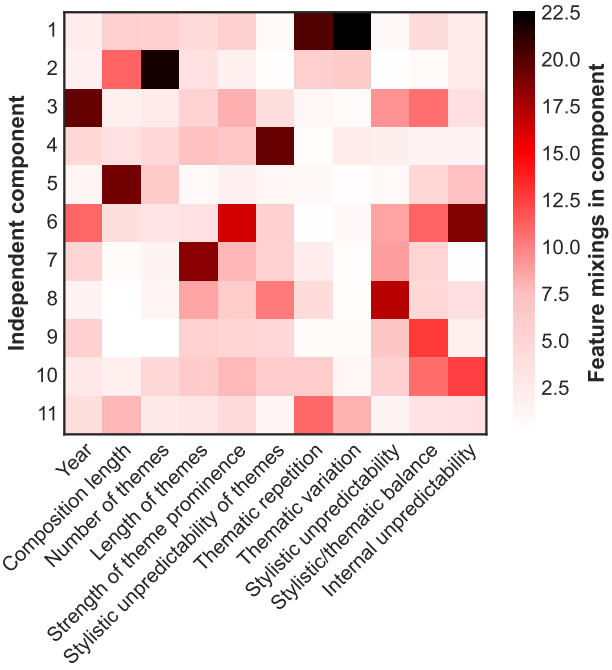


Figure 4.4: Independent mixings for pitch interval measures in output components of ICA.

influenced each independent component, as shown in Figure 4.4.

Five of the statistically independent components produced had only one main contributor, indicating that they were largely 'noise-free', with limited interaction with any of the other measures. These were: (3) *composition year*, (5) *composition length*, (7) *length of themes*, (4) *stylistic unpredictability of themes*, and (8) *stylistic unpredictability* (although the latter was not quite as independent as the others, sharing a relationship with the other measure of style). *Number of themes* maintained its connection with *composition length*, but with some degree of independence, shown in the fourth component. The first component listed strongly linked together *thematic repetition* and *thematic variation*. *Internal unpredictability* and *strength of theme prominence* (that used the same STM to identify themes) both contributed to component six.

Aside from knowing that several of our measures are independent of all others, the case of the connection between *thematic repetition* and *thematic variation* suggests that if there is more repetition within a work, it will tend to undergo embellishment and variation (at least in the case of pitch and for the works in this corpus).

Example compositions

The performance of the model measures in capturing appropriate structural variation can be illustrated by example compositions taken from the extremes

of the measures *thematic repetition*, *thematic variation*, the *stylistic unpredictability* of thematic material, and the *internal unpredictability* of each composition. Based on the underlying hypotheses of this thesis—that thematic structures are perceptible through the statistical regularities that they form, these four measures hold a particular importance when capturing the effects of large-scale thematic structure; these repetition-based measures were also found to explain the largest amount of variance in the PCA. The five compositions at the extremes of each of these measures are listed in Table 4.2.

The greater extreme end of the *thematic repetition* measure can be characterised well by the *bourrée* of Bach Cello Suite No. 4 (one of the shortest works in the dataset), that contained two detected thematic candidates and several exact repetitions of them, and the far more substantial Schubert *Impromptu*, in which seven thematic candidates were detected, each of which repeated frequently and exactly throughout the composition. Such compositions have clear statements of theme—produced by rigid structural blocks, between which there is little sharing of material—as well as containing highly frequent repetition of material with little variation. Example compositions at the lower end of this scale had little repeated material detected. For some compositions, this lack of repeated thematic material may be a feature of their style; for example, the 'impressionistic' compositional style used by Debussy in the *Arabesque* frequently involved the introduction of thematic material not necessarily closely related to that preceding it—and revisits little (Potter, 2003, p. 144; Grout et al., 2010, p. 792). In other cases, a low proportion of repeated material may have been due to the loss of information that can occur with the necessary monophonic manipulation of compositions—for example, affecting the Beethoven Piano Sonata listed.

The measure of *thematic variation* aimed to capture how far repeated material within a composition strayed from the original thematic candidates identified. A continuation of the effects of simple theme-statement and exact repetition compositions (as described for the measure of *thematic repetition*) was also seen at the lower end of this measure—little actual variation of material occurs. At the end of greater variation, example compositions were, for the most part, those that contained substantial development sections, providing greater opportunity to embellish upon previously stated themes, and those of a more fantasy-like composition style.

The works that obtained the lowest values for *stylistic unpredictability*, and so were the most stylistically predictable or congruent, were those in which repeated material mostly followed the conventions of style, as taken from the corpus.[9] This is illustrated very clearly in the Chopin *Etude*, a composition that

[9]The Bartók *Bagatelle* may seem to be something of an anomaly in this high stylistic predictability category as it is one of the compositions in the corpus that strays furthest from tonality. However, for the purpose of completeness it is included in this list. Its appearance can be ac-

Table 4.2: Example Compositions From the Extremes (Top and Bottom Five) of *Thematic Repetition*, *Thematic Variation*, *Stylistic Unpredictability*, and *Internal Unpredictability* for the Pitch Interval Representation

Composition	Movt.	Composer	Year[a]	Length	Value
		Thematic repetition			
String Quartet in C major, Op.50 No.2	2	Haydn, Joseph	1787	417	.06
Deux Arabesques, L.66	1	Debussy, Claude	1890	615	.07
Piano Sonata No.29, Op.106	3	Beethoven, Ludwig van	1817	1259	.08
String Quartet in E-flat major, Op.76 No.6	1	Haydn, Joseph	1796	372	.08
Recorder Sonata in F major, HWV 369	2	Handel, George Frideric	1712	574	.09
String Quartet in E-flat major, Op.33 No.2	4	Haydn, Joseph	1781	758	.84
Violin Sonata No.3, Op.45	1	Grieg, Edvard	1887	1239	.85
String Quartet in D major, Op.71 No.2	4	Haydn, Joseph	1793	731	.85
Vier Impromptus, D.899	1	Schubert, Franz	1827	1197	.86
Cello Suite No.4, BWV 1010	6	Bach, Johann Sebastian	1717	104	.89
		Thematic variation			
Cello Suite No.1, BWV 1007	1	Bach, Johann Sebastian	1717	654	0.29
String Quartet in B-flat major, Op.1 No.1	1	Haydn, Joseph	1757	302	0.30
Piano Sonata No.29, Op.106	3	Beethoven, Ludwig van	1817	1259	0.30
Ballade, Op.24	–	Grieg, Edvard	1900	2524	0.30
Violin Sonata No.2 in A major, Op.12	2	Beethoven, Ludwig van	1797	400	0.30
Piano Sonata in C minor, D.958	4	Schubert, Franz	1828	2583	1.28
Cello Sonata, Op.5 No.1	2	Beethoven, Ludwig van	1796	2178	1.29
Legends, Op.59	6	Dvořák, Antonín	1881	472	1.31
Violin Sonata No.1 in D major, Op.12	1	Beethoven, Ludwig van	1797	1644	1.34
Piano Sonata No.23, Op.57	1	Beethoven, Ludwig van	1804	2212	1.46
		Stylistic unpredictability			
Piano Sonata No.13, Op.27 No.1	3	Beethoven, Ludwig van	1800	226	1.54
Piano Sonata No.12, Op.26	3	Beethoven, Ludwig van	1800	443	1.63
Etudes, Op.10	2	Chopin, Frédéric	1829	766	1.76
14 Bagatelles, Op.6	14	Bartók, Béla	1908	647	1.79
String Quartet in C major, Op.33 No.3	4	Haydn, Joseph	1781	636	1.82
Cello Suite No.5, BWV 1011	4	Bach, Johann Sebastian	1717	216	4.56
Cello Suite No.1, BWV 1007	6	Bach, Johann Sebastian	1717	252	4.74
Piano Sonata, Op.7	4	Grieg, Edvard	1865	1360	5.30
6 Klavierstücke, Op.118	1	Brahms, Johannes	1893	224	5.31
5 Morceaux de fantaisie, Op.3	2	Rachmaninoff, Sergei	1892	459	5.39
		Internal unpredictability			
Vier Impromptus, D.899	4	Schubert, Franz	1827	2300	1.07
Mazurkas, Op.30	5	Chopin, Frédéric	1835	276	1.23
Piano Sonata No.20, Op.49 No.2	1	Beethoven, Ludwig van	1795	1528	1.38
Recueil d'Impromptus, Op.32	1	Alkan, Charles-Valentin	1849	621	1.41
Mazurkas, Op.30	1	Chopin, Frédéric	1835	468	1.48
8 Klavierstücke, Op.76	1	Brahms, Johannes	1871	339	4.35
String Quartet in D major, Op.1 No.3	3	Haydn, Joseph	1757	159	4.48
Préludes, Book 1	10	Debussy, Claude	1909	112	4.64
Nocturne No.10, Op.99	–	Fauré, Gabriel	1908	383	4.64
Violin Concerto, Op.61	2	Elgar, Edward	1905	139	5.24

[a]Where only a range of composition dates are known, earliest years in range are given.

almost entirely consists of chromatic scale passages. At the opposite end of this measure, example compositions were all still tonal (the entire corpus was)—possibly the biggest factor of style in the model. For this measure, it does not necessarily follow that the entire compositions themselves were *stylistically unpredictable*, it is simply that the thematic material and its derived repetitions were stylistically novel in the context of the corpus.

The final measure for which example compositions were explored was that of *internal unpredictability*. Unlike the three previous measures discussed, this measure applies to all material within a composition, rather than that identified through the repetition-detection process. In some ways, the output of this measure can be seen as similar to that of a combination of *thematic repetition* and *thematic variation* (with a closer relationship to the former); by examining the examples at the lesser extreme of this measure, it can be seen that compositions contain large amounts of repetition. However, this repetition has a slightly different nature to that identified for *thematic repetition*. The size of material repeated differed dramatically, often repeating short passages varied enough not to explicitly belong to a thematic candidate, but not novel enough to constitute new themes themselves (this effect was only a subtle one, compositions at this extreme still scored highly in *thematic repetition*).

4.4.2 The rhythmic domain

It is apparent that the rhythmic representations of the musical surface behave substantially differently to those of pitch with respect to thematic structure. The nature of rhythm, particularly within the compositions of this corpus of western-classical tonal music, is such that temporal representations face a number of challenges in capturing this type of musical structure. The effects of this were limited in the pitch-based analysis presented in the preceding portion of this chapter because, in many cases within the corpus, it is evident that pitch takes precedence over rhythm in this particular matter.

There are two fundamental problems for the model of thematic structure when trying to capture large-scale structure in the rhythmic domain: firstly, it is quite possible for pieces to be rhythmically-isochronous (or isorhythmic), such that an entire composition can consists only of notes of a single duration (or a single rhythmic pattern). For such compositions (for example, many of the Bach Cello Suites), the entirety of the structurally-important thematic information must, therefore, belong to the pitch domain, leaving no structure to model. While the level of rhythmic isochrony is variable across the corpus, it

counted for by three reasons: (1) due to the skylining process, a fair proportion of stylistically unconventional material is lost—opposing pitch classes are often used simultaneously in the two hands; (2) the resulting main thematic candidate detected is relatively stylistically predictable; and (3) it has a low level of repeated material detected which undergoes little variation, meaning all thematic material detected is still predictable.

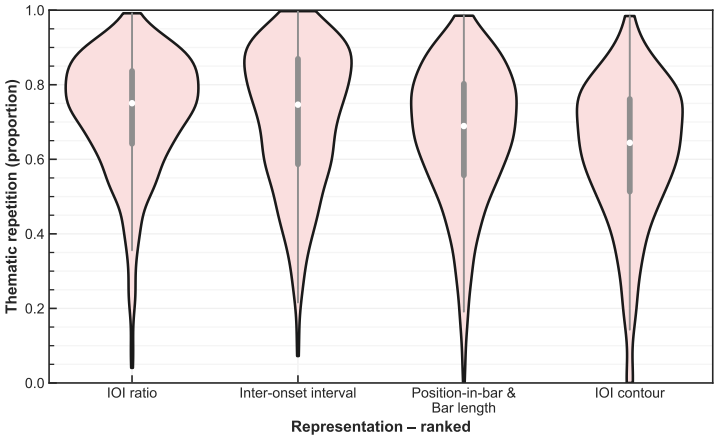


Figure 4.5: Distributions for the measure of *thematic repetition* across the corpus for all rhythmic representations, ranked in order of median value.

is present to such an extent as to mask rhythmic effects of thematic structure. The second problem leads on from the first—how should the best performing rhythmic representation be chosen? The heuristic of choosing the representation that captures the most repetition (*i.e.*, that used when selecting the pitch interval representation) is no longer productive; when a single duration is repeated relentlessly, rhythmically-isochronous works contain a maximum of repetition while containing the minimum of structural content.

Despite these limitations, some effects of large-scale thematic structure in the rhythmic domain can be observed tentatively in the following analysis.

In this rhythmic domain, *thematic repetition* was still used to inform the choice of representation. Compositions with isochronous rhythms have the same potential to influence all three of the representations based on inter-onset interval (those using position-in-bar capture fewer repetitions for this material, only allowing repeats at the bar level, rather than every note); any differences in the amounts of thematic repetition between compositions must be due to the more desirable complex rhythmic structure. Figure 4.5 shows the distributions of *thematic repetition* across the corpus for these rhythmic viewpoints. It can be seen for all representations that a large number of works appear to have been composed almost entirely of repeated rhythmic material, these cases were those with only one note duration repeated. As its distribution shows, the inter-onset interval representation contained the fewest works with low repetition, making it a reasonable choice of viewpoint for further analysis.

The respective pairwise correlations for measure results for using the inter-onset interval representation are given in Table 4.3. Summary statistics for these measures, alongside those using interval, and the correlations between the two representations are given in Table 4.4. Differing particularly from the results for interval were the mean amounts of *thematic repetition*, which, due to isochron-

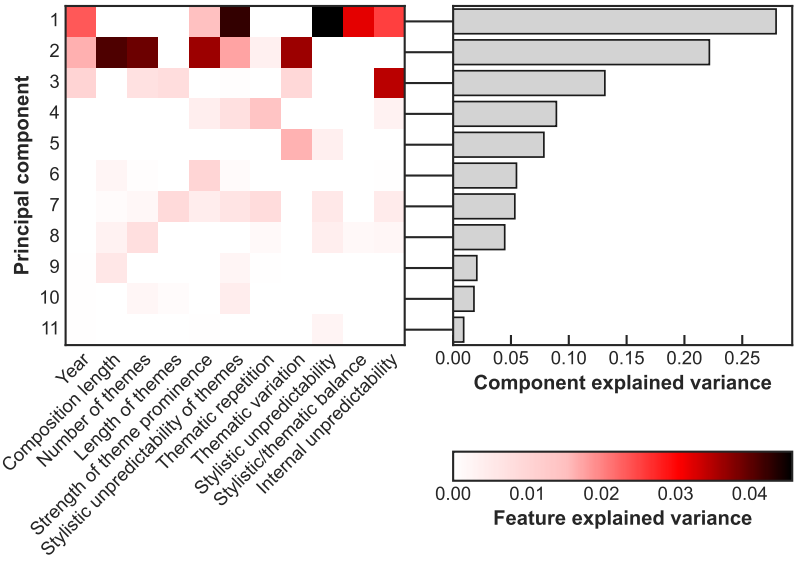


Figure 4.6: Overall explained variances for each measure using inter-onset interval and total explained variance for each output component in PCA.

ous material, was higher for inter-onset interval, and both the *stylistic unpredictability* of thematic material and *internal unpredictability*, which were lower (*i.e.*, more predictable), also likely due to the large amount of rhythmically-isochronous material.

Principal Component Analysis

As with the analysis of measures using the pitch interval representation, the amounts of variance in compositions that could be explained when using the inter-onset interval representation was explored using a PCA. For this PCA, the analysis had as its input all 11 measures (representation-independent measures of *composition year* and *composition length* included), to output 11 principal components. Figure 4.6 displays the output of this PCA; the first five components alone account for 80% of the total variance. The inner weightings of measures for each component are shown.

The results of this analysis showed that far fewer of the measures were orthogonal to each other when using this rhythmic representation, compared to those for pitch in Figure 4.3. The first three components contained large amounts of the explained variance, with the first two containing mixtures across several measures. The first identified principal component contained the strongest weightings from the *stylistic unpredictability of themes*, thematic material's *stylistic unpredictability*, *thematic/stylistic balance*, and to a lesser extent, *composition year* and *internal unpredictability*. The combination of these measures would suggest a grouping based on the stylistic properties of compositions. *Internal unpredictability*, though included here was more strongly-weighted in

Table 4.3: Correlations for Measures of Thematic Structure Using Inter-Onset Interval. Pearson's *r*, *n* = 623

	Measure	1	2	3	4	5	6	7	8	9	10
1	*Composition year*	–									
2	*Composition length*	.05	–								
3	*Number of themes*	.06	**.73**	–							
4	*Length of themes*	−.14	−.04	−.07	–						
5	*Strength of theme prominence*	.24	.36	.23	−.37	–					
6	*Stylistic unpredictability of themes*	.36	−.12	−.14	−.44	.49	–				
7	*Thematic repetition*	−.17	.13	.21	.05	−.01	−.28	–			
8	*Thematic variation*	.07	.44	.42	−.23	.22	.13	−.02	–		
9	*Stylistic unpredictability*	.26	−.22	−.29	−.17	.13	**.64**	−.45	−.01	–	
10	*Thematic/stylistic balance*	.15	−.15	−.34	−.14	.21	.33	−.25	−.15	**.66**	–
11	*Internal unpredictability*	.32	−.24	−.07	−.04	−.11	.42	−.37	.07	.40	−.22

Table 4.4: Summary Statistics and Correlations for Measures of Thematic Structure Between Representations. Pearson's *r*, *n* = 623

	Pitch interval		Inter-onset interval		
Measure	*M*	*SD*	*M*	*SD*	*r*
Number of themes	2.48	1.98	4.01	2.80	**.66**
Length of themes	17.00	3.52	17.59	3.05	**.61**
Strength of theme prominence	1.85	0.86	0.92	0.42	.12
Stylistic unpredictability of themes	3.64	0.64	2.13	0.90	.16
Thematic repetition	0.42	0.19	0.61	0.22	.08
Thematic variation	0.64	0.28	0.85	0.31	.37
Stylistic unpredictability	2.91	0.53	1.08	0.58	.13
Thematic/stylistic balance	0.65	0.12	0.28	0.15	.22
Internal unpredictability	2.77	0.63	1.68	0.42	**.59**

the third component, its partial inclusion in the first may be due to the STMs accruing stylistic knowledge within each composition, an effect that may be increased by (highly-stylistic) rhythmic isochrony.

The second principal component identified contained weightings of measures of theme-detection and variation (with a notable absence of *thematic repetition*), and the third solely that of *internal unpredictability*. The grouping together of these measures, alongside the absences of explained variance attributable only to individual measures (as seen in the pitch PCA in Figure 4.3), was likely due to the effects of rhythmically isochronous material. The presence of this material reduces the amounts of variance available in these measures; this was particularly true in the case of *thematic repetition*, which did not substantially weight any of the components.

Independent Component Analysis

Again, as with the analysis of pitch interval measures, an ICA was used to examine the statistical independence between measures using the inter-onset interval representation. The output of this analysis is shown in Figure 4.7. Unlike the corresponding analysis for the pitch representation, the output of the inter-onset interval ICA shows far fewer independent measures. Instead, there was an apparent overlapping of pairs of measures: *strength of theme prominence* and *stylistic unpredictability of themes*; *composition length* and *number of themes*; *length of themes* and *strength of theme prominence*; *thematic repetition* and *thematic variation* (twice); and *stylistic unpredictability* and *thematic/stylistic balance*. The pattern of this output suggests that there was a lack of independence between all of these measures; again, this may be due to the common effect of isochrony across measures.

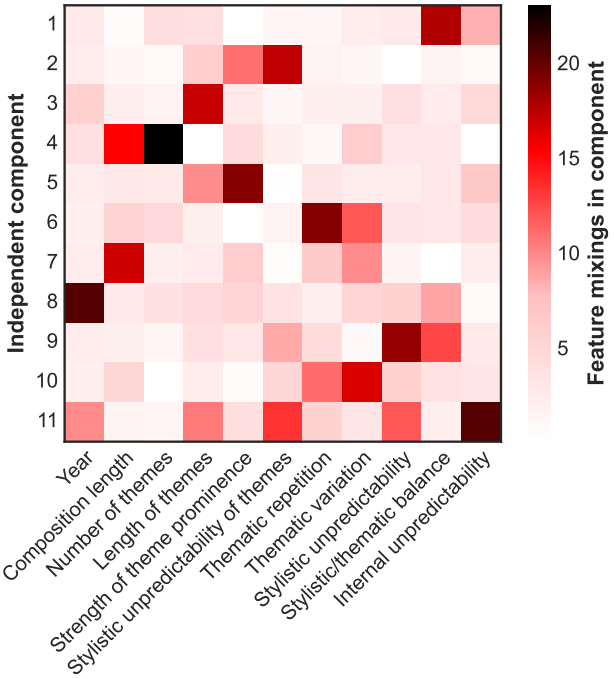


Figure 4.7: Independent mixings for inter-onset interval measures in output components of ICA.

Example compositions

The example works taken from the extremes of the measures of *thematic repetition*, *thematic variation*, *stylistic unpredictability*, and *internal unpredictability* are given in Table 4.5. The effects of rhythmically-isochronous compositions can be seen to dominate one half of each of these measures. For *thematic repetition*, these works appear as those consisting entirely of repetition (all five of the high-repetition compositions contain greater than 98% repeated material when using a rhythmic representation). For *thematic variation*, it was those with a lower value—those that underwent little variation due to all material appearing identical in this representation. For *stylistic unpredictability*, it was those at the lower end (more stylistically predictable). For *internal unpredictability*, it was those compositions with very low mean information content values.

It should also be noted that nearly rhythmically-isochronous works can have an additional contrary effect on the characterisations made by *thematic repetition*. For example, the Prelude of Bach's Cello Suite No. 3 gained the least amount of modelled repetition. The inter-onset interval representation of this work consisted entirely of sixteenth-notes, except for its opening and ending; one theme was detected at the opening, the rhythm of which is substantially different to the remainder.

However, despite the influence of these works, the converse extreme of each measure, in some cases, can still identify useful explanatory example composi-

Table 4.5: Example Compositions From the Extremes (Top and Bottom Five) of *Thematic Repetition*, *Thematic Variation*, *Stylistic Unpredictability*, and *Internal Unpredictability* for the Inter-Onset Interval Representation

Composition	Movt.	Composer	Year	Length	Value
		Thematic repetition			
Cello Suite No.3, BWV 1009	1	Bach, Johann Sebastian	1717	970	.07
String Quartet in C major, Op.50 No.2	2	Haydn, Joseph	1787	417	.09
Suite No.1 in A major, HWV 426	4	Handel, George Frideric	1720	960	.13
On the Seashore (A Memory)	–	Smetana, Bedřich	1861	304	.16
String Quartet in G major, Op.33 No.5	2	Haydn, Joseph	1781	417	.22
Cello Suite No.4, BWV 1010	1	Bach, Johann Sebastian	1717	803	.988
String Quartet in B minor, Op.33 No.1	4	Haydn, Joseph	1781	784	.989
Etudes, Op.10	5	Chopin, Frédéric	1829	944	.990
Cello Suite No.1, BWV 1007	1	Bach, Johann Sebastian	1717	654	.995
Etudes, Op.10	2	Chopin, Frédéric	1829	766	.997
		Thematic variation			
String Quartet in E-flat major, Op.76 No.6	2	Haydn, Joseph	1796	193	0.34
String Quartet in E-flat major, Op.33 No.2	3	Haydn, Joseph	1781	352	0.36
Six Etudes pour piano, Op.52	3	Saint-Saëns, Camille	1877	1309	0.36
Piano Sonata in A major, D.959	2	Schubert, Franz	1828	441	0.36
String Quartet in F major, Op.50 No.5	4	Haydn, Joseph	1787	821	0.36
Piano Sonata No.1, Op.1	1	Brahms, Johannes	1852	1677	1.76
Piano Sonata in A minor, D.845	1	Schubert, Franz	1825	1864	1.79
Piano Sonata No.7, Op.10 No.3	1	Beethoven, Ludwig van	1797	1926	1.79
Lieder ohne Worte, Op.19b	5	Mendelssohn, Felix	1829	850	1.89
Piano Sonata No.31, Op.110	1	Beethoven, Ludwig van	1821	1058	2.00
		Stylistic unpredictability			
Etudes, Op.10	2	Chopin, Frédéric	1829	766	0.06
Cello Suite No.1, BWV 1007	1	Bach, Johann Sebastian	1717	654	0.07
Cello Suite No.2, BWV 1008	3	Bach, Johann Sebastian	1717	728	0.09
Cello Suite No.4, BWV 1010	7	Bach, Johann Sebastian	1717	942	0.12
Piano Sonata No.22, Op.54	2	Beethoven, Ludwig van	1804	2374	0.13
Nocturne No.12 in E minor, Op.107	–	Fauré, Gabriel	1916	615	2.93
String Quartet in G major, Op.33 No.5	2	Haydn, Joseph	1781	417	3.03
Piano Sonata No.32, Op.111	2	Beethoven, Ludwig van	1821	2135	3.86
Suite No.1 in A major, HWV 426	4	Handel, George Frideric	1720	960	4.65
Violin Sonata No.4 in A minor, Op.23	1	Beethoven, Ludwig van	1800	1454	5.23
		Internal unpredictability			
Cello Suite No.6, BWV 1012	5	Bach, Johann Sebastian	1717	350	0.68
Cello Suite No.4, BWV 1010	5	Bach, Johann Sebastian	1717	810	0.69
Cello Suite No.6, BWV 1012	3	Bach, Johann Sebastian	1717	1278	0.70
Vier Impromptus, D.899	4	Schubert, Franz	1827	2300	0.70
Mazurkas, Op.30	5	Chopin, Frédéric	1835	276	0.72
Préludes, Book 1	10	Debussy, Claude	1909	112	2.72
Album, Op.72	5	Saint-Saëns, Camille	1884	235	2.72
String Quartet in C major, Op.50 No.2	2	Haydn, Joseph	1787	417	2.79
Miroirs	2	Ravel, Maurice	1903	356	2.85
Violin Concerto, Op.61	2	Elgar, Edward	1905	139	3.72

tions. For the measure of *thematic repetition*, some compositions with low repetition often favour diversions to material of an almost purely stylistic nature—which, in this representation are largely isorhythmic—rather than repetitions and variations of thematic material. Examples of this can be seen in the Haydn String Quartet movements listed. The higher extreme of *thematic variation* can be seen to contain mostly sonata-form movements, containing large amounts of variation due both to their relatively large length, and to the nature in which material is developed in sonata form. The compositions that were most *stylistically unpredictable* were often those that contained relatively little-used rhythms, when compared to the rest of the corpus. For example, the 9/16 time signature of Beethoven's Piano Sonata No. 32, second movement, produced rhythmic sequences (within the inter-onset interval representation that is strict with respect to absolute duration) that rarely occurred elsewhere. Finally, some compositions with high values of *internal unpredictability*, such as the work by Ravel, contained wide varieties of rhythmic patterns throughout the composition that were not substantially repeated.

4.5 Summary

To investigate the behaviour of the model described in Chapter 3, a corpus of 623 monophonic western-classical works was created, within which there is inherent variation of large-scale thematic structure between compositions. The analysis of the model had the aim of exploring the extent to which the quantitative measures of thematic structure, theme detection and general work properties could account for variation within the corpus.

The analysis of the model output when applied to the corpus was investigated separately for pitch and rhythmic representations. For both representations, the analysis first considered the relative performance of different representations of the musical surface within each domain. The analysis then examined the ability of measures to explain variance inherent in the corpus. To do so, principal component and independent component analyses were used to understand the nature of the measures. Example compositions from the extremes of the four primary model measures were assessed to further illustrate the properties of this variation.

As an initial step in the evaluation of the model of Chapter 3, this analysis provides useful information for the behavioural studies that make up the remainder of this thesis. Firstly, the corpus created and described here provides the ability to computationally sample experimental stimuli, based on their modelled properties. This aids the experiment designs as it can provide the assurance that stimuli provide sufficient variance in the domain of interest. Secondly, the comparison of the performance of musical representations allowed for the

most-suitable pitch and rhythm representations to be identified; these representations are then used in the modelling of experimental data, allowing for the separate effects of pitch and rhythm to be examined without requiring a process of representation selection for each experiment. Finally, this corpus analysis first identifies the patterns of results produced by the presence of purely rhythmically-isochronous compositions, an important consideration when interpreting the output of the model in the rhythmic domain.

Chapter 5

Modelling Small-Scale Thematic Structure

5.1 Overview

This chapter presents the findings of a first behavioural experiment that aimed to provide some preliminary validation of the modelling of thematic structure, based on the statistical learning of structurally-important features, described in Chapter 3. While the primary focus of this thesis is on the perception on thematic structures on the large-scale, in this initial experiment it was important to examine whether measures could predict perception on a local level before extending the enquiry to large-scale thematic structures in the following parts of the thesis. The experiment tested the abilities of participants to perceive a relationship between pairs of themes and repetition phrases, identified by the theme and repetition-detection processes described in Chapter 3, within the monophonic compositions of the corpus created in Chapter 4.

In many ways, the task of this experiment follows on from those reviewed in Chapter 2 that also considered the perceived similarity between passages of music (Lamont & Dibben, 2001; Ziv & Eitan, 2007), with the added dimension of computationally-chosen stimuli and analysis of the statistical and stylistic features of excerpts. The experimental question used differs from an explicit judgement of similarity (for example, as used by Lamont & Dibben, 2001); in this experiment, participants were asked to rate the extent to which they believed the two musical excerpts belonged to the same original work or not. While each pair of excerpts presented were from the same composition in all cases, this question was chosen so as to reflect the fact that pairs could be judged according to both intra-opus and stylistic properties.

The chapter is ordered as follows. The elements of computational modelling are set out. Due to the small-scale focus of this experiment, the complete set of measures introduced in Chapter 3 are not all applicable; three model-

derived measures specific to this experiment are described. Then follows the hypotheses, methods and results of the experiment, with a final discussion of the findings and implications of for the perception of thematic structure.

5.2 Modelling

The measures of thematic structure introduced in Chapter 3 and tested in Chapter 4 were designed to characterise statistical elements across complete pieces of music. The measures were formed using different configurations of IDyOM variable-order Markov models, using different training domains. IDyOM, given a training sequence, estimates likelihoods of the occurrence of note-events in a sequence, smoothing between models of different orders (Pearce, 2005). These likelihood estimates were converted into a value of information content—the unpredictability of a note-event occurring, given the training sequence.

To allow for measures to be created based only on the thematically-relevant material within a composition, Chapter 3 presented a method of identifying possible novel themes, and repetitions related to such themes, based on patterns of *internal unpredictably* (STM information content) within a composition.

5.2.1 Measures

For this experiment, three measures were used to characterise the theme–phrase pairs according to different hypothesised statistical learning mechanisms. Due to the focus on small time-scales in this initial behavioural study, measures were simpler and fewer than the primary measures of Chapter 3 from which they were derived. Pairs could differ in two ways: (1) in their thematic similarity to each other (*i.e.*, the extent to which one was predictable, given the other), or (2) in the extent to which they were predictable, given stylistic norms internalised during long-term prior musical experience. Additionally, the pair together could be characterised by their stylistic predictability. By understanding these properties on a local scale, we can learn how they may function during a composition, where frequent judgments of similarity, both thematic and stylistic, may be made between occurring themes and repetitions.

While the model of thematic structure used to generate these experiment measures (and the IDyOM components underpinning it) can be applied to many different representations of a musical surface, this experiment and its analyses were limited to two—the sequential interval between pitches (in semitones), and the inter-onset intervals between note events. These two representations covered both melodic pitch and rhythmic domains, providing an exploration of the independent effects of these domains on the perception of

thematic structure, while still allowing the evaluation of the measures to be tractable. The measures used in this experiment therefore had two forms, one using each representation.

Dissimilarity. The first of these differences was quantified in an experiment measure of *dissimilarity*. This measure, the singular instance of Chapter 3's *thematic variation*, used the information-theoretic measure of compression distance (Li et al., 2003; Pearce & Müllensiefen, 2017) between the theme and the subsequent repetition. The distance $D(r|t)$ between a theme, t, and a later repetition, r (length l), is calculated as the normalised summed information content of the note-events of the repetition phrase, given a model trained on the corresponding theme, the measure of *dissimilarity* used in this experiment is also related closely to a composition's *internal unpredictability*.

$$D(r_1^l \mid t_1^k) = \frac{1}{l} \sum_{i=1}^{l} h_t(r_i)$$

Stylistic difference. To characterise differences between theme and repetition *stylistic unpredictability*, the stylistic information content of each was modelled using an IDyOM LTM, trained on a corpus of western tonal melodies (described in Chapter 4). The absolute difference between *stylistic unpredictability* information contents for the two sequences was used to form the measure of experiment measure of *stylistic difference*. For a theme, t_1^k, of k events and repetition, r_1^l, of l events, where h_{ltm} is the LTM information content:

$$Y(t_1^k, r_1^l) = \Big| \frac{1}{k} \sum_{i=1}^{k} h_{\text{ltm}}(t_i|t_1^{i-1}) - \frac{1}{l} \sum_{i=1}^{l} h_{\text{ltm}}(r_i|r_1^{i-1}) \Big|$$

Mean stylistic unpredictability. To test whether the overall *stylistic unpredictability* of a pair would influence the perception of a relationship between the themes and repetitions, a second derivative of *stylistic unpredictability* was also used. *Mean stylistic unpredictability* quantified how predictable each pair was, combined, using the grand mean of *stylistic unpredictability* of themes and repetitions.

$$S(t_1^k, r_1^l) = \frac{1}{2}\Big(\frac{1}{k} \sum_{i=1}^{k} h_{\text{ltm}}(t_i|t_1^{i-1}) + \frac{1}{l} \sum_{i=1}^{l} h_{\text{ltm}}(r_i|r_1^{i-1})\Big)$$

5.3 The present experiment

This first behavioural experiment aimed to provide some preliminary validation of the modelling of thematic structure presented in Chapter 3; it aimed to test

the extent to which measures based on statistical learning within a composition could explain listeners' perception of relationships between passages of music. It was important to examine whether the model-based measures could predict perception on a local level before extending the enquiry to large-scale thematic structures in the subsequent experiments of this thesis—the testing of which introduces multiple complications. This experiment specifically examined the effects of *dissimilarity* (the individual component of *thematic variation*), *stylistic difference* and *mean stylistic unpredictability* (both derived from *stylistic unpredictability*) on the extent to which pairs of stimuli, each consisting of a theme and a later thematic repetition, were perceived as being derived from the same piece of music

For this experiment, it was hypothesised that all three of the measures of thematic structure would be able to account significantly for participants' perception of thematic relationships in some way. This general hypothesis would provide evidence that thematic structure can be perceived in a systematic way, and that this perception can be simulated in terms of psychological mechanisms of statistical learning.

More specifically, three hypotheses were advanced for the relationships between participant ratings and the experiment measures:

1. That *dissimilarity* would have the strongest effect on participant responses, versus the stylistic measures, with stimulus pairs showing high *dissimilarity* more likely to be judged as unrelated

2. That *stylistic difference* would be the measure of second-most importance, with pairs of high *stylistic difference* being more likely to be perceived as unrelated, particularly when combined with high *dissimilarity* not immediately identifying the two as related (*i.e.*, two *dissimilar* phrases could still be perceived as originating from the same work if they were close in style)

3. That overall *mean stylistic unpredictability* would be associated with a tendency to perceive the excerpts as unrelated—if the stimuli were stylistically novel to a listener, their ability to judge the magnitude of distance of the other measures would be reduced.

Additionally, for this experiment, measures of participants' musical backgrounds were recorded. It was hypothesised that participants with greater musical training, and so having a greater prior exposure to music and its stylistic conventions, would increase the impact of the two stylistic measures on ratings. We made no specific hypotheses as to the effects of musical background on the intra-opus *dissimilarity* measure.

5.4 Methods

5.4.1 Participants

Forty participants were recruited to participate through the online platform Prolific.[1] No exclusion criteria were applied other than a required first language of English and normal or corrected to normal hearing. Participants had a mean age of 31.98 years (*SD* = 10.76) and 24 were women. Participants were of eight nationalities, with 30 of UK nationality. There were no prerequisites on musical training; 20 participants reported having received some formal musical training on at least one instrument, of which four reported training for more than 10 years.

5.4.2 Stimuli

Stimuli were selected from a large corpus of western-classical tonal melodies (presented in Chapter 4) using the output of the theme and repetition detection models presented in Chapter 3. Stimuli took the form of pairs of a theme and a later phrase from the same composition identified by the model as being related (or belonging) to that theme. These theme–phrase pairs varied in three experiment measures: (1) *dissimilarity*, the normalised compression distance between the two stimuli; (2) *stylistic difference*, the absolute difference in LTM information content between the two stimuli; and (3) *mean stylistic unpredictability*, the mean LTM information content of the pair.

Theme–phrase pairs were selected from the full set of possible pairs extracted from the corpus by the model. Outliers more distant than 2 SD from the mean for each measure were removed. Phrases that, in their entirety, were exact matches with any sub-part of the theme were discarded; for the paradigm used, these phases would not return useful information. To avoid any possible overlap of material between stimuli, selection was limited to one phrase per theme and one theme per composition—the first theme was used in all cases. Following these constraints, 100 theme–phrase pairs were selected at random from the subset of 6,687 available.[2]

Audio files were generated for the selected stimuli in a piano timbre with a uniform loudness using MuseScore[3] notation software. Original pitches, rhythms and tempos were preserved.

In the selected stimuli, themes had a mean duration of 8.32 seconds (*SD* = 4.92) and phrases had a mean duration of 4.93 seconds (*SD* = 1.50).

[1] https://www.prolific.co/

[2] Modelling code, stimuli and data for this experiment can be found at https://osf.io/zmqh2/.

[3] https://musescore.org/

5.4.3 Procedure

The experiment was implemented using the jsPsych[4] JavaScript library—with a custom designed task—to be conducted online through the participant's web browser (de Leeuw, 2015).

The experiment consisted of two parts. For the first (the more substantial task, taking around 30 minutes for completion), participants were presented with all 100 of the theme–phrase pairs (theme first, separated by a short gap) in a randomised order. Stimulus pairs were presented in four equal blocks with a 30-second mandatory rest period in-between each. Participants were not informed of the theme–phrase relationship between the pairs of stimuli; instead they were instructed that pairs may, or may not, be excerpts from two different compositions. Participants were asked to give a rating answering the question 'to what extent do these two excerpts sound like they are from the same piece?' Ratings were given on a moveable slider between 'same piece' (returning a value of 0) and 'different pieces' (returning 100). Participants were required to listen fully to both excerpts before submitting a rating. Additionally, participants were asked to indicate if the piece corresponding to either excerpt was known to them.

In the second part of the experiment, participants gave their responses to the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire for musical sophistication (Müllensiefen et al., 2014).

5.4.4 Statistical analysis

Ratings for stimuli were treated differently in five separate analyses. For all but the final analysis, the effects of measures in the pitch and rhythmic representations are presented separately.

The first analysis was categorical, comparing participant responses between high and low values of each experiment measure. Ratings for theme–phrase pairs were split into two at the median of each of the three measures, giving high/low *dissimilarity* (dissimilar/similar), high/low *stylistic difference* (stylistically distant/close) and high/low *mean stylistic unpredictability* (stylistically unpredictable/predictable) separately for both pitch and onset models. For each participant, ratings for stimuli in each cell of the resulting factorial design were averaged. A three-way repeated-measures ANOVA was used for each representation to test main and interaction effects of the three categorical measures on participant ratings.

The remaining analyses were continuous, the second using correlation coefficients (Pearson's r) between the mean participant results for each theme–phrase pair and the pair's corresponding value for *dissimilarity*, *stylistic differ-*

[4] https://www.jspsych.org/

ence and *mean stylistic unpredictability* in each representation. Third, a multiple linear regression analysis was used, for each representation, to test the combined abilities of *dissimilarity*, *stylistic difference*, and *mean stylistic unpredictability* to predict the ratings of theme–phrase pairs. Fourth, a multiple linear regression analysis compared the effects of measures between pitch and onset representations on participant ratings.

In the final analysis, participant responses to Gold-MSI questions were aggregated by summing participant responses (accounting for question phrasing) to produce a measure of *general sophistication* for each participant, and aggregated for question subsets relating to participants' *perceptual abilities* and *musical training*. Correlation coefficients were also calculated between mean ratings for each participant and their aggregated Gold-MSI scores, and between linear regression slopes for individual participants in each measure and Gold-MSI scores.

5.5 Results

The results are presented individually for each analysis as follows: (1) a summary of participant ratings, their distributions, and Gold-MSI scores; (2) the ANOVA for interactions between ratings and model measures; (3) correlation and multiple linear regression analyses of relationships between ratings and model measures; (4) the multiple linear regression analysis comparing measures from models using pitch and rhythm representations; and (5) the analysis of correlations between ratings and model measures, and between participants' Gold-MSI scores, slopes and their ratings.

All 40 participants returned ratings for all 100 theme–phrase stimuli pairs and all gave full responses to the Gold-MSI questionnaire. Participants gave ratings of the extent to which the two phrases belonged to the same composition with a mean of 47.36 (SD = 34.29), using 98% (SD = 3.73) of the scale on average.

After aggregation of responses to the Gold-MSI self-report questionnaire into scores for each participant, participants had a mean score for *general sophistication* (scale range 18–126) of 53.30 (SD = 17.19), a mean score for *perceptual abilities* (out of a possible scale range of 9–63) of 38.48 (SD = 6.30), and a mean score for *musical training* (scale range 7–49) of 14.53 (SD = 9.58).

5.5.1 Categorical analysis

For the pitch interval and inter-onset interval representations, stimuli were categorised as being either 'high' or 'low' in each of their three model measures, split at the median. Participants' ratings were averaged for the stimuli in the eight combinations of categories produced, giving one value per participant in

Table 5.1: Descriptive Statistics of the Ratings for Stimuli in the Factorial Experimental Conditions of *Dissimilarity* (high, low), *Stylistic Difference* (high, low), and *Mean Stylistic Unpredictability* (high, low)

Stylistic difference	*Mean stylistic unpredictability*	*Dissimilarity*			
		High		Low	
		M	*SD*	*M*	*SD*
		Pitch interval			
High	High	59.09	7.27	43.60	12.82
	Low	60.67	8.93	42.20	11.64
Low	High	61.60	14.13	34.08	10.30
	Low	50.38	8.46	33.48	7.82
		Inter-onset interval			
High	High	63.15	7.98	28.62	11.21
	Low	66.18	11.00	37.12	9.12
Low	High	60.11	9.20	37.81	9.85
	Low	59.08	15.15	27.43	9.51

each condition. Table 5.1 summarises the ratings for these categories.

A three-way ANOVA (repeated-measures) was used to test the interaction between the three categorical measures and participant ratings for each representation. For pitch interval, individually all three measures showed significant main effects: *dissimilarity*, $F(1, 39) = 230.40$, $p < .001$, $\eta_p^2 = .59$; *stylistic difference*, $F(1, 39) = 34.06$, $p < .001$, $\eta_p^2 = .14$; and *mean stylistic unpredictability*, $F(1, 39) = 14.41$, $p < .001$, $\eta_p^2 = .03$. For all three measures, mean ratings were higher (*i.e.*, indicating different pieces) for the 'high' category. There was a significant two-way interaction between *dissimilarity* and *stylistic difference,* $F(1, 39) = 9.72$, $p < .01$, $\eta_p^2 = .03$, such that the effect of *stylistic difference* was greater when *dissimilarity* was low. There was also a significant interaction between *stylistic difference* and *mean stylistic unpredictability*, $F(1, 39) = 12.79$, $p < .001$, $\eta_p^2 = .03$, such that the effect of *mean stylistic unpredictability* was greater when *stylistic difference* was low. There was a significant three-way interaction between all model measures, $F(1, 39) = 13.18$, $p < .001$, $\eta_p^2 = .04$, such that when *dissimilarity* was high and *stylistic difference* low, there was a greater effect of *mean stylistic unpredictability* (see Figure 5.1).

For inter-onset interval, significant effects were shown for individual measures of *dissimilarity,* $F(1, 39) = 870.89$, $p < .001$, $\eta_p^2 = .76$ and *stylistic difference*, $F(1, 39) = 7.13$, $p < .01$, $\eta_p^2 = .03$, but not for *mean stylistic unpredictability*. For both significant measures, mean ratings were higher for the 'high' category. There was a significant two-way interaction between *dissimilarity* and *stylistic difference,* $F(1, 39) = 5.87$, $p = .02$, $\eta_p^2 = .02$, such that the effect of *stylistic difference* was greater when *dissimilarity* was high—the converse relationship to that of pitch interval. There was also a significant interaction between *stylistic difference* and *mean stylistic unpredictability*, $F(1, 39) = 33.16$, $p < .001$, $\eta_p^2 = .11$,

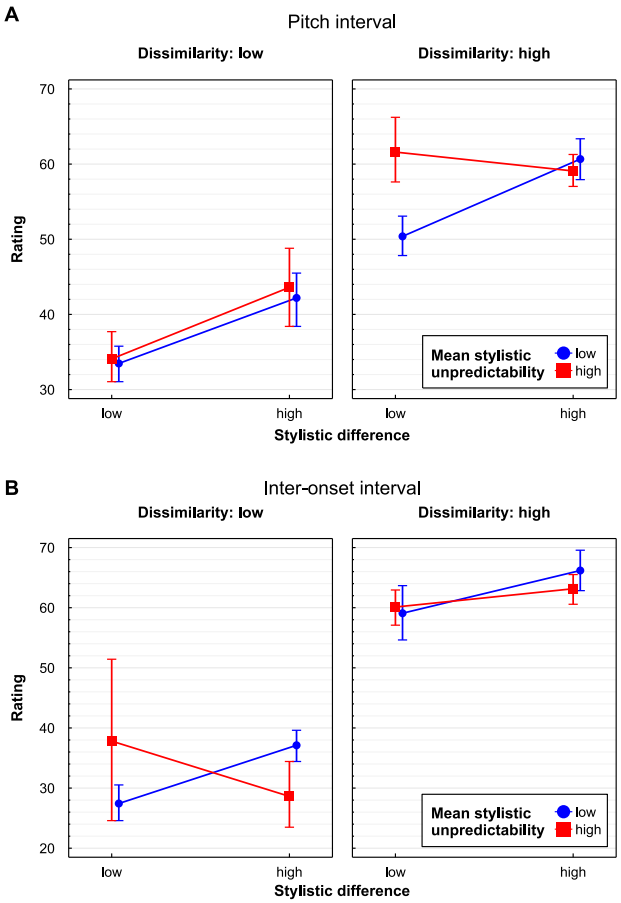


Figure 5.1: Interaction effects of categorical measures on mean participant ratings for (A) pitch interval and (B) inter-onset interval. Note: Error bars show 95% confidence intervals.

Table 5.2: Pearson's *r* Correlations Between Experiment Measures and Participant Ratings

Variable	*M*	*SD*	1	2	3
Pitch interval					
1 *Dissimilarity*	0.88	0.27	–		
2 *Stylistic difference*	0.73	0.50	.29 ••	–	
3 *Mean stylistic unpredictability*	3.20	0.53	.04	.09	–
Ratings (stimuli means)	47.36	21.69	.58 •••	.37 •••	.05
Inter-onset interval					
1 *Dissimilarity*	1.56	1.19	–		
2 *Stylistic difference*	1.18	1.13	.23 •	–	
3 *Mean stylistic unpredictability*	1.93	0.81	.04	.25 •	–
Ratings (stimuli means)	47.36	21.69	.56 •••	.27 ••	.17

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

such that when *stylistic difference* was low, pairs with low *mean stylistic unpredictability* would be given lower ratings (*i.e.*, indicating the same parent piece), with the inverse true at high *stylistic difference*. A significant three-way interaction between all model measures, $F(1, 39) = 13.84$, $p < .001$, $\eta_p^2 = .05$, shows that the interaction between *stylistic difference* and *mean stylistic unpredictability* is greatest when *dissimilarity* is low (see Figure 5.1).

5.5.2 Continuous analyses

Participants' ratings for each stimulus were averaged to give a mean rating for each theme–phrase pair. These mean ratings were then analysed using the three model measures in each representation as predictors. For both pitch interval and inter-onset interval, the mean ratings had a significant positive correlations both with *dissimilarity* ($r(98) = .58$, $p < .001$ and $r(98) = .56$, $p < .001$, respectively) and with *stylistic difference* ($r(98) = .37$, $p < .001$ and $r(98) = .27$, $p < .01$) but had no correlation with *mean stylistic unpredictability*, ($r(98) = .05$, $p = .64$ and $r(98) = .17$, $p = .08$). The extent to which measures themselves were correlated is given in Table 5.2.

Using mixed-effects multiple linear regression, ratings were analysed using the three measures as predictor variables, accounting for the random effects of participant and stimulus pair. Due to the over-fitting of data when using the maximal random effects structure, only random intercepts were included. Interaction effects were included based on the significant interactions found in the preceding categorical analysis, Summaries for analyses for both measures are shown in Table 5.3. For pitch interval, *dissimilarity* was found to be a significant predictor, with greater *dissimilarity* predicting higher ratings that pairs were not thought to come from the same piece, $\beta^* = 0.32$. The pitch interval model accounted for 41.86% of the total variance in the data; participant inter-

cepts $SD = 5.59$, stimulus intercepts $SD = 17.04$. No significant interactions were found between the variables in their continuous form. For inter-onset interval, all three measures were significant: *dissimilarity*, $\beta^* = 0.34$; *stylistic difference*, $\beta^* = 0.17$; and *mean stylistic unpredictability*, $\beta^* = 0.09$. The inter-onset interval had one significant interaction between the two stylistic measures, as the *stylistic difference* increased the effect of *mean stylistic unpredictability* decreased. The inter-onset interval model accounted for 41.81% of the total variance in the data; participant intercepts $SD = 5.59$, stimulus intercepts $SD = 16.67$.

A final mixed-effects linear regression analysis was used to test the relative ability of all measures across both representations to predict participant's ratings, again with random intercepts of participant and stimulus. The one interaction that was supported in its continuous form, between inter-onset interval *stylistic difference* and *mean stylistic unpredictability* was included. As shown in Table 5.4, *dissimilarity* in both pitch interval and inter-onset interval was found to be a significant predictor of ratings ($\beta^* = 0.24$, $\beta^* = 0.24$), with higher *dissimilarity* corresponding with higher ratings. Additionally, pitch interval *stylistic difference* was found to have a significant effect ($\beta^* = 0.10$), as well as inter-onset interval *mean stylistic unpredictability* ($\beta^* = 0.10$). The combined representation model accounted for 41.69% of the total variance in the data; participant intercepts $SD = 5.59$, stimulus intercepts $SD = 14.42$.

To test whether participant's musical backgrounds influenced ratings, Gold-MSI scores were correlated both with the mean ratings for each participant and with coefficients (slopes) from simple linear models predicting each participant's ratings from each of the model measures in each of the representations. No significant correlation was found between any of the three scores of *general sophistication*, *perceptual abilities* and *musical training*, and participants' mean ratings or slopes for the pitch interval model. A significant correlation was found between *general sophistication* and slopes using inter-onset interval *stylistic difference,* $r(38) = .37$, $p = .02$, such that ratings from participants with higher Gold-MSI scores were more influenced by the *stylistic difference* of pairs.

5.6 Discussion

This first behavioural experiment sought to provide some preliminary validation of the modelling of thematic structure, based on the statistical learning of structurally-important features, described in Chapter 3. Understanding how these features interact and influence the perception of musical phrases presented in isolation as immediate pairwise comparisons of thematic material was an important first step to understanding their effects when integrated into music over far longer timespans. For repetition to be identified, the process of comparing incoming musical material to existing themes already heard and stored

Table 5.3: Mixed-Effects Linear Regression Analyses Predicting Participants' Belongingness Ratings by Experiment Measures for Each Pitch and Rhythmic Representation, Accounting for Participant and Stimulus Differences

Predictor	β	*SE*	*df*	*t*	*p*	
Pitch interval						
Intercept	12.75	24.65	93.24	0.52	.61	
Dissimilarity	42.12	12.01	93.00	3.51	<.001	•••
Stylistic difference	2.78	61.89	93.00	−0.05	.96	
Mean stylistic unpredictability	−3.08	6.71	93.00	−0.46	.65	
Dissimilarity : *Stylistic difference*	−0.94	57.82	93.00	−0.02	.99	
Stylistic difference : *Mean stylistic unpredictability*	4.29	19.23	93.00	0.22	.82	
Dissimilarity : *Stylistic difference* : *Mean stylistic unpredictability*	−0.11	17.31	93.00	−0.01	.99	
Inter-onset interval						
Intercept	3.69	8.26	95.12	0.47	0.65	••
Dissimilarity	14.06	3.34	93.00	4.21	<.001	•••
Stylistic difference	26.78	9.37	93.00	2.86	<.01	••
Mean stylistic unpredictability	8.99	3.02	93.00	2.98	<.01	••
Dissimilarity : *Stylistic difference*	−8.51	5.27	93.00	−1.62	.11	
Stylistic difference : *Mean stylistic unpredictability*	−8.34	3.20	93.00	−2.61	.01	•
Dissimilarity : *Stylistic difference* : *Mean stylistic unpredictability*	2.56	1.73	93.00	1.48	.14	

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

Table 5.4: Mixed-Effects Linear Regression Analyses Predicting Participants' Belongingness Ratings by Experiment Measures Across Both Pitch and Rhythmic Representations, Accounting for Participant and Stimulus Differences

Predictor	β	*SE*	*df*	*t*	*p*	
Intercept	−11.43	11.34	93.12	−1.01	.32	
Pitch interval						
Dissimilarity	30.86	6.41	92.00	4.81	<.001	•••
Stylistic difference	6.63	3.28	92.00	2.02	.05	•
Mean stylistic unpredictability	−0.24	3.02	92.00	−0.08	.94	
Inter-onset interval						
Dissimilarity	6.87	1.38	92.00	4.97	<.001	•••
Stylistic difference	8.02	4.10	92.00	1.96	.05	
Mean stylistic unpredictability	7.46	2.70	92.00	2.76	.01	••
Stylistic difference : *Mean stylistic unpredictability*	−2.84	1.46	92.00	−1.95	.05	

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

in memory must happen many times throughout the course of listening to a composition. For the statistically-learned features discussed in this thesis to be able to influence the perception of such large-scale repetition across an entire composition, they must first influence individual instances of repetition within the music. Therefore, the rationale of the experiment presented in this chapter was to test the model-based measures of thematic structure in the microcosm provided by single pairwise comparisons of musical passages from the same piece.

Due to the nature of this experimental paradigm, only a subset of the primary model measures discussed in Chapter 3 were applicable—*dissimilarity* (derived from *thematic variation*), *stylistic difference* and *mean stylistic unpredictability* (both derived from *stylistic unpredictability*). Of the remaining measures, *thematic repetition* was not meaningful in this context, and *internal unpredictability*—insofar as its functional aspect in this paradigm is the *unpredictability* of the phrase given the theme—is encompassed by *dissimilarity* between the two musical passages.

As hypothesised, the results of this experiment showed that the three experiment measures, using both pitch interval and inter-onset interval representations, had a significant influence on listeners' perception of theme–repetition relationships, operationalised as the extent to which the two excerpts sounded like they came from the same piece of music. The measure of *dissimilarity* between the theme and repetition had the greatest effect on this relationship, such that—as hypothesised—high *dissimilarity* resulted in pairs being less likely to be judged as belonging to the same piece. This finding was confirmed for pitch and rhythm variants separately, and when measures from both representations were combined in analyses together. Given that the task is closely related to one providing direct similarity ratings between musical passages, and the measure of *dissimilarity* is calculated as the compression distance between the pair of excerpts, this result provides further evidence that compression distance accurately simulates perceived similarity between pairs of melodies (Pearce & Müllensiefen, 2017).

Compression distance provided a direct, encapsulated measure of dissimilarity that was found to be independent of any relationship the excerpts had with other music or musical styles. This effect of *dissimilarity* provides some support for the importance of intra-opus statistical learning in the perception of thematic relationships in music. The evidence that these stimulus pairs can be judged as being related solely through the unpredictability of one, given the other, supports similar relationships being perceived in the same way when pairs are situated in large pieces of music.

However, the results also suggest that there was a secondary factor influencing the extent to which listeners perceived two musical passages as coming

from the same piece; evidence was found supporting the case that judgements were also dependent on differences between theme and repetition congruence with western stylistic norms, internalised in memory via statistical learning during long-term prior musical experience. Independent of any effect of *dissimilarity*, increasing the pitch interval *stylistic difference* between the two passages decreased the likelihood that they would be considered to belong to the same piece (as illustrated in Figure 5.1).

The results also uncovered evidence of a more nuanced role of style in the perception of these small-scale thematic structures. In addition to the effects found for *stylistic difference*, the *mean stylistic unpredictability* of pairs of themes and repetitions together showed a more subtle influence; this suggested that, when the two passages showed little *stylistic difference*, greater overall incongruence with western-musical norms resulted in listeners being more likely to perceive the two passages as coming from different compositions, perhaps on the basis of overall stylistic unfamiliarity.

The results of this experiment provided little evidence that the musical backgrounds of listeners had any influence their judgements of relationships between theme–repetition pairs. Only one significant difference was found, with participants' *general sophistication* influencing the importance of the *stylistic difference* measure for rhythm only. This finding provides some tentative support that listeners' musical backgrounds, through their greater musical training and so greater exposure to musical conventions, had some influence on their sensitivity to stylistic elements. The lack of any other significant findings based on musical background should perhaps be treated with some caution; it was not a primary aim of this experiment to explicitly compare musically-trained and non-trained participants, and so the participants have a narrower range of musical expertise and background than would otherwise be chosen. However, the lack of any effect of musical background on any intra-opus measures for this task corroborates the findings of Lamont and Dibben (2001) and Ziv and Eitan (2007), both of which explicitly tested for—and did not find—effects of musical training on similarity judgement tasks.

5.7 Summary

The experiment presented in this chapter provides an important initial corroboration of the theory and accompanying computational measures presented in Chapter 3, providing evidence in support of statistical learning as an underlying psychological mechanism for the perception of thematic structure. Specifically, the three measures used in this experiment provided evidence that intra-opus compression distance between two thematic excerpts, and stylistic regularity acquired through long-term statistical learning, influenced how listeners per-

ceived relationships between the subsections of music making up an overall piece. This experiment, however, is limited to testing these measures on a short timescale, without the repetition of material that typically occurs in music. The local effects revealed in this experiment are not necessarily generalisable, on their own, to the perception of thematic structure on the much larger timescales found in musical compositions.

Chapter 6

Modelling Large-Scale Repetition and Unity

6.1 Overview

The previous chapter reported an experiment providing an initial behavioural test of some of the concepts core to the understanding of thematic structure as a statistical process on a local scale, the present chapter presents a experiment that provides substantial testing of our theory and its model on the timescales needed for large-scale structure. The experiment had two primary objectives—to test the ability of listeners to identify large-scale repetition within a composition, and to test their ability to perceive the structural unity of the composition as a whole—both being highly-important indicators of large-scale thematic structure in music. To investigate the first of these, a task was used in which participants were presented long monophonic compositions (with durations of approximately 2 minutes) and asked, at four given moments in the later portion of the composition, whether material at that point constituted a repetition of material from earlier within the composition. The task investigating the second of these—the perception of unity—used a task following those used in the previous work of Tan and Spackman (2005), Tan et al. (2006) and Lalitte and Bigand (2006); participants were asked to rate the given composition on the extent of its perceived structural unity. These two tasks serve as important indicator functions for the more general perception of thematic structure—the perception of which is difficult to experimentally measure directly (see Chapter 2 for a review of similar ratings of structure in previous experimental work).

As with the previous experimental chapter, this chapter first presents the relevant modelling concepts specific to this experiment, before presenting the experiment's procedure, results and a discussion of its findings.

6.2 Modelling

Chapter 3 of this thesis proposed that statistical learning could provide a plausible underlying mechanism for the perception of thematic structure. According to this proposal, large-scale thematic structures are perceived through the implicit recognition of statistical regularities learned through both exact and inexact repetition and variation of material. The model of the perception of thematic structure, presented in that chapter, employed the probabilistic modelling of IDyOM to calculate a range of measures quantifying different properties of a composition's thematic structure.

For a given composition, three configurations of IDyOM were used, differing in the music used for training:

1. As a short-term model (STM), learning incrementally from an initially empty state within a given musical sequence, representing a listener's short-term acquisition of statistical knowledge about repeated structure within an individual piece of music (see Chapter 3.3)

2. As a long-term model (LTM), in which it is trained on a separate set of musical sequences, representing long-term learning of the statistical structure of a musical style, before being applied to predicting the notes of a musical composition

3. As a theme-trained model (TTM), trained on each theme identified in a composition (one TTM per theme) and then used to predict the remainder of that composition (see Chapter 3.4)

All IDyOM models, however they were configured for training, computed a conditional probability estimate for each note-event in a composition when trained, respectively, on the preceding portion of the composition (dynamically), on a corpus of melodies, or on the possible themes identified within the composition. Probability estimates were converted into information content values, $h = -\log_2 p$, giving a representation of the unpredictability of the event, given the event's context and the respective model training.

6.2.1 Measures

The measures introduced in Chapter 3 represented different hypotheses about the psychological mechanisms underlying perception of thematic structure. The experiment reported in the present chapter provided an empirical comparison of four of these measures—*internal unpredictability*, *thematic repetition*, *thematic variation*, and *stylistic unpredictability*. These measures were adapted to the two tasks of the experiment, with terms used to distinguish between

whether measures were applied to the entirety of a given composition, or characterising properties up to a given moment within the composition.

All of the measures used in this experiment were used in two variants, each using a different representation of the musical surface: (1) using the sequential interval between pitches, and (2) using the inter-onset intervals between note-events. These two representations covered both melodic pitch and rhythmic domains, providing an exploration of the independent effects of these domains on the perception of thematic structure. Of these, the pitch interval representation was prioritised in stimulus selection to avoid the complications identified in Chapter 4 when modelling thematic structure with rhythmic representations; it is viable (and relatively common) in western-classical music for compositions to be extensively rhythmically isochronous.

Internal unpredictability

Internal unpredictability. For the present experiment, the measure of *internal unpredictability* was used as defined in Chapter 3, specifically referring to the mean STM information content of note-events in a composition.

Internal unpredictability at moment. The same dynamically-trained STMs were also applied to the entire part of a composition before the beginning of a given moment of interest, characterising the *internal unpredictability* of that section.

Internal unpredictability of moment. Additionally, a related measure was used to quantify how predictable the phrase at the moment itself was, given the preceding portion of the composition.[1]

Identification of thematic material

While *internal unpredictability* was based on all events in a composition (or part thereof), the remaining three measures applied only to thematically-relevant material—material relating to any number of *thematic candidates* (*i.e.*, potential themes) identified by the model within the musical piece. The theme detection method implemented in Chapter 3 aimed to identify thematic candidates in a cognitively-plausible manner; this constraint necessitated themes to be detected dynamically as a composition unfolded. Potential themes were identified

[1] As this phrase does not appear in isolation to the listener, but rather with some context of the melody immediately preceding it, this measure was practically calculated as the mean *internal unpredictability* (which is dynamically trained throughout the melody) for the note-events of the moment phrase. This method also accounted for any further learning taking place within the phrase itself.

within a composition through a process of theme detection based on patterns in *internal unpredictability*.

Thematic candidates were identified as the onset of substantially novel pitch interval material, based on sequential comparisons of unpredictability of musical phrases with that of the preceding material within the composition. Two thresholds dictated the magnitude of change in unpredictability needed for a phrase to be identified as a thematic candidate, chosen with the intention of providing robust identification for a wide range of compositions: A phrase was declared a thematic candidate if: (1) its mean unpredictability was greater than the cumulative mean of the material preceding it by more than half a standard deviation; and (2) there was an absence of highly predictable events, such that the event with the lowest unpredictability was greater by one cumulative standard deviation than that of the preceding phrase. Composition beginnings were considered implicitly to be thematic candidates. Given the difficulty of identifying the precise length of a theme, all thematic candidates were taken as two phrases long.

It should be noted that this method identified thematic candidates purely on the basis of their intra-opus novelty and so does not correspond entirely to the notion of theme in music analysis. True themes may be considered to possess additional properties that add to their perceptual salience. Having noted this, for brevity, we refer to the thematic candidates as themes for the remainder of this research.

Thematic repetition. The identified themes were used as training for TTMs, with a separate model created for each theme identified. Each TTM was trained on the theme and then used to predict the remainder of the composition commencing at that theme's onset. The note-by-note information contents thus generated were divided into two clusters using Gaussian Mixture modelling, with the lower (more predictable given the TTM) cluster labelled 'thematic material'. *Thematic repetition* reflects the proportion of events in a musical stimulus that are labelled as thematic material by any of the TTMs.

Thematic repetition at moment. As with the other four primary measures, the proportion of repetition identified in the material preceding a given moment was also characterised.

Thematic variation. The extent to which the thematic material varied from its corresponding theme was quantified when clustering TTM values at the phrase level. The information-theoretic measure of compression distance (Li et al., 2003; Pearce & Müllensiefen, 2017)—the normalised summed information content of the notes making up a phrase, given a TTM trained on the corresponding

theme—was used to give the *dissimilarity* between each phrase and its corresponding theme. These values were averaged for each melody to give a measure of how far thematic material developed from its parent theme; higher *thematic variation* would indicate greater divergence between the thematic material and the parent themes within a composition.

Thematic variation at moment. *Thematic variation at moment* quantified the extent to which thematic material identified before a moment varied from its parent themes.

Stylistic unpredictability. The stylistic content of the thematic material was modelled using an LTM trained on a corpus of western tonal music (as described in Chapter 4). The mean LTM information content for thematic note-events was used to provide a measure of the *stylistic unpredictability* of the thematic material.

Stylistic unpredictability at moment. The *stylistic unpredictability* of the thematic material occurring before a given moment was also calculated.

6.3 The present experiment

The experiment presented in this chapter aimed to test the extent to which these measures could account for listeners' perception of thematic structure over large timescales. By testing these measures, we had three general aims: (1) to examine the extent to which listeners can perceive differences in thematic structure across a range of compositions; (2) to compare different structural properties of music that may facilitate the perception of these structures; and (3) to assess the hypothesis put forward by this thesis, that the perception of thematic structures relies on psychological processes of statistical learning.

This experiment tested the influence of the four model measures on two different indicators of thematic structure—first, the ability of participants to identify musical material at specified moments as being a repetition of material that appeared earlier in the piece (the recognition task), and second, the perception of a composition's structural unity (the unity task). For the recognition task, the four primary measures were applied to the preceding portion of a composition before each recognition–moment and, additionally, the unpredictability of the moment itself within the composition was investigated.

It was hypothesised that each of these measures of thematic structure would possess some ability to influence participants' performance in the tasks. Corroboration of this general hypothesis would provide evidence that large-scale thematic structure can be perceived in a systematic way, and that this perception can

be simulated in terms of psychological mechanisms of statistical learning.

Specifically, two hypotheses were proposed for the effects of experiment measures on participants' abilities to recognise repetition and perceive differences in structural unity: (1) that material with low *internal unpredictability* would be recognised as thematically-related repetition; and (2) that low *internal unpredictability*, high *thematic repetition*, or low *stylistic unpredictability* would each increase the ability of participants to perceive repeated thematic structure and, therefore, increase the perceived sense of unity over an entire composition.

For this experiment, additional measures of participants' musical backgrounds were recorded, hypothesising that increased musical training—through greater exposure to music and increased learning of stylistic conventions—would increase the impact of stylistic measures on perception of thematic structure. We made no specific hypotheses as to the effects of musical background on intra-opus learning.

6.4 Methods

6.4.1 Participants

Forty participants were recruited using Prolific; no exclusion criteria were applied other than a required first language of English and normal or corrected to normal hearing. Participants had a mean age of 33.58 years (SD = 14.03) and 21 were women. Participants were of nine nationalities, with 25 of UK nationality. There were no prerequisites on musical training: 27 participants reported having received some formal musical training on at least one instrument, of which three reported training for more than 10 years.

6.4.2 Stimuli

Stimuli for this experiment consisted of 40 two-minute-long melodies.[2] Melodies were taken from the corpus of complete western tonal monophonic compositions, trimming over-length compositions and disregarding those shorter than the target stimulus length.

The possibility that a composition may be entirely rhythmically isochronous posed several obstacles for selecting stimuli and identifying repetitions of musical material for this experiment. For this reason, pitch interval was the only representation used in the stimulus selection process.

For each of the four model measures, compositions more distant than 3 SD from the mean were removed and 40 melodies were randomly sampled from the remaining pool. To minimise disruption to the perception of large-scale

[2]Details of the compositions used are given in Table B.1 of Appendix B

structure by local effects of the stimuli ending abruptly, endings were manually identified within an additional 15-second window; endings were identified (with descending order of preference) at either a section ending marked in the score, before a substantial gap in the melody, or at a strong phrase ending—*i.e.*, with a perfect cadence. Audio files were generated for the selected stimuli in a piano timbre with a uniform loudness using Apple's GarageBand software with original pitches, rhythms and tempos preserved.[3]

Within the second minute of each stimulus, four phrases were selected for use in the recognition task. These phrases were selected so as to span a wide range of *internal unpredictability* (mean *unpredictability* for the phrase) when compared to the cumulative median *unpredictability* for the stimulus. To ensure this, phrases were selected iteratively, alternating between high—above the cumulative median—and low—below the median—states (with the initial state randomised); at each iteration, the phrase furthest from the median in the state direction was selected. Phrases within two seconds of those already identified were excluded in order to minimise potential disruption caused by moments being presented in quick succession. This process was repeated within the second minute of each stimulus until four phrases were identified for each. These selected phrases are referred to as 'recognition–moments' or simply 'moments'.

The selected stimuli had a mean duration of 127.37 seconds (SD = 5.12). Recognition–moments had a mean duration of 2.85 seconds (SD = 1.74).

6.4.3 Procedure

An online experiment was created using the jsPsych JavaScript library. The experiment comprised a first part containing the main experimental task (taking around 50 minutes for completion) and a second part of the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire for musical sophistication (Müllensiefen et al., 2014).

Four sets of 20 stimuli each were created, distributed randomly, with each stimulus present in exactly two sets. Participants heard one set each—10 hearing each set—with order of presentation randomised individually for each participant.

Individual trials contained a single stimulus melody and involved two tasks, the first a recognition memory task at four moments while listening to the melody, and the second a rating of coherent unity for the melody in its entirety.

For the recognition task, a colour-coded visual indicator was used. A red indicator notified participants that they should just listen to the stimulus, an amber indicator gave notice that a recognition–moment was approaching, and a green indicator identified the passage of music for which the task was to be con-

[3] Modelling code, stimuli and data for this experiment can be found at https://osf.io/e86ys/.

ducted. Participants were asked to state whether they thought they had heard the exact musical passage playing at the indicated moment (while the green indicator was displayed) at any point previously in the stimulus. Responses were to be given as quickly as possible after the indicator had changed back to red using the 'y' (*i.e.*, heard before) and 'n' (not heard before) keys. Four such recognition–moments occurred during the second half of each stimulus, as described above.

After listening to each stimulus, participants were asked to provide a rating of the level of structural unity they perceived in the stimulus. Specifically, they were asked to rate 'the extent to which you think the different parts of the piece unify into a coherent whole' on a continuous scale from 'not very unified' (returning a value of 0) to 'very unified' (returning 100). Participants were also given the following guidance (adapted from the experimental materials used in Tan & Spackman, 2005):

> For a unified piece, even though there may be many different ideas in it, the music still sounds like one well-integrated, whole, single composition. A piece that is not unified, on the other hand, is one that sounds like unconnected fragments of music that do not seem to belong together, so that they do not hold together as one well-integrated, whole, single piece of music.

6.4.4 Statistical analysis

Data collected for this experiment contained three levels of detail: first, responses for each recognition–moment; second, unity ratings for each stimulus; and third, Gold-MSI questionnaire responses for each participant. Each of these levels was analysed in turn for both pitch interval and inter-onset interval representations.

The independent variables for the analyses consisted of model-based characterisations of the stimuli both at the level of the entire stimulus and at the level of specific moments. At the stimulus level, melodies were characterised in terms of the four basic model measures described above: (1) *internal unpredictability*; (2) *thematic repetition*; (3) *thematic variation*; and (4) *stylistic unpredictability*. In addition to the 'high' or 'low' categorisation from the selection process for each moment (and an analogous categorisation made using inter-onset interval[4]), moments were also characterised by five experiment measures. Firstly, the *unpredictability of moment* gave the mean unpredictability for the phrase when trained on the preceding portion of the stimulus. Furthermore, moments

[4]Due to the effect of completely isorhythmic content for a small number of stimuli (the factor that excluded this representation from use in moment selection) there was a small discrepancy between numbers of 'high' and 'low' unpredictability categorisations for inter-onset interval.

were also characterised in terms of the following properties of the melody up to the start of the recognition–moment (corresponding to the four stimulus-level characteristics): the *unpredictability at moment*, the *thematic repetition at moment*, the *thematic variation at moment*, and the *stylistic unpredictability at moment*. Figure 6.1 illustrates how these measures apply to the stimuli. Separate measures for pitch interval and inter-onset interval were calculated for stimulus-level and moment-level measures.

For the recognition task, data for pitch and rhythm models were first analysed using a chi-squared test comparing participant responses to the moments' high/low *internal unpredictability*. Chi-squared tests were then used to test responses separately against binary classifications for the remaining five measures for each representation, splitting each at the median value for that measure across all recognition–moments. In all cases missing responses were dropped. Secondly, recognition–moment data were analysed using mixed-effects logistic regression analyses with participants' responses as the dependent variable. The independent variables were *unpredictability of moment* and the four 'at moment' measures for both pitch interval and inter-onset interval representations.

Analysis of participant unity ratings firstly tested for correlation (Pearson's r) between ratings and the four interval model measures for each representation. Multiple linear regression analyses were then used to examine whether the four measures were significant predictors of unity ratings for each representation.

Participant responses to Gold-MSI questions were aggregated to produce measures of *general sophistication*, *perceptual abilities*, and *musical training* for each participant. To examine the effect of participants' musical backgrounds on their recognition responses, a multiple logistic regression analysis was conducted predicting responses from Gold-MSI scores. The correlation between Gold-MSI scores and both mean unity rating per participant and slopes from simple linear regression models between ratings and measures was also examined.

6.5 Results

The results for this experiment are presented individually for each level of data collected for both pitch interval and inter-onset interval representations—at the levels of (1) recognition–moments, (2) entire stimuli, and (3) participant Gold-MSI scores.

6.5.1 Recognition–moments

Each of the 40 participants heard 20 melodies, each containing four recognition–moments, with 40 melodies across all participants. In total, 1,642

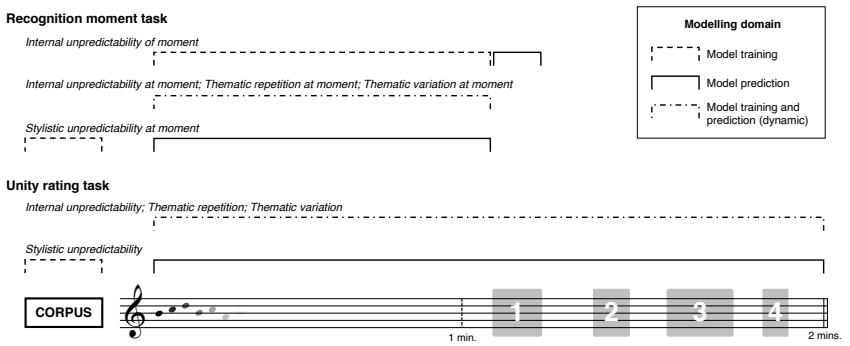


Figure 6.1: Illustrative training and prediction domains for experiment measures. *Recognition moment task:* For a given recognition moment (shown here for the first moment but applied to all four), domains used for model training only are shown with dashed lines and model prediction only with a solid line; domains in which both training and prediction are applied dynamically are shown with dash-dotted lines. *Unity rating task:* As previous task, dash-dotted lines illustrate measures based on training and prediction within a stimulus composition; *stylistic unpredictability* is computed from a model trained on a corpus and used to predict events in the composition.

responses were given for moments being heard before and 1,514 for moments not heard before, with 44 missing responses. Responses were aggregated to give the proportion of 'heard before' to 'not heard before' responses for each stimulus moment (missing responses were not included). Correlations were found for corresponding measures between pitch and rhythmic representations; measures involving intra-opus training showed this positive correlation, only *stylistic unpredictability at moment* did not.

For Chi-squared tests, proportions were rounded to produce a single binary response for each moment. Classification of the moment according to *internal unpredictability* (high or low), relative to that of the melody before the moment, was found to have a significant effect on responses for both pitch interval, $\chi^2(1, N = 80) = 34.25$, $p < .001$, and inter-onset interval, $\chi^2(1, N = 80) = 32.33$, $p < .001$. For binary classifications of the remaining five measures Table 6.1, *unpredictability of moment* was found to be significant for pitch interval, $\chi^2(1, N = 80) = 30.64$, $p < .001$, and inter-onset interval, $\chi^2(1, N = 80) = 18.24$, $p < .001$, and *unpredictability at moment* was found to be significant for pitch interval only, $\chi^2(1, N = 80) = 12.08$, $p < .001$. For all three of these measures, participants were more likely to mark high *unpredictability* moments as not being heard before. An additional significant effect of *thematic repetition at moment*, for inter-onset interval only was found $\chi^2(1, N = 80) = 4.22$, $p = .04$. The remaining measures were not found to be significant for either

Table 6.1: Descriptive Statistics for High and Low Categories of Individual Experiment Measures

Measure	High		Low	
	M	*SD*	*M*	*SD*
Pitch interval				
Unpredictability of moment	4.78	1.03	1.00	0.55
Unpredictability at moment	3.10	0.47	1.70	0.42
Thematic repetition at moment	0.55	0.09	0.30	0.08
Thematic variation at moment	0.91	0.15	0.46	0.09
Stylistic unpredictability at moment	3.55	0.37	2.65	0.26
Inter-onset interval				
Unpredictability of moment	2.23	0.80	0.83	0.36
Unpredictability at moment	1.65	0.25	1.11	0.30
Thematic repetition at moment	0.81	0.10	0.40	0.16
Thematic variation at moment	1.06	0.20	0.59	0.14
Stylistic unpredictability at moment	1.78	0.62	0.62	0.24

pitch or rhythm.

Using mixed-effects logistic regression, recognition–moment responses were analysed using the five pitch interval measures (in continuous form) as predictor variables. The model accounted for the random effects of participants and moments (nested in their respective stimulus), however, due to the overfitting to data when using the maximal random-effects structure, these differences where modelled using random intercepts only. As shown in Table 6.2, predictor *unpredictability of moment* was found to be highly significant. Higher values (increased unpredictability) favoured the response 'not heard before'; an increase in 1SD of *unpredictability of moment* was associated with an odds ratio of 2.78 for 'not heard before'. Participant intercepts had $SD = 0.62$, moment intercepts $SD = 1.02$. For a logistic regression model predicting responses using inter-onset interval measures with random intercepts of participant and moment (see Table 6.2), a similar significant effect of *unpredictability of moment* was found. An increase in 1SD of *unpredictability of moment* was associated with an odds ratio of 3.13 for the response of 'not heard before'. Participant intercepts had $SD = 0.62$, moment intercepts $SD = 1.02$.

To test both representations together, a mixed-effects logistic regression using all measures from both, as well as accounting for the random intercepts of participants and stimulus moments, was used to predict recognition response proportions. As shown in Table 6.3, both pitch interval and inter-onset interval *unpredictability of moment* were significant predictors, with increases of 1 SD in each being associated with odds ratios of 1.85 and 2.04, respectively, for the response 'not heard before'. An AIC score of 3495.00 for this combined model showed a better goodness of fit to the data than those using only pitch measures, 3525.40, or rhythmic measures, 3517.40 (where the rhythmic meas-

Table 6.2: Mixed-Effects Logistic Regression Analyses Predicting Recognition–Moment Responses by Experiment Measures for Each Pitch and Rhythmic Representation, Accounting for Participant and Stimulus Differences

Predictor	β	*SE*	*z*	*p*	
Pitch interval					
Intercept	2.08	0.76	2.72	<.01	••
Unpredictability of moment	−0.49	0.05	−10.44	<.001	•••
Unpredictability at moment	−0.25	0.13	−1.90	.06	
Thematic repetition at moment	−0.02	0.72	−0.23	.98	
Thematic variation at moment	−0.10	0.42	−0.24	.81	
Stylistic unpredictability at moment	0.05	0.20	0.24	.81	
Inter-onset interval					
Intercept	2.25	0.59	3.78	<.001	•••
Unpredictability of moment	−1.21	0.12	−10.44	<.001	•••
Unpredictability at moment	−0.20	0.27	−0.72	.47	
Thematic repetition at moment	−0.68	0.45	−1.51	.13	
Thematic variation at moment	0.59	0.38	1.53	.13	
Stylistic unpredictability at moment	−0.06	0.15	−0.42	.67	

•• $p < .01$; ••• $p < .001$

ure model showed a better fit than the pitch one). Participant intercepts had *SD* = 0.62, moment intercepts *SD* = 0.89.

6.5.2 Unity ratings

All 40 participants returned ratings of unity for all 20 melodies presented to them. Across the 40 total stimuli, unity was rated with a mean of 57.50 (*SD* = 27.88) and the mean scale usage by participants was 82% (*SD* = 15.43). Unity ratings were combined by producing an average for each stimulus. As with the recognition task measures, corresponding measures between representations were positively correlated, excluding *stylistic unpredictability.*

Correlations between mean unity ratings and the four pitch interval model measures showed a highly significant correlation between *internal unpredictability* and unity, $r(38) = -.61$, $p < .001$. No significant correlation was found for the remaining three measures; *thematic repetition*, $r(38) = -.03$, $p = .86$; *thematic variation*, $r(38) = -.11$, $p = .50$; and *stylistic unpredictability*, $r(38) = -.02$, $p = .89$. Correlations between unity ratings and inter-onset interval measures showed the same pattern of results, with a highly significant correlation found with *internal unpredictability*, $r(38) = -.78$, $p < .001$, and no significant correlations found for *thematic repetition*, $r(38) = .11$, p = .48; *thematic variation*, $r(38) = -.11$, $p = .50$; and *stylistic unpredictability*, $r(38) = -.05$, $p = .06$.

Using mixed-effects linear regression, unity ratings were analysed using the four measures as predictor variables, accounting for random intercepts of participant and stimulus. As shown in Table 6.4, for both representations the predictor of *internal unpredictability* accounted for a significant proportion of vari-

Table 6.3: Mixed-Effects Logistic Regression Analyses Predicting Recognition–Moment Responses by Experiment Measures Across Both Pitch and Rhythmic Representations, Accounting for Participant and Stimulus Differences

Predictor	β	*SE*	*z*	*p*	
Intercept	2.46	0.80	3.06	<.01	••
Pitch interval					
Unpredictability of moment	−0.30	0.05	−5.85	<.001	•••
Unpredictability at moment	−0.02	0.15	−0.16	.87	
Thematic repetition at moment	0.34	0.68	0.51	.61	
Thematic variation at moment	−0.21	0.40	−0.52	.60	
Stylistic unpredictability at moment	0.05	0.20	0.26	.80	
Inter-onset interval					
Unpredictability of moment	−0.76	0.13	−5.91	<.001	•••
Unpredictability at moment	−0.27	0.32	−0.83	.41	
Thematic repetition at moment	−0.70	0.47	−1.50	.13	
Thematic variation at moment	0.67	0.39	1.73	.08	
Stylistic unpredictability at moment	−0.13	0.15	−0.88	.38	

•• $p < .01$; ••• $p < .001$

ance, with higher *internal unpredictability* corresponding to lower perceived unity for a melody ($\beta^* = 0.30$ and $\beta^* = 0.38$, respectively). The pitch interval model accounted for 32% of the total variance in the data and the inter-onset interval model accounted for 31% of the total variance in the data. Participant intercepts varied for pitch interval $SD = 9.58$ and inter-onset interval $SD = 9.71$. Stimulus intercepts varied for pitch interval $SD = 9.56$ and inter-onset interval $SD = 6.62$.

As the sole predictor of significance in the regression models for both representations, the relative ability of *internal unpredictability* to account for variance in participant ratings can be compared between representations (still accounting for stimulus and participant random effects). Pitch interval *internal unpredictability* accounted for 31% of variance in mean ratings ($\beta = -12.08$, $df = 38.28$, $t = -4.64$, $p < .001$) and inter-onset interval *internal unpredictability* accounted for 31% of variance, ($\beta = -22.73$, $df = 37.74$, $t = -7.88$, $p < .001$) as shown in Figure 6.2. However, the strong correlation between the measures of internal unpredictability for pitch and rhythm, $r(38) = .72$, $p < .001$, makes it difficult to ascertain which has the stronger effect.

6.5.3 Gold-MSI scores

After averaging of Gold-MSI responses into scores for each participant, participants had a mean score for *perceptual abilities* (out of a possible scale range of 9–63) of 40.32 ($SD = 6.16$), a mean score for *musical training* (scale range 7–49) of 16.53 ($SD = 8.66$), and a mean score for *general sophistication* (scale range 18–126) of 56.60 ($SD = 17.48$).

Multiple logistic regression was used to test for the effects of musical soph-

Table 6.4: Mixed-Effects Linear Regression Analyses Predicting Stimulus Unity Ratings by Experiment Measures for Each Pitch and Rhythmic Representation, Accounting for Participant and Stimulus Differences

Predictor	β	*SE*	*df*	*t*	*p*	
		Pitch interval				
Intercept	88.87	12.41	36.15	7.16	<.001	•••
Internal unpredictability	−12.82	2.77	35.28	−4.63	<.001	•••
Thematic repetition	−3.87	11.72	35.01	−0.33	.74	
Thematic variation	−4.30	8.01	35.14	−0.54	.60	
Stylistic unpredictability	3.30	3.28	35.13	1.00	.32	
		Inter-onset interval				
Intercept	99.73	7.53	37.83	13.24	<.001	•••
Internal unpredictability	−23.40	3.49	35.13	−6.70	<.001	•••
Thematic repetition	−7.94	6.43	35.25	−1.24	.26	
Thematic variation	2.34	5.78	34.65	0.41	.69	
Stylistic unpredictability	−0.89	2.40	37.00	−0.37	.71	

••• $p < .001$

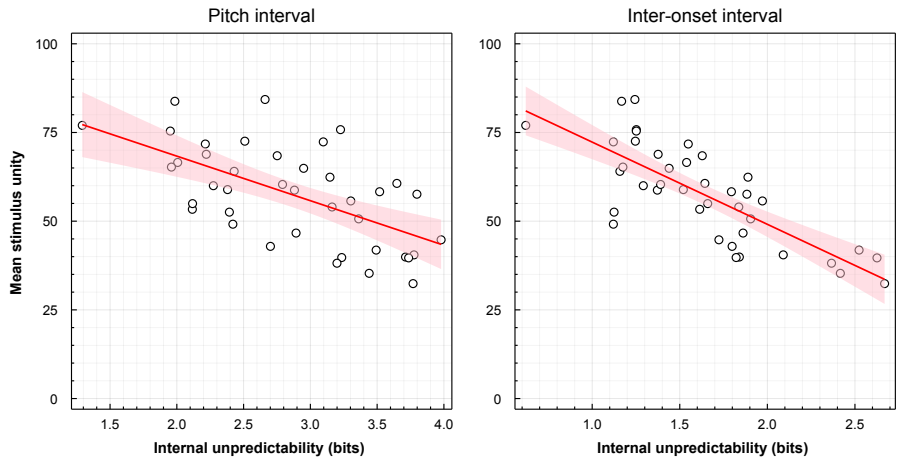


Figure 6.2: Fitted linear relationships of *internal unpredictability* predicting mean stimulus ratings using pitch interval and inter-onset interval representations. Note: Shaded regions show 95% confidence intervals.

Table 6.5: Correlation with Gold-MSI Scores for Participant Mean Ratings and Slopes Predicting Participants' Unity Ratings by Model Measures for Each Pitch and Rhythmic Representation

Variable	*M*	*SD*	*General sophistication*	*Perceptual abilities*	*Musical training*
Mean ratings	57.50	11.30	.07	−.14	.03
			Pitch interval		
Slopes[a]					
Internal unpredictability	−11.85	9.36	−.09	−.37 •	−.07
Thematic repetition	−5.87	33.04	−.17	−.10	−.12
Thematic variation	−10.65	22.86	−.41 ••	−.08	−.20
Stylistic unpredictability	−1.37	9.78	−.05	−.10	−.01
			Inter-onset interval		
Slopes[a]					
Internal unpredictability	−22.67	13.76	−.14	−.34 •	−.11
Thematic repetition	5.27	22.77	−.23	−.07	−.13
Thematic variation	−6.49	20.06	−.34 •	−.19	−.32 •
Stylistic unpredictability	−7.18	8.50	.06	−.26	.01

[a]Slopes for linear regressions predicting each participant's unity ratings from a model measure

• $p < .05$; •• $p < .01$

istication scores on participants' recognition–moment responses. Participants' *perceptual abilities* were found to have a weakly significant effect, $b = -0.02$, $z(3155) = -2.14$, $p = .03$. The other scores were not significant predictors: *general sophistication*, $b = 0.004$, $z(3155) = 1.27$, $p = .20$; and *musical training*, $b = -0.002$, $z(3155) = -0.39$, $p = .70$.

Correlations between Gold-MSI scores and mean unity ratings for each participant were tested; no significant correlations were found. Slope coefficients from linear models predicting each participant's unity ratings from model measures values for stimuli were also tested against Gold-MSI scores (shown in Table 6.5). For both representations, significant correlations were found between *general sophistication* and individual slopes for measures of *thematic variation*, and between *perceptual abilities* and slopes for measures of *internal unpredictability*. Additionally, participant slopes for inter-onset interval *thematic variation* was found to have a significant correlation with *musical training*.

6.6 Discussion

The experiment presented in this chapter aimed to test listeners' abilities to perceive two highly-important indicators of large-scale thematic structure in music—their ability to identify repetitions of thematic material and their ability to perceive structural unity. In testing these abilities, this experiment sought to examine the extent to which statistically-learned features could account for

listeners' perception of these two properties and, by extension, the extent to which statistical learning can allow large-scale thematic structures to be perceived.

While the first behavioural experiment, presented in Chapter 5, sought to provide some initial evaluation of the theoretical modelling of thematic structures based on the statistical learning of structural features on a small scale, the experiment of the current chapter extended the scope to cover longer musical passages (of around two minutes duration). With the full musical contexts available in these longer stimuli, Chapter 3's model measures could be applied in the same form as they were hypothesised to operate in real-world musical listening—quantifying the *internal unpredictability*, *thematic repetition*, *thematic variation* and *stylistic unpredictability* of whole compositions, and of substantial portions of compositions before given moments. The analysis of this experiment focused on how knowledge acquired dynamically via statistical learning throughout the entire piece could influence the perception of both repetition and unity.

The findings of this experiment presented a striking contrast between the influence of the four primary modelled features. Across both recognition–moment and unity tasks, there was strong evidence of an influence of *internal unpredictability* (and its associated measures) across both pitch and rhythm representations.

Specifically, the results from the recognition–moment task provided convincing evidence that intra-opus *internal unpredictability* had an effect on the perception of thematic repetition within compositions. The results indicated that this effect was realised in two ways, however, with some differences between pitch and rhythmic representations. First, if a passage of music was predictable, given the music that had occurred before it, it was likely to be perceived as a repetition for both pitch and rhythmic variants of the measure of *internal unpredictability of moment* (how predictable a moment was, given the portion of the composition before it). Second, and to a lesser extent, if the music occurring before a passage in question was itself predictable (*i.e.*, low *internal unpredictability at moment*), the passage was more likely to be perceived as a repetition, with the chief evidence for this measure found when using the pitch representation. These findings supported our hypotheses regarding the relationship between *internal unpredictability* and repetition recognition.

The difference between pitch and rhythmic representations for the results of the repetition recognition task may be due, in part, to the presence of rhythmically isochronous stimuli; such stimuli would offer maximum repetition with minimum variation of inter-onset interval, increasing the likelihood that a passage would be perceived as having been heard earlier in the piece.

The results for ratings of structural unity revealed a similar pattern of effects;

there was a strong effect of *internal unpredictability* on the perception of structural unity for both pitch interval and inter-onset interval representations. This result corroborated our hypothesis that compositions that were internally predictable due to repetition of material would be perceived as having a stronger sense of unity.

For this experiment, however, the extent to which measures based on the isolation of thematic material influenced participants' perception of thematic structure was less apparent, with an absence of any significant effect of these measures found. We suggest two possible causes for the differences between *internal unpredictability* and the two thematic measures. Firstly, we should consider potential differences between the way in which themes (or rather, thematic candidates) are detected in the model, and how they may be perceived in real-world listening. The model identifies potential themes as substantially-novel material, working incrementally through a composition from beginning to end. While this is consistent with traditional understandings of 'theme', it most likely represents something of a simplification of perceptual theme identification. The extent to which this may the case is hard to discern, given the relatively little empirical research on theme perception. Any such discrepancies between the theme-detection model and perception may add noise within the measures relying on theme detection—*thematic repetition* and *thematic variation*—whereas *internal unpredictability* is computationally simpler and does not rely on theme detection.

However, a second explanation presents itself. In addition to being computationally simpler, *internal unpredictability* is also applicable to the entirety of a composition's material, whereas *thematic repetition* and *thematic variation* apply only to the material identified as being thematic. From this, we can infer that material identified by the model as thematic is not the sole contributor to perception of thematic structure. Instead, the results suggest that *any* material that becomes predictable through repetition, no matter how insignificant, can contribute to a constantly accumulating perception of thematic structure. This provides important indicators as to how repetition contributes to perception of large-scale structure. Relatively precise repetition of important material, the focus of most empirical investigations of repetition perception to date, may be secondary to the effects of any repetition at all, no matter how small in scale or approximate. In this light, we can trace the significant effect of *dissimilarity* in the experiment of the previous chapter to the significant effect of *unpredictability of moment* in the recognition–moment task of the current experiment, rather than to *thematic variation*.

The extent to which listeners' musical backgrounds influence how they perceive thematic structure is still somewhat uncertain. In both tasks, the majority of Gold-MSI score comparisons to participant responses did not yield any sig-

nificant relationships, with three exceptions. First, there was a significant effect of *perceptual abilities* as a predictor of recognition–moment identification. This appears to indicate that listeners with lower *perceptual abilities* were more likely overall to consider the recognition–moments as repetitions, perhaps reflecting less accurate encoding and therefore less finely discriminated recognition of familiar material. Second, participants with greater overall *general sophistication* (and, additionally, *musical training* in the case of rhythm) showed a greater sensitivity to *thematic variation*, with greater sophistication associated with a stronger negative association between *thematic variation* and unity. Third, participants with higher scores of *perceptual ability* had greater sensitivity to the *internal unpredictability* of stimuli, such that greater *perceptual ability* was associated with a stronger negative effect of *internal unpredictability* on unity.

6.7 Summary

This chapter reported the second behavioural experiment of the thesis, testing the perception of two highly-important indicators of large-scale thematic structure. The experiment tested participants' abilities to identify given moments as being repetitions of earlier thematic material and their judgements of the sense of structural unity of compositions. As this experiment tested for these features in long monophonic compositions (with durations of approximately 2 minutes), the measures of thematic structure of Chapter 3 were evaluated in the same form as they were hypothesised to operate in real-world musical listening.

Overall, the findings of this experiment provide evidence that large-scale thematic structure can be readily perceived by listeners, reflected both by an ability to identify the internal elements of structure over a relatively long timespan, and by an ability to distinguish differences in inherent structural unity of different compositions. Furthermore, it was possible to predict differences in perception of large-scale thematic structure between stimuli as a function of experimental measures. Importantly, the significant influence of *internal unpredictability* in accounting for these effects supports our hypothesis that statistical learning is a plausible psychological mechanism allowing large-scale thematic structure to be perceived. These findings provide some important corroboration of overarching hypothesis of this thesis—that thematic structures can be perceptible through the structural regularities they form.

Chapter 7

Modelling Large-Scale Continuation

7.1 Overview

In this chapter, an experiment is presented with the objective of testing the extent to which statistical learning could account for listeners' perception of large-scale continuation. In this experiment, participants heard substantial context portions of compositions, followed by three passages of possible continuations; participants were asked to rank the continuations based on the extent to which they believed each best continued the given context. For stimulus compositions, plausible continuations were computationally generated, based on predictability arising from training based on either a context composition, its themes, or from the corpus of Chapter 4, using both pitch interval and inter-onset interval representations. Measures based on those of Chapter 3 were compared to participants' responses in this task.

This task was chosen for its ability to act as an important indicator of the effects of large-scale thematic structure. In a manner similar to the well-established probe-tone paradigm's purpose of inspecting listeners' perception of pitch, given their exposure to a particular musical context, the continuation task allows us to examine the features learned by listeners in a context on the timescales needed to investigate the effects of large-scale structure.

In addition to contributing to the cumulative evidence of the relevance of intra-opus statistical learning to the perception of thematic structure through the experimental testing of another indicator, this experiment expands on the scope of the preceding behavioural experiments presented in this thesis in two ways. First, it increased the duration of stimulus compositions used, testing whether the general patterns of effect found in Chapter 6 are still present over yet larger timescales, using compositions both of 2 and 4 minutes long; and it provides the initial introduction and testing of techniques for the modelling

of thematic structure in polyphonic compositions of two voices (the theory of which was discussed in Chapter 3).

7.2 Modelling

The thesis hypothesis that the perception of large-scale thematic structure is facilitated by the learning of statistical regularities within compositions is subjected to continued testing in the experiment reported in this chapter. As with all the behavioural experiments of this thesis, the computational measures of thematic structure presented in Chapter 3 were adapted to produce measures specific to the current task.

As with the measures used in the previous two experiments, measures for this experiment use different configurations of IDyOM variable-order Markov models, using different training domains. IDyOM, given a training sequence, estimates likelihoods of the occurrence of note-events in a sequence, smoothing between models of different orders (Pearce, 2005). As detailed in Chapter 3, the likelihood estimates were converted into a value of information content, or unpredictability. For a given composition, the training sequence may be taken from within the composition itself (STM; using dynamic training of the model)—modelling its *internal unpredictability*—trained only on themes extracted from within the composition (TTM), or trained on a large corpus of compositions (LTM)—modelling its *stylistic unpredictability*.

Unlike the measures used in the empirical research presented in this thesis so far, this experiment expands the use of modelling to cover polyphonic compositions (limited to perceptually segregated two-voices; the assumptions and discussion of this technique is given in Chapter 3). Used here in its simplest form—only ever predicting note-events in a monophonic continuation (even if training material is polyphonic), only non-dynamically trained models are needed. For example, to model the note-events of a monophonic continuation, given a polyphonic context, the voices of the context can be used to train the model separately.

7.2.1 Measures

Five measures were constructed, each presenting a competing hypothesis as to the ways in which continuations may be ranked. Each measure characterised continuations based on the statistical learning of different categories of musical material. For each, an IDyOM model with a different training domain was used to estimate the unpredictability of note-events in a given continuation, averaging to produce a value in that measure for each continuation. The measures all shared the same hypothesised directionality—from low to high—that lower-

scoring continuations would be considered more desirable, according to that particular measure.

All of the measures used in this experiment were dependent on a specific representation of the musical surface, such that they were each used twice; with pitch interval and inter-onset interval representations to separately examine effects of pitch and temporal domains.

Measures were used to characterise properties of a musical passage (a continuation) that could follow from a large excerpt of a composition (a context).

Composition unpredictability. To quantify how predictable a continuation may be within its composition (*i.e.*, as an extension to the context preceding it), for each composition, models were trained on the context section and used to predict the note-events of the possible continuations. According to this measure, continuations that received lower values were more closely related to the material of the context than their counterparts. For the present experiment, this measure most closely embodies the effects of intra-opus *internal unpredictability* for the paradigm used.

Late-composition unpredictability. Likewise, a second measure also considered predictability relating to the material from the context section of the composition. In this measure, however, only the second half of the context (by number of note-events) was used in the training of models. When compared to the first, this measure hypothesised some memory constraint on learning during the context section, such that only material present later in a composition informed the comparative ranking of continuations.

Thematic unpredictability. As discussed in Chapters 3, 4 and 6, we also considered that all material might not be equal in its ability to influence the perception of structure. Using the methods for detecting possible themes within compositions, described in Chapter 3, themes were identified for each composition. These themes were then used to train a single model for that composition and estimate the unpredictability of the possible continuations.

Continuation unpredictability. The first three measures introduced above each characterises a way in which a continuation can be judged, based on the preceding context; however, any possible influence of the internal unpredictability of the continuations themselves should also be considered. For this measure, an IDyOM STM was used to dynamically train and predict note-events for each continuation individually.

Stylistic unpredictability. To account for any influence of stylistic conventions on the perceived fit of continuations, a measure was used to encapsulate rankings based on the relative *stylistic unpredictability* of continuations. For each continuation, information content was calculated using a model trained on a corpus of 600 monophonic melodies (described in Chapter 4).

7.3 The present experiment

The experiment presented in this chapter was designed to test the extent to which statistical learning within musical compositions could account for listeners' choice of continuations. Judgements based on the statistical features embodied in the described measures would indicate the importance of statistical learning in the perception of large-scale thematic structure. Furthermore, as each measure was implemented using two variants—using pitch interval and inter-onset interval representations of the musical surface—the comparative importance of the pitch and rhythmic domains in the perception of large-scale structure was explored.

In addition to the testing of another aspect of large-scale structure, the current experiment expanded on the scope of the experiment of Chapter 6 in two further areas. By expanding the durations of stimuli used, this experiment aimed to test for any potential differences occurring when the effects of large-scale structure found in the previous experiment were scaled up to structures of twice the duration. The current experiment also aimed to test the expansion of modelling techniques to include probabilistic modelling of polyphonic stimuli. Additionally, this experiment aimed to provide some replication of the relationship between *internal unpredictability* and ratings of unity found in the previous chapter—testing for relationships between the current measures and participants' perceived sense of structural unity.

While each of these measures presented their own competing internal hypotheses—that participants would rank continuations for a single composition from low to high values of information content—we hypothesised that measures of predictability based on features learned during the context section would have the greatest effects on participants' rankings. Of these three measures, *composition unpredictability* was hypothesised to be the most important because the related measure of *internal unpredictability* was found be significant for the indicators of thematic structure used in the experiment of Chapter 6. Likewise, based on the findings of the previous experiments, we hypothesised that measures using the pitch representation would be able to account for a greater portion of the variance in participants' responses than those of rhythm.

Two stimulus category conditions were introduced in this experiment: (1) a length condition of 'short' compositions (2-minutes long, the same as used

in the previous experiment) and 'long' compositions of twice the length; and (2) a texture condition of 'monophonic' and 'polyphonic' compositions. For both of these conditions, we hypothesised that no differences would be found within them. In particular, in the length category, this hypothesis states that participants would have the ability to perceive the effects of thematic structures regardless of the composition duration.

For the unity ratings given by participants for compositions, we followed the hypothesis given in Chapter 6, that the measures of *internal unpredictability* would most closely predict ratings.

For this experiment, we also recorded measures of participants' musical backgrounds. We hypothesised that participants with increased musical training would show an increase in the impact of *stylistic unpredictability* on rankings of continuations, facilitated by their greater exposure to music and so increased implicit learning of stylistic conventions.[1]

7.4 Methods

7.4.1 Participants

Eighty participants were recruited to participate using the Prolific online recruiting platform. Participants had a mean age of 41.00 (SD = 12.51), 43 were men and 37 were women. Participants were required to have English as their first language and to have normal, or corrected to normal, hearing. All participants were residents of the UK or the USA. No recruitment criteria were enforced on participants' level of musical training; 48 participants reported having some musical training on at least one instrument, with 12 of those reporting more than 10 years of training.

7.4.2 Stimuli

Stimuli for this experiment consisted of 40 context passages of music, each with three possible continuations. Stimuli were adapted from 40 Western-classical compositions, belonging to either a 'monophonic' or 'polyphonic' (of two voices) texture category, and to either a 'short'—of approximately 2 minutes—or 'long'—approximately 4 minutes—length category.[2] Compositions spanned three centuries (1703–1934), featuring works by 20 composers. The 20 monophonic compositions were selected from the set used in the experiment of Chapter 6, which, in turn, were sampled from a large corpus of western-classical

[1] Although we have hypothesised for an effect of musical background, based on collected Gold-MSI scores, it should be noted that participants were not selected based on their musical experiences; this was not the primary goal of this research. The scores of musical training have an additional value in precisely characterising the samples for purposes of replication.

[2] Details of the compositions used are given in Table B.2 of Appendix B

tonal melodies (described in Chapter 4). The 20 polyphonic melodies consisted of western-classical instrumental duets. Compositions that were longer than their respective length category were shortened; endings that produced the minimal structural disruption were manually identified within an additional 15-second window: (1) at either a section ending marked in the score (2) before a substantial gap in the melody, or (3) at a strong phrase ending.

Short closing portions (approximately 10 seconds long) of the trimmed compositions were identified, at strong phrase endings, as the basis for the generation of possible continuations. Continuations were monophonic only, with the voice containing the predominant melodic material during the continuation section selected for those compositions of the 'polyphonic' category.

Probabilistic generation methods were used to create a range of plausible continuations that favoured the effects of statistical learning from a range of different materials. Using the procedure of Pearce (2005) (Chapter 9) for producing novel melodies from IDyOM Markov models, edits were made to the existing material of the continuation sections by sampling from the distributions of trained models using the Metropolis–Hastings algorithm (Robert & Casella, 2004). Through this process, a continuation could be made more predictable, given the material used to train the model.

To generate three different continuations for each composition, for each new continuation two generation passes were made to modify the original according to a different combination of training materials. For each pass, the training domains of models were sampled randomly (excluding duplicates between continuations for that composition) from either: (1) the context section of the composition; (2) the themes identified within the context (using the theme detection method described in Chapter 3, $M = 2.05$ themes per context); (3) a large corpus of western-classical tonal melodies (described in Chapter 4, excluding compositions used as stimuli); or (4) none—making no changes. The model for each pass used either the pitch interval or inter-onset interval representation (randomly selected), modifying the pitch or duration of a continuation's notes, respectively. No edits were made within the final phrase of each continuation, allowing all continuations for a single context to end consistently, removing any influence of a continuation's sense of final closure on participants' rankings. A mean of 24% of events were changed from the originals, per continuation.

The resulting stimuli[3] contained contexts with a mean duration of 111.64 seconds for the 'short' category ($SD = 4.51$; of which, 'monophonic' $M = 112.93$, $SD = 4.63$, and 'polyphonic' $M = 110.36$, $SD = 10.70$), and 227.50 seconds for those in the 'long' category ($SD = 8.35$; of which, 'monophonic' $M = 224.08$, $SD = 4.42$, and 'polyphonic' $M = 230.91$, $SD = 10.09$). 'Short' continuations had

[3]Modelling code, stimuli and data for this experiment can be found at https://osf.io/kru4x/.

a mean duration of 10.54 seconds ($SD = 5.74$; of which, 'monophonic' $M = 12.45$, $SD = 5.79$, and 'polyphonic' $M = 8.63$, $SD = 4.13$), and 'long' continuations had a mean duration of 12.16 seconds ($SD = 4.67$; of which, 'monophonic' $M = 13.82$, $SD = 5.15$, and 'polyphonic' $M = 10.50$, $SD = 3.47$).

Audio files for contexts and continuations were generated using MuseScore notation software at a uniform loudness, preserving original pitches, rhythms, and tempos. 'Monophonic' stimuli were rendered using a piano timbre; to aid the perceived separation between voices, 'polyphonic' stimuli were rendered with a piano timbre for the first voice and a marimba timbre for the second.

7.4.3 Procedure

The task design was implemented using the jsPsych JavaScript library (de Leeuw, 2015) for online use in a web-browser. The experiment consisted of two parts: first, the primary experimental continuation task (taking approximately 20 minutes); and second, the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire for musical sophistication (Müllensiefen et al., 2014).

Each participant received four trials of the continuation task, with stimuli covering the four different stimulus category combinations, and presented in a randomised order. For each trial, participants were first played the stimulus context. After the context, three possible continuations were presented in a randomised order, with a 1 second gap before each; participants heard each continuation only once. The context and continuation being played was indicated to participants onscreen, with continuations labelled from 'A' to 'C'. Participants were instructed to rank the continuations based on the extent to which they would best continue the given context.

After each continuation task, participants were asked to rate the composition—the context and their top-ranked continuation—on its level of structural unity (from 0 to 100), defined as the extent to which the music still sounded like one well-integrated, whole, single composition (giving participants the same extended definition of unity as used in the experiment of Chapter 6).

7.4.4 Statistical analysis

Data collected during this experiment belonged to three categories: (1) continuation rankings, for each trial conducted by each participant; (2) unity ratings for each composition heard by each participant; and (3) Gold-MSI scores for each participant, aggregated from their questionnaire responses. The statistical analyses for this experiment examined each of these in turn.

Within the data collected for the continuation task, differences between continuations at participants' chosen rankings were examined. First, differences in the composition texture and length categories were tested. Mann–Whitney U

tests were used to test for differences in levels of inter-participant agreement between categories. For each measure, independent t-tests were used to test for significant differences between categories using the values of the measure for participants' first-choice continuations. Second, to determine the effects of individual measures on participant rankings, individual one-way repeated-measures (within participant and stimulus) ANOVAs were used to compare measure values of continuations between ranks. Third, the closeness of continuations to their unmodified version was tested—a Friedman test compared the number of edits used to create continuations positioned at each rank. Fourth, correlations between measure values for continuations at the first rank were tested. Finally, an ordinal logistic regression (Christensen, 2022) was used to predict the rankings of continuations using combined measures from both representations.

Analysis of unity ratings tested for correlations (Pearson's r) between ratings and the non-centred measures of participants' chosen continuations for each representation.

Participant responses to Gold-MSI questions were aggregated to produce measures of *general sophistication*, *perceptual abilities*, and *musical training* for each participant. To examine the effect of participants' musical backgrounds on their orderings, correlations between participants' Gold-MSI scores and mean measure values were tested. A multiple linear regression analysis was used to examine the ability of Gold-MSI scores to predict participant mean unity ratings.

7.5 Results

The results for this experiment are presented in turn for each of the three types of data collected—first, analysis of continuation ranks, second, analysis of the unity ratings given for each composition, and third, participants' Gold-MSI scores.

7.5.1 Continuation rankings

Each of the 80 participants completed the continuation ranking task, providing a ranking of three possible continuations each for four different composition contexts. Each participant gave responses for stimuli covering the four different stimulus length and texture category combinations; eight ranking responses were recorded for each composition, for a total of 320 rankings.

Each continuation was modelled using the five measures described at the beginning of this chapter. As the primary interest of these analyses was to test for differences between continuations with different properties, values for each

continuation were centred to the mean for the three continuations belonging to the same parent composition. During modelling, alphabet sizes were kept constant between all compositions within the same representation, allowing measure values to be directly comparable between compositions.

Comparison of stimulus categories

A score of agreement between participants for each stimulus composition was calculated as the proportion of participants who selected the most popular continuation, in a given rank. Agreement ranged from a maximum of 1 to a minimum possible score of 0.375 for the eight participants ranking continuations for each composition. At the first rank, participants had a mean agreement of 0.58 ($SD = 0.17$), at the second rank 0.52 ($SD = 0.13$), and at the third rank 0.52 ($SD = 0.11$).

Differences in the agreement scores between compositions belonging to different texture and length categories were tested using Mann–Whitney U tests. No significant difference in participant agreement was found between 'monophonic' ($M = 0.61$) and 'polyphonic' ($M = 0.56$) categories ($U = 227$, $z = 0.74$, $p = .47$). No significant difference was found between 'short' ($M = 0.60$) and 'long' ($M = 0.57$) categories ($U = 186$, $z = 0.39$, $p = .71$).

For each measure, for each representation, independent t-tests were used to test for differences in composition-centred measure values for participants' first-choice continuations (averaged per stimulus) between stimuli in texture and length categories. As shown in Table 7.1, no measure was found to show a significant difference between 'monophonic' and 'polyphonic' categories. For the length category, shown in Table 7.2, a significant difference between 'short' and 'long' stimuli was found for the pitch interval measure of *continuation unpredictability*, with participants' chosen continuations more internally predictable for 'short' category compositions than 'long'. The subsequent analyses made use of the combined data across categories.

Comparison of participant's rankings

Individual one-way repeated-measures (within participant and stimulus) ANOVAs were used to test for differences in measure values between participants' chosen continuations at each rank. Table 7.3 displays the results of these tests for each measure, for both pitch interval and inter-onset interval representations. The closely related measures of *composition unpredictability* and *late-composition unpredictability* were found to present highly significant differences between ranks, for both pitch interval and inter-onset interval representations. Paired-samples t-tests were used to make post hoc comparisons between ranks against a Bonferroni-adjusted alpha value of .017. For pitch-

Table 7.1: T-Test Comparisons Between 'Monophonic' and 'Polyphonic' Stimulus Categories for Each Measure, Using Participants' First-Choice Continuations

	Monophonic		Polyphonic				
Measure	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*
	Pitch interval						
Composition unpredictability	−0.21	0.48	−0.06	0.31	32.53	−1.21	.24
Late-composition unpredictability	−0.20	0.43	−0.05	0.29	33.14	−1.32	.19
Thematic unpredictability	−0.13	0.34	−0.10	0.37	37.64	−0.27	.79
Continuation unpredictability	−0.05	0.14	−0.01	0.20	33.49	−0.78	.44
Stylistic unpredictability	−0.03	0.20	0.04	0.16	36.96	−1.22	.23
	Inter-onset interval						
Composition unpredictability	−0.19	0.46	−0.15	0.44	37.94	−0.33	.74
Late-composition unpredictability	−0.25	0.61	−0.14	0.45	34.66	−0.71	.49
Thematic unpredictability	0.01	1.15	−0.12	0.40	23.55	0.48	.64
Continuation unpredictability	0.02	0.15	0.01	0.27	30.23	0.21	.84
Stylistic unpredictability	0.02	0.11	−0.01	0.14	36.30	0.69	.49

Table 7.2: T-Test Comparisons Between 'Short' and 'Long' Stimulus Categories for Each Measure, Using Participants' First-Choice Continuations

	Short		Long					
Measure	*M*	*SD*	*M*	*SD*	*df*	*t*	*p*	
	Pitch interval							
Composition unpredictability	−0.15	0.34	−0.11	0.47	34.45	−0.27	.79	
Late-composition unpredictability	−0.15	0.32	−0.11	0.42	35.56	−0.31	.76	
Thematic unpredictability	−0.20	0.43	−0.02	0.23	29.35	−1.70	.10	
Continuation unpredictability	−0.09	0.19	0.03	0.13	33.94	−2.34	.03	•
Stylistic unpredictability	−0.04	0.20	0.05	0.16	35.87	−1.49	.15	
	Inter-onset interval							
Composition unpredictability	−0.12	0.45	−0.21	0.44	37.97	0.63	.53	
Late-composition unpredictability	−0.15	0.60	−0.24	0.47	36.09	0.52	.60	
Thematic unpredictability	0.11	1.12	−0.21	0.43	24.46	1.18	.25	
Continuation unpredictability	0.05	0.29	−0.02	0.09	22.49	0.92	.36	
Stylistic unpredictability	0.02	0.16	−0.00	0.08	27.95	0.50	.62	

• $p < .05$

based variants, significant differences were present between the first rank and the lower two (for *composition unpredictability*, $t(627.42) = -3.72$, $p < .001$ and $t(624.24) = -3.12$, $p = .002$, respectively, and for *late-composition unpredictability*, $t(627.72) = -3.78$, $p < .001$ and $t(626.10) = -3.54$, $p < .001$, respectively), with continuations in the first rank being more predictable according to these measures. For the rhythmic versions of these measures, significant differences were present between the first rank and the last, and between the last two ranks (for *composition unpredictability*, $t(579.72) = -5.17$, $p < .001$ and $t(600.18) = -3.79$, $p < .001$, respectively, and for *late-composition unpredictability*, $t(584.59) = -5.02$, $p < .001$ and $t(606.55) = -3.70$, $p < .001$, respectively), with more predictable continuations placed in higher ranks. A similar significant difference was found between ranks for the pitch interval measure of *thematic unpredictability* (between ranks 1 and 2, $t(627.45) = -2.47$, $p = .014$, between ranks 1 and 3, $t(627.45) = -2.47$, $p = .014$; a similar, non-significant effect was found for the equivalent inter-onset interval measure).

In a similar manner, the closeness of continuations to the originals from which they were modified was tested between ranks. The number of edits used to generate stimulus continuations, for a given representation, was compared across ranks using non-parametric Friedman tests. A significant effect of closeness to original continuations was found for the pitch interval representation ($\chi^2(2) = 13.74$, $p < .001$), with continuations with fewer edits significantly more likely to be placed in the first rank than the lower two. No corresponding effect was found when using inter-onset interval ($\chi^2(2) = 2.06$, $p = .36$).

Relative importance of measures

The extent to which measures were correlated with each other was examined using the participants' first-choice continuations, averaged across participants to provide a single value for each stimulus, for each measure. Table 7.4 shows the Pearson's *r* correlations between all measure combinations, for both representations. Measures that involved training material from all or part of the context section (*composition unpredictability*, *late-composition unpredictability* and *thematic unpredictability*) had a high positive correlation within each representation, but not for corresponding measures between representations. Additionally, a significant correlation was found between *continuation unpredictability* and *stylistic unpredictability*.

Using mixed-effects ordinal logistic regression, measures (centred to composition means) for both representations were used to predict participants' rankings, accounting for differences between participants and stimuli continuations. Due to the close relationship between the measures of *composition unpredictability* and *late-composition unpredictability* (as shown in the correlation

Table 7.3: Repeated-Measures ANOVAs Testing for Differences Between Continuation Ranks in Individual Measures

Measure	Rank 1		Rank 2		Rank 3							
	M	*SD*	*M*	*SD*	*M*	*SD*	*df*	*SS*	*F*	η^2	*p*	
Pitch interval												
Composition unpredictability	−0.13	0.68	0.08	0.77	0.05	0.79	2	8.46	5.07	.02	.007	••
Late-composition unpredictability	−0.13	0.62	0.07	0.70	0.06	0.71	2	7.88	5.72	.02	.003	••
Thematic unpredictability	−0.11	0.70	0.03	0.80	0.08	0.81	2	6.33	3.56	.01	.03	•
Continuation unpredictability	−0.03	0.32	−0.01	0.33	0.04	0.34	2	0.85	2.61	.01	.07	
Stylistic unpredictability	0.00	0.34	−0.01	0.37	0.01	0.40	2	0.08	0.19	.00	.83	
Inter-onset interval												
Composition unpredictability	−0.17	0.82	−0.07	0.88	0.24	1.14	2	28.78	10.40	.03	<.001	•••
Late-composition unpredictability	−0.20	0.99	−0.08	1.08	0.28	1.36	2	38.81	9.74	.03	<.001	•••
Thematic unpredictability	−0.05	1.24	−0.11	1.11	0.16	1.31	2	13.29	2.96	.00	.05	
Continuation unpredictability	0.01	0.36	0.00	0.32	−0.02	0.36	2	0.14	0.39	.01	.68	
Stylistic unpredictability	0.01	0.27	0.00	0.25	−0.01	0.27	2	0.07	0.34	.00	.71	

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

Table 7.4: Pearson's *r* Correlations Between Experiment Measures for Participants' First-Choice Continuations

	Measure	1	2	3	4	5	6	7	8	9	10
					Pitch interval						
1	*Composition unpredictability*	–									
2	*Late-composition unpredictability*	.99 •••	–								
3	*Thematic unpredictability*	.42 ••	.44 ••	–							
4	*Continuation unpredictability*	−.05	−.04	.03	–						
5	*Stylistic unpredictability*	−.09	−.09	.05	.69 •••	–					
					Inter-onset interval						
6	*Composition unpredictability*	−.13	−.09	.07	.14	.07	–				
7	*Late-composition unpredictability*	−.08	−.04	.13	.19	.20	.95 •••	–			
8	*Thematic unpredictability*	−.18	−.17	−.15	−.05	−.36 •	.57 •••	.54 •••	–		
9	*Continuation unpredictability*	.15	.15	.19	−.01	−.12	.07	.12	.41 ••	–	
10	*Stylistic unpredictability*	.22	.21	.09	−.0	−.07	−.21	−.17	.16	.87 •••	–

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

Table 7.5: Mixed-Effects Ordinal Logistic Regression Analysis Predicting Participants' Rankings Using Combined Measures From Both Representations, Accounting for Participant and Continuation Differences

Predictor	β	*SE*	*z*	*p*	
Pitch interval					
Late-composition unpredictability	0.62	0.23	2.67	<.01	••
Thematic unpredictability	−0.22	0.20	−1.09	.27	
Continuation unpredictability	0.52	0.33	1.59	.11	
Stylistic unpredictability	−0.33	0.29	−1.13	.26	
Inter-onset interval					
Late-composition unpredictability	0.69	0.13	5.21	<.001	•••
Thematic unpredictability	−0.38	0.13	−2.93	<.01	••
Continuation unpredictability	−0.40	0.39	−1.05	.30	
Stylistic unpredictability	0.47	0.50	0.94	.35	

•• $p < .01$; ••• $p < .001$

analyses above) presenting a potential and undesirable source of collinearity within the regression model, separate regression analyses were conducted; each combined one of these two measures with the remaining three, using variants for both representations together as predictors. The best-fitting model, using *late-composition unpredictability* is shown in Table 7.5 (with an AIC score of 2054.94, compared to 2058.75 for the model using *composition unpredictability*).

Significant effects were found for *late-composition unpredictability* for both pitch interval and inter-onset interval, such that the probability of a continuation being placed in a better rank (*i.e.*, closer to Rank 1) increased as the information content in these measures decreased. For pitch, an increase of 1SD was associated with an odds ratio of 2.38 for a better rank; for rhythm, an increase of 1SD was associated with an odds ratio of 1.23 for a better rank. This effect is illustrated in Figure 7.1. Conversely, rhythmic *thematic unpredictability* displayed a smaller, but still significant effect, in the opposite direction. Lower values in this measure increased the probability that a continuation would be placed in the bottom rank—an increase of 1SD was associated with an odds ratio of 0.47 for a better rank.

7.5.2 Unity ratings

All 80 participants returned a rating of structural unity for each composition they heard. Across all 320 ratings, unity was rated with a mean of 59.26 (*SD* = 26.99). As participants were explicitly instructed to base their ratings on the whole of the context and their top-ranked continuation, unity ratings were combined to produce a mean rating for each different continuation (109 of the continuations were placed in the first rank by at least one participant).

Using non-centred measure values (allowing for testing of effects between

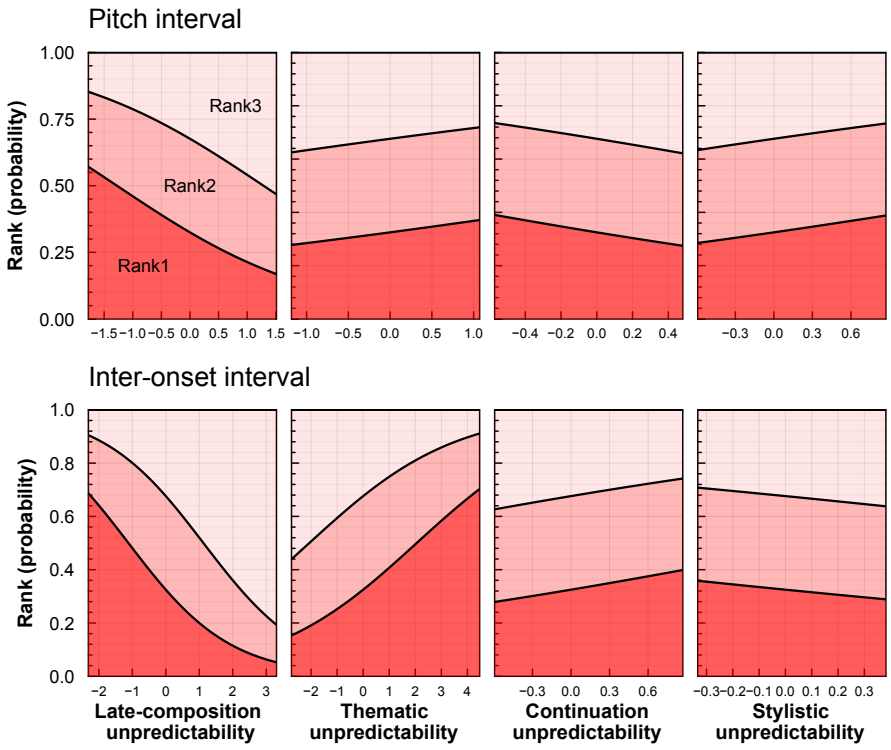


Figure 7.1: Results of ordinal logistic regression predicting participants' rankings using *late-composition unpredictability*, *thematic unpredictability*, *continuation unpredictability* and *stylistic unpredictability* together for both pitch interval and inter-onset interval representations. Subplots show probabilities of a continuation being in Rank 1, 3 or 3, given a measure information content value. For each subplot, all other measures are held at zero (*i.e.*, at their means).

stimuli, as well as between individual sets of continuations), correlations between measures and ratings were tested. No measure using the pitch interval representation was found to have a significant correlation with unity ratings (*composition unpredictability*, $r(107) = -0.04$, $p = .67$; *late-composition unpredictability*, $r(107) = -0.03$, $p = .73$; *thematic unpredictability*, $r(107) = 0.04$, $p = .66$; *continuation unpredictability*, $r(107) = -0.11$, $p = .25$; and *stylistic unpredictability*, $r(107) = -0.08$, $p = .43$). Significant negative correlations with unity were found for inter-onset interval measures of *continuation unpredictability* ($r(107) = -0.40$, $p < .001$) and *stylistic unpredictability* ($r(107) = -0.35$, $p < .001$), with no significant effects of the remaining measures (*composition unpredictability*, $r(107) = -0.07$, $p = .48$; *late-composition unpredictability*, $r(107) = -0.05$, $p = .59$; and *thematic unpredictability*, $r(107) = -0.06$, $p = .53$).

7.5.3 Gold-MSI scores

After aggregating Gold-MSI responses into scores for each participant. Participants had a mean score for *general sophistication* (out of a possible scale range 18–126) of 52.36 ($SD = 17.76$), a mean score for *perceptual abilities* (scale range 9–63) of 39.98 ($SD = 6.65$), and a mean score for *musical training* (scale range 7–49) of 14.98 ($SD = 9.45$).

Composition-centred measures were aggregated by taking means of each participant's first-choice continuations. The correlations between Gold-MSI scores and aggregated measures are shown in Table 7.6. A sole significant positive correlation was found between participants' *perceptual abilities* and pitch interval *late-composition internal unpredictability*.

The ability of Gold-MSI scores to predict participants' mean unity ratings was tested using a multiple linear regression analysis. As shown in Table 7.7, no score managed to significantly predict ratings.

7.6 Discussion

The experiment presented in this chapter aimed to test the extent to which statistical learning could account for listeners' perception of large-scale continuation. This task of identifying suitable continuations was chosen for its ability to act as an important indicator of the effects of large-scale thematic structure. In a manner similar to the well-established probe-tone paradigm's ability elicitation of pitch expectations, given exposure to a particular musical context, the continuation task allows us to examine the features learned by listeners in a context on the time-scales needed to investigate the effects of large-scale structure. Five computational measures, each using both pitch interval and inter-onset interval representations, were used to characterise competing statistically-learned

Table 7.6: Pearson's *r* Correlations Between Participants' Gold-MSI Scores and Mean Measure Values for Their First-Choice Continuations

Measure	*General Sophistication*	*Perceptual abilities*	*Musical training*
Pitch interval			
Composition unpredictability	.01	.20	.03
Late-composition unpredictability	.05	.22 •	.07
Thematic unpredictability	−.06	.14	−.06
Continuation unpredictability	.14	.12	.21
Stylistic unpredictability	.14	.03	.18
Inter-onset interval			
Composition unpredictability	.04	−.12	.03
Late-composition unpredictability	.05	−.12	.06
Thematic unpredictability	−.07	−.14	−.09
Continuation unpredictability	.04	.12	−.06
Stylistic unpredictability	.06	.17	−.01

• $p < .05$

Table 7.7: Linear Regression Analyses Predicting Participants' Mean Unity Ratings by Their Gold-MSI Scores

Predictor	β	*SE*	*t*	*p*	
Intercept	48.58	11.33	4.30	<.001	•••
General sophistication	0.05	0.22	0.24	.81	
Perceptual abilities	0.38	0.40	0.95	.34	
Musical training	−0.47	0.32	−1.48	.14	

••• $p < .001$

features of hypothesised importance in this task; measures were derived from the theoretical modelling of thematic structures based on the statistical learning of structural features, discussed in Chapter 3.

The findings of the continuation task of this experiment provide strong evidence that statistical learning has an important role in the perception of large-scale continuation. In particular, the results showed significant evidence that intra-opus statistical learning influenced participants' perception of large-scale continuation, as hypothesised. Measures involving training on material taken from the context section of compositions were found to differ significantly between participants' rankings. This effect, in general, appears to be a robust one; the three measures of *composition unpredictability*, *late-composition unpredictability*, and *thematic unpredictability* (taking training from the context, the second half of the context, and the themes identified within the context, respectively) showed significant influence over rankings in the ANOVA of individual measures (Table 7.3), for both representations, and pitch *thematic unpredictability* and rhythm *late-composition unpredictability* showed similar effects when measures were combined in the ordinal regression analysis (Table 7.5).

More specifically, however, the precise differences between these measures—and so the particular features that may be used by listeners in this task—are not so clearly defined. The three measures are, by their nature, closely related. The results of the correlation analyses showed that *composition unpredictability* and *late-composition unpredictability* were extremely-highly correlated (Table 7.4); this is likely due to the repetition of earlier material being great enough within the context second-halves that any additional learning afforded by the inclusion of complete compositions was not required. This possibly indicates that listeners do not need to form perfect memories of the occurrence of events within large compositions for such statistical learning of thematic structures to take place (this is tentatively further supported by the fact that the regression model containing *late-composition unpredictability* provided a better fit to the rating data than the model containing *composition unpredictability*, although the difference in AIC was small). This redundancy in training material for intra-opus learning is further supported by the significance of rhythmic *thematic unpredictability*, requiring learning based on far less information to judge effects of structure. These findings provide some confirmation of our hypothesis, developed from the findings of the experiment of Chapter 6, of the importance of *composition unpredictability* over the other context-based measures.

The three context-based measures discussed so far showed mostly consistent effects between pitch interval and inter-onset interval representations, producing similar effects in the analysis, although not being significantly correlated across representations. However, for the remaining measures of *continuation*

unpredictability and *stylistic unpredictability*, differences were found between variants for each representation. As discussed previously in this thesis, a likely cause of this disparity between representations is the relatively frequent occurrence of rhythmically-isochronous compositions in western music.

When comparing the differences between representations, limitations of the continuation generation process should be considered. Unlike the modification of notes in continuations based on pitch, modifications based on inter-onset interval have the potential to disrupt metrical patterns outside of the immediately changed note-event. However, this disruption is mitigated, firstly, by the short length and fragmentary nature of the continuations (often containing only a small number of bars), and secondly, as generation models were trained only on metrically complete material, these changes must still be rhythmically predictable (*i.e.*, modifications that produce unusual rhythmic patterns are highly unlikely). No evidence was found that participants favoured continuations that were rhythmically close to the originals (containing the original metrical patterns) over the modified version, whereas a significant effect was found for pitch interval edits.

The findings of this experiment provide some validation for the expansion of the methods used to model statistical learning from monophonic to polyphonic material. As hypothesised, no significant differences were found between compositions based on their assigned texture category, with no difference between measures when using monophonic and polyphonic techniques.

Likewise, few differences were found between compositions in 'short' and 'long' categories; a sole significant difference emerged for the pitch interval measure of *continuation unpredictability* (Table 7.2). This difference may be caused by the duration of continuations, with the longer continuations of the 'long' composition category containing a greater opportunity for variation. Aside from this measure, the lack of a significant difference between these categories confirms our hypothesis that participants would have the ability to perceive the effects of thematic structure, regardless of composition duration.

The task of rating the perceived unity of compositions (contexts plus participants' first-rank continuations) sought to replicate the link between *internal unpredictability* and perceived structural unity found in the experiment of Chapter 6. However, no significant effect was found between ratings and any of the context-based measures; this included the measure of *composition unpredictability*, directly related to a compositions *internal unpredictability*. There are several potential reasons why a comparable effect was not found, the most likely being that, due to the priorities of participants in completing the main continuation task, the unity task was too far removed from that of the previous experiment, in which all of the composition (*i.e.*, the context and chosen continuation) was heard together without being interrupted. This conclusion,

perhaps, also gains some support from the measures that were significantly correlated with ratings (inter-onset interval *continuation unpredictability* and *stylistic unpredictability*); these effects may suggest participants were only using properties of the individual continuations, independent of their contexts, to inform their ratings. An additional possibility is that any other effects may have been confounded if some participants were to treat the unity rating instead as a confidence rating in their rankings.

The results for participants' Gold-MSI scores provided little evidence of any influence of musical background on their first-rank choices. Only the single measure of pitch interval *late-composition unpredictability* was found to be significantly correlated with any score. For this finding *late-composition unpredictability* was positively correlated with participants *perceptual abilities*; this would imply that participants with greater *perceptual abilities* were more likely to choose continuations that were more unpredictable, based on this measure. No significant effect was found between Gold-MSI scores and unity ratings. No evidence was found to support the hypothesised connection between musical background and stylistic features of music. The lack of evidence in support of this relationship in the continuation task may be due to the ranking nature of the task, with participants prioritising intra-opus considerations regardless of the level of stylistic conformity.

7.7 Summary

The experiment presented in this chapter contributes to the testing of the underlying hypothesis of this thesis—that large-scale thematic structures can be perceived through the statistical regularities that they form—and the testing of computational methods informed by this theory. This experiment had the objective of testing listeners' judgements of large-scale continuations, given a long musical context, with plausible continuations generated according to predictability due to composition, theme and stylistic considerations. This experiment expanded on the scope of the previous experiment presented testing the perception of large-scale structure in two ways. First, it increased the duration of stimulus compositions used (and so tested whether the significant general findings of Chapter 6 hold for longer structures), using compositions both of 2 and 4 minutes long. Second, it introduced and tested techniques (discussed in Chapter 3) for the modelling of thematic structure in polyphonic compositions of two voices.

The findings for this experiment corroborate those of the previous experiment, finding significant evidence in support of the importance of *internal unpredictability*, or in this experiment *context unpredictability*, in the perception of large-scale thematic structures. The results showed that continuations

that contributed to lowering a composition's overall *internal unpredictability* (*i.e.*, continuations that were predictable given the context composition) were favoured.

Chapter 8

Modelling Large-Scale Order

8.1 Overview

In this chapter, both the final behavioural experiment and a computational experiment are presented. These experiments aim to contribute to the converging understanding of thematic structure built-up throughout the course of this thesis through the investigation of the perception of large-scale order. Testing the ability of listeners to perceive large-scale order is central to understanding the cognition of thematic structure; the order of musical segments is one of the key specifications of large-scale forms (Whittall, 2001), and through order repeated material can be progressively developed and distributed throughout a composition.

The majority of experimental research into the perception of large-scale structures (reviewed in Chapter 3) used experimental designs in which multiple versions of a composition would be created based on its internal segmentation, with participants choosing or rating versions on various perceptual scales (either within or between subjects). There have, however, been several attempts to adopt a puzzle-based task, in which participants are supplied with the segments of the composition and are tasked with placing them into what they believe is the most appropriate order (Granot & Jacoby, 2011, 2012; Tillmann et al., 1998). Certainly, this task is further removed from the real-world listening conditions provided by the more-established experimental paradigm, however, in doing so the experimenter gains the potential to have a far greater insight into listeners' abilities and sensitivity to musical properties. Specifically in the case of thematic structure, as noted by Granot and Jacoby (2011, 2012) in their motivation for employing such a paradigm, the puzzle task has the advantage that the assumption that the original composition order should inherently and solely be preferred is avoided.

However, puzzle-task experiments have the disadvantage that their output is more complex to analyse; although the combinatorial mathematics means

that if orders do match between participants it is highly unlikely to be due to chance, determining the more likely partial effects is difficult. Additionally, due to the greater experimental control afforded participants, there is an increased risk of no meaningful conclusions being able to be drawn at all.

The first experiment of this chapter adapted the experimental paradigm of Granot and Jacoby (2011); participants were tasked with ordering the segments of monophonic and two-voice polyphonic compositions (with complete compositions of approximately 2 and 4 minutes in duration). One alteration was made to the task—participants were provided with the original starting segments. This information provided to participants works to make the task more tractable and removes a potential source of random noise among responses. The analysis of this experiment has an advantage over those of previous puzzle experiments—the computational modelling allows for individual participant orders to be characterised based on features of thematic and tonal structure. A Monte Carlo approach was used to test for significant differences between participants' orders and those of a large randomly-generated set.

Using these same methods that characterise orders and their differences from chance, a second, computational experiment is presented in this chapter in which modelled properties of original composition orders are analysed, alongside example orders taken from the extremes of the randomly-generated set.

8.2 Modelling

Our over-arching hypothesis for this research is that the perception of large-scale thematic structure is facilitated by the learning of statistical regularities within a composition. We continue to apply this hypothesis to the experiments presented in this chapter. As with the previous experiments of this thesis, computational measures tailored to the specific experimental paradigm were used to test the effects of statistical learning—as well as other plausible features—on the given task.

The principles, detailed in Chapter 3, of modelling thematic structure using Prediction by Partial Matching (PPM) models were applied again here. PPM is a variable-order Markov modelling technique that estimates likelihoods for the occurrence of note-events within a symbolic sequence, given the number of occurrences of subsequences of varying size within a training sequence, smoothing between models of different orders (Bunton, 1997; Cleary & Witten, 1984). The likelihood estimates are converted into a value of information content, $h = -\log_2 p$, or the unpredictability of the event. In this research, for a given composition, the training sequence may be taken from within the composition itself (often necessitating dynamic training of the model), modelling the its *internal unpredictability*, or the model may be trained on a large corpus of compositions,

modelling its *stylistic unpredictability*. Unlike the PPM models used in the previous chapters that possess perfect memory for all events, an additional PPM version is also used that is memory-constrained to a certain duration before a given event (Harrison et al., 2020).[1]

8.2.1 Measures

A series of measures were constructed to characterise a given segment order for a composition in terms of features of hypothesised importance. As well as modelling the effects of inter and intra-opus statistical learning, measures sought to quantify other features that may influence a given order, such as tonal considerations, or the closeness of an order to that occurring in the original composition.

In many cases, measures were dependant on a specific representation of the musical surface. These measures were used in two forms, using pitch interval and inter-onset interval to separately examine effects of pitch and temporal domains.

Internal unpredictability

As with the other experiments presented in this thesis, the effects of intra-opus statistical learning are central to our investigation. Two measures were constructed to characterise the *internal unpredictability* of a newly ordered composition for a given representation. This was calculated as the mean information content across note-events in all segments, using individual models for each segment that were trained on the preceding segments in the order, as well as being dynamically trained within itself.

This method differs from the measures of *internal unpredictability* used in the previous experiments, that—more simply—used information content values from a single model trained dynamically within a composition. The resulting difference between such an approach, if it were applied to the current experiment (joining all segments together before dynamically training a model), and the method used is that the presented method excludes n-grams from spanning section boundaries.

The two measures using this method differed in the memory constraints applied to their training.

Internal unpredictability (perfect). The first of these measures used PPM models with perfect memory (such as possessed by IDyOM in the previous experiments), so that all observations carried equal weight.

[1] It is for this reason that the modelling in this chapter does not use IDyOM, but rather is based on the PPM-decay implementation of Harrison et al. (2020). As only one-dimensional representations of the musical surface are used, the output of the two implications should be in agreement.

Internal unpredictability (buffer). The second of these measures used a memory-constrained variant of PPM (Harrison et al., 2020); only events that occurred within a fixed buffer period before a given note-event contributed to the information content of that event. The length of this buffer period can be considered a free parameter, with the optimal duration for the stimuli to be determined. For the present experiments, the length of this buffer period was optimised to the segmented compositions used, maximising the variance of *internal unpredictability (buffer)* between possible orderings of a composition. This allowed the measure to describe the maximum amount of variation in orders and to provided the largest possible distinction to *internal unpredictability (perfect)* (*i.e.*, with an infinitely long buffer).

Information flow

The above measures give an absolute value of orders' *internal unpredictability*; however, these measures do not capture the positioning of segments due to relative differences in internal unpredictability.

Internal unpredictability flow. A measure of *internal unpredictability flow* was used to characterise these relative differences as the mean absolute difference between information content means for segments, trained as described for *internal unpredictability (perfect)*. Orders with segments closer in overall unpredictability scored lower in this measure.

Stylistic unpredictability flow. Likewise, a measure was used to encapsulate ordering based on relative *stylistic unpredictability* of segments. Segment information contents were calculated using a model trained on a corpus of 600 monophonic melodies.

Variation

Variation. Unlike the measures of variation and repetition used in the previous experiments of this thesis, we cannot apply the same techniques of theme detection to the ordering paradigm. An alternative measure of *variation* was used. To characterise the extent to which the segments of an order progress in a linear manner (*i.e.*, each development of material takes it further from its original instance). Evidence suggests that the information-theoretic concept of *compression distance* can provide a cognitively plausible measure of dissimilarity between musical sequences (Pearce & Müllensiefen, 2017), where the directional compression distance between two sequences, $D(x|y)$, is calculated as the normalised summed information content of the latter sequence (of length k), given a PPM model trained on the former.

$$D(x|y) = \frac{1}{k}\sum_{i=1}^{k} h_{m_y}(x_i)$$

For an ordering of segments, s_1^n, for each segment the minimum compression distance to those preceding it was taken and these distances averaged across the order.

$$V(s_1^n) = \frac{1}{n-1}\sum_{i=2}^{n} \min\{D(s_i|s_1), \ldots, D(s_i|s_{i-1})\}$$

As orders were exhaustive, with each segment appearing once, and only once, an overall lower mean compression distance indicates that segments were ordered such that variation of themes progressed more linearly throughout the order.

Tonality

The rearranging of a composition's segments, as used in this paradigm, disrupts both large-scale thematic and tonal structures alike. While this thesis is primarily concerned with the effects of thematic structures, the effects of tonal structure need to be taken into consideration.

Tonal distance. A measure was used that summarised the extent to which an order conformed to the western-classical norm of favouring modulation to more closely related keys. Keys were identified for the beginnings and endings of segments (the first and final thirds, based on duration) using the Krumhansl and Schmuckler algorithm (Krumhansl, 1990). The distance of transitions between keys was measured as the number of steps around a circle of fifths between adjacent key signatures, with an additional penalty of 1 added for any further move to a relative major or minor. The mean of these distances was used as the measure of *tonal distance*.[2]

Global features

To account for the effects of closure on orders, two measures were used that modified measures of *tonal distance* and *variation* to apply only to the relationship between starting and ending segments.

Global tonality. To measure tonal closure, the circle of fifths distance described above was applied between to opening keys of the first segments and

[2]While it is possible that multiple modulations can be present in a segment—resulting in a larger than expected distance between keys in adjacent segments—the fact that all ordering are exhaustive maintains the premise of this measure within the context of a given set of segments.

the closing keys of the last. Orders with more closely tonally related openings and endings gained a lower score.

Global distance. The compression distance, dependant on pitch interval or inter-onset interval representation, between first and last segments in an orders were used as a test for thematic closure—the extent to which these sections contained closely related material.

8.2.2 Closeness to original orders

Finally, two measures were used to quantify the relationship between a given order and a specific target order—for example, between an order and the original order of segments from that composition. There are several ways in which closeness between orders can be judged; the absolute positions of segments can be compared (*i.e.*, that segments in the same position match), or the extent to which smaller matching subsequences exist between the two can be tested.

Dissimilarity. Compression distances were calculated between segments at corresponding positions in the given and target (or original) orders (dependent on representation). Distances were averaged across all positions to give a measure of overall *dissimilarity*.[3]

Edit distance. As with analysis of orders used by Granot and Jacoby (2012), Levenshtein (1966) edit distance was used as a metric to compare two orders based on the minimum number of operations required to transform one order to the other. Three types of operation were permitted: (1) insertion of a value into the sequence; (2) deletion of a value; or (3) substitution of one value for another. Orders with a lower distance contained a greater number of subsequences in common. For the purposes of the present experiments, as multiple segments within a composition may be highly similar, or even identical, no distinction between highly-similar segments was made when calculating this measure. Using the pairwise compression distances between segments, segments were treated as equivalent if the distance between them was less than 4 SD from the mean of compression distances comparing segments to themselves.

8.2.3 Monte Carlo simulation of orders

The measures, as described above, each provide a way to quantitatively describe a segment order for a specific composition and to compare it to other orderings

[3] Not to be confused with the measure of the same name specific to the experiment of Chapter 5.

that use the same segments. For these measures to be useful as a means of evaluating the wider effects of large-scale order, they need two further properties—they need to have the ability to be generalisable between compositions, and they need to provide a measure of how unexpected such an order may be.

To achieve this, a Monte Carlo approach was used, generating random orders for each composition (maintaining the fixed first segment constraint). For compositions with large numbers of segments (nine or greater), 20,000 random orders were created; otherwise, the exhaustive set of permutations was used (i.e., for compositions where all possible permutations of segments were lower than this value). Measure values were calculated for each of the randomly generated orders and the means and standard deviations were obtained for each composition, to which any given future order could be standardised. These standardised measure scores provided a measure of the distance from chance for an order (*i.e.*, the distance from zero), and appropriately, adjusts scores to the ranges available within combinations, allowing measures to be compared between orders using different compositions.

8.3 The present experiments

The research presented in this chapter aimed to examine the importance of large-scale structural order in music. The investigation is divided into two experiments—the first, a behavioural experiment testing the abilities of participants to order segmented compositions and how they perceive unity, the second, modelling the order of segments in their original compositions and exploring the orders that can produce extremes in these measures.

Experiment 1 aimed to test the extent to which large-scale structural order can be perceived by listeners, firstly through the ability of participants to recreate original orders of segmented compositions, secondly by the extent to which the modelled measures can explain participants' orders. Of these measures, we particularly aimed to test the effects of features based on statistical learning. Additionally, Experiment 1 aimed to test for relationships between the properties of orders as reflected by the computational measures and participants' perceived sense of structural unity.

For this experiment, we hypothesised that participants' chosen orders would be significantly related to those of the original compositions (using the *dissimilarity* and *edit distance* measures). Secondly, we hypothesised that participants would prioritise certain features in their orderings—that orders would be more internally and stylistically predictable than chance orders (*i.e.*, having lower values of *internal unpredictability (perfect)*, *internal unpredictability (buffer)*, *internal unpredictability flow* and *stylistic unpredictability flow*), and would more closely follow tonal convention than chance orders. Based on the find-

ings of the previous three behavioural experiments of this thesis, of these measures we hypothesised that measures of *internal unpredictability* would have the strongest effect on orders given.

For the unity ratings given by participants to their orders, we followed the hypothesis given in Chapter 6, that the measures of *internal unpredictability* would most closely predict ratings.

For Experiment 1, we also recorded measures of participants' musical backgrounds, hypothesising that the greater exposure to music and its stylistic conventions that comes with greater musical training would aid participants in their recreation of compositions' original orders and favour approaches based on *stylistic unpredictability*.

While the aims of this research are chiefly concerned with the behavioural elements of the first experiment, Experiment 2 aimed to explore further the properties of the modelled measures. Firstly, the task of composers, in their creation of the original compositions from which segments were taken, shares some similarity with the task given to participants; they are both producing musical material following some organisational scheme. The techniques used to model and test participants' orders were, therefore, also applied to the orders in which the segments appeared in the original compositions, with the aim of understanding the features that may have influenced the organisation within them. Secondly, as the Monte Carlo method, described above, made it possible to sample orders from different points across the range of measures, examples of orders that can produce minimum and maximum values in each of the measures were examined for illustrative purposes.

For the original orders, it was hypothesised that these orders would prioritise the same measures as those of the participant, that orders would be more predictable and more their segments more closely tonally related than chance orders.

8.4 Experiment 1

8.4.1 Methods

Participants

Eighty participants were recruited to participate in the experiment using the Prolific online platform. Participants had a mean age of 37.73 (SD = 12.81), 47 were women and 33 were men. Participants were required to have English as their first language and to have normal, or corrected to normal, hearing. All participants were residents of the UK or the USA. No recruitment criteria were enforced on participants' level of musical training; 47 participants reported having some musical training on at least one instrument, with seven of those reporting

more than 10 years of training.

Stimuli

Stimuli consisted of segmented versions of 40 compositions (using the same compositions as used in the experiment of Chapter 7).[4] Compositions belonged to either a 'monophonic' or 'polyphonic' (of two voices) texture category, and to either a 'short' or 'long' length category (of approximately 2 or 4 minutes, respectively). The 20 monophonic compositions were selected from a large corpus of western-classical tonal melodies (described in Chapter 4); the 20 polyphonic melodies consisted of western-classical instrumental duets. In cases where compositions were longer than their respective length category, endings that produced the minimal structural disruption were manually identified within an additional 15-second window.

Possible appropriate boundaries for segmentation were manually identified for each composition (in a descending order of preference) at section boundaries marked in the score, at substantial changes of melodic content, or otherwise at strong phrase endings. Segmentation targeted 7 to 10 segments for 'short' compositions and 10 to 15 segments for 'long' compositions of approximately similar duration. To discourage participants from attempting to complete the task by matching small surface features between segment endings and beginnings, segments were modified to end at their cadences. Removed material consisted of passing notes and other stylistic figurations (such as scales) that did not contain any extra harmonic or melodic function other than to lead into the next section.

The resulting stimuli[5] contained a mean of 8.20 segments per 'short' composition (SD = 0.95) and 12.65 segments per 'long' composition (SD = 1.60). 'Short' segments had a mean duration of 15.09 seconds (SD = 4.83; of which, 'monophonic' M = 14.95, SD = 4.85, and 'polyphonic' M = 15.24, SD = 4.60). 'Long' segments had a mean duration of 18.89 seconds (SD = 5.56; of which, 'monophonic' M = 18.94, SD = 5.98, and 'polyphonic' M = 18.84, SD = 5.14).

Audio files were generated for each segment using MuseScore notation software at a uniform loudness, preserving original pitches, rhythms and tempos. 'Monophonic' stimuli were rendered using a piano timbre; to aid the perceived separation between voices, 'polyphonic' stimuli were rendered with a piano timbre for the first voice and a marimba timbre for the second.

[4] Details of the compositions used are given in Table B.2 of Appendix B

[5] Modelling code, stimuli and data for this experiment can be found at https://osf.io/4zs2j/.

Procedure

The experiment was conducted online using a web browser-based task, implemented using the jsPsych JavaScript library (de Leeuw, 2015). The experiment consisted of two parts—the first containing the main experimental ordering task (taking around 35 minutes for completion), the second containing the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire for musical sophistication (Müllensiefen et al., 2014).

For the ordering task, each participant received two trials. All texture and length categories were covered between the two trials, such that a participant would receive either a 'monophonic' and 'short' category trial followed by a 'polyphonic' and 'long' one, or a 'polyphonic' and 'short' category trial followed by a 'monophonic' and 'long' one. Due to the vastly higher number of permutations possible for segment sets in the longer category, those in the easier 'short' category were always first to be presented. Following this structure, compositions were assigned at random to participants, with each composition used exactly four times.

As shown in Figure 8.1, for a given trial of the ordering task, the segments of an individual stimulus composition were represented onscreen as a series of solid-coloured circles, randomly distributed within the central portion of the screen, with each segment having its own circle of a different colour. Segment circles could be double-clicked to play from the beginning or stop playing the respective segment,[6] and could be freely dragged around the window. Participants were asked to arrange the segments within an indicated response area in the order they thought made 'the most coherent piece of music'. Additionally, participants were advised to not try to assemble their orders based on matching segment endings to beginnings (which would be a very difficult task). The first segment of the order was given within the response area (immovable and coloured black). An additional option was given to play through all segments in the response area in the order in which they were positioned. Completion of the task was limited to 10 and 20 minutes for 'short' and 'long' categories, respectively, with a further minute then allowed for participants to submit their final orderings.

For each trial, after the ordering task participants were asked to rate the composition—in the order they had produced—on its level of structural unity (from 0 to 100), defined as the extent to which the music still sounded like one well-integrated, whole, single composition (see Chapter 6 for a more detailed description of this question).

[6] Each segment would always start playing from its beginning, as with the modified segment endings, this behaviour was implemented to discourage matching surface features across segments.

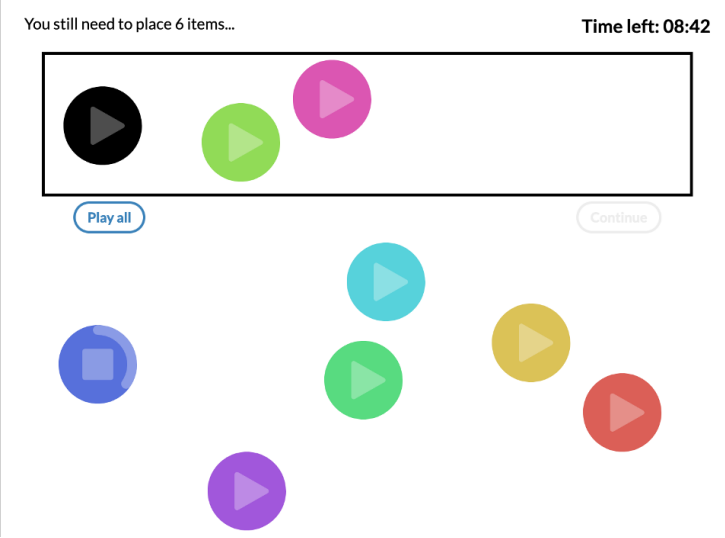


Figure 8.1: Interface used in ordering task trials. Segments of an individual stimulus composition were represented onscreen as solid-coloured circles. Segment circles could be double-clicked to play from the beginning or stop playing the respective segment. Participants were asked to arrange the segments within an indicated response area in the given time.

Statistical analysis

Data collected during this experiment belonged to three categories: (1) segment orders for each composition, for each participant; (2) unity ratings for each composition, for each participant; and (3) Gold-MSI scores for each participant, aggregated from their questionnaire responses. The statistical analyses for this experiment examined each of these in turn.

Values for each measure for individual participant orders were calculated; segments were modelled in their orders and standardised by composition to the means and standard deviations from the Monte Carlo set of randomised orders. For measures dependant on a specific representation of music, both pitch interval and inter-onset interval variants were analysed separately.

Differences between measures for orders belonging to compositions of different categories were examined. Independent t-tests were used to test for significant differences between measures for stimuli in the 'monophonic' and 'polyphonic' categories, and between 'short' and 'long' categories. To determine the effects of individual measures on participant orders, one-sample t-tests were used to test the significance of distances between each measure and the means of the Monte Carlo set of orders (the distance from zero for the standardised measures). The relationships between measures for participants' orders were then examined; first, by testing the correlations between measures, and second, using an exploratory principal component analysis (PCA) to describe the unique variance that could be explained by each measure (R-Core-Team, 2020).

Analysis for unity ratings split ratings at the mid-point of the scale to form

high and low-unity categories. The differences between these categories were tested using t-tests for each measure.

Participant responses to Gold-MSI questions were aggregated to produce measures of *general sophistication*, *perceptual abilities*, and *musical training* for each participant. To examine the effect of participants' musical backgrounds on their orderings, correlations between participants' Gold-MSI scores and mean measure values were tested. A multiple linear regression analysis was used to examine the ability of Gold-MSI scores to predict participant mean unity ratings.

8.4.2 Results

The results for this experiment are presented in turn for each of three types of data collected: participants' orders, their unity ratings of the ordered segments, and their Gold-MSI scores.

Participants' orders

Each of the 80 participants completed the ordering task for two sets of segments covering all texture and length categories. All participants successfully submitted an order for each of their trials, giving 160 orders, four for each different composition segment set. For each order provided by participants, all measures described in Section 8.2 were calculated, using both interval and inter-onset interval representations where relevant. Measures were then standardised to Monte Carlo means and standard deviations for each composition. This provided standardised measures, using both representations for each order.

For calculation of the *internal unpredictability (buffer)* measure, an optimal buffer duration was found for the stimuli used. The buffer duration parameter was optimised to maximise the variance between the information contents of randomly generated orders of a composition, averaged across all compositions, with the values for each representation combined. A smaller set of Monte Carlo orders was used than those used in the main analyses; 100 orders were generated for each composition. Figure 8.2 shows the relationship between buffer duration and the amount of variance between orders for both representations. An optimal duration of 30.01 seconds was found.

Participant agreement was calculated using measures of *edit distance* (normalised by number of segments), the proportion of edits needed to transform one order to another, and *dissimilarity* (pitch and rhythmic), the dissimilarity between segments in corresponding positions, between orders within each composition. Means in these measures were calculated when comparing each order to all others using that stimulus. Participants had a mean *edit distance* agreement of 0.53 (SD = 0.07), indicating the average proportion of edits needed to

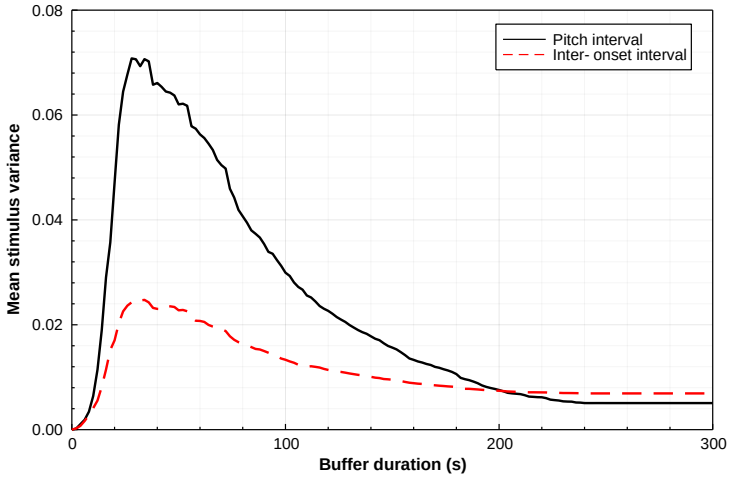


Figure 8.2: Amounts of variance between 100 randomly generated orders for segments of each composition, modelled using *internal unpredictability* with varying durations of buffer.

Table 8.1: Agreement Between Participant Orders or Stimulus Categories

	Monophonic		Polyphonic		Short		Long	
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Edit distance	0.54	0.06	0.52	0.08	0.49	0.07	0.57	0.06
Dissimilarity (PI)	0.76	0.15	0.78	0.15	0.71	0.16	0.83	0.12
Dissimilarity (IOI)	2.13	2.24	1.41	0.74	1.64	1.41	1.904	1.96

transform one order to another. Table 8.1 gives the level of agreement between participant orders for each of the texture and length stimulus categories.

For each measure, for each representation, independent t-tests were used to test for differences in participant responses between stimuli in texture and length categories. As shown in Table 8.2, only a single measure showed a significant difference between 'monophonic' and 'polyphonic' categories—that of *global distance* using the pitch interval representation. Chosen ending segments were closer (in compression distance) to given starting segments for 'monophonic' category stimuli. Table 8.3 shows tests between measures for 'short' and 'long' categories. Significant differences were found between categories for *internal unpredictability (buffer)* in both representations—in both cases 'long' stimuli were more predictable—and for *internal unpredictability flow* for pitch interval only—'long' stimuli had a large average difference in unpredictability between segments. The subsequent analyses make use of the combined data across categories.

Individual t-tests were used to test whether measures were significantly different to chance for participants' orders. Table 8.4 displays the result of these tests for both representations. The measures *dissimilarity*, for both representations, and *edit distance* were found to be highly significant, with participants' or-

Table 8.2: T-Test Comparisons Between 'Monophonic' and 'Polyphonic' Stimulus Categories for Each Measure, Using Participants' Orders

Measure	Monophonic		Polyphonic					
	M	*SD*	*M*	*SD*	*df*	*t*	*p*	
Pitch interval								
Internal unpredictability (perfect)	−0.26	1.07	−0.30	1.06	157.99	0.21	.83	
Internal unpredictability (buffer)	−0.56	1.36	−0.48	1.24	156.82	−0.42	.67	
Internal unpredictability flow	0.13	1.06	0.20	1.03	157.86	−0.38	.71	
Stylistic unpredictability flow	−0.43	1.10	−0.30	0.99	156.16	−0.78	.43	
Variation	−0.16	0.97	−0.09	1.04	157.03	−0.42	.67	
Global distance	−0.16	1.13	0.19	0.83	145.16	−2.24	.03	•
Dissimilarity	−0.31	1.26	−0.45	1.14	156.51	0.74	.46	
Inter-onset interval								
Internal unpredictability (perfect)	0.12	1.02	−0.06	1.04	157.95	1.09	.28	
Internal unpredictability (buffer)	−0.46	1.39	−0.38	1.21	155.00	−0.37	.71	
Internal unpredictability flow	−0.10	1.02	−0.12	1.11	156.86	0.10	.92	
Stylistic unpredictability flow	−0.47	1.32	−0.58	1.27	157.80	0.53	.60	
Variation	−0.15	1.12	−0.13	1.10	157.98	−0.13	.90	
Global distance	−0.16	1.02	0.02	1.00	157.96	−1.14	.26	
Dissimilarity	−0.32	1.17	−0.34	1.14	157.86	0.13	.90	
Other								
Tonal distance	−0.10	1.16	−0.21	0.98	153.53	0.63	.53	
Global tonality	−0.15	0.98	−0.20	0.94	157.80	0.35	.73	
Edit distance	−0.29	1.23	−0.48	1.14	157.18	1.02	.31	

• $p < .05$

Table 8.3: T-Test Comparisons Between 'Short' and 'Long' Stimulus Categories for Each Measure, Using Participants' Orders

Measure	Short		Long					
	M	*SD*	*M*	*SD*	*df*	*t*	*p*	
Pitch interval								
Internal unpredictability (perfect)	−0.22	1.01	−0.34	1.12	156.37	0.72	.47	
Internal unpredictability (buffer)	−0.27	1.18	−0.78	1.36	154.73	2.53	.01	•
Internal unpredictability flow	−0.01	1.01	0.34	1.06	157.70	−2.11	.04	•
Stylistic unpredictability flow	−0.23	0.98	−0.51	1.09	156.28	1.70	.09	
Variation	−0.12	0.97	−0.13	1.04	157.06	0.09	.93	
Global distance	0.07	1.04	−0.04	0.97	157.39	0.67	.50	
Dissimilarity	−0.39	1.16	−0.38	1.24	157.31	−0.07	.95	
Inter-onset interval								
Internal unpredictability (perfect)	−0.10	1.11	0.16	0.94	153.81	−1.63	.11	
Internal unpredictability (buffer)	−0.21	1.07	−0.63	1.47	144.56	2.09	.04	•
Internal unpredictability flow	−0.13	1.04	−0.09	1.09	157.59	−0.21	.84	
Stylistic unpredictability flow	−0.38	1.28	−0.67	1.29	157.99	1.42	.16	
Variation	−0.21	0.99	−0.07	1.22	151.66	−0.79	.43	
Global distance	−0.00	1.03	−0.14	0.99	157.77	0.86	.39	
Dissimilarity	−0.37	1.20	−0.30	1.11	157.06	−0.40	.69	
Other								
Tonal distance	−0.09	1.07	−0.22	1.07	157.99	0.81	.42	
Global tonality	−0.23	0.95	−0.11	0.97	157.92	−0.83	.41	
Edit distance	−0.47	1.15	−0.29	1.21	157.60	−0.98	.33	

• $p < .05$

Table 8.4: T-Tests for Measures Testing Distance of Participants' Orders From Chance

Measure	*M*	*SD*	*df*	*t*	*d*	*p*	
			Pitch interval				
Internal unpredictability (perfect)	−0.28	1.06	159	−3.31	−.26	0.001	••
Internal unpredictability (buffer)	−0.52	1.30	159	−5.08	−.40	<.001	•••
Internal unpredictability flow	0.17	1.04	159	2.01	.16	0.05	•
Stylistic unpredictability flow	−0.37	1.05	159	−4.42	−.35	<.001	•••
Variation	−0.13	1.00	159	−1.58	−.12	0.12	
Global distance	0.01	1.00	159	0.19	.01	0.85	
Dissimilarity	−0.38	1.20	159	−4.03	−.32	<.001	•••
			Inter-onset interval				
Internal unpredictability (perfect)	0.03	1.03	159	0.37	.03	0.71	
Internal unpredictability (buffer)	−0.42	1.30	159	−4.11	−.32	<.001	•••
Internal unpredictability flow	−0.11	1.06	159	−1.33	−.11	0.18	
Stylistic unpredictability flow	−0.52	1.29	159	−5.13	−.41	<.001	•••
Variation	−0.14	1.11	159	−1.56	−.12	0.12	
Global distance	−0.07	1.01	159	−0.87	−.07	0.38	
Dissimilarity	−0.33	1.15	159	−3.65	−.29	<.001	•••
			Other				
Tonal distance	−0.15	1.07	159	−1.83	−.14	0.07	
Global tonality	−0.17	0.96	159	−2.27	−.18	0.02	•
Edit distance	−0.38	1.18	159	−4.07	−.32	<.001	•••

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

ders significantly closer to composition original orders than those generated by chance (as shown in Figure 8.3). Using *edit distance* values without Monte Carlo standardisation but rather normalised by number of segments, edits to 62% of segments, on average, were needed to completely transform a participant order to that of an original composition

Measures of inter-opus statistical learning were found to be highly significant: pitch interval *internal unpredictability (perfect)*, and *internal unpredictability (buffer)* for both representations; participant orders were found to be more predictable than chance orders. Information-flow-based measures showed significant effects for both *stylistic unpredictability flow*, in both representations, and pitch interval *internal unpredictability flow*, where participants favoured smaller differences in predictability between segments than would be expected by chance. While no significant effect was found for the measure of *tonal distance*, the simpler *global tonality* measure was found to be significant, with participants choosing endings in keys more closely related to those of the starting segments.

The extent to which measures for participant orders were correlated with each other is shown in Table 8.5. While the full results for this analysis are given in the table, it is worth noting the results for measures of closeness to original orders. These measures were found to positively correlate to each other,

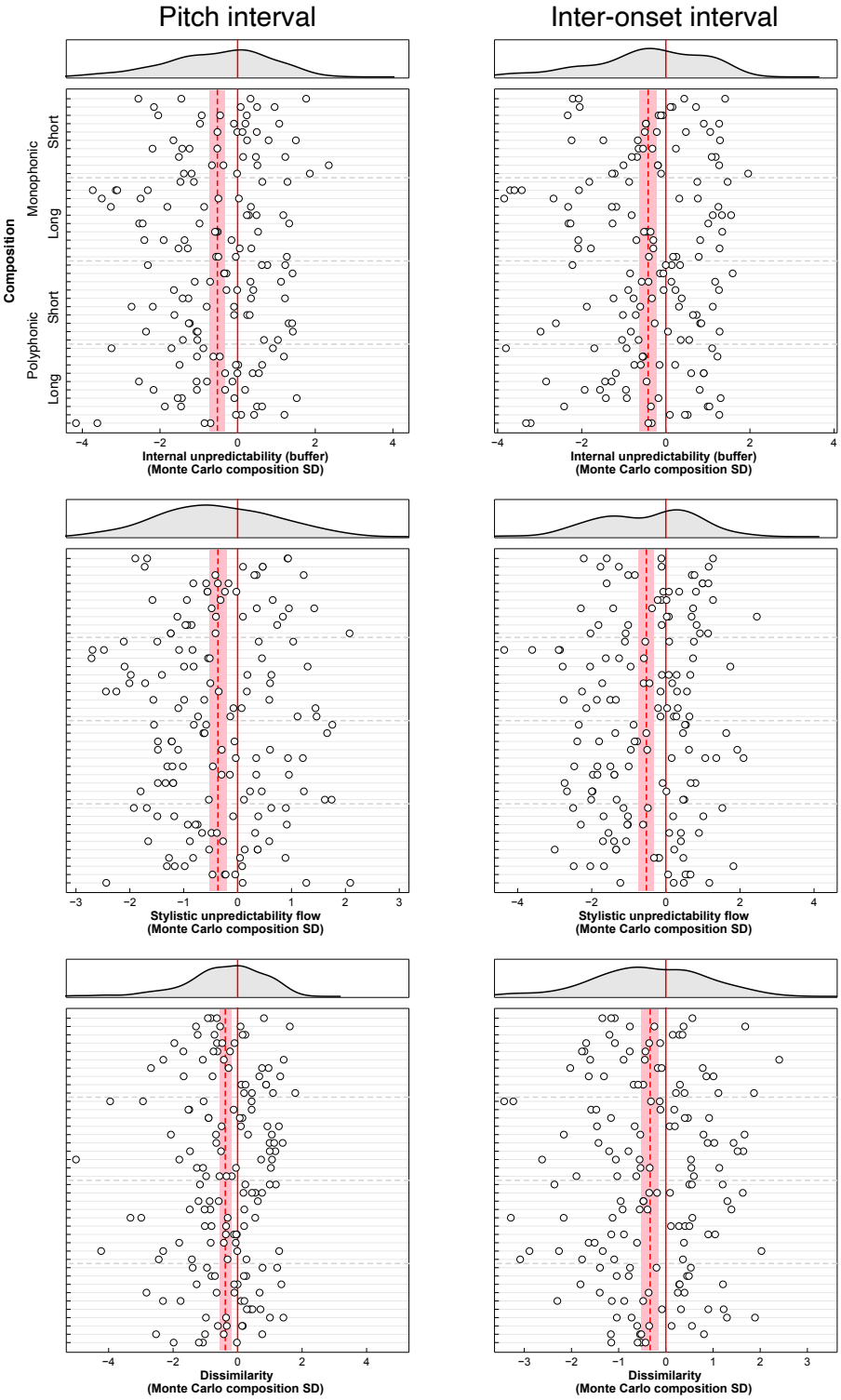


Figure 8.3: Values of participants' ordered segments, calculated for the measures of *internal unpredictability (buffer)*, *stylistic unpredictability flow* and *dissimilarity* for both pitch interval and inter-onset interval representations, standardised to Monte Carlo means and SDs. Solid lines show Monte Carlo composition means, dashed lines shows means of orders, with shaded region showing confidence intervals for t-tests of Table 8.4.

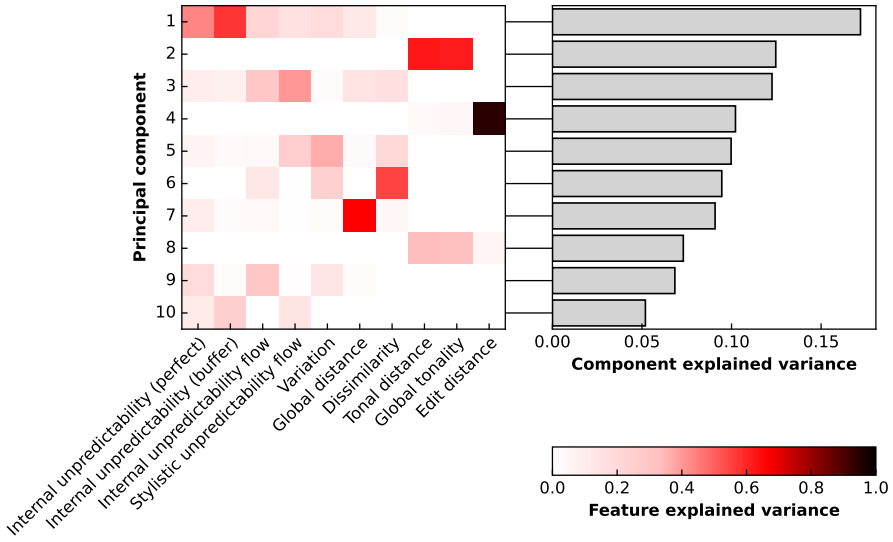


Figure 8.4: Explained variances and measure contributions for components of PCA for pitch interval.

between *edit distance* and both representations of *dissimilarity*. The only additional correlation found to one of these measures was between inter-onset interval *dissimilarity* and *variation*, however, the later measure was not found to be significantly different to chance in the prior t-test analysis.

A PCA was used to examine relationships between measures using the pitch interval representation (additionally including measures that were not representation specific) by geometrically reducing them into a smaller set of orthogonal components. The variance explained by each of these new components, and the contribution of individual measures to each of them, is displayed in Figure 8.4. A summary of the dominant features of the first seven of the identified components (that together account for 80.67% of the explained variance) can be given as follows: (1) a component of *internal unpredictability*, accounting for 17.20% of the variance in the data; (2) tonality, exclusively made up of *tonal distance* and *global tonality* measures, accounting for 12.47%; (3) information flow, however, with some weighting of similarity measures, accounting for 12.26%; (4) *edit distance*, accounting for 10.22%; (5) a more mixed component encompassing elements of *variation*, *stylistic unpredictability flow*, and *dissimilarity*, accounting for 9.98%; (6) *dissimilarity*, accounting for 9.46%; and (7) *global distance*, accounting for 9.08%.

Likewise, a PCA was used to examine relationships between measures using the inter-onset interval representation (also including non-representation-specific measures), shown in Figure 8.5. The first six components, accounting for a combined 73.45% of the variance, can be summarised as: (1) a component mixing *internal unpredictability (buffer)* and *stylistic unpredictability flow*, with only a limited effect of the other *internal unpredictability* measures, ac-

Table 8.5: Pearson's *r* Correlations Between Experiment Measures for Participants' Orders

	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
							Pitch interval										
1	*Internal unpredictability (perfect)*	–															
2	*Internal unpredictability (buffer)*	.31 •••	–														
3	*Internal unpredictability flow*	−.25 ••	−.17 •	–													
4	*Stylistic unpredictability flow*	−.02	.35 •••	.03	–												
5	*Variation*	.19 •	.17 •	.01	.07	–											
6	*Global distance*	−.12	−.08	.17 •	−.02	−.07	–										
7	*Dissimilarity*	.00	.10	.02	.06	.06	.02	–									
							Inter-onset interval										
8	*Internal unpredictability (perfect)*	.41 •••	.24 ••	−.15	−.00	.24 ••	−.09	.02	–								
9	*Internal unpredictability (buffer)*	.3 •••	.82 •••	−.16 •	.37 •••	.25 ••	−.07	.09	.23 ••	–							
10	*Internal unpredictability flow*	−.06	.03	.40 •••	.11	.10	.15	.07	.04	.02	–						
11	*Stylistic unpredictability flow*	.11	.54 •••	−.12	.38 •••	.19 •	−.02	.04	.10	.55 •••	.32 •••	–					
12	*Variation*	.09	.13	−.07	.10	.15	.05	.07	−.05	.16 •	.04	.11	–				
13	*Global distance*	−.05	−.02	.08	.04	−.08	.61 •••	−.03	−.18 •	−.03	−.02	.03	.07	–			
14	*Dissimilarity*	−.07	.09	−.02	.07	.06	.02	.78 •••	.08	.07	.03	.08	.19 •	.03	–		
							Other										
15	*Tonal distance*	.12	.21 ••	−.0	.12	.00	.14	.05	.11	.25 ••	−.03	.14	−.09	.01	.07	–	
16	*Global tonality*	−.14	−.11	.12	.02	−.08	.08	.06	−.04	−.14	.09	−.03	−.01	.05	.03	−.25 ••	–
17	*Edit distance*	.05	−.03	.10	−.04	.05	.06	.67 •••	.03	−.01	.02	−.04	.11	.05	.52 •••	.06	.05

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

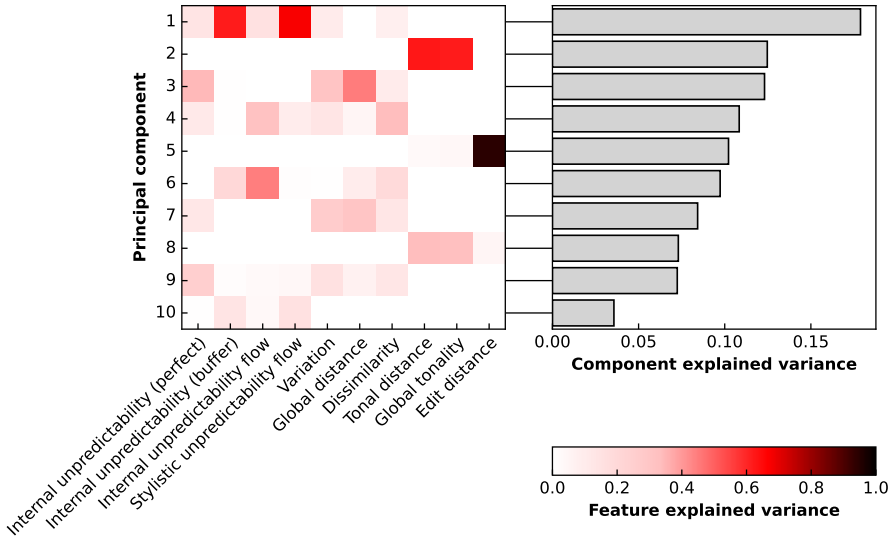


Figure 8.5: Explained variances and measure contributions for components of PCA for inter-onset interval.

counting for 17.88% of explained variance; (2) tonality, accounting for 12.47%; (3) *global distance*, *internal unpredictability (perfect)*, and *variation*, accounting for 12.31%; (4) *internal unpredictability flow* and *dissimilarity*, accounting for 10.84%; (5) *edit distance*, accounting for 10.22%; (6) another element of *internal unpredictability flow*, accounting for 9.73%.

Unity ratings

All 80 participants returned a rating of structural unity for each of their orders. Across all 160 orders, unity was rated with a mean of 55.65 (*SD* = 23.66).

As ratings were distinctly bimodal in their distribution around the scale mid-point of 50, orders were split into low- and high-unity categories. Table 8.6 gives the results for t-tests comparing high- and low-unity orders for each measure. No measure was found to significantly influence the unity category in which an order was placed.

Gold-MSI scores

After aggregating Gold-MSI responses into scores for each participant, participants had a mean score for *general sophistication* (out of a possible scale range 18–126) of 51.85 (*SD* = 18.72), a mean score for *perceptual abilities* (scale range 9–63) of 39.35 (*SD* = 7.64), and a mean score for *musical training* (scale range 7–49) of 13.82 (*SD* = 8.76).

Measures were aggregated by taking the mean for each participant's orders. The correlations between Gold-MSI scores and aggregated measures are shown in Table 8.7. Significant positive correlations were found for pitch interval *in-*

Table 8.6: Results of Comparison T-Tests Between Participants' High and Low Ratings of Orders' Unity for Each Measure

Measure	Low unity		High unity				
	M	*SD*	*M*	*SD*	*df*	*t*	*p*
Pitch interval							
Internal unpredictability (perfect)	−0.28	1.07	−0.28	1.06	129.79	−0.01	.99
Internal unpredictability (buffer)	−0.47	1.26	−0.55	1.33	135.26	0.41	.68
Internal unpredictability flow	0.15	1.07	0.17	1.03	126.61	−0.13	.90
Stylistic unpredictability flow	−0.51	0.93	−0.28	1.11	145.98	−1.42	.16
Variation	0.00	1.08	−0.21	0.95	117.13	1.26	.21
Global distance	−0.03	0.97	0.04	1.03	134.85	−0.42	.67
Dissimilarity	−0.31	1.08	−0.43	1.27	145.00	0.65	.51
Inter-onset interval							
Internal unpredictability (perfect)	0.15	1.02	−0.05	1.04	131.70	1.19	.24
Internal unpredictability (buffer)	−0.53	1.32	−0.35	1.28	127.23	−0.83	.41
Internal unpredictability flow	−0.14	1.04	−0.09	1.08	133.04	−0.28	.78
Stylistic unpredictability flow	−0.51	1.35	−0.53	1.26	123.67	0.07	.94
Variation	−0.07	1.17	−0.18	1.07	121.19	0.60	.55
Global distance	−0.14	0.95	−0.03	1.05	139.58	−0.69	.49
Dissimilarity	−0.36	1.03	−0.32	1.23	146.39	−0.23	.82
Other							
Tonal distance	−0.13	0.97	−0.17	1.13	143.64	0.22	.83
Global tonality	−0.17	0.94	−0.17	0.97	132.87	0.01	.99
Edit distance	−0.30	1.11	−0.43	1.23	139.28	0.73	.46

ternal unpredictability (buffer) and participants' *perceptual abilities*, and inter-onset interval *internal unpredictability (buffer)* and their level of *musical training*. Significant negative correlations were found between inter-onset interval *dissimilarity* and scores of participants' *general sophistication* and *musical training*.

The ability of Gold-MSI scores to predict participants' mean unity ratings was tested using a multiple linear regression analysis. As shown in Table 8.8, no score managed to significantly predict ratings.

8.4.3 Summary

While the findings and implications of this experiment are discussed in detail in the general discussion below, alongside the second experiment, the key findings are summarised here. Firstly, this experiment provided significant evidence that listeners are sensitive to the large-scale ordering of musical material. This was shown through the ability of participants to reconstruct the original orders of compositions significantly better than chance would suggest. This closeness to original orders existed both in the absolute positioning of segments within the order, and in the ability to correctly match smaller subsequences of the originals. The studies of Granot and Jacoby (2011, 2012) produced similar findings, where participant orders were closer to originals than chance would suggest (measured using edit distance).

Table 8.7: Pearson's *r* Correlations Between Participants' Gold-MSI Scores and Mean Measure Values for Their Orders

Measure	*General Sophistication*	*Perceptual abilities*	*Musical training*
Pitch interval			
Internal unpredictability (perfect)	.13	.16	.09
Internal unpredictability (buffer)	.19	.23 •	.21
Internal unpredictability flow	−07	.01	−07
Stylistic unpredictability flow	−03	.15	.03
Variation	.13	.05	.14
Global distance	−06	.01	−05
Dissimilarity	−19 .	.03	−14
Inter-onset interval			
Internal unpredictability (perfect)	.13	.10	.10
Internal unpredictability (buffer)	.16	.19	.26 •
Internal unpredictability flow	.03	.08	.01
Stylistic unpredictability flow	.01	−00	.02
Variation	−02	−03	.02
Global distance	−09	−04	.00
Dissimilarity	−24 •	−07	−22 •
Other			
Tonal distance	−14	−04	−05
Global tonality	−05	−03	−09
Edit distance	−14	−07	−08

• $p < .05$

Table 8.8: Linear Regression Analyses Predicting Participants' Mean Unity Ratings by Their Gold-MSI Scores

Predictor	β	*SE*	*t*	*p*	
Intercept	56.80	11.79	4.82	<.001	•••
General sophistication	0.01	0.23	0.06	.96	
Perceptual abilities	−0.15	0.40	−0.38	.70	
Musical training	0.30	0.40	0.74	.46	

••• $p < .001$

The findings of this experiment also provide evidence that statistical learning has an important role in the perception of large-scale order. The three measures that modelled possible effects of intra-opus learning—the measures of *internal unpredictability (perfect)*, *internal unpredictability (buffer)*, *internal unpredictability flow*—were found to be significantly lower than chance for participant orders when using the pitch interval representation. *Internal unpredictability (buffer)* was found to be significantly lower than chance when using the inter-onset interval representation.

Analysis of Gold-MSI scores of participants provides some limited evidence of an influence of musical background on orderings. Two measures were found to be significantly correlated with Gold-MSI scores. The findings suggest that participants with greater *perceptual abilities* or *musical training* returned orders that were closer to chance for *internal unpredictability (buffer)* (for pitch interval in inter-onset interval, respectively). A significantly negative correlation was found between inter-onset interval *dissimilarity* and scores of *general sophistication* and *musical training*; orders given by participants with higher scores were closer to original segment orders than those of participants scoring lower in these categories.

None of the measures were found to have a significant effect on ratings of unity of ordered compositions.

8.5 Experiment 2

The second experiment reported in this chapter aimed to explore further the role of large-scale order in music through the same computational measures as the first. While the first experiment focused on the orderings provided by participants in a behavioural experiment, the second investigated the properties of compositions in the orders in which they were originally produced. The process of composition shares some similarity with the task given to participants. They are both producing musical material following some organisational scheme and for both this ordering is based, in part, on their past musical experiences. Like the participants a composer must also decide on an ordering for the sections of a composition.

The measures described in Section 8.2 were also applied to the orders in which the segments appeared in the original compositions, with the aim of understanding the features that may have influenced the organisation within them. Additionally, the Monte Carlo standardisation made it possible to sample orders from different points across the range of measures, and so produce examples of orders from the extremes of each of the measures

8.5.1 Methods

Stimuli

This experiment used the same stimuli as in Experiment 1. Stimuli consisted of segmented versions of 40 compositions. Compositions belonged to either a 'monophonic' or 'polyphonic' (of two voices) texture category, and to either a 'short' or 'long' length category (of approximately 2 or 4 minutes, respectively). The 20 monophonic compositions were selected from a large corpus of western-classical tonal melodies; the 20 polyphonic melodies consisted of western-classical instrumental duets.

Stimuli used the same segmentations as in Experiment 1. Possible appropriate boundaries for segmentation were identified for each composition. Stimuli contained a mean of 8.20 segments per 'short' composition (SD = 0.95) and 12.65 segments per 'long' composition (SD = 1.60). 'Short' segments had a mean duration of 15.09 seconds (SD = 4.83; of which, 'monophonic' M = 14.95, SD = 4.85, and 'polyphonic' M = 15.24, SD = 4.60). 'Long' segments had a mean duration of 18.89 seconds (SD = 5.56; of which, 'monophonic' M = 18.94, SD = 5.98, and 'polyphonic' M = 18.84, SD = 5.14).

Statistical analysis

For the first analysis of this experiment, segments were examined in the orders they occurred in the original compositions. Values for each measure of these orders were calculated, with measures standardised to the means and standard deviations for each composition from the Monte Carlo set of randomised orders. For measures dependant on a specific representation of music, both pitch interval and inter-onset interval variants were analysed separately.

As with the analysis of participant orders in Experiment 1, first, differences between measures for orders belonging to compositions of different categories were examined. T-tests were used to test for significant differences between measures for stimuli in the 'monophonic' and 'polyphonic' categories, and between 'short' and 'long' categories. Likewise, to determine the effects of individual measures on the original orders, one-sample t-tests were used to test the significance of distances between each measure and the means of the Monte Carlo set of orders. Correlations between measures when calculated using the original composition orders were tested.

The second analysis in this experiment used the Monte Carlo set of randomised orders—for each composition, the lower of either the exhaustive permutations of all segments, or 20,000 randomly generated orders. Four example compositions were selected, covering each of the possible category combinations. Examples for high and low scoring orders were selected as the minimum

Table 8.9: T-Test Comparisons Between 'Monophonic' and 'Polyphonic' Stimulus Categories for Each Measure, Using Original Orders

Measure	Monophonic		Polyphonic				
	M	*SD*	*M*	*SD*	*df*	*t*	*p*
	Pitch interval						
Internal unpredictability (perfect)	0.07	1.19	−0.26	0.87	34.83	1.00	.32
Internal unpredictability (buffer)	−0.22	1.60	−0.13	1.35	36.94	−0.19	.85
Internal unpredictability flow	−0.20	1.27	−0.35	1.25	37.99	0.38	.71
Stylistic unpredictability flow	0.25	1.33	0.24	1.13	37.05	0.02	.98
Variation	−0.23	0.76	−0.26	0.98	35.77	0.14	.89
Global distance	0.11	0.92	0.28	0.66	34.32	−0.66	.51
	Inter-onset interval						
Internal unpredictability (perfect)	0.27	0.85	−0.02	0.92	37.76	1.03	.31
Internal unpredictability (buffer)	−0.36	1.22	0.11	1.37	37.50	−1.14	.26
Internal unpredictability flow	0.15	0.92	−0.07	1.29	34.23	0.61	.54
Stylistic unpredictability flow	0.10	1.08	−0.23	1.18	37.75	0.95	.35
Variation	−0.13	0.79	−0.52	1.12	34.00	1.29	.20
Global distance	−0.02	0.95	0.11	0.85	37.55	−0.48	.64
	Other						
Tonal distance	−0.69	1.24	−0.12	0.81	32.63	−1.71	.10
Edit distance	−0.27	1.13	−0.24	1.12	37.99	−0.10	.92

and maximum orders within the Monte Carlo set for each measure, for both representations.

8.5.2 Results

Original orders

There were 40 compositions with segments in their original orders used in this analysis.

As with the behavioural experiment, for each measure, for each representation, independent t-tests were used to test for differences in the standardised measures between compositions in texture and length categories. As shown in Table 8.9, no measure was found to be significantly different for original orders between 'monophonic' and 'polyphonic' categories. Likewise, when comparing 'short' and 'long' composition categories, no measure was found to be significant—shown in Table 8.10. A non-significant effect of *global distance* was shown when using pitch interval.

For each measure, for each representation, t-tests were used to test for significant differences between measure values for original orders and those of the large randomly generated set. The outcomes of these tests are displayed in Table 8.11. A sole significant effect was found for the measure of *tonal distance*—modelling the extent to which an order used more distantly related keys between sections. Key relations were found to be significantly closer between the segments of the original order than would be expected by chance orderings.

Table 8.10: T-Test Comparisons Between 'Short' and 'Long' Stimulus Categories for Each Measure, Using Original Orders

Measure	Short		Long				
	M	*SD*	*M*	*SD*	*df*	*t*	*p*
	Pitch interval						
Internal unpredictability (perfect)	−0.26	1.00	0.07	1.08	37.83	−0.98	.33
Internal unpredictability (buffer)	−0.36	1.42	0.00	1.52	37.80	−0.79	.44
Internal unpredictability flow	−0.04	1.08	−0.51	1.39	35.81	1.18	.24
Stylistic unpredictability flow	−0.08	1.37	0.57	0.97	34.13	−1.73	.09
Variation	−0.40	0.82	−0.09	0.90	37.65	−1.13	.27
Global distance	0.43	0.81	−0.04	0.72	37.50	1.91	.06
	Inter-onset interval						
Internal unpredictability (perfect)	0.02	0.91	0.22	0.87	37.89	−0.72	.47
Internal unpredictability (buffer)	−0.19	1.24	−0.06	1.39	37.53	−0.31	.76
Internal unpredictability flow	0.26	1.28	−0.18	0.88	33.73	1.27	.21
Stylistic unpredictability flow	−0.16	1.17	0.03	1.11	37.89	−0.51	.61
Variation	−0.51	1.08	−0.14	0.85	35.98	−1.20	.24
Global distance	0.26	0.96	−0.17	0.79	36.72	1.53	.13
	Other						
Tonal distance	−0.45	0.80	−0.37	1.31	31.58	−0.21	.83
Edit distance	−0.03	1.02	−0.48	1.17	37.30	1.28	.21

The *variation* measure, for both pitch interval and inter-onset interval representations was the only other to approach significance for the original orders; this non-significant effect showed *variation* lower for original orders than those of chance.

The extent to which measures for original orders were correlated with each other is shown in Table 8.12.

Example orders

The large number of orders generated for Monte Carlo approach used in these experiments provided a comprehensive set of orders across the full ranges of each measure, from which useful examples could be sampled. Four compositions were selected from the available stimuli, one for each category combination: (A) 'monophonic' and 'short', taken from Handel, Recorder Sonata in A minor, HWV 362, fourth movement; (B) 'monophonic' and 'long', taken from Beethoven, Piano Sonata No. 15, Op. 28, second movement; (C) 'polyphonic' and 'short', the fourth duo from Glière, 12 Duos for 2 Violins, Op. 49; and (D) 'polyphonic' and 'long', taken from Mozart, Duo for Violin and Viola, K.423, first movement. For the purpose of interpreting orders given for these compositions, scores of the segmented compositions are given in Appendix C.

The orders from the Monte Carlo set that produced the minimum and maximum values were found for each measure, excluding those of closeness to original orders and those that only compare starting and ending segments. These example orders are given in Table 8.13.

Table 8.11: T-Tests for Measures Testing Distance of Original Orders From Chance

Measure	*M*	*SD*	*df*	*t*	*d*	*p*	
Pitch interval							
Internal unpredictability (perfect)	−0.09	1.04	39	−0.58	−.09	.57	
Internal unpredictability (buffer)	−0.18	1.46	39	−0.78	−.12	.44	
Internal unpredictability flow	−0.28	1.25	39	−1.40	−.22	.17	
Stylistic unpredictability flow	0.25	1.21	39	1.28	.20	.21	
Variation	−0.26	0.86	39	−1.95	−.28	.06	
Global distance	0.20	0.79	39	1.57	.25	.12	
Inter-onset interval							
Internal unpredictability (perfect)	0.12	0.88	39	0.87	.14	.39	
Internal unpredictability (buffer)	−0.13	1.30	39	−0.61	−.10	.54	
Internal unpredictability flow	0.04	1.11	39	0.21	.03	.83	
Stylistic unpredictability flow	−0.07	1.13	39	−0.37	−.06	.71	
Variation	−0.31	.03	39	−1.90	−.33	.07	
Global distance	0.05	0.90	39	0.32	.05	.75	
Other							
Tonal distance	−0.41	1.07	39	−2.41	−.38	.02	•
Global tonality	−0.25	1.11	39	−1.44	−.23	.16	

• $p < .05$

Qualitatively, several general trends can be seen. For the three measures of *internal unpredictability*, elements of inter-measure consistency can be seen. Patterns of small subsequences are repeated between orders for each measure, for example, in composition A between minimum pitch interval *internal unpredictability (buffer)* and inter-onset interval *internal unpredictability (perfect)*. This pattern becomes closer when considering segments that may similar-enough so as to be freely interchangeable with each other. Due to this effect of segment interchangeability, it can be seen that measures using inter-onset interval bear a closer resemblance to each other than those of pitch interval. While none of these example compositions are rhythmically-isochronous, this effect is still most likely due to the reduced variation between segments in this representation.

For these measures of *internal unpredictability*, this close similarity between segments can be seen to influence example orders in another way; they show a marked tendency to group together segments that are highly-similar (or in many cases identical) early in the ordering. This effect can be seen most strongly in orders for composition D, in which segment 10, the exact repeat of the opening segment is placed directly after it for both pitch and rhythm variants of the first two measures. This can be seen as a direct consequence of the minimisation of *internal unpredictability*; repeating material before the model's training material has diversified will minimise the possible unpredictability overall. It is also particularly interesting to note that this behaviour seems to be closely analogous to that of the 'early repetition' identified as a compositional strategy by

Table 8.12: Pearson's *r* Correlations Between Experiment Measures for Original Orders

	Measure	1	2	3	4	5	6	7	8	9	10	11	12	13	14
						Pitch interval									
1	*Internal unpredictability (perfect)*	–													
2	*Internal unpredictability (buffer)*	.28	–												
3	*Internal unpredictability flow*	−.16	−.40 ••	–											
4	*Stylistic unpredictability flow*	−.03	.33 •	−.24	–										
5	*Variation*	.07	.08	.05	.17	–									
6	*Global distance*	−.36 •	−.30	.15	−.19	−.29	–								
						Inter-onset interval									
7	*Internal unpredictability (perfect)*	.16	−.01	.17	−.05	.17	−.07	–							
8	*Internal unpredictability (buffer)*	.04	.80 •••	−.21	.24	.09	−.08	.00	–						
9	*Internal unpredictability flow*	.08	−.19	.49 ••	−.03	.22	−.00	.12	.00	–					
10	*Stylistic unpredictability flow*	.05	.39 •	.03	.32 •	−.11	−.02	.01	.54 •••	.42 ••	–				
11	*Variation*	.05	.17	−.13	.19	.36 •	−.15	.10	.10	.08	.24	–			
12	*Global distance*	−.18	−.12	.34 •	−.20	−.10	.45 ••	−.16	−.08	.11	.03	−.20	–		
						Other									
13	*Tonal distance*	−.12	−.19	−.07	−.05	−.08	.13	.00	−.12	.15	−.03	−.00	.08	–	
14	*Global tonality*	−.30	−.34 •	.18	.01	−.11	.19	.17	−.32 •	−.09	−.22	−.33 •	.17	−.01	–

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

Table 8.13: Minimum and Maximum Randomly Generated Orders for Each Measure, for Both Representations, for Four Example Compositions

Stimulus	Measure	Minimum		Maximum	
		Order	Value	Order	Value
		Pitch interval			
A	*Internal unpredictability (perfect)*	1, 5, 2, 4, 7, 3, 9, 6, 8	−2.82	1, 8, 3, 5, 6, 2, 7, 4, 9	2.71
	Internal unpredictability (buffer)	1, 5, 4, 2, 6, 3, 7, 8, 9	−3.53	1, 5, 7, 4, 8, 3, 9, 2, 6	2.35
	Internal unpredictability flow	1, 8, 2, 7, 9, 6, 5, 4, 3	−3.15	1, 4, 2, 5, 7, 6, 8, 3, 9	2.90
	Stylistic unpredictability flow	1, 4, 5, 2, 8, 7, 3, 6, 9	−3.56	1, 7, 5, 3, 8, 9, 2, 6, 4	2.05
	Variation	1, 6, 4, 2, 7, 5, 8, 9, 3	−0.99	1, 8, 5, 2, 7, 9, 3, 4, 6	1.94
B	*Internal unpredictability (perfect)*	1, 3, 11, 2, 9, 4, 12, 8, 5, 6, 13, 7, 10	−3.68	1, 6, 10, 4, 8, 12, 7, 3, 11, 2, 13, 9, 5	2.87
	Internal unpredictability (buffer)	1, 2, 9, 12, 3, 11, 4, 13, 5, 6, 10, 8, 7	−4.89	1, 11, 7, 4, 12, 8, 3, 10, 5, 9, 13, 6, 2	1.96
	Internal unpredictability flow	1, 6, 10, 4, 13, 8, 12, 7, 3, 5, 11, 2, 9	−3.03	1, 9, 11, 3, 13, 2, 5, 4, 8, 6, 10, 7, 12	3.96
	Stylistic unpredictability flow	1, 9, 5, 6, 2, 13, 7, 8, 10, 11, 3, 4, 12	−3.78	1, 4, 6, 12, 5, 10, 2, 9, 13, 11, 8, 3, 7	2.49
	Variation	1, 9, 7, 12, 2, 5, 3, 10, 11, 6, 8, 4, 13	−1.80	1, 10, 13, 12, 6, 8, 4, 5, 11, 2, 7, 3, 9	2.69
C	*Internal unpredictability (perfect)*	1, 8, 5, 6, 9, 4, 7, 2, 3	−2.44	1, 2, 3, 7, 4, 6, 5, 9, 8	2.01
	Internal unpredictability (buffer)	1, 3, 8, 9, 2, 7, 4, 5, 6	−3.13	1, 4, 5, 3, 9, 6, 2, 8, 7	2.06
	Internal unpredictability flow	1, 2, 7, 4, 5, 3, 9, 6, 8	−2.72	1, 8, 9, 3, 7, 2, 5, 6, 4	2.77
	Stylistic unpredictability flow	1, 8, 3, 7, 6, 9, 5, 4, 2	−3.91	1, 2, 3, 9, 8, 6, 7, 4, 5	1.83
	Variation	1, 9, 2, 3, 8, 4, 5, 6, 7	−2.13	1, 6, 7, 4, 2, 8, 9, 3, 5	2.25
D	*Internal unpredictability (perfect)*	1, 10, 3, 9, 12, 11, 5, 2, 6, 7, 14, 13, 4, 8, 15	−2.68	1, 6, 7, 15, 8, 4, 2, 14, 11, 5, 13, 9, 3, 10, 12	2.55
	Internal unpredictability (buffer)	1, 10, 7, 9, 13, 4, 6, 14, 5, 15, 8, 3, 12, 11, 2	−5.95	1, 13, 12, 2, 6, 9, 10, 7, 5, 8, 11, 4, 15, 14, 3	2.06
	Internal unpredictability flow	1, 6, 14, 7, 8, 4, 15, 9, 12, 11, 5, 13, 10, 3, 2	−3.54	1, 12, 3, 2, 11, 15, 14, 9, 10, 7, 13, 4, 6, 5, 8	3.40
	Stylistic unpredictability flow	1, 10, 8, 6, 7, 5, 14, 2, 15, 3, 9, 12, 4, 13, 11	−3.83	1, 3, 8, 9, 2, 11, 7, 14, 15, 13, 6, 12, 10, 4, 5	2.64
	Variation	1, 5, 4, 8, 12, 15, 14, 13, 11, 2, 9, 6, 7, 3, 10	−2.11	1, 7, 9, 10, 6, 14, 3, 2, 8, 11, 15, 13, 5, 12, 4	2.42

Table 8.2: *Cont.*

		Inter-onset interval			
A	*Internal unpredictability (perfect)*	1, 5, 4, 2, 7, 6, 3, 8, 9	−3.52	1, 9, 6, 5, 8, 3, 7, 4, 2	2.25
	Internal unpredictability (buffer)	1, 4, 7, 2, 3, 5, 6, 8, 9	−3.29	1, 9, 5, 7, 8, 2, 6, 4, 3	2.44
	Internal unpredictability flow	1, 2, 8, 3, 9, 5, 4, 6, 7	−2.53	1, 4, 5, 2, 3, 6, 8, 7, 9	2.82
	Stylistic unpredictability flow	1, 4, 9, 8, 7, 3, 6, 2, 5	−3.59	1, 2, 8, 3, 9, 5, 4, 6, 7	1.96
	Variation	1, 3, 4, 9, 6, 5, 8, 7, 2	−1.30	1, 2, 9, 7, 4, 5, 3, 8, 6	1.80
B	*Internal unpredictability (perfect)*	1, 8, 13, 9, 6, 7, 5, 2, 12, 11, 3, 4, 10	−3.56	1, 10, 3, 2, 13, 9, 6, 11, 4, 5, 8, 7, 12	2.28
	Internal unpredictability (buffer)	1, 2, 9, 12, 3, 11, 4, 13, 5, 6, 10, 8, 7	−4.08	1, 5, 3, 6, 11, 13, 12, 2, 7, 10, 9, 8, 4	2.22
	Internal unpredictability flow	1, 4, 13, 2, 7, 9, 3, 11, 5, 10, 12, 8, 6	−3.32	1, 10, 5, 7, 2, 6, 4, 12, 3, 9, 8, 13, 11	2.33
	Stylistic unpredictability flow	1, 11, 3, 9, 4, 2, 6, 8, 10, 12, 7, 5, 13	−3.41	1, 7, 4, 6, 9, 12, 2, 10, 11, 8, 3, 5, 13	2.82
	Variation	1, 3, 13, 10, 12, 4, 11, 5, 2, 7, 9, 8, 6	−2.41	1, 2, 8, 7, 13, 12, 3, 9, 6, 5, 4, 11, 10	1.53
C	*Internal unpredictability (perfect)*	1, 5, 3, 8, 6, 9, 2, 4, 7	−2.78	1, 7, 4, 2, 6, 9, 3, 5, 8	2.59
	Internal unpredictability (buffer)	1, 5, 8, 6, 3, 7, 4, 9, 2	−3.17	1, 6, 7, 8, 9, 5, 2, 4, 3	2.88
	Internal unpredictability flow	1, 7, 4, 9, 6, 2, 5, 8, 3	−3.34	1, 8, 5, 3, 7, 6, 9, 2, 4	2.52
	Stylistic unpredictability flow	1, 8, 3, 5, 6, 9, 2, 7, 4	−3.20	1, 6, 7, 5, 4, 9, 2, 3, 8	1.94
	Variation	1, 3, 4, 9, 2, 5, 6, 7, 8	−1.92	1, 6, 7, 8, 2, 9, 4, 5, 3	2.11
D	*Internal unpredictability (perfect)*	1, 10, 8, 4, 9, 12, 3, 7, 11, 13, 6, 5, 2, 14, 15	−3.43	1, 14, 6, 4, 15, 2, 3, 7, 10, 5, 8, 9, 13, 11, 12	3.77
	Internal unpredictability (buffer)	1, 10, 5, 15, 3, 14, 8, 9, 6, 11, 2, 13, 4, 12, 7	−5.15	1, 9, 10, 11, 3, 13, 6, 12, 4, 5, 14, 7, 2, 15, 8	2.77
	Internal unpredictability flow	1, 6, 8, 15, 9, 13, 11, 5, 14, 4, 10, 2, 7, 3, 12	−4.22	1, 10, 4, 14, 2, 11, 15, 3, 8, 5, 12, 9, 7, 6, 13	3.40
	Stylistic unpredictability flow	1, 10, 6, 8, 11, 2, 4, 15, 7, 14, 12, 3, 13, 9, 5	−3.73	1, 11, 6, 14, 2, 12, 10, 5, 8, 3, 15, 13, 7, 9, 4	2.52
	Variation	1, 15, 6, 7, 3, 4, 14, 12, 5, 2, 11, 13, 8, 10, 9	−1.90	1, 12, 7, 13, 8, 4, 11, 14, 10, 9, 5, 6, 15, 3, 2	2.45
		Other			
A	*Tonal distance*	1, 2, 6, 8, 4, 9, 5, 3, 7	−2.31	1, 3, 5, 9, 4, 6, 2, 7, 8	2.32
B	*Tonal distance*	1, 13, 6, 8, 5, 7, 9, 3, 11, 10, 4, 2, 12	−4.06	1, 4, 6, 12, 5, 13, 11, 9, 7, 10, 3, 8, 2	2.71
C	*Tonal distance*	1, 4, 3, 8, 9, 7, 5, 2, 6	−2.82	1, 8, 9, 4, 5, 3, 7, 6, 2	3.02
D	*Tonal distance*	1, 10, 15, 3, 2, 5, 9, 4, 11, 14, 8, 13, 12, 6, 7	−4.60	1, 11, 7, 3, 2, 8, 4, 14, 15, 5, 6, 13, 9, 12, 10	2.70

Huron (2013). The converse of this effect can also be observed, though somewhat harder to distinguish, in the maximum-inducing orders for these measures. For example, in composition B, maximum orders of *internal unpredictability* measures jump between different thematic ideas at each segment.

Although fragments of the original sequential orders can be seen, it is clear that none of these measures, at either extreme, are substantially recreating the originals—that the originals are not best performing in any of these measures.

8.6 Discussion

The two experiments presented in this chapter together provide evidence of the importance of large-scale order in music. More specifically, the results also provide evidence that perception of the way in which sections of musical material are ordered is influenced by the dynamic learning of statistical properties within compositions. Both experiments tested whether computational measures of thematic structure, covering both pitch and rhythm, influence the ordering of musical material. Experiment 1 did so using orders provided by participants in a behavioural study; Experiment 2 did so by examining original and theoretically optimal orders.

The results of Experiment 1 provided significant evidence that listeners are sensitive to the large-scale ordering of musical material. One way that this was shown is through the ability of participants to reconstruct the original orders of segments significantly better than chance, as characterised by both the measures of *dissimilarity* between segments in corresponding positions (for both pitch and rhythm representations) and of the overall *edit distance* between orders. The joint significance of these measures makes this finding a robust one. The closeness of participant orders to originals existed both in the absolute positioning of segments within the order, and in the ability to correctly match smaller subsequences of the originals.

However, the sensitivity to original orderings was not the only significant influence on the perception of order supported by the findings of this experiment. The results for the other modelling measures tested showed effects on orderings that were independent of any relationship with the segment order present in the original compositions. In particular, the values for participant orders in the measures of *internal unpredictability (perfect)*, *internal unpredictability (buffer)*, *internal unpredictability flow*, *stylistic unpredictability flow*, and *global tonality* were all significantly lower than would be expected by chance (*i.e.*, participant orders were more predictable, or more closely related) for least one representation, as hypothesised for these measures. These effects were not significantly different between compositions that lasted around 2 minutes and those lasting around 4 minutes.

This suggested independence between measures of closeness to original orders and other significant measures was first indicated in the analysis of correlations between measures (see Table 8.5), with no significant correlations to otherwise significant measures found. Another small indication that participants' orders were partly motivated by features independent of original orders is that no significant effect for these measures was found in the analysis of original orders, however, with the caveat that tests on original orders are limited to a much smaller sample size than those of the behavioural experiment.

This separation was also shown in Experiment 1 when using PCA to combine measures into a smaller number of orthogonal components. *Edit distance* was found to be highly independent from other measures; *dissimilarity*, while differing in amounts of interaction with other measures between representations, was still mostly independent. Aside from these two measures, the PCAs suggest that there are several orthogonal factors present in participants' orders—particularly in the pitch interval representation (the limitations that arise with inter-onset interval are discussed below). The first of which being an effect of *internal unpredictability* (both *perfect* and *buffer* variants).

These findings provide strong evidence that statistical learning has an important role in the perception of large-scale order, as hypothesised. The three measures that modelled possible effects of intra-opus learning—the measures of *internal unpredictability (perfect)*, *internal unpredictability (buffer)*, *internal unpredictability flow*—were found to be significantly lower than chance for participant orders when using the pitch interval representation; *internal unpredictability (buffer)* was found to be significantly lower than chance when using the inter-onset interval representation. The results also suggest that there was a significant stylistic component, providing some evidence that inter-opus statistical learning plays an additional role in the perception or order. These findings suggest that participants were, at least in part, assembling their orders so as to minimise unpredictability—both of the whole composition, and on a segment-to-segment basis.

Comparison between the performances of *internal unpredictability (perfect)* and *internal unpredictability (buffer)* measures can provide some insight into the role of memory in the perception of structural elements on these larger times-scales. As can be seen in the optimisation process finding the buffer duration (Figure 8.2), restricting the memory of the model can greatly increase the amount of variance that can be captured between orders; subsequently, the buffer-limited version out-performs the perfect-memory version in explaining variance in participants' orders (see Table 8.4). The two measures are, however, still highly related. These results suggest that, for intra-opus statistical to allow listeners to perceive the effects of large-scale orders, only a sensitivity to more local statistical regularities may be needed. It is interesting to note that the op-

timal buffer value for the stimuli used in these experiments of 30 seconds is a little under double that of the mean segment duration (17 seconds).

The effect of tonal properties presents one of the clearest differences between those orders chosen by participants and those of the composers' originals. *Tonal distance* was the only measure, out of all those tested, that was significantly lower than chance for original orders; furthermore, it was found not to be significantly different from chance for participant orders. Interestingly, this result would suggest that, while the composers used in this set of tonal music largely followed the convention of modulation to related keys, participants were not as sensitive to this convention.

However, it is not necessarily the case that participants made no consideration to tonal features at all. The behavioural experiment does provide evidence that participants attached some importance to *global tonality*—that the first and last segments were in closely-related keys—that persisted even in the 'long' stimuli category. The fact that a similar effect of this measure is lacking for the original orders may be due to the cropping of works to meet the 2 and 4 minute categories. A possible explanation for the discrepancy between *tonal distance* and *global tonality* for participants is that the preference for low *internal unpredictability*, as discussed above, obscured ordering based on segment-by-segment tonal relationships, but that participants' sensitivity to tonality was just strong enough to influence the single global relationship.

As so far discussed, the results of the behavioural experiment show participant orders had both significant effects of measures of closeness to original orders and, independently, measures of statistically learned features. However, the apparent lack of a significant relationship between any of the significant measures and the measures of *dissimilarity* and *edit distance* (with no correlation between measures, and little shared explained variance in the PCA) poses a question as to the features that allowed participants to continue to achieve this closeness to original orders. One possible explanation is that, despite the steps taken to discourage such an approach to the task, the matching of very local continuities between segment ends and starts still had some influence over participants' orders. However, were this to be the case, it cannot account for the significant impact of the measures that were independent of the original orders, so would have limited effect. Alternatively, it is possible that many of the significant measures contributed in a small way to partially recreate originals—of too weak an effect to have been significant with the tests used. There is an argument for this scenario in that *dissimilarity* was found not to be purely orthogonal to all other dimensions in the PCAs, however, *edit distance* largely was so. The final possibility is that, due to the high complexity of music as a stimulus, participants were using other features that are beyond the scope of our modelling here.

It is clear from the results for these experiments that the pitch and rhythmic domains, while similar in many aspects, do present some significant differences. This is most noticeable where measures that are significant when modelled using pitch interval are not so when using inter-onset interval. These differences are largely restricted to measures of *internal unpredictability—internal unpredictability (perfect)* and *internal unpredictability flow*, in particular. As discussed in the previous experiments of this thesis, a likely cause of this disparity is the relatively high frequency of rhythmically-isochronous compositions in western music. Specifically, for the current experiments, little variance in rhythmic material is possible between different orders of segments from compositions that are highly (or even partly) rhythmically isochronous. The more general point is that pitch structure is the dominant carrier of thematic structure in the musical styles included here. This has the effect of reducing the abilities of measures to distinguish between orders when modelled using inter-onset interval.

In order to replicate the link between *internal unpredictability* and perceived structural unity found in Chapter 6, the behavioural experiment of this chapter included a unity rating task for participants' orders. However, no significant effect was found between ratings and any of the modelled measures—including those of *internal unpredictability*. There are several possible explanations as to why this effect was not replicated. Firstly, the unity task in this experiment is restricted by the requirements needed for the ordering task—such as requiring far more time for each trial, reducing the overall number of responses that could be gathered—it is possible it is lacking sufficient statistical power for any effect to become apparent. Secondly, it could be that the cognitive, problem-solving nature of the ordering task disrupted the more immediate perception of unity when ratings were subsequently taken. Finally, it is possible that the task had become too far removed from that of the experiment of Chapter 6, such that participants' abilities to judge unity were too disrupted by the fragmented nature of the task. This may also be compounded were participants to treat the rating—despite the instructions—as a confidence rating in their given orders.

The Gold-MSI scores of participants in the first experiment provide some limited evidence of an influence of musical background on orderings. Two measures were found to be significantly correlated with Gold-MSI scores—*internal unpredictability (buffer)* and *dissimilarity*. For the first, the findings suggest that participants with greater *perceptual abilities* or *musical training* returned orders that were closer to chance for that measure (for pitch interval in inter-onset interval, respectively); this would seem to imply that the participants who score more highly on these scores were prioritising other methods of ordering. These participants may be more sensitive to segments' original orders. A significantly negative correlation was found between inter-onset inter-

val *dissimilarity* and scores of *general sophistication* and *musical training*; orders given by participants with higher scores were closer to original segment orders than those of participants scoring lower in these categories. No significant effect was found between Gold-MSI scores and unity ratings. No evidence was found to support the hypothesised connection between musical background and stylistic features of music.

8.7 Summary

In this chapter, the final experiments of this thesis were presented. These experiments aimed to investigate properties of the large-scale ordering of material within a composition. The first experiment observed the responses of participants using a puzzle paradigm previously employed by Granot and Jacoby (2011), participants were tasked with ordering segmented compositions in the order they judged to produce the most coherent composition. A Monte Carlo approach was used to test for significant differences between participants' orders and those of chance, based on measures derived from those of the model of Chapter 3, extensions of these measures to cover effects of memory limitation and the optimal information flow, closeness of orders to those of the original compositions, and considerations of tonal structure. Compositions used in this experiment differed in two category conditions. Compositions were evenly split between monophonic and two-voice polyphonic categories (using the techniques for the dynamic modelling of polyphonic discussed in Chapter 3), and between 'short' and 'long' categories (approximately 2 and 4 minutes, respectively). The second experiment applied to same procedure to original composition orders and explored example orders from the extremes of each measure.

The results of the behavioural experiment provided significant evidence that listeners are sensitive to the large-scale ordering of musical material, producing support for the importance of several measures. Firstly, participant orders were found to be significantly closer to those of originals than would occur by chance (measures of *dissimilarity* in both representations and *edit distance*). Secondly, measures of *internal unpredictability (perfect)*, *internal unpredictability (buffer)*, *internal unpredictability flow*, *stylistic unpredictability flow*, and *global tonality* were all significantly lower than chance orders in at least one representation. The findings of this chapter contribute to the converging evidence across this thesis in support of the perception of elements of thematic structure through the statistical learning of compositional regularities.

Chapter 9

Conclusions

9.1 Overview

The primary goal of this thesis was to develop an improved understanding of how large-scale thematic structures—the organisation of material within a musical composition—can be perceived. Whether, and if so, how, these structures may be perceived provides an interesting psychological problem, combining many aspects of memory, pattern recognition, and similarity judgement. However, strong experimental evidence supporting the perception of large-scale thematic structures (reviewed in Chapter 2) remains limited, often arising from difficulties in measuring and disrupting their perception.

In order to address some of the shortcomings identified within the literature, Chapter 3 sought to provide a concrete specification of the cognitive processes involved in the perception of large-scale thematic structure. The chapter described a probabilistic computational model of the perception of thematic structure, based on the hypothesis that thematic structures can be perceived through the statistical regularities they form over the course of a composition. This chapter introduced IDyOM (Pearce, 2005), a computational model of auditory expectation, which provided a platform for the modelling of thematic structure, then summarised the main components of the model itself—the processes of theme and repetition detection—and, finally, presented four model-based measures to characterise properties of thematic structure for any given composition: (1) the intra-opus *internal unpredictability* of a composition; (2) *thematic repetition*—the proportion of material in a composition identified by the model as repetition of identified themes; (3) *thematic variation*—the extent to which this repetition develops from its parent themes; and (3) the *stylistic unpredictability* of this identified material. Given this foundation, the remainder of the research reported in this thesis was then concerned with the empirical evaluation of this model, and the overarching hypothesis on which it was based (Table 3.1 gave an overview of the relationship between the four primary model

measures and those used in the behavioural experiments).

Chapter 4 provided the first step in this evaluation, presenting a corpus analysis based on the measures of the model. A corpus of 623 complete monophonic compositions was created and described in this chapter—not only for use in the analysis, but also as a basis for stimulus selection in the subsequent behavioural experiments. The analysis aimed to demonstrate that the measures of hypothesised importance to the perception of thematic structure varied systematically when applied to the corpus, reflecting the inherent variation of structure present within the corpus itself. Through this analysis, the common output patterns, ranges, and relationships between measures were explored, with the four model measures of importance found to be able to substantially explain variance within the corpus. In addition to comparing measures, the corpus analysis also served to identify the representations of musical surface best-suited to modelling features of thematic structure in the pitch and rhythmic domains; representations of pitch interval and inter-onset interval, respectively, were identified and subsequently using in the modelling of the following behavioural experiments.

The first behavioural experiment, presented in Chapter 5, provided some preliminary validation of concepts from the model of Chapter 3. This experiment was the only one reported in this thesis not to test measures directly on large-scale compositions; it was important to examine first whether measures could predict perception on a local level in this initial experiment, before extending the enquiry to large-scale thematic structures in the subsequent experiments. The experiment tested the abilities of participants in perceiving a relationship between pairs of themes and repetition phrases, identified by the model within the monophonic compositions of Chapter 4's corpus. This task aimed to understand how modelled features interacted and influenced responses in isolation as an important first step to understanding their effects when integrated into compositions over far longer timespans. Measures were found to correspond significantly with participants' responses in both pitch and rhythm domains; this effect was found to exist primarily in intra-opus features, but also in the stylistic relationship between pairs.

Chapter 6 presented the second behavioural experiment of the thesis, testing the perception of two highly-important indicators of large-scale thematic structure; the experiment tested participants' abilities to identify given moments as being repetitions of earlier thematic material and their judgements of the sense of structural unity of compositions. As this experiment tested for these features in monophonic compositions of substantial length (with durations of approximately 2 minutes), the measures of thematic structure were evaluated in the same form as they were hypothesised to operate in real-world musical listening. The findings of this experiment presented a striking contrast

between the influence of the four measures. Across both tasks, the results of this experiment provided strong evidence of an influence of *internal unpredictability*, across both pitch and rhythm representations. *Internal unpredictability*, both of a given moment and of the composition preceding it, was found to aid the correct identification of repeated material, and compositions with lower *internal unpredictability* were found to have a significantly stronger sense of unity. These findings provided some corroboration of the overarching hypothesis of this thesis—that thematic structures can be perceptible through the structural regularities they form.

The influence of statistical features on the perception of large structures was further tested in the experiment presented in Chapter 7. This experiment aimed to test listeners' judgements of large-scale continuation, as a means of examining the importance and nature of statistically learned features in this task. For stimulus compositions, plausible continuations were computationally generated, based on predictability from the composition, its themes, or from the corpus of Chapter 4, using both pitch interval and inter-onset interval representations. Measures based on those of Chapter 3 were compared to participants' responses in this task. The findings of this experiment provided robust evidence that intra-opus statistically-learned elements had an important role in influencing the perception of these continuations. This experiment expanded on the scope of the experiment of Chapter 6 in two ways: it increased the duration of stimulus compositions used (and so the scale of possible structures), using compositions both of 2 and 4 minutes long; and it introduced and tested techniques (discussed in Chapter 3) for the modelling of thematic structure in polyphonic compositions of two voices.

The final behavioural experiment was presented in Chapter 8. This experiment aimed to contribute to the converging understanding of thematic structure built-up throughout the course of these experiments by testing the perception of large-scale order. Using a puzzle paradigm previously employed by Granot and Jacoby (2011), participants were tasked with ordering segmented compositions in the order they judged to produce the most coherent composition. As with the stimuli of the previous experiment, both monophonic and polyphonic compositions were used, with complete compositions of 2 and 4 minutes in length. A Monte Carlo approach was used to test the significance of participants' orders based on measures derived from those of the model of Chapter 3, alongside additional measures characterising tonality and closeness to original orderings. The results of this experiment provided significant evidence that listeners are sensitive to the large-scale ordering of musical material. Participants' orderings were found both to be closer to the original orders than chance would suggest and favoured significantly lower *internal unpredictability* in both the pitch and rhythmic domains.

9.2 Research outcomes

9.2.1 Intra-opus statistical learning of thematic structure

Taken together, the findings of the behavioural experiments of Chapters 5, 6, 7 and 8, in their evaluation of the measures of thematic structure of Chapter 3, provide converging evidence that the perception of large-scale thematic structures can be accounted for by the dynamic learning of statistical regularities within musical compositions. Table 9.1 summaries the key findings for measures of thematic structure. All four experiments found that the *internal unpredictability* of material (or the experiment-specific measure that most-closely encapsulated its effects) consistently presented the strongest influence on their respective tasks. As listeners' perception of structure cannot be tested directly, each experiment focused on different experimental indicators that thematic structures were perceived; experiments tested the abilities of listeners to perceive structural unity, large-scale repetition, continuation, and large-scale order. The ability of *internal unpredictability* to have a significant effect on all of these indicators makes this finding a robust one.

Testing the perception of structural unity, Chapter 6 found strong evidence that the sense of structure unity of a composition is directly related to its *internal unpredictability*, with compositions that contain a greater amount of repetition, and so have higher intra-opus predictability, perceived as being more unified. For completeness, it should, however, be noted that the experiments of Chapters 7 and 8 were unable to replicate this finding though the collection of unity ratings after completion of their respective tasks. As discussed in these chapters, this lack of significance was most likely due to the prioritisation of the primary experimental tasks resulting in the unity rating task being too far removed from the original (with the additional possibility participants erroneously treated this rating as an expression of their confidence in their responses to the main task). This finding of the ability of listeners to perceive large-scale unity builds on those of Tan and Spackman (2005), Tan et al. (2006) and Lalitte and Bigand (2006).

Testing the perception of large-scale repetition, the other task in the experiment reported in Chapter 6 found strong evidence that listeners are able to perceive repetition within compositions over large time-scales—building on the findings of Margulis (2012). The experiment found effects of *internal unpredictability* for both pitch and rhythm, such that phrases that were predictable given a preceding musical context were more likely to be perceived as having been heard before, especially when the preceding context was itself predictable.

Analysis of the properties of continuations chosen by participants in the experiment of Chapter 7 showed that continuations that contributed to lowering a composition's overall *internal unpredictability* were favoured.

Table 9.1: Summary of Empirical Evidence Supporting Experiment Measures Used in This Thesis (based on Table 3.1).
Measures found to be significant shown in bold. P and R indicates the domain in which an effect was found; pitch interval and inter-onset interval, respectively.

Measure	Experiment 1 (Chapter 5)	Experiment 2 (Chapter 6)	Experiment 3 (Chapter 7)	Experiment 4 (Chapter 8)[a]
Internal unpredictability	***(Dissimilaity)***	***Internal unpredictability* (P,R)** Compositions with lower unpredictability were rated as more unified ***Internal unpredictability of moment* (P,R)** Phrases with lower unpredictability within a composition were more likely to be perceived as repetition *Internal unpredictability at moment*	***Composition unpredictability* (P,R)** Continuations with lower unpredictability, given the preceding composition, were more likely to be chosen ***Late-composition unpredictability* (P,R)** Continuations with lower unpredictability, given the second half of the composition, were more likely to be chosen *(Continuation unpredictability)*	***Internal unpredictability (perfect)* (R)** Participant orders were more internally unpredictable than chance orders ***Internal unpredictability (buffer)* (P,R)** Participant orders were more internally unpredictable than chance orders ***Internal unpredictability flow* (P)** Participants favoured orders with smaller differences in unpredictability between segments
Thematic repetition	–[b]	*Thematic repetition* *Thematic repetition at moment*	***Thematic unpredictability* (P)** Continuations that were more predictable, given a composition's themes, were preferred	–[b]
Thematic variation	***Dissimilarity* (P,R)** Pairs with low *dissimilarity* were rated as being more closely related	*Thematic variation* *Thematic variation at moment*	*(Thematic unpredictability)*	*(Variation)*
Stylistic unpredictability	***Stylistic difference* (R)** Pairs that were stylistically similar were rated as being more closely related ***Mean stylistic unpredictability* (R)** Pairs that were stylistically more predictable were rated as being more closely related	*Stylistic unpredictability* *Stylistic unpredictability at moment*	*Stylistic unpredictability*	***Stylistic unpredictability flow* (P,R)** Participants favoured orders with smaller differences in stylistic unpredictability between segments

[a] Experiment used additional measures not listed in this table
[b] Theme-detection modelling not applicable to the paradigms used in these experiments

Testing the sensitivity of listeners to the large-scale ordering of compositions, the behavioural experiment presented in Chapter 8 built on the findings of Granot and Jacoby (2011, 2012). This experiment demonstrated that the overall *internal unpredictability* of ordered compositions has a significant influence on the perception of structure—that participants aimed to produce orders that minimised *internal unpredictability*.

9.2.2 Towards a cognitive model of large-scale thematic structure

Given the empirical evidence gathered throughout the research presented in this thesis (see Table 9.1), the concept of a cognitive model of the perception of large-scale thematic structure can be reviewed. The evidence gives strong and robust support for a model based on the internal unpredictability of compositions (it would also suggest that processes of discrete theme and related repetition detection are of lesser importance) .

When we listen to music, our auditory sensory input contains patterns that recur both exactly and approximately at a range of different timescales. As we detect and recognise repeated patterns, our brains form schemas or mental representations of these patterns, based on dynamic online statistical learning. These schemas help us anticipate what will happen next in the music and enable us to make predictions based on our previous experiences with the composition.

As a part of this pattern detection and prediction process, musical structures made from the repetition of material give the brain an opportunity to process structured musical inputs as unified units. This may give the potential to simplify cognitive load, making it easier to process complex auditory information in music and discern yet larger structures.

Although not covered directly by this thesis, we can hypothesise how this model of thematic structure can be applied to hierarchical musical structures. Repetition, as it is often nested hierarchically, can allow for relationships to be perceived between different levels of structures. The same properties of structural unity based on internal predictability that were investigated over whole compositions can be applied to sections in a structural hierarchy, with high internal predictability indicating more unified sections.

9.2.3 Pitch and rhythm domains

In all four behavioural experiments reported in this thesis, there was a large amount agreement between measures using representations of pitch and rhythm. For measures found to have the strongest effects in each experiment, effects were often found when using either pitch interval or inter-onset interval

versions: in Chapter 5, *dissimilarity* in both predicted responses; in Chapter 6, the predictability of repetitions and the ability to predict unity ratings were significant for both representations; in Chapter 7, significant effects were found in both representations for measures based on predictability, given the context of the compositions; and, in Chapter 8, effects of closeness of orders to original versions, memory limited *internal unpredictability*, and style were consistent across representations.

However, outside of these cases, a substantial disparity was found in the ways in which pitch and rhythm influenced thematic structure. Broadly, these differences should be treated with some caution; as first identified in Chapter 4, and discussed throughout this thesis, these differences may likely be due to the ability of western-classical compositions to be rhythmically-isochronous—compositions consisting of notes of only a single duration (or pairs of durations, for example, in a swung pattern). The presence of such compositions may exaggerate the influence of certain features when testing using representations of rhythm through their maximum repetition and minimum variation. However, rhythmically isochronous compositions are relatively frequent in western music, suggesting such compositions should not be unfamiliar to listeners or disrupt their perceptual processing of thematic structure. Instead, it seems likely that the psychological representations involved in the cognitive process of perceiving thematic structure vary based on their contextual relevance, as has been shown in research investigating the perception of other musical features (Prince, 2014; Prince et al., 2009).

The nature of the stimuli across all of these experiments provides some insight into the relationship between pitch and rhythm representations over different timescales. The correlation between composition-trained (*i.e.*, intra-opus, non-stylistic) measures across representations suggests these properties become more closely correlated as the length of excerpt they are applied to increases—barring any effects of isochrony. The cause of this correlation may be due to the repetition characterised by these measures. Repeated material usually maintains at least some of its pitch and rhythmic content between repetitions; as longer excerpts can afford more repetition—and increased repetition increases predictability—when material is repeated, predictability in both domains increases together.

9.2.4 Musical background

There is one further way in which the four experiments were in agreement—little consistent evidence was found that supported listeners' musical backgrounds being able to influence performance in any of the tasks. For each behavioural experiment, alongside the listening task, participants gave their re-

sponses to questions from the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire for musical sophistication (Müllensiefen et al., 2014). Three scores from the questionnaire were used: (1) *general sophistication*; (2) *perceptual abilities*; and (3) *musical training*. Effects of scores were often found to be contradictory between experiments.

Participants with higher scores of *general sophistication* were found to be influenced more by stylistic differences in Chapter 5, more likely to show stronger negative effects of *thematic variation* on perception of unity in Chapter 6, and prioritised rankings similar to original orders to a lesser extent than other participants in Chapter 8. Participants with higher scores of *perceptual abilities* were found to possess a greater sensitive to *internal unpredictability* in perception of unity in Chapter 6, but conversely, provided orders in Chapter 8 that were closer to chance in memory limited *internal unpredictability* than other participants. Participants with higher scores of *musical training* also provided orders that were closer to chance for *internal unpredictability (buffer)* and prioritised rankings similar to original orders to a lesser extent than other participants in Chapter 8.

It should, however, be cautioned that testing the effects of musical background was not the primary aim of this research and, as such, participants were not recruited to capture a wide range of musical abilities. The conclusions that may be drawn are, therefore, limited without further targeted confirmatory research.

9.2.5 Computational modelling of music

Although the primary aim of the research presented in this thesis is to better understand processes of music perception and cognition, both the computational methods used and the empirical findings can offer some insight into the wider computational modelling of music. Several elements of the modelling used in this thesis have some overlap with common tasks in the field of music information retrieval, such as measures of similarity, repetition identification and structure analysis.

Even though the modelling of this thesis is based on the first time listening to music, and so models must operate linearly as a work develops in order to be function as models of cognition—a constraint that does not need to apply to MIR approaches—this work may be more widely applicable. In this case, we consider general (non-cognitive) approaches to form an idealised or 'perfect' understanding of music, as may be achieved by multiple hearings of a composition, perfect memory and musical training.

In this research, the similarly relationships between passages of music were modelled using the information-theoretic metric of compression distance. In

particular, in the experiment presented in Chapter 5, the task used is closely related to one providing direct similarity ratings between musical passages. The findings of this experiment provide strong empirical validation that compression distance accurately simulates perceived similarity between pairs of melodies, supporting the findings of Pearce and Müllensiefen (2017).

The research presented in this thesis, can provide some insights useful for computational methods of segmentation and musical structure analysis. The empirical support it has found for the hypothesis that thematic structure can be perceived through the statistical regularities caused by repetition highlights the central role played by repetition in the perception of structure. This provides some support for methods that use repetitions to find structural boundaries (Nieto et al., 2020; Rodríguez-López & Volk, 2015); the recognition task of Chapter 5 shows that repetitions can be perceived and this perception is in part due to their relationship to the wider perception of structure in a composition. The experiment of Chapter 8 provides some insight into the ways in which repetition can inform the order of segments, such as favouring the placement of more closely related segments closer together earlier in a composition.

However, as discussed above, the findings of this thesis also consistently showed measures of *internal unpredictability* out-performing measures that sought to only use identified thematic repetitions. This finding indicated that all repetition (as captured by *internal unpredictability*) could aid the perception of structure, rather than larger, specifically-identified repetitions. In the area of MIR automatic segmentation, this finding provides particular support to probabilistic methods that identify boundaries through patterns in unpredictability or entropy (Juhász, 2004; Lattner et al., 2015; Pearce, Müllensiefen & Wiggins, 2010), particularly if they aim to provide a more cognitively-relevant focus. As presented in Chapter 3 (and testing in Chapters 7 and 8), the development of techniques for the modelling of *internal unpredictability* in polyphonic music may now provide opportunities to expand these information-theoretic segmentation approaches outside of monophonic contexts.

9.3 Limitations and future directions

As discussed in Chapter 1, a number of general limitations were placed on the scope of the research presented in this thesis. Perhaps the foremost among these are those applied to the computational modelling of thematic structure, presented in Chapter 3 and evaluated throughout the thesis. As a computational account of the real-world perception of large-scale thematic musical structure, the model presented does possess some limitations. For the most part, these limitations were brought about through the necessity to reduce the processes involved to a form that is tractable with the methods in existence, and to avoid unjustified

assumptions about cognitive processes for which empirical evidence is lacking. The model is, of course, a simplification of human perception and cognition, and of its parallels in music theory and analysis. The limitations faced by this model can be broadly attributed to three areas: (1) constraints on the types of musical information the model deals with—particularly the limitation to monophonic material in Chapters 4, 5 and 6, and limitation to two-voice polyphony in Chapters 7 and 8; (2) limitations on the ways in which music can be represented, and how multiple such representations can be combined across domains; and (3) limitations resulting from the selection of parameters within the model and its constituent components.

Materials used in the modelling of thematic structure were constrained either to monophonic compositions, or those containing only two independent voices (and meeting the other assumptions given in Chapter 3). The reduction of existing compositions, particularly into monophonic versions leads to a loss of information—particularly when compared with the original works from which the melodies were extracted. For example, in the earlier example of the Mozart Piano Sonata given in Chapter 3, some occurrences of themes in traditional analyses rely on harmonic or cadential cues that are not present in the melody. However, we caution that, while illustrative, the melody should not be considered as being directly the same work as the original composition from which it was extracted. Instead, we maintain that these melodies of the corpus described in Chapter 4, with selections used as stimuli in all the experiments, can function as valid compositions in their own right, albeit with different auditory information available. While the monophonic constraint was maintained during the initial portions of this research to avoid the complications of processing thematic material across many polyphonic layers, the first steps towards the modelling of intra-opus thematic structure in polyphonic music were made in the final two experiments. The modelling of statistical learning in polyphonic material introduced in this thesis is, however, still subject to several stringent assumptions (see Chapter 3). In this thesis, the modelling of polyphony was limited to carefully chosen compositions of two voices, far from representing the full experience available in polyphonic music. For these modelling techniques to be more generally applicable, accurate modelling of auditory streams is needed to extract all perceptually salient voices from the polyphonic texture.[1]

For the modelling presented in this thesis, as with much symbolic modelling of music, limitations also stem from how the music can be meaningfully represented; in particular, how to represent and combine the separate domains of pitch and rhythm. The analysis of these two domains in this research maintains an amount of separation between them, allowing for the relative effects of

[1] Multiple successful models of voice separation have been designed; however, further research is needed to investigate the extent to which such voices are perceptible to listeners.

thematic structure to be compared within measures of a single representation, then the relative effects across domains. However, as frequently noted, the ability of compositions to contain only a single uniform pattern places limitations on the utility of the rhythmic modelling of thematic structure; the combination of the two domains with equal weighting (*i.e.*, their combination into a single linked representation of both the pitch and rhythmic surface) would still propagate this issue. The route for the future combination of representations in the modelling of thematic structure likely exists in the development of a selection process of the appropriate representations for a given composition, in some cases including rhythmic representations and excluding (or down-weighting) them for others. The exact nature of this mechanism requires further development and experimental exploration.

Finally, in order for the model to make classifications on phrase boundaries, thematic candidates and thematic repetitions, the model requires certain conditions to be met. These conditions exist as free parameters within the model—for example, the number of phrases that are extracted to make up a theme. While values for these parameters are informed by theory or statistics, the true optimum will vary between compositions, styles and listeners. As a result, these mechanisms have the potential to miss or misclassify their elements on a small number of occasions, adding noise to the modelling data. For example, the phrase boundaries identified by the *Grouper* algorithm may differ from those perceived by a listener, or a theme may not be identified due to it narrowly missing the novelty criteria. The cognitive modelling of theme identification, in particular, presents a useful area for future investigation; there is little behavioural research into the musical conditions and features that allow listeners to identify themes within compositions. Such research is needed for the optimisation of these remaining free parameters.

Aside from these limitations needed to constrain the aspects of computational modelling used in this thesis, limitations to the scope of its application in behavioural evaluation were also necessitated, any of which would be fruitful avenues for future investigation.

First, the study of thematic structure included in this thesis was limited to the domain of western-classical tonal music. This limitation was a continuation of that of the vast majority of preceding literature and computational tools, which largely borrow terminology from, or seek to test, western music theory. However, the cognitive processes hypothesised and modelled in this thesis need not be constrained to western-classical music; the foundational principles of repetition and similarity have evidence of being musically universal (Nettl, 2010). While the model in the form presented in this thesis contains assumptions specific to western music—for example, in the process of phrase detection—its adaptation and evaluation in wider cultural domains would provide many op-

portunities for future work to better understand the perception of large-scale structures.

Although all four of the behavioural experiments collected and analysed information on their participants' musical abilities and backgrounds, the specific testing of effects of level of musical ability on the perception of thematic structure was not their priority. However, through the targeted recruitment of participants and the design of appropriate experiments, the testing of effects of musicality and musical trained on the perception of structure would provide a useful area of possible future research.

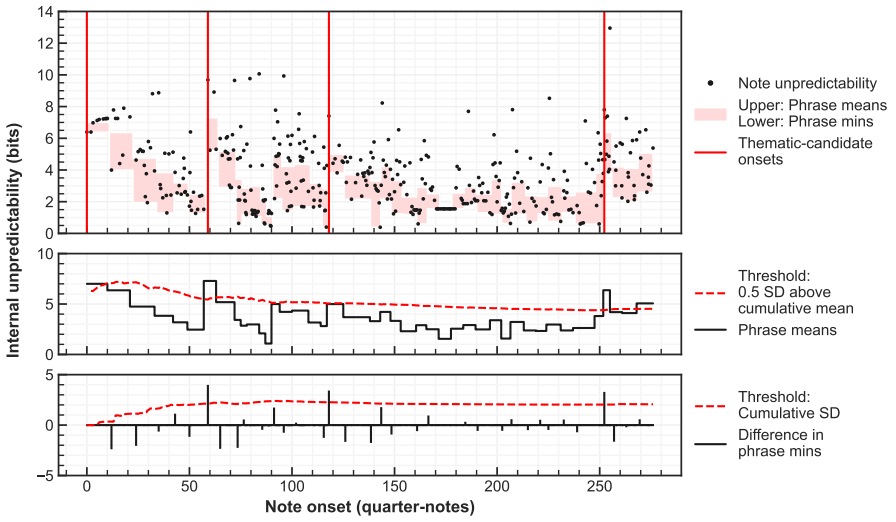
Throughout the course of these behavioural experiments, the scale of structures tested were incrementally increased from the local scale, to compositions of 2 minutes, to those with durations of 4 minutes in the final two experiments. Excluding the study focusing on the effect over small timescales, these lengths were chosen as being long enough both to contain large-scale musical structure and to provide an appropriate amount of context material for the tasks used, but short enough to collect sufficient experimental data in a reasonable experimental session. However, while there do exist many western-classical compositions of these lengths, they are certainly on the shorter end of the spectrum of composition length. The findings of this thesis are, therefore, limited when scaling up from these compositions to those of quarter- or half-hour durations, or longer; perception of thematic structure of such large composition lengths remains to be investigated in future experimental research.

Finally, when taken together, the results of these experiments provide converging evidence that the perception of large-scale thematic structures can be accounted for by the dynamic learning of statistical regularities within compositions. These consistent effects found across experiments, each testing for a different aspect of thematic structure, make the overall findings of this thesis robust ones; future research into the perception of large-scale thematic structure should continue this approach, providing insight into the many different ways in which the effects of thematic structures in musical compositions may be perceived.

Appendix A

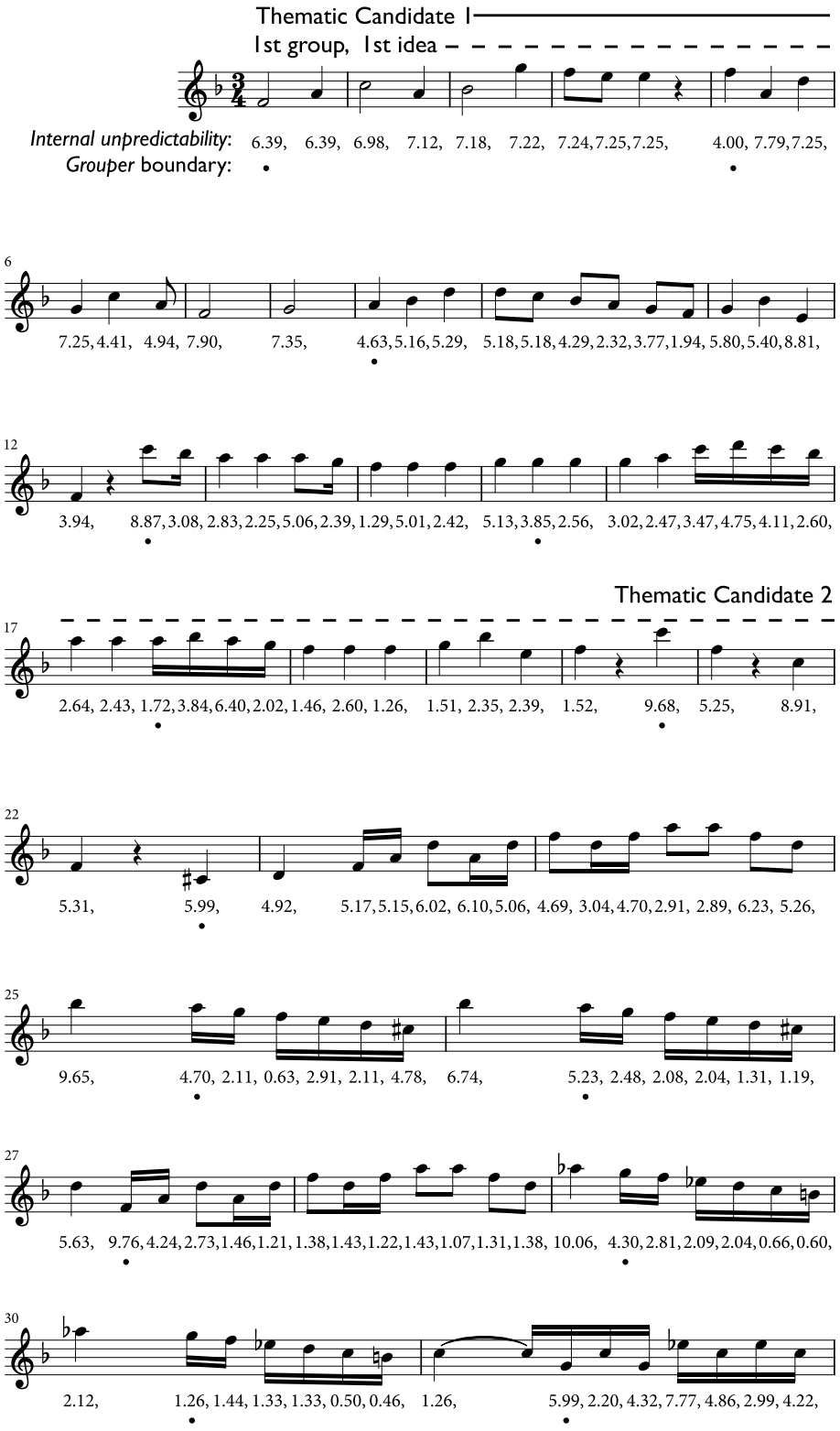
Thematic-Candidate Detection Example

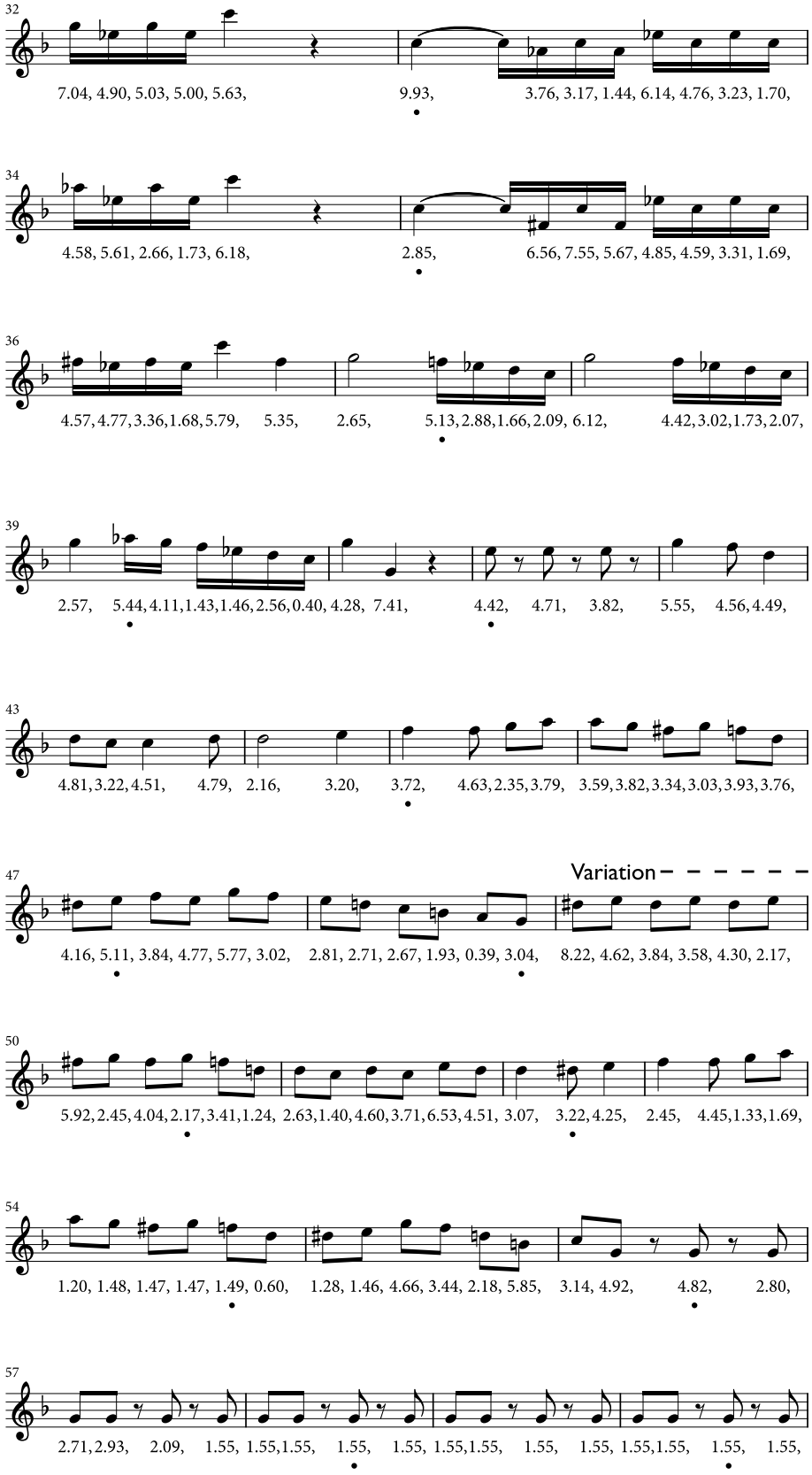
This appendix contains an annotated score for the thematic-candidate detection given in Chapter 3. The four thematic candidates identified by the model (also shown in Figure 3.5) are marked by brackets with solid lines. Sections of the score identified by the analyses of Beach (1994) and Kinderman (2006) are marked by brackets with dashed lines. The IDyOM STM *internal unpredictability* values when using the pitch interval representations, as well as phrase boundaries detected by *Grouper* are shown below notes.

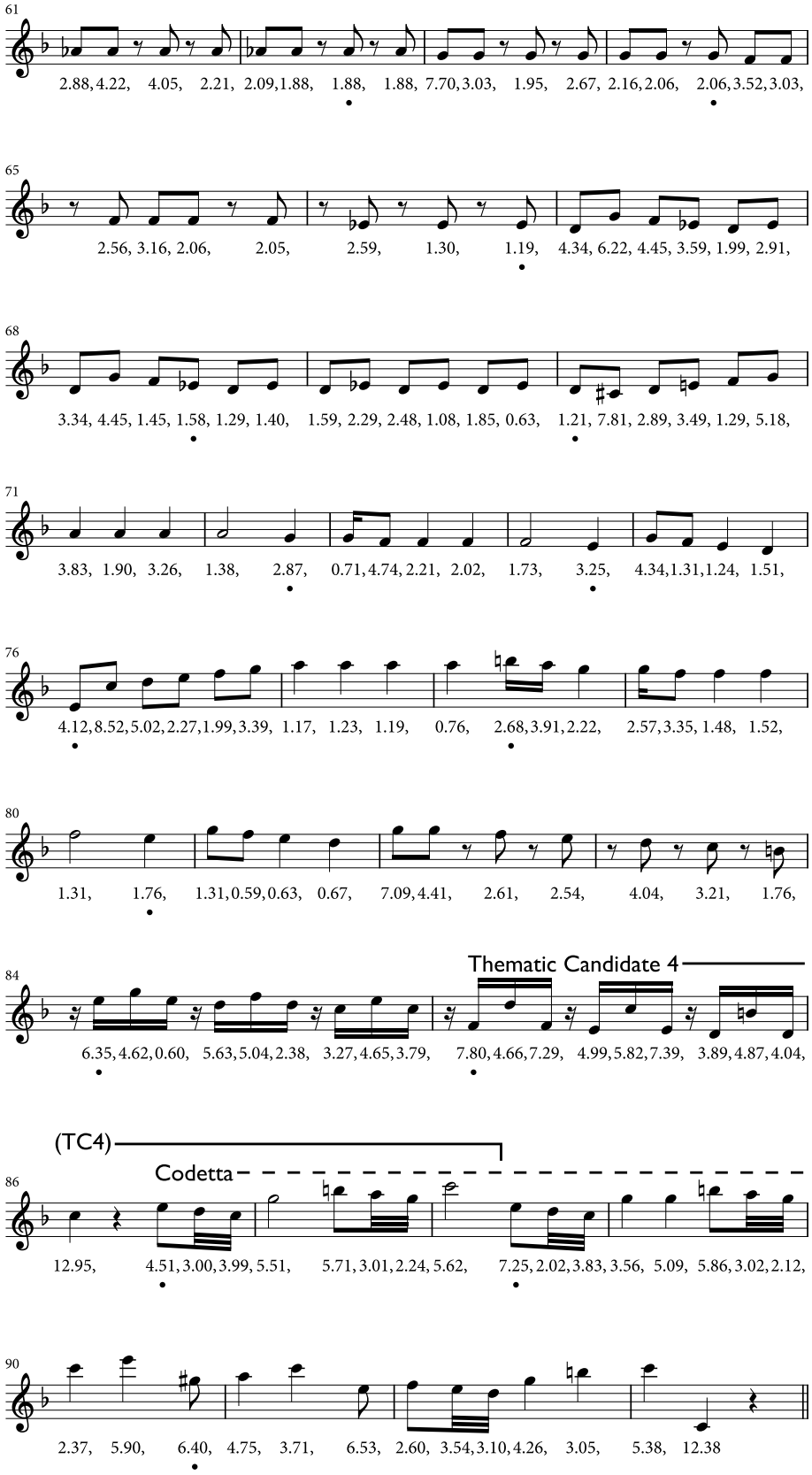


[Copy of Figure 3.4]: Theme detection for Mozart K. 332, first movement, exposition. Four onsets were identified at 0, 60, 121 and 253 quarter-notes, based on pitch interval internal unpredictability. Thematic candidates were identified by phrase mean information contents being greater than a threshold of a 0.5 SD above the cumulative mean (subplot 2), and having a difference in phrase minimums greater than 1 SD (subplot 3).

Mozart, Piano Sonata No. 12, K.332, first movement, bars 1–93




32
7.04, 4.90, 5.03, 5.00, 5.63,
9.93,
3.76, 3.17, 1.44, 6.14, 4.76, 3.23, 1.70,
34
4.58, 5.61, 2.66, 1.73, 6.18,
2.85,
6.56, 7.55, 5.67, 4.85, 4.59, 3.31, 1.69,
36
4.57, 4.77, 3.36, 1.68, 5.79,
5.35,
2.65,
5.13, 2.88, 1.66, 2.09,
6.12,
4.42, 3.02, 1.73, 2.07,
39
2.57,
5.44, 4.11, 1.43, 1.46, 2.56, 0.40,
4.28,
7.41,
4.42,
4.71,
3.82,
5.55,
4.56, 4.49,
43
4.81, 3.22, 4.51,
4.79,
2.16,
3.20,
3.72,
4.63, 2.35, 3.79,
3.59, 3.82, 3.34, 3.03, 3.93, 3.76,
47
Variation
4.16, 5.11, 3.84, 4.77, 5.77, 3.02,
2.81, 2.71, 2.67, 1.93, 0.39, 3.04,
8.22, 4.62, 3.84, 3.58, 4.30, 2.17,
50
5.92, 2.45, 4.04, 2.17, 3.41, 1.24,
2.63, 1.40, 4.60, 3.71, 6.53, 4.51,
3.07,
3.22, 4.25,
2.45,
4.45, 1.33, 1.69,
54
1.20, 1.48, 1.47, 1.47, 1.49, 0.60,
1.28, 1.46, 4.66, 3.44, 2.18, 5.85,
3.14, 4.92,
4.82,
2.80,
57
2.71, 2.93,
2.09,
1.55,
1.55, 1.55,
1.55,
1.55,
1.55, 1.55,
1.55,
1.55,
1.55, 1.55,
1.55,
1.55,


61
2.88, 4.22, 4.05, 2.21, 2.09, 1.88, 1.88, 1.88, 7.70, 3.03, 1.95, 2.67, 2.16, 2.06, 2.06, 3.52, 3.03,
65
2.56, 3.16, 2.06, 2.05, 2.59, 1.30, 1.19, 4.34, 6.22, 4.45, 3.59, 1.99, 2.91,
68
3.34, 4.45, 1.45, 1.58, 1.29, 1.40, 1.59, 2.29, 2.48, 1.08, 1.85, 0.63, 1.21, 7.81, 2.89, 3.49, 1.29, 5.18,
71
3.83, 1.90, 3.26, 1.38, 2.87, 0.71, 4.74, 2.21, 2.02, 1.73, 3.25, 4.34, 1.31, 1.24, 1.51,
76
4.12, 8.52, 5.02, 2.27, 1.99, 3.39, 1.17, 1.23, 1.19, 0.76, 2.68, 3.91, 2.22, 2.57, 3.35, 1.48, 1.52,
80
1.31, 1.76, 1.31, 0.59, 0.63, 0.67, 7.09, 4.41, 2.61, 2.54, 4.04, 3.21, 1.76,
84
Thematic Candidate 4
6.35, 4.62, 0.60, 5.63, 5.04, 2.38, 3.27, 4.65, 3.79, 7.80, 4.66, 7.29, 4.99, 5.82, 7.39, 3.89, 4.87, 4.04,
(TC4)
Codetta
86
12.95, 4.51, 3.00, 3.99, 5.51, 5.71, 3.01, 2.24, 5.62, 7.25, 2.02, 3.83, 3.56, 5.09, 5.86, 3.02, 2.12,
90
2.37, 5.90, 6.40, 4.75, 3.71, 6.53, 2.60, 3.54, 3.10, 4.26, 3.05, 5.38, 12.38

Appendix B

Experiment Stimuli Compositions

This appendix contains lists of compositions used in the experiments presented in Chapters 6, 7 and 8. Table B.1 gives the details of the 40 compositions used in the experiment of Chapter 6. These were all monophonic compositions taken from the corpus of Chapters 4, adjusted to be approximately 2 minutes in duration. Table B.2 gives the details of the 40 compositions used in the experiments of Chapter 7 and Chapter 8. These compositions belonged to a texture category of monophonic compositions (taken from those of Table B.1) and polyphonic compositions of two voices. Compositions belonged to a length category of 'short' compositions of approximately 2 minutes duration, and 'long' compositions of approximately 4 minutes.

Table B.1: Stimulus Compositions Used in the Experiment of Chapter 6

Composer	Work	Movt.	Year[a]
Alkan, Charles-Valentin	Gigue et air de ballet, Op.24	2	[1844]
Bach, Johann Sebastian	Cello Suite No.3, BWV 1009	1	1717
Bach, Johann Sebastian	Cello Suite No.5, BWV 1011	1	1717
Bartók, Béla	2 Romanian Dances, Op.8a	1	1910
Beethoven, Ludwig van	Piano Sonata No.7, Op.10 No.3	4	1797
Beethoven, Ludwig van	Piano Sonata No.10, Op.14 No.2	1	1798
Beethoven, Ludwig van	Piano Sonata No.12, Op.26	4	1800
Beethoven, Ludwig van	Piano Sonata No.13, Op.27 No.1	1	1800
Beethoven, Ludwig van	Piano Sonata No.15, Op.28	2	1801
Beethoven, Ludwig van	Piano Sonata No.17, Op.31 No.2	3	1801
Beethoven, Ludwig van	Violin Sonata No.1 in D major, Op.12 No.1	1	1797
Brahms, Johannes	Ballades, Op.10	2	1854
Brahms, Johannes	Cello Sonata No.1, Op.38	3	1862
Brahms, Johannes	Scherzo, Op.4	–	1851
Chopin, Frédéric	Mazurkas, Op.33	3	1837
Chopin, Frédéric	Mazurkas, Op.68	4	1826
Chopin, Frédéric	Waltzes, Op.69	2	1829
Dvořák, Antonín	Legends, Op.59	3	1881
Fauré, Gabriel	Thème et variations, Op.73	–	1895
Fauré, Gabriel	Valse Caprice No.4 in A-flat, Op.62, Op.62	–	1894
Grieg, Edvard	Piano Sonata, Op.7	3	1865
Handel, George Frideric	Fantasia in C	–	1703
Handel, George Frideric	Recorder Sonata in A minor, Op.1 No.4	4	1712
Haydn, Joseph	String Quartet in B-flat major, Op.55 No.3	4	1788
Haydn, Joseph	String Quartet in C major, Op.64 No.1	4	1790
Haydn, Joseph	String Quartet in C major, Op.76 No.3	3	1796
Haydn, Joseph	String Quartet in E-flat major, Op.64 No.6	4	[1791]
Haydn, Joseph	String Quartet in F major, Op.17 No.2	1	1771
Haydn, Joseph	String Quartet in G major, Op.17 No.5	3	1771
Haydn, Joseph	String Quartet in G major, Op.77 No.1	3	1799
Liszt, Franz	Aus der Ungarischen Krönungsmesse	–	1869
Mozart, Wolfgang Amadeus	Piano Sonata No.7, K.309	2	1777
Mozart, Wolfgang Amadeus	Piano Sonata No.7, K.309	3	1777
Mozart, Wolfgang Amadeus	Piano Sonata No.12, K.332	2	1783
Mozart, Wolfgang Amadeus	Piano Sonata No.17, K.570	2	1789
Mozart, Wolfgang Amadeus	Piano Sonata No.18, K.576	3	1789
Saint-Saëns, Camille	6 Etudes, Op.111	6	1892
Schubert, Franz	4 Impromptus, D.935	2	1827
Schubert, Franz	4 Impromptus, D.899	4	1827
Vaughan Williams, Ralph	Suite for Viola & Small Orchestra	1	1934

[a]Where composition year is unknown, first publication year is given

Table B.2: Stimulus Compositions Used in the Experiments of Chapters 7 and 8

Category					
Texture	Length	Composer	Work	Movt.	Year[a]
Monophonic	Long	Alkan, Charles-Valentin	Gigue et air de ballet, Op.24	1	[1844]
Monophonic	Long	Beethoven, Ludwig van	Piano Sonata No.15, Op.28	2	1801
Monophonic	Long	Brahms, Johannes	Cello Sonata No.1, Op.38	3	1862
Monophonic	Long	Debussy, Claude	Deux Arabesques, CD 74	2	1890
Monophonic	Long	Dvořák, Antonín	Legends, Op.59	3	1881
Monophonic	Long	Handel, George Frideric	Fantasia in C, HWV 490	–	1703
Monophonic	Long	Haydn, Joseph	String Quartet in C major, Hob.III:65	4	1790
Monophonic	Long	Haydn, Joseph	String Quartet in F major, Hob.III:26	1	1771
Monophonic	Long	Mozart, Wolfgang Amadeus	Piano Sonata No.7, K.309	2	1777
Monophonic	Long	Mozart, Wolfgang Amadeus	Piano Sonata No.18, K.576	3	1789
Monophonic	Short	Beethoven, Ludwig van	Piano Sonata No.4, Op.7	3	1796
Monophonic	Short	Brahms, Johannes	Ballades, Op.10	2	1854
Monophonic	Short	Brahms, Johannes	Scherzo, Op.4	–	1851
Monophonic	Short	Fauré, Gabriel	Thème et variations, Op.73	–	1895
Monophonic	Short	Grieg, Edvard	Piano Sonata, Op.7	3	1865
Monophonic	Short	Handel, George Frideric	Recorder Sonata in A minor, HWV 362	4	1712
Monophonic	Short	Haydn, Joseph	String Quartet in E-flat major, Hob.III:64	4	[1791]
Monophonic	Short	Mozart, Wolfgang Amadeus	Piano Sonata No.17, K.570	2	1789
Monophonic	Short	Schubert, Franz	4 Impromptus, D.935	2	1827
Monophonic	Short	Vaughan Williams, Ralph	Suite for Viola & Small Orchestra	1	1934

Table A.2: *Cont.*

Polyphonic	Long	Bach, Johann Sebastian	Duetto No.2 in F major, BWV 803	–	[1739]
Polyphonic	Long	Bach, Johann Sebastian	Duetto No.4 in A minor, BWV 805	–	[1739]
Polyphonic	Long	Bériot, Charles-Auguste de	No.1 from 3 Concertant Duets, Op.57	1	[1847]
Polyphonic	Long	Glière, Reinhold	Douze Duos pour 2 Violons, Op.49	1	1909
Polyphonic	Long	Mozart, Wolfgang Amadeus	Duo for Violin and Viola, K.423	1	1783
Polyphonic	Long	Sibelius, Jean	Duo for Violin and Viola	–	1891
Polyphonic	Long	Spohr, Louis	No.1 from 3 Grand Duos for 2, Op.39	1	[1820]
Polyphonic	Long	Spohr, Louis	No.1 from 3 Grand Duos for 2, Op.39	2	[1820]
Polyphonic	Long	Spohr, Louis	No.2 from 3 Concertant Duos, Op.67	1	[1828]
Polyphonic	Long	Stamitz, Carl Philipp	No.2 from 6 Duos, Op.34	1	
Polyphonic	Short	Bach, Johann Sebastian	Duetto No.1 in E minor, BWV 802	–	[1739]
Polyphonic	Short	Bach, Johann Sebastian	Duetto No.3 in G major, BWV 804	–	[1739]
Polyphonic	Short	Bach, Johann Sebastian	Invention in F minor, BWV 780	–	1723
Polyphonic	Short	Beethoven, Ludwig van	Duo for 2 Flutes, WoO 26	1	1792
Polyphonic	Short	Bériot, Charles-Auguste de	No.3 from 3 Concertant Duets, Op.57	2	[1847]
Polyphonic	Short	Fuchs, Robert	20 Duos, Op.55	4	[1896]
Polyphonic	Short	Glière, Reinhold	Douze Duos pour 2 Violons, Op.50	2	1909
Polyphonic	Short	Glière, Reinhold	Douze Duos pour 2 Violons, Op.51	4	1909
Polyphonic	Short	Spohr, Louis	No.2 from 3 Concertant Duos, Op.67	1	[1828]
Polyphonic	Short	Telemann, Georg Philipp	No.2 from 3 Duets	4	[1728]

[a]Where composition year is unknown, first publication year is given

Appendix C

Four Example Compositions

This appendix contains the segmented versions of the four example compositions for Chapter 8, Experiment 2. Each Composition is presented with the segment orders that gained minimum and maximum values within the large Monte Carlo set of randomised orders.

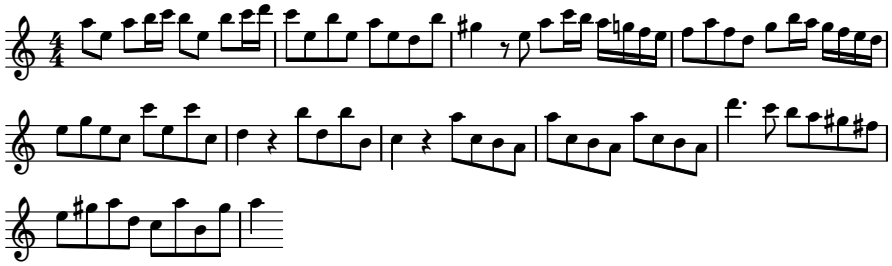
Composition A

Handel, Recorder Sonata in A minor, HWV 362, fourth movement
'Monophonic' and 'short' categories

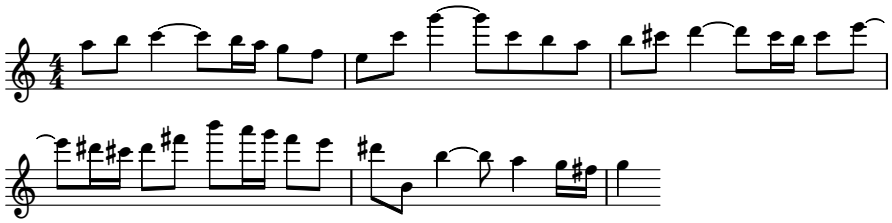
	Minimum	
Measure	Order	Value
Pitch interval		
Internal unpredictability (perfect)	1, 5, 2, 4, 7, 3, 9, 6, 8	−2.82
Internal unpredictability (buffer)	1, 5, 4, 2, 6, 3, 7, 8, 9	−3.53
Internal unpredictability flow	1, 8, 2, 7, 9, 6, 5, 4, 3	−3.15
Stylistic unpredictability flow	1, 4, 5, 2, 8, 7, 3, 6, 9	−3.56
Variation	1, 6, 4, 2, 7, 5, 8, 9, 3	−0.99
Inter-onset interval		
Internal unpredictability (perfect)	1, 5, 4, 2, 7, 6, 3, 8, 9	−3.52
Internal unpredictability (buffer)	1, 4, 7, 2, 3, 5, 6, 8, 9	−3.29
Internal unpredictability flow	1, 2, 8, 3, 9, 5, 4, 6, 7	−2.53
Stylistic unpredictability flow	1, 4, 9, 8, 7, 3, 6, 2, 5	−3.59
Variation	1, 3, 4, 9, 6, 5, 8, 7, 2	−1.30
Other		
Tonality	1, 2, 6, 8, 4, 9, 5, 3, 7	−2.31

Measure	Maximum Order	Maximum Value
Pitch interval		
Internal unpredictability (perfect)	1, 8, 3, 5, 6, 2, 7, 4, 9	2.71
Internal unpredictability (buffer)	1, 5, 7, 4, 8, 3, 9, 2, 6	2.35
Internal unpredictability flow	1, 4, 2, 5, 7, 6, 8, 3, 9	2.90
Stylistic unpredictability flow	1, 7, 5, 3, 8, 9, 2, 6, 4	2.05
Variation	1, 8, 5, 2, 7, 9, 3, 4, 6	1.94
Inter-onset interval		
Internal unpredictability (perfect)	1, 9, 6, 5, 8, 3, 7, 4, 2	2.25
Internal unpredictability (buffer)	1, 9, 5, 7, 8, 2, 6, 4, 3	2.44
Internal unpredictability flow	1, 4, 5, 2, 3, 6, 8, 7, 9	2.82
Stylistic unpredictability flow	1, 2, 8, 3, 9, 5, 4, 6, 7	1.96
Variation	1, 2, 9, 7, 4, 5, 3, 8, 6	1.80
Other		
Tonal distance	1, 3, 5, 9, 4, 6, 2, 7, 8	2.32

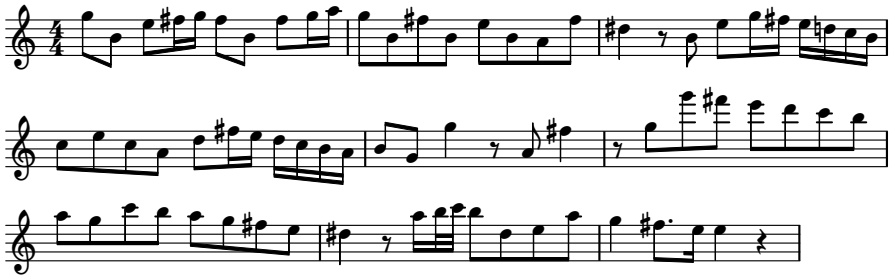
Segment 1



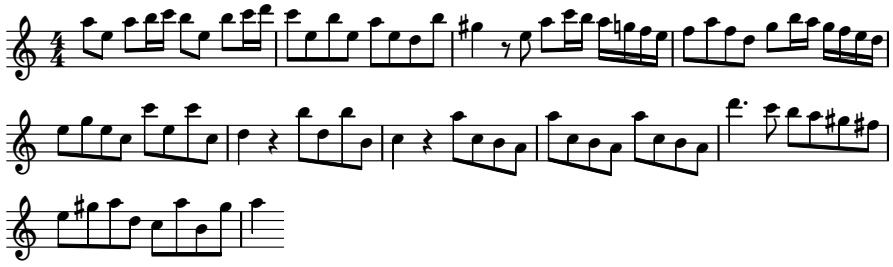
Segment 2



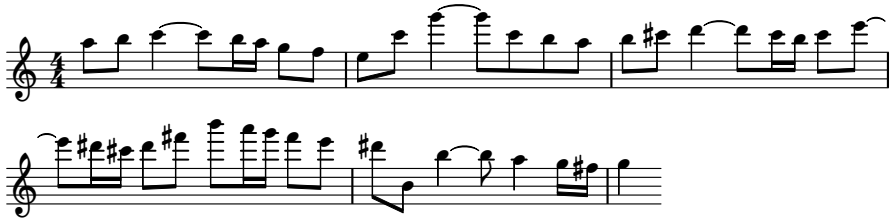
Segment 3



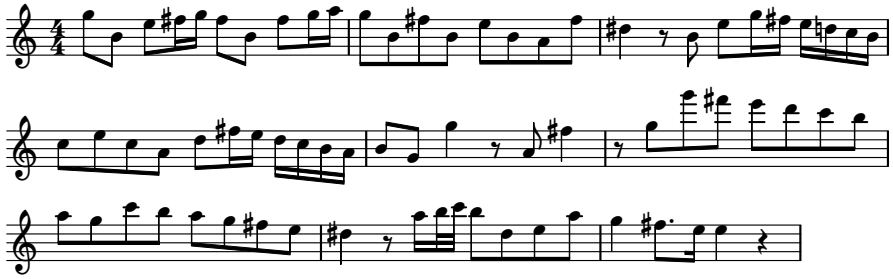
Segment 4



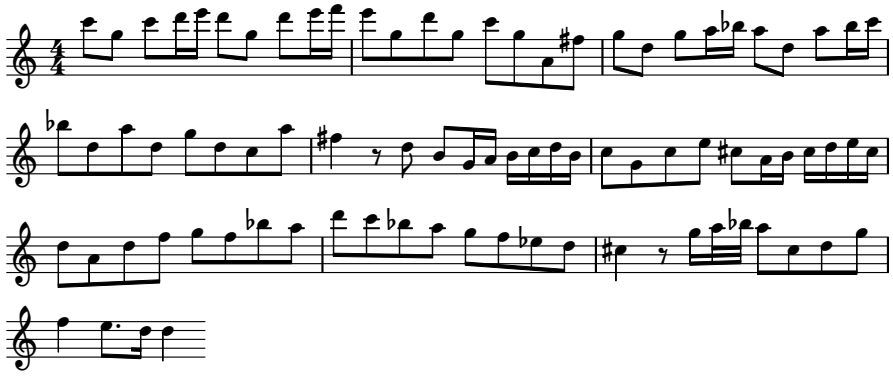
Segment 5



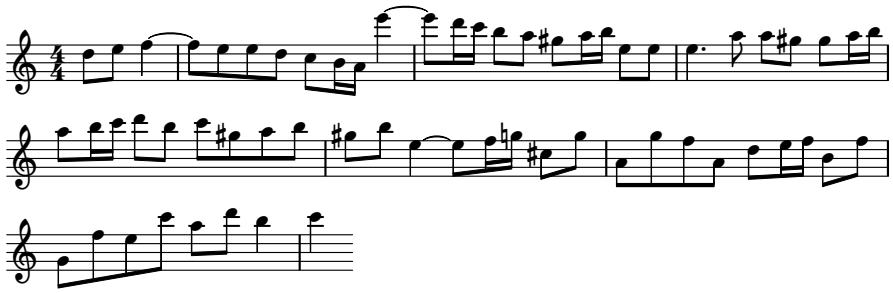
Segment 6



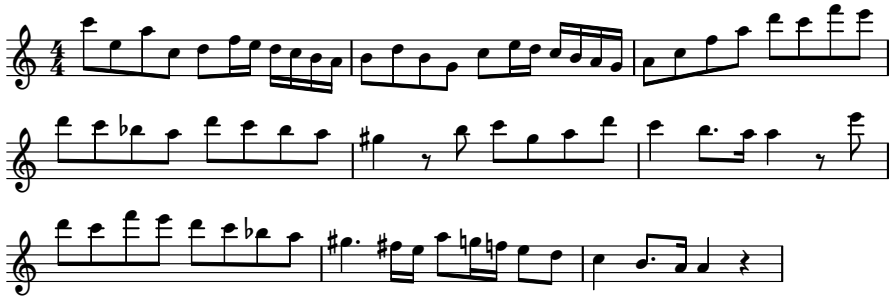
Segment 7



Segment 8



Segment 9



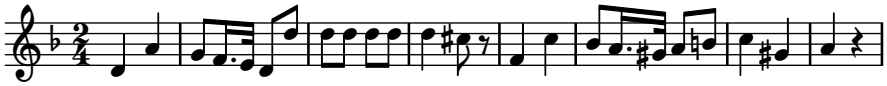
Composition B

Beethoven, Piano Sonata No. 15, Op. 28, second movement
'Monophonic' and 'long' categories

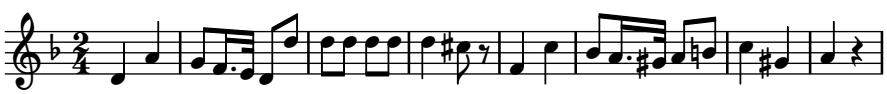
Measure	Minimum Order	Minimum Value
	Pitch interval	
Internal unpredictability (perfect)	1, 3, 11, 2, 9, 4, 12, 8, 5, 6, 13, 7, 10	−3.68
Internal unpredictability (buffer)	1, 2, 9, 12, 3, 11, 4, 13, 5, 6, 10, 8, 7	−4.89
Internal unpredictability flow	1, 6, 10, 4, 13, 8, 12, 7, 3, 5, 11, 2, 9	−3.03
Stylistic unpredictability flow	1, 9, 5, 6, 2, 13, 7, 8, 10, 11, 3, 4, 12	−3.78
Variation	1, 9, 7, 12, 2, 5, 3, 10, 11, 6, 8, 4, 13	−1.80
	Inter-onset interval	
Internal unpredictability (perfect)	1, 8, 13, 9, 6, 7, 5, 2, 12, 11, 3, 4, 10	−3.56
Internal unpredictability (buffer)	1, 2, 9, 12, 3, 11, 4, 13, 5, 6, 10, 8, 7	−4.08
Internal unpredictability flow	1, 4, 13, 2, 7, 9, 3, 11, 5, 10, 12, 8, 6	−3.32
Stylistic unpredictability flow	1, 11, 3, 9, 4, 2, 6, 8, 10, 12, 7, 5, 13	−3.41
Variation	1, 3, 13, 10, 12, 4, 11, 5, 2, 7, 9, 8, 6	−2.41
	Other	
Tonality	1, 13, 6, 8, 5, 7, 9, 3, 11, 10, 4, 2, 12	−4.06

Measure	Maximum Order	Maximum Value
	Pitch interval	
Internal unpredictability (perfect)	1, 6, 10, 4, 8, 12, 7, 3, 11, 2, 13, 9, 5	2.87
Internal unpredictability (buffer)	1, 11, 7, 4, 12, 8, 3, 10, 5, 9, 13, 6, 2	1.96
Internal unpredictability flow	1, 9, 11, 3, 13, 2, 5, 4, 8, 6, 10, 7, 12	3.96
Stylistic unpredictability flow	1, 4, 6, 12, 5, 10, 2, 9, 13, 11, 8, 3, 7	2.49
Variation	1, 10, 13, 12, 6, 8, 4, 5, 11, 2, 7, 3, 9	2.69
	Inter-onset interval	
Internal unpredictability (perfect)	1, 10, 3, 2, 13, 9, 6, 11, 4, 5, 8, 7, 12	2.28
Internal unpredictability (buffer)	1, 5, 3, 6, 11, 13, 12, 2, 7, 10, 9, 8, 4	2.22
Internal unpredictability flow	1, 10, 5, 7, 2, 6, 4, 12, 3, 9, 8, 13, 11	2.33
Stylistic unpredictability flow	1, 7, 4, 6, 9, 12, 2, 10, 11, 8, 3, 5, 13	2.82
Variation	1, 2, 8, 7, 13, 12, 3, 9, 6, 5, 4, 11, 10	1.53
	Other	
Tonal distance	1, 4, 6, 12, 5, 13, 11, 9, 7, 10, 3, 8, 2	2.71

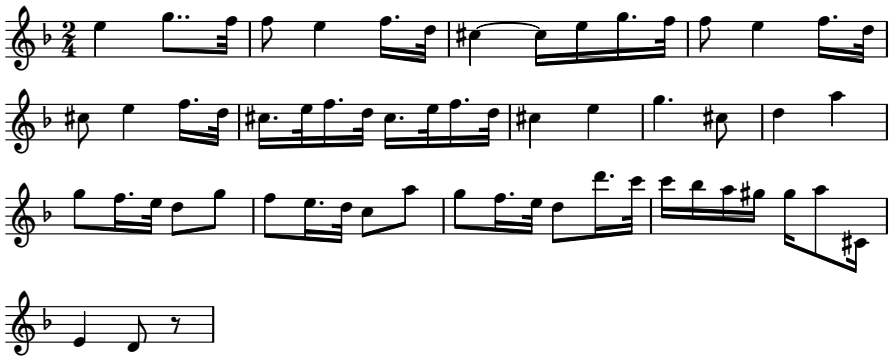
Segment 1



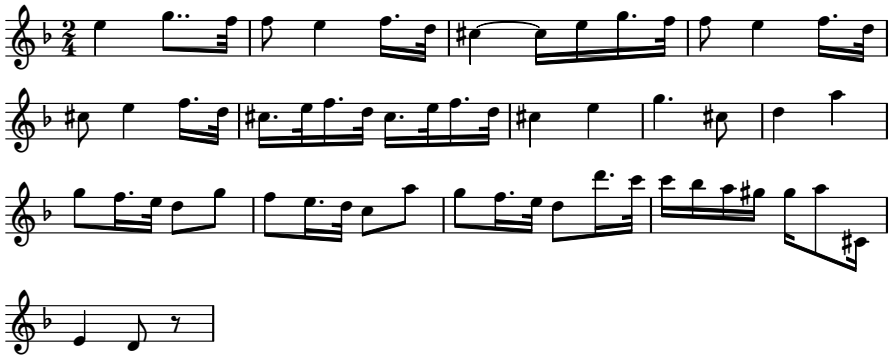
Segment 2



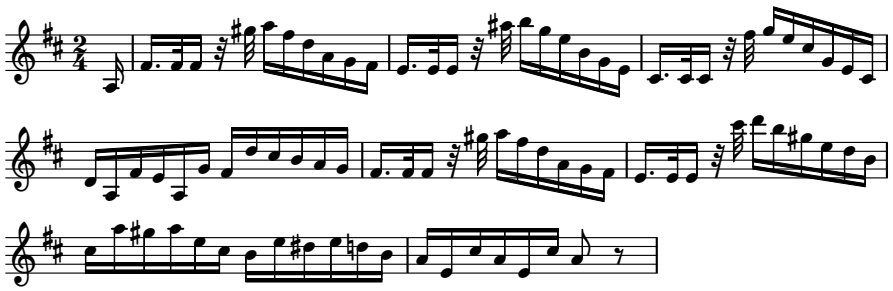
Segment 3



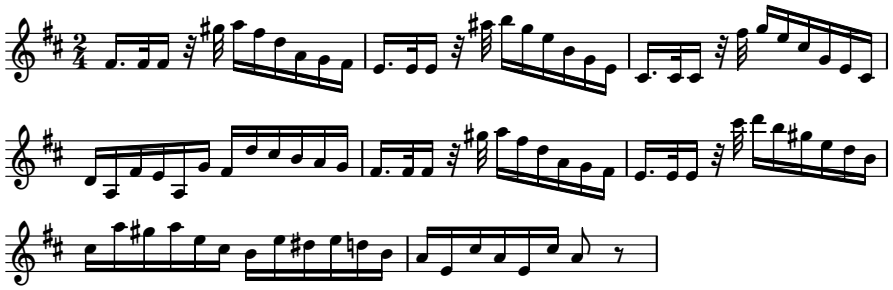
Segment 4



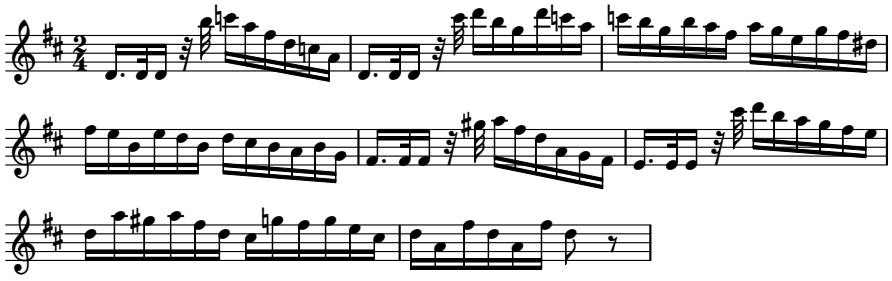
Segment 5



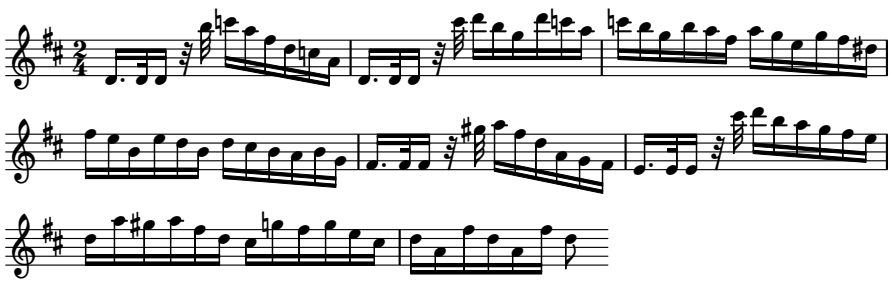
Segment 6



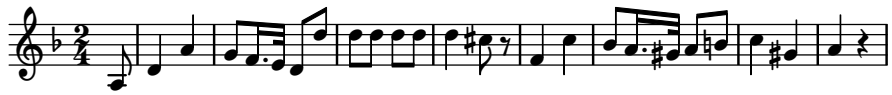
Segment 7



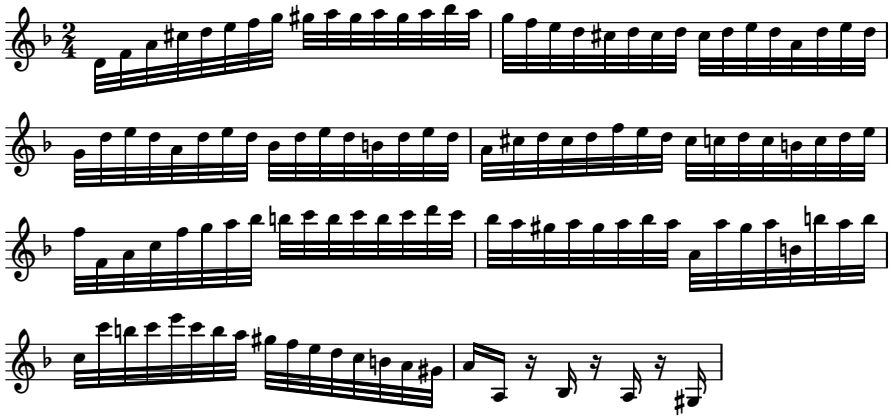
Segment 8



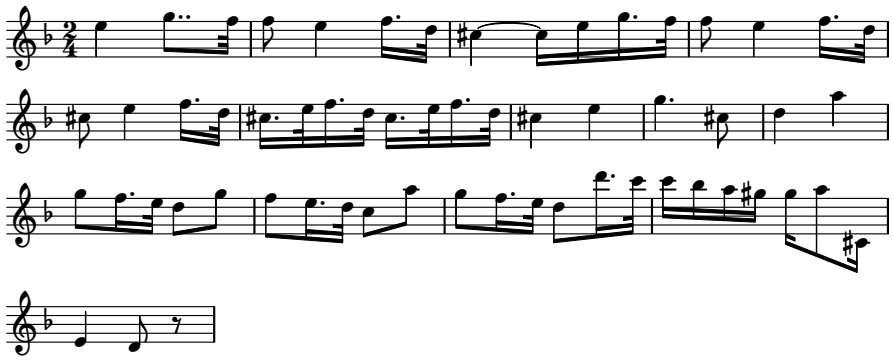
Segment 9



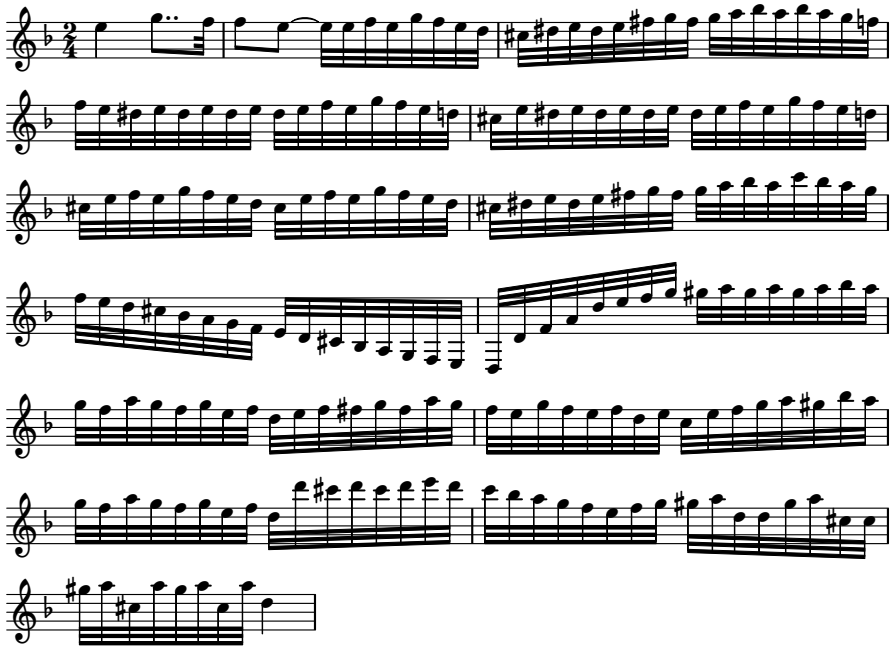
Segment 10



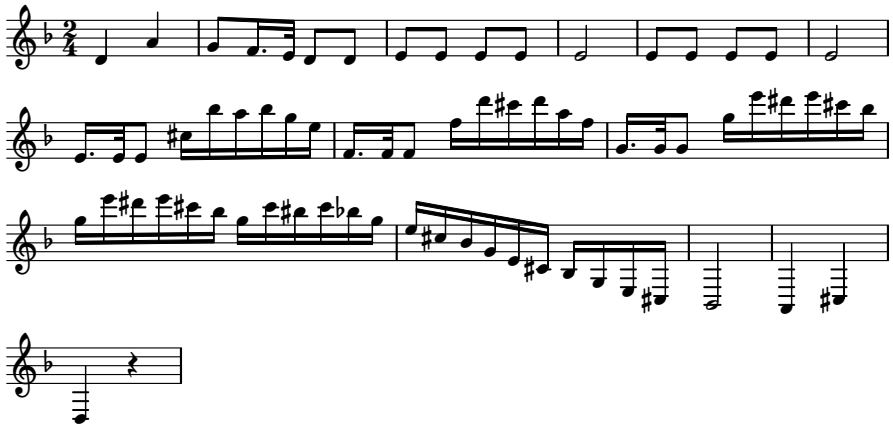
Segment 11



Segment 12



Segment 13



Composition C

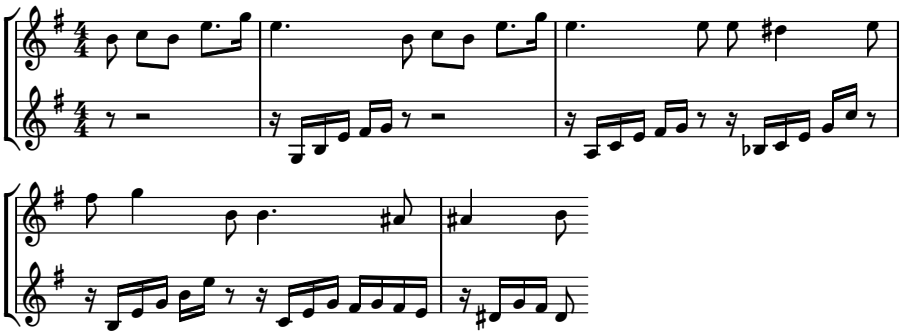
Glière, 12 Duos for 2 Violins, Op. 49, fourth duo

'Polyphonic' and 'short' categories

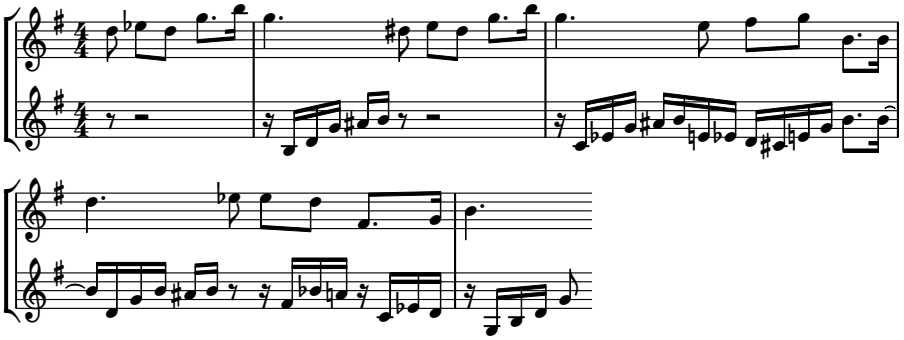
Measure	Minimum Order	Minimum Value
Pitch interval		
Internal unpredictability (perfect)	1, 8, 5, 6, 9, 4, 7, 2, 3	−2.44
Internal unpredictability (buffer)	1, 3, 8, 9, 2, 7, 4, 5, 6	−3.13
Internal unpredictability flow	1, 2, 7, 4, 5, 3, 9, 6, 8	−2.72
Stylistic unpredictability flow	1, 8, 3, 7, 6, 9, 5, 4, 2	−3.91
Variation	1, 9, 2, 3, 8, 4, 5, 6, 7	−2.13
Inter-onset interval		
Internal unpredictability (perfect)	1, 5, 3, 8, 6, 9, 2, 4, 7	−2.78
Internal unpredictability (buffer)	1, 5, 8, 6, 3, 7, 4, 9, 2	−3.17
Internal unpredictability flow	1, 7, 4, 9, 6, 2, 5, 8, 3	−3.34
Stylistic unpredictability flow	1, 8, 3, 5, 6, 9, 2, 7, 4	−3.20
Variation	1, 3, 4, 9, 2, 5, 6, 7, 8	−1.92
Other		
Tonality	1, 4, 3, 8, 9, 7, 5, 2, 6	−2.82

Measure	Maximum Order	Maximum Value
Pitch interval		
Internal unpredictability (perfect)	1, 2, 3, 7, 4, 6, 5, 9, 8	2.01
Internal unpredictability (buffer)	1, 4, 5, 3, 9, 6, 2, 8, 7	2.06
Internal unpredictability flow	1, 8, 9, 3, 7, 2, 5, 6, 4	2.77
Stylistic unpredictability flow	1, 2, 3, 9, 8, 6, 7, 4, 5	1.83
Variation	1, 6, 7, 4, 2, 8, 9, 3, 5	2.25
Inter-onset interval		
Internal unpredictability (perfect)	1, 7, 4, 2, 6, 9, 3, 5, 8	2.59
Internal unpredictability (buffer)	1, 6, 7, 8, 9, 5, 2, 4, 3	2.88
Internal unpredictability flow	1, 8, 5, 3, 7, 6, 9, 2, 4	2.52
Stylistic unpredictability flow	1, 6, 7, 5, 4, 9, 2, 3, 8	1.94
Variation	1, 6, 7, 8, 2, 9, 4, 5, 3	2.11
Other		
Tonal distance	1, 8, 9, 4, 5, 3, 7, 6, 2	3.02

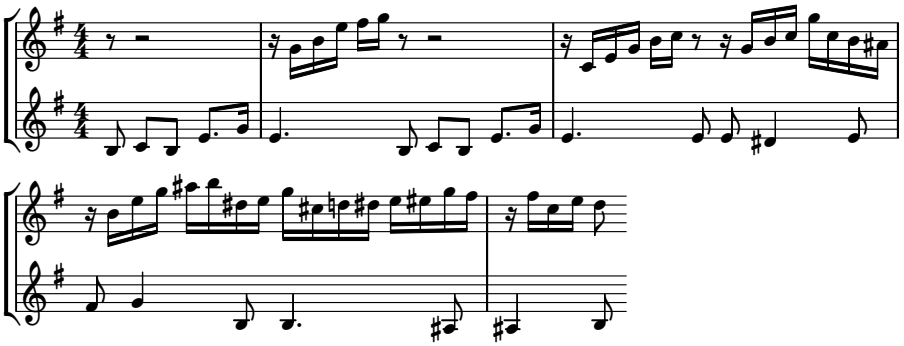
Segment 1



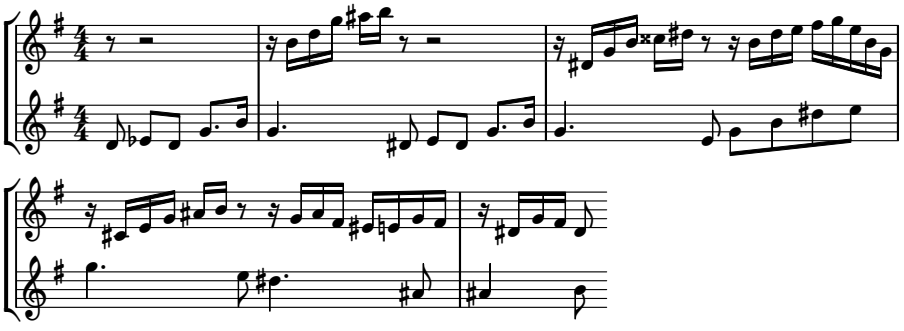
Segment 2



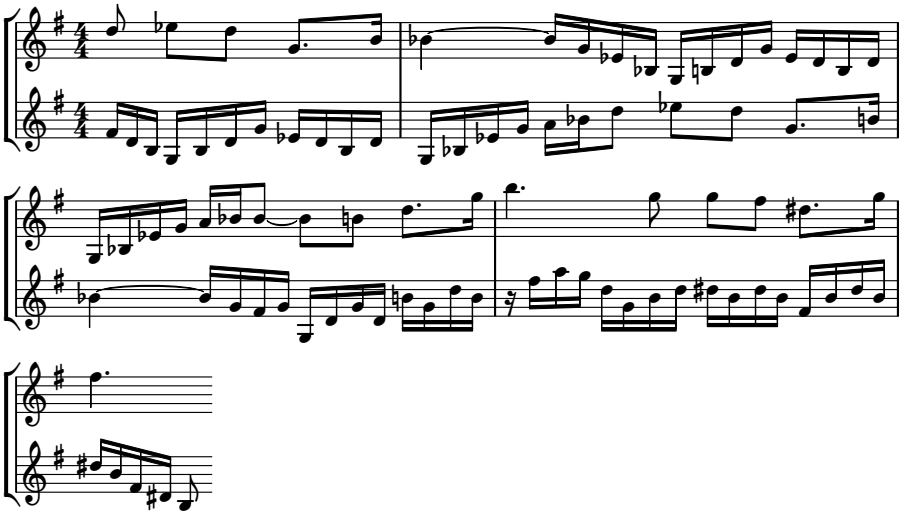
Segment 3



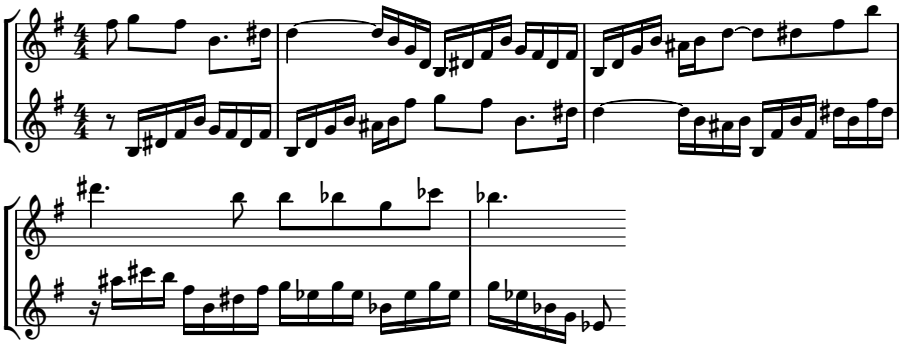
Segment 4



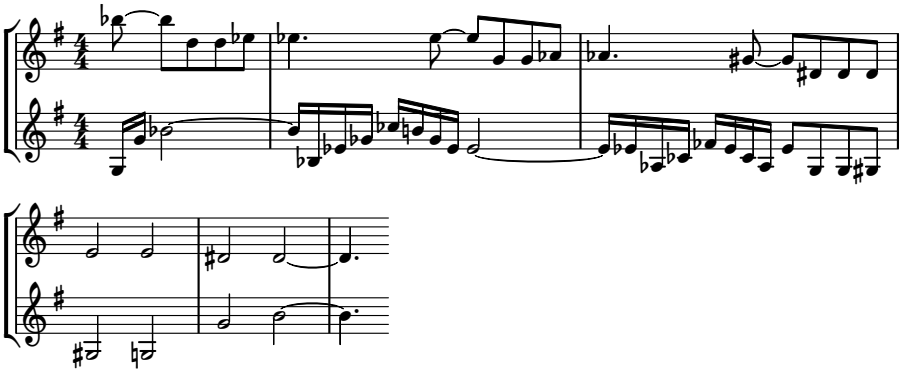
Segment 5



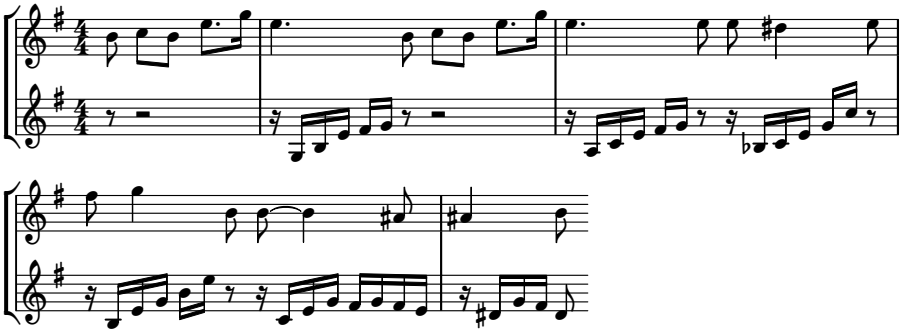
Segment 6



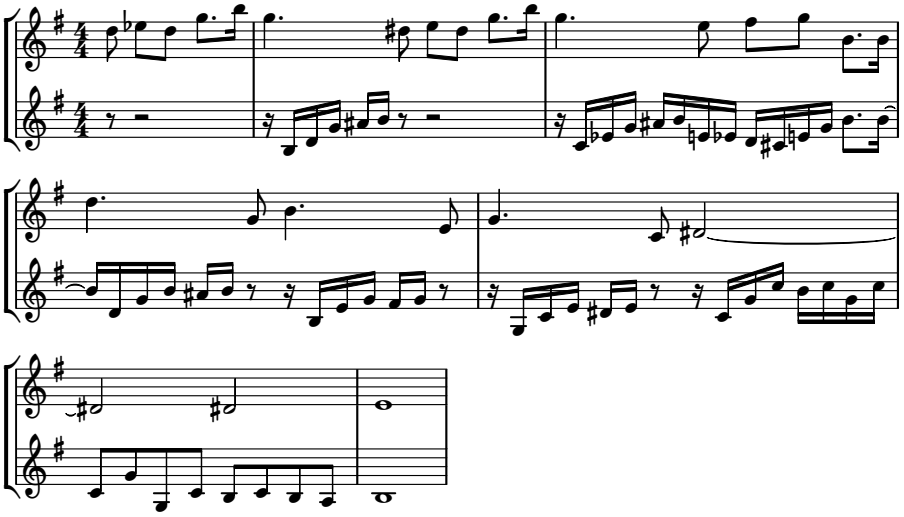
Segment 7



Segment 8



Segment 9



Composition D

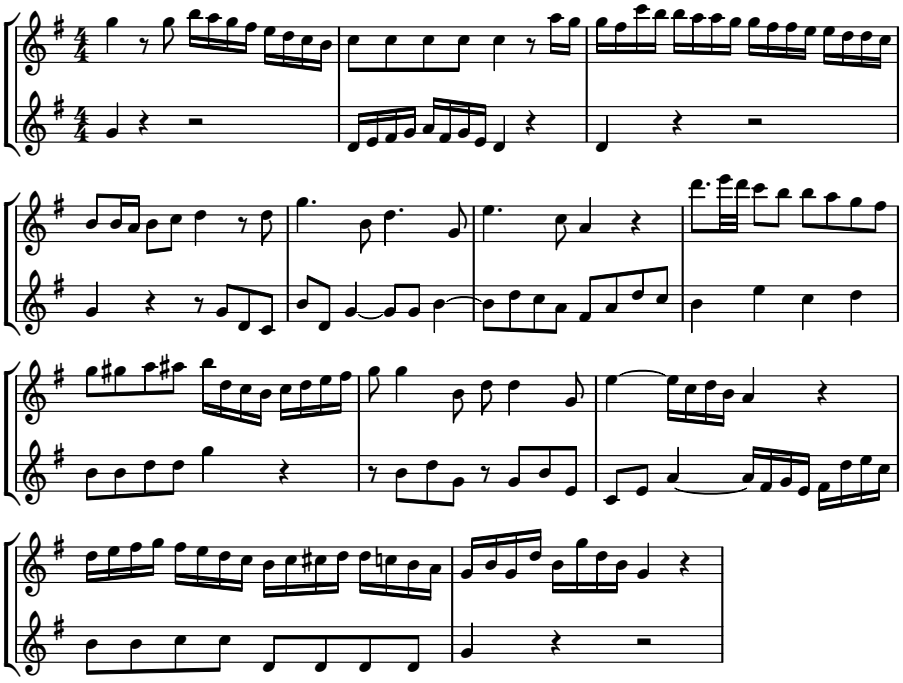
Mozart, Duo for Violin and Viola, K.423, first movement

'Polyphonic' and 'long' categories

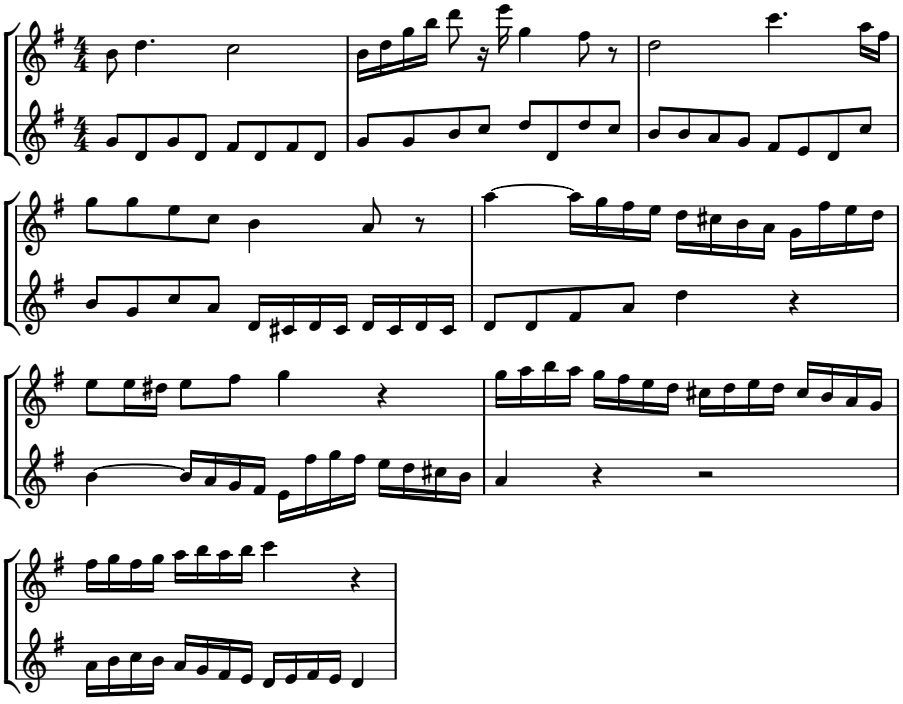
Measure	Minimum Order	Value
	Pitch interval	
Internal unpredictability (perfect)	1, 10, 3, 9, 12, 11, 5, 2, 6, 7, 14, 13, 4, 8, 15	−2.68
Internal unpredictability (buffer)	1, 10, 7, 9, 13, 4, 6, 14, 5, 15, 8, 3, 12, 11, 2	−5.95
Internal unpredictability flow	1, 6, 14, 7, 8, 4, 15, 9, 12, 11, 5, 13, 10, 3, 2	−3.54
Stylistic unpredictability flow	1, 10, 8, 6, 7, 5, 14, 2, 15, 3, 9, 12, 4, 13, 11	−3.83
Variation	1, 5, 4, 8, 12, 15, 14, 13, 11, 2, 9, 6, 7, 3, 10	−2.11
	Inter-onset interval	
Internal unpredictability (perfect)	1, 10, 8, 4, 9, 12, 3, 7, 11, 13, 6, 5, 2, 14, 15	−3.43
Internal unpredictability (buffer)	1, 10, 5, 15, 3, 14, 8, 9, 6, 11, 2, 13, 4, 12, 7	−5.15
Internal unpredictability flow	1, 6, 8, 15, 9, 13, 11, 5, 14, 4, 10, 2, 7, 3, 12	−4.22
Stylistic unpredictability flow	1, 10, 6, 8, 11, 2, 4, 15, 7, 14, 12, 3, 13, 9, 5	−3.73
Variation	1, 15, 6, 7, 3, 4, 14, 12, 5, 2, 11, 13, 8, 10, 9	−1.90
	Other	
Tonality	1, 10, 15, 3, 2, 5, 9, 4, 11, 14, 8, 13, 12, 6, 7	−4.60

Measure	Maximum Order	Value
	Pitch interval	
Internal unpredictability (perfect)	1, 6, 7, 15, 8, 4, 2, 14, 11, 5, 13, 9, 3, 10, 12	2.55
Internal unpredictability (buffer)	1, 13, 12, 2, 6, 9, 10, 7, 5, 8, 11, 4, 15, 14, 3	2.06
Internal unpredictability flow	1, 12, 3, 2, 11, 15, 14, 9, 10, 7, 13, 4, 6, 5, 8	3.40
Stylistic unpredictability flow	1, 3, 8, 9, 2, 11, 7, 14, 15, 13, 6, 12, 10, 4, 5	2.64
Variation	1, 7, 9, 10, 6, 14, 3, 2, 8, 11, 15, 13, 5, 12, 4	2.42
	Inter-onset interval	
Internal unpredictability (perfect)	1, 14, 6, 4, 15, 2, 3, 7, 10, 5, 8, 9, 13, 11, 12	3.77
Internal unpredictability (buffer)	1, 9, 10, 11, 3, 13, 6, 12, 4, 5, 14, 7, 2, 15, 8	2.77
Internal unpredictability flow	1, 10, 4, 14, 2, 11, 15, 3, 8, 5, 12, 9, 7, 6, 13	3.40
Stylistic unpredictability flow	1, 11, 6, 14, 2, 12, 10, 5, 8, 3, 15, 13, 7, 9, 4	2.52
Variation	1, 12, 7, 13, 8, 4, 11, 14, 10, 9, 5, 6, 15, 3, 2	2.45
	Other	
Tonal distance	1, 11, 7, 3, 2, 8, 4, 14, 15, 5, 6, 13, 9, 12, 10	2.70

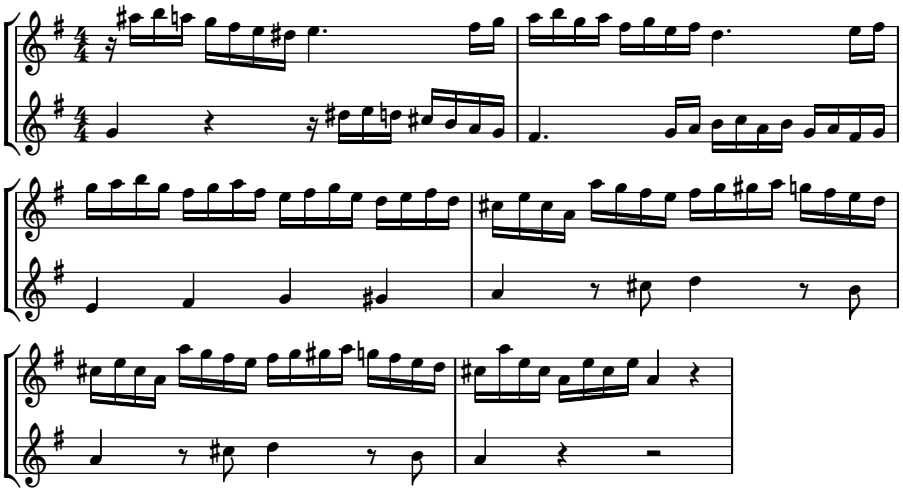
Segment 1



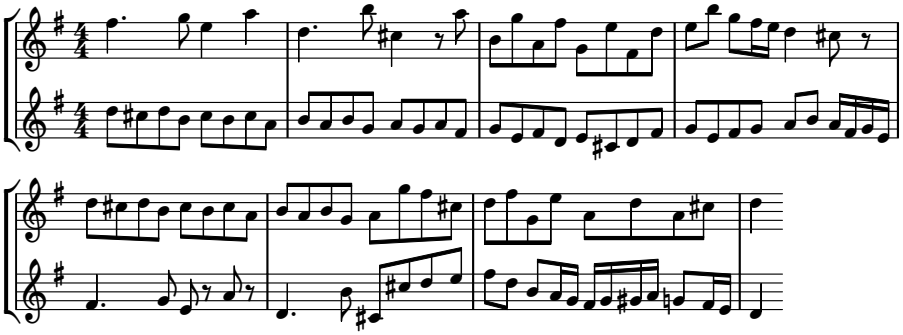
Segment 2



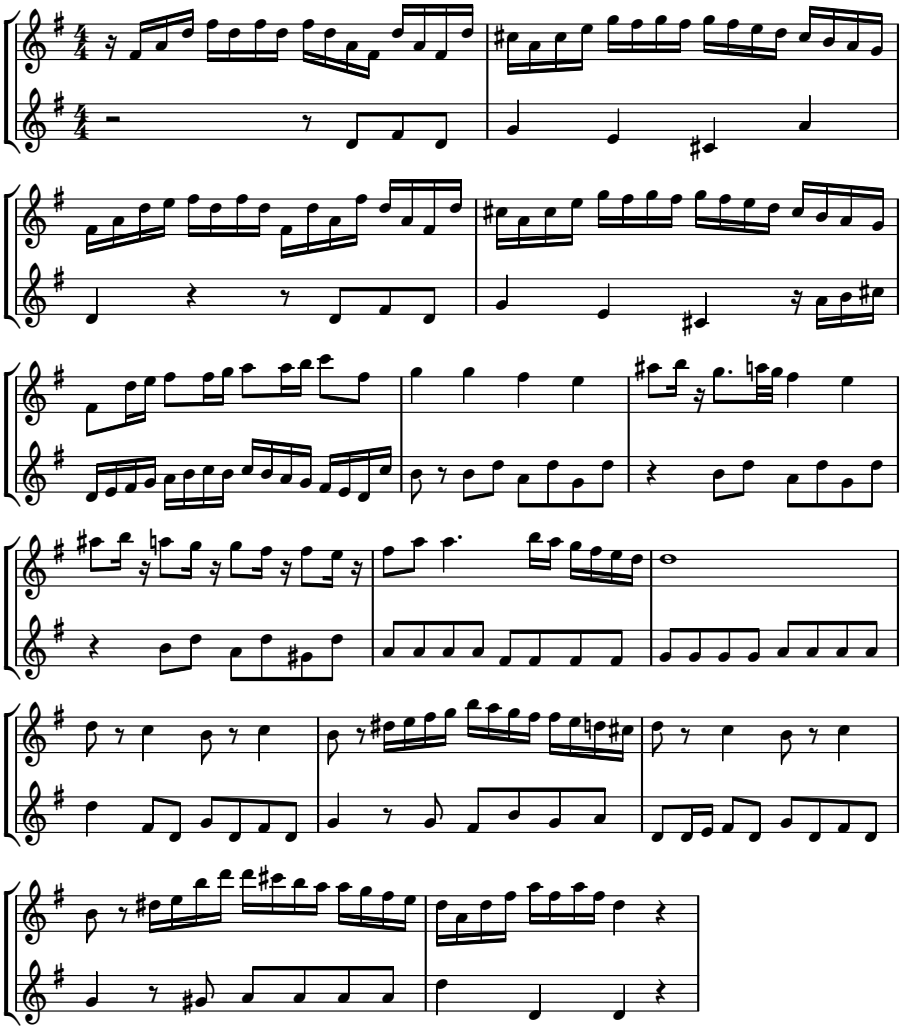
Segment 3



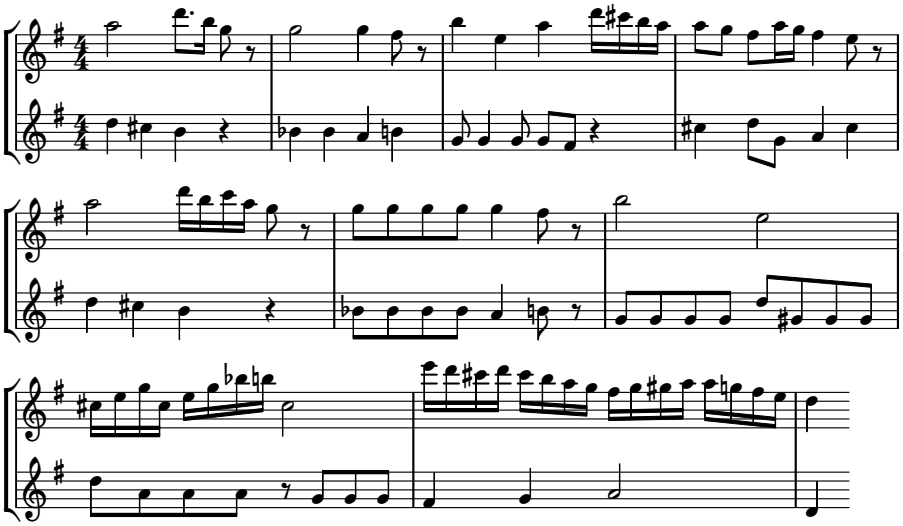
Segment 4



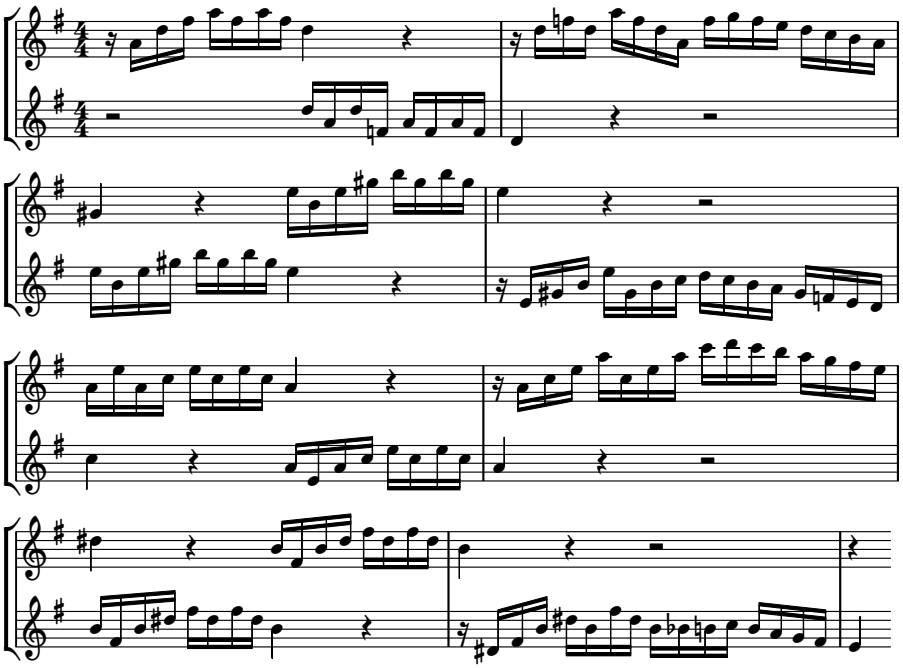
Segment 5



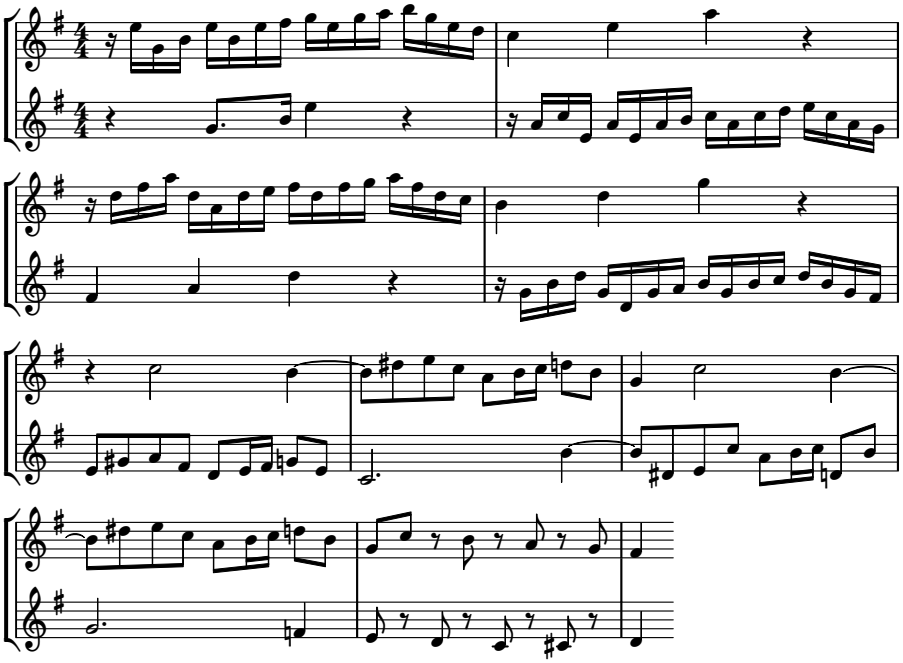
Segment 6



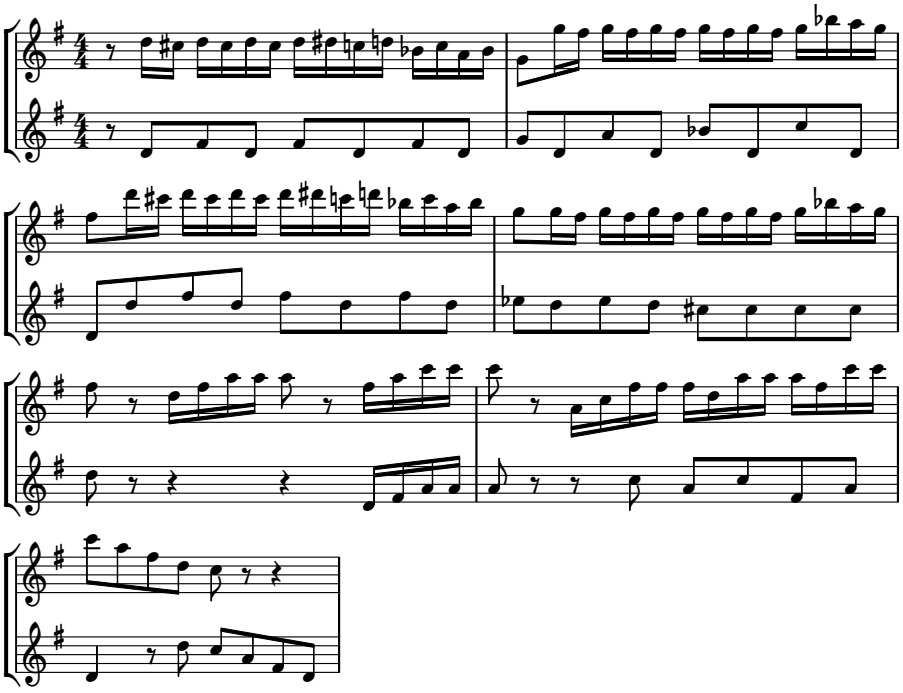
Segment 7



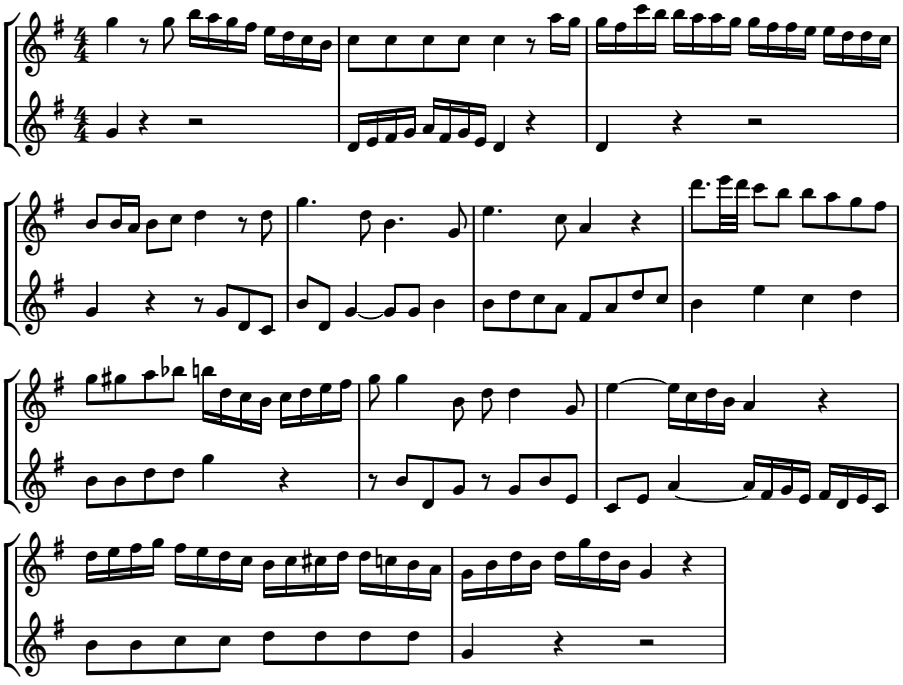
Segment 8



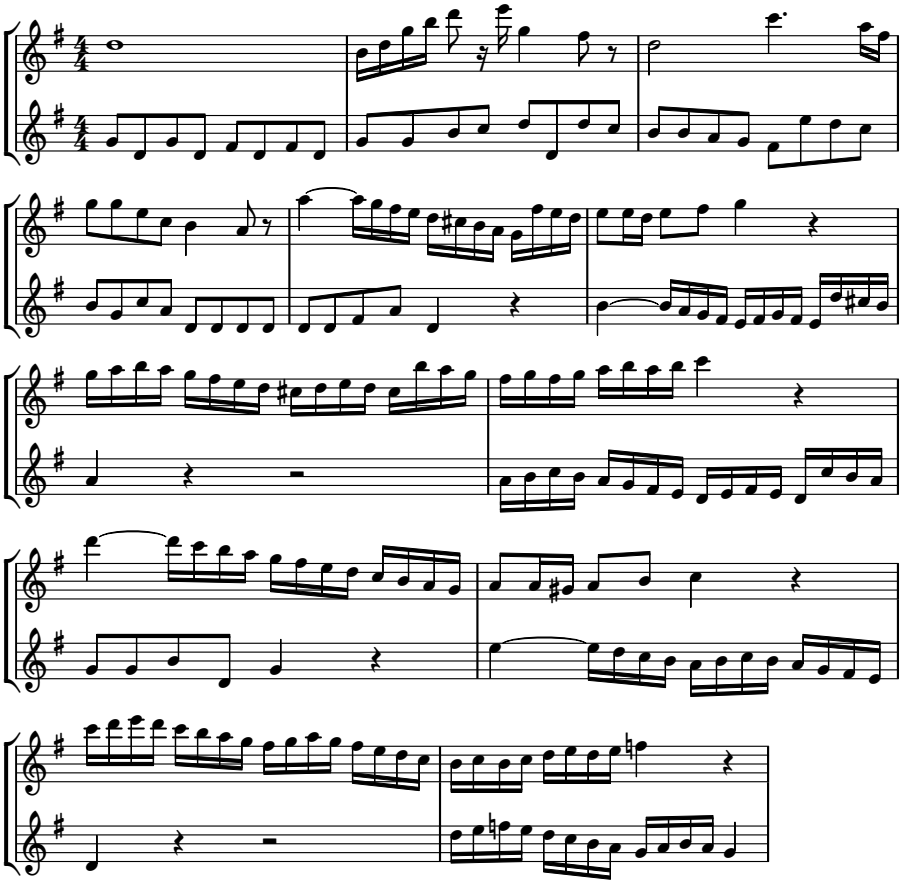
Segment 9



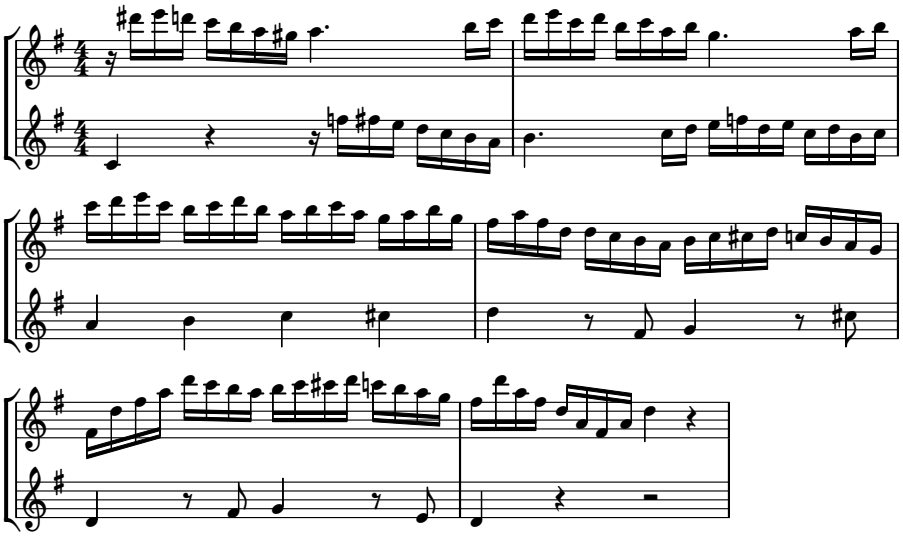
Segment 10



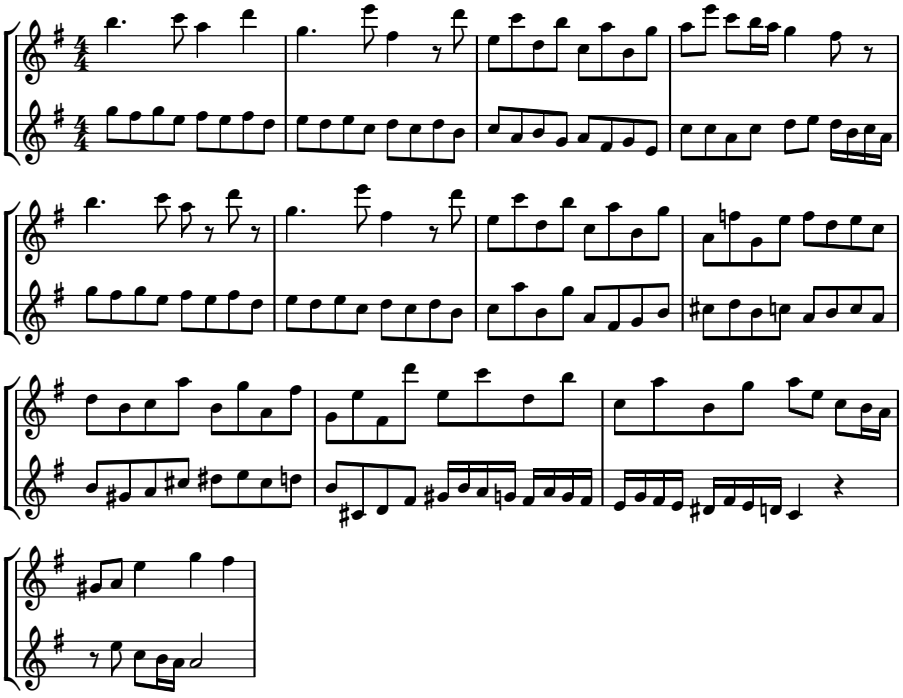
Segment 11



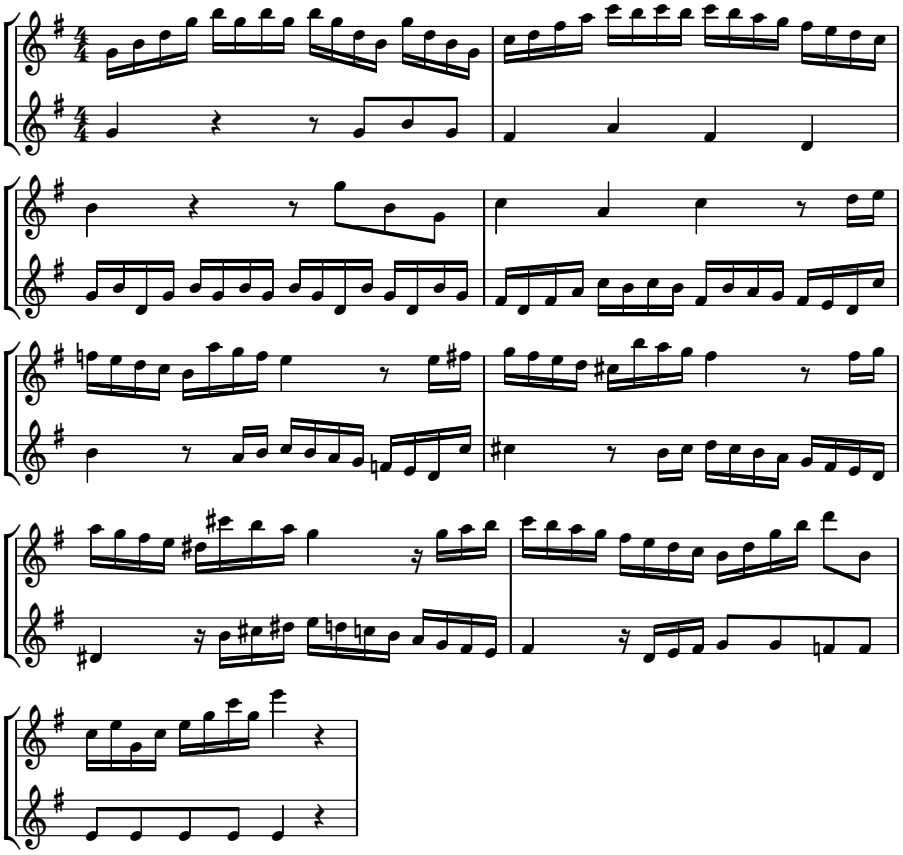
Segment 12



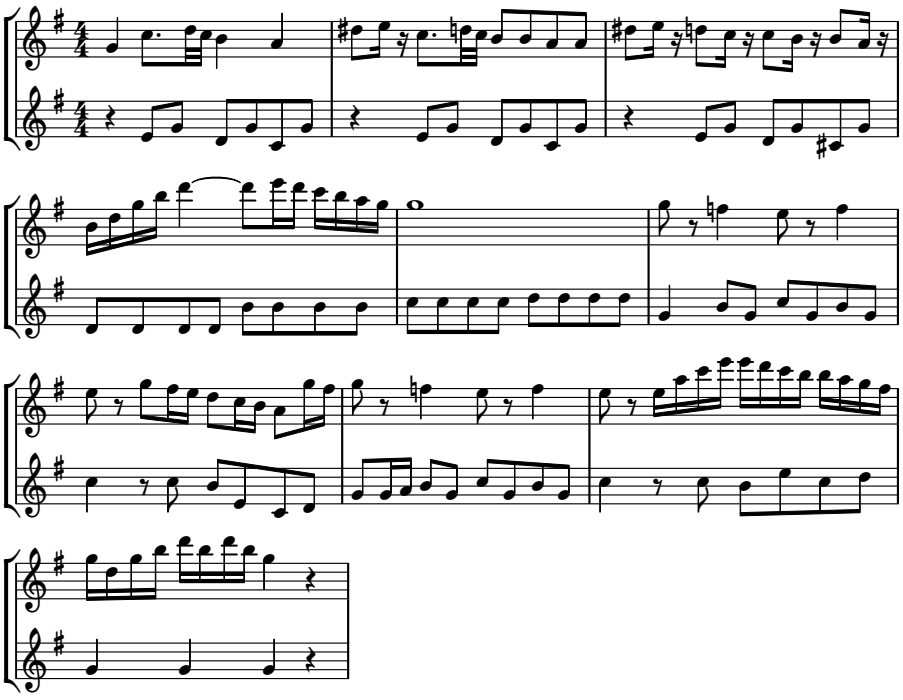
Segment 13



Segment 14



Segment 15



References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*. https://doi.org/10.1002/wics.101

Agres, K., Abdallah, S., & Pearce, M. T. (2018). Information-theoretic properties of auditory sequences dynamically influence expectation and memory. *Cognitive Science*, *42*, 43–76. https://doi.org/10.1111/cogs.12477

Agus, T. R., & Pressnitzer, D. (2013). The detection of repetitions in noise before and after perceptual learning. *The Journal of the Acoustical Society of America*, *134*, 464–473. https://doi.org/10.1121/1.4807641

Ahlbäck, S. (2007). Melodic similarity as a determinant of melody structure. *Musicae Scientiae*, *11*, 235–280. https://doi.org/10.1177/102986490701100110

Allanbrook, W. J. (1992). Two threads through the labyrinth: Topic and process in the first movement of K.332 and K.333. In W. J. Allanbrook, J. M. Levy & W. P. Mahrt (Eds.). Pendragon Press.

Aloupis, G., Fevens, T., Langerman, S., Matsui, T., Mesa, A., Nuñez, Y., Rappaport, D., & Toussaint, G. (2006). Algorithms for computing geometric measures of melodic similarity. *Computer Music Journal*, *30*, 67–76. https://doi.org/10.1162/comj.2006.30.3.67

Aschenbrenner, K. (1985). Coherence in music. Springer, Dordrecht. https://doi.org/https://doi.org/10.1007/978-94-009-5327-7_18

Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 501–515. https://doi.org/10.1037/0096-1523.6.3.501

Bartlett, J. C., & Dowling, W. J. (1988). Scale structure and similarity of melodies. *Music Perception*, *5*, 285–314. https://doi.org/10.2307/40285401

Batt, R. (1987). Comments on "The effects of instrumentation, playing style, and structure in the 'Goldberg Variations' by Johann Sebastian Bach". *Music Perception*, *5*, 207–213. https://doi.org/10.2307/40285393

Beach, D. (1994). The initial movements of Mozart's Piano Sonatas K. 280 and K. 332: Some striking similarities. *Intégral*, *8*, 125–146.

Beghin, T. (2014). Recognizing musical topics versus executing rhetorical figures. In D. Mirka (Ed.). Oxford University Press.

Bell, T. C., Cleary, J. G., & Witten, I. H. (1990). *Text compression*. Prentice Hall.

Benetos, E. (2015). A SyMMuS workshop on audio-symbolic music similarity modelling.

Bianco, R., Harrison, P. M. C., Hu, M., Bolger, C., Picken, S., Pearce, M. T., & Chait, M. (2020). Long-term implicit memory for sequential auditory patterns in humans. *eLife*, *9*, 1–6. https://doi.org/10.7554/eLife.56073

Bohak, C., & Marolt, M. (2009). Calculating similarity of folk song variants with melody-based features. *Proceedings of the International Conference on Music Information Retrieval*.

Boot, P., Volk, A., & de Haas, W. B. (2016). Evaluating the role of repeated patterns in folk song classification and compression. *Journal of New Music Research*, *45*, 223–238. https://doi.org/10.1080/09298215.2016.1208666

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT Press.

Bunton, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, *40*, 76–93. https://doi.org/10.1093/comjnl/40.2_and_3.76

Buteau, C., & Mazzola, G. (2008). Motivic analysis according to Rudolph Réti: Formalization by a topological model. *Journal of Mathematics and Music*, *2*, 117–134. https://doi.org/10.1080/17459730802518292

Cambouropoulos, E. (2001). The Local Boundary Detection Model (LBDM) and its application in the study of expressive timing. *Proceedings of the International Computer Music Conference, Havana*, 17–22.

Cambouropoulos, E. (2006). Musical parallelism and melodic segmentation: A computational approach. *Music Perception*, *23*(3), 249–268. https://doi.org/10.1525/mp.2006.23.3.249

Cambouropoulos, E. (2008). Voice and stream: Perceptual and computational modeling of voice separation. *Music Perception*, *26*. https://doi.org/10.1525/mp.2008.26.1.75

Cambouropoulos, E. (2009). How similar is similar? *Musicae Scientiae*, *13*, 7–24. https://doi.org/10.1177/102986490901300102

Caplin, W. E. (2001). *Classical Form: A theory of formal functions for the instrumental music of Haydn, Mozart, and Beethoven*. Oxford University Press, USA.

Cenkerová, Z., Hartmann, M., & Toiviainen, P. (2018). Crossing phrase boundaries in music. In A. Georgaki & A. Andreopoulou (Eds.).

Chen, T.-P., & Su, L. (2017). Discovery of repeated themes and sections with pattern clustering. *Music Information Retrieval Evaluation eXchange (MIREX 2017)*.

Cheung, V. K. M., Harrison, P. M. C., Meyer, L., Pearce, M. T., Haynes, J. D., & Koelsch, S. (2019). Uncertainty and surprise jointly predict musical pleasure and Amygdala, Hippocampus, and Auditory Cortex activity. *Current Biology*, *29*, 4084–4092.e4. https://doi.org/10.1016/j.cub.2019.09.067

Christensen, R. H. B. (2022). *Ordinal: Regression models for ordinal data*.

Cleary, J. G., & Teahan, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, *40*, 67–75. https://doi.org/10.1093/comjnl/40.2_and_3.67

Cleary, J. G., & Witten, I. H. (1984). Data compression using adaptive coding and partial string matching. *IEEE Transactions on Communications*, *32*, 396–402. https://doi.org/10.1109/TCOM.1984.1096090

Collins, T., Arzt, A., Flossmann, S., & Widmer, G. (2013). SIARCT-CFP: improving precision and the discovery of inexact musical patterns in point-set representations. *Proceedings of the 14th International Society for Music Information Retrieval Conference*, 549–554.

Conklin, D. (2010). Discovery of distinctive patterns in music. *Intelligent Data Analysis*, *14*, 547–554. https://doi.org/10.3233/IDA-2010-0438

Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*, 51–73. https://doi.org/https://doi.org/10.1080/09298219508570672

Cook, N. (1987). The perception of large-scale tonal closure. *Music Perception*, *5*, 197–205. https://doi.org/10.2307/40285392

Cross, I. (1998). Music analysis and music perception. *Music Analysis*, *17*, 3. https://doi.org/10.2307/854368

Davis, S. (2006). Implied polyphony in the solo string works of J. S. Bach: A case for the perceptual relevance of structural expression. *Music Perception*, *23*, 423–446. https://doi.org/10.1525/mp.2006.23.5.423

de Carvalho Jr, A., & Batista, L. (2012). SMS identification using PPM, psychophysiological concepts, and melodic and rhythmic elements. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*, 1–12. https://doi.org/10.3758/s13428-014-0458-y

Deliège, I. (2007). Similarity relations in listening to music: How do they come into play? *Musicae Scientiae*, *11*, 9–37. https://doi.org/10.1177/102986490701100102l

Deliège, I., Mélen, M., Stammers, D., & Cross, I. (1996). Musical schemata in real-time listening to a piece of music. *Music Perception*, *14*, 117–159. https://doi.org/10.2307/40285715

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Downie, J. S. (2004). The scientific evaluation of Music Information Retrieval systems: Foundations and future. *Computer Music Journal*, *28*, 12–23. https://doi.org/10.1162/014892604323112211

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology*, *29*, 247–255. https://doi.org/10.1250/ast.29.247

Downie, J. S., Ehmann, A. F., Bay, M., & Jones, M. C. (2010). The Music Information Retrieval Evaluation eXchange: Some observations and insights. In Z. W. Raś & A. A. Wieczorkowska (Eds.). Springer. https://doi.org/10.1007/978-3-642-11674-2_5

Drabkin, W. (2001a). Motif. *Oxford Music Online*.

Drabkin, W. (2001b). Theme. *Oxford Music Online*.

Egermann, H., Pearce, M. T., Wiggins, G. A., & McAdams, S. (2013). Probabilistic models of expectation violation predict psychophysiological emotional responses to live concert music. *Cognitive, Affective, & Behavioral Neuroscience*, *13*, 533–553. https://doi.org/https://doi.org/10.3758/s13415-013-0161-y

Eitan, Z., & Granot, R. Y. (2008). Growing oranges on Mozart's apple tree: "Inner form" and aesthetic judgment. *Music Perception*, *25*, 397–418. https://doi.org/10.1525/mp.2008.25.5.397

Epstein, D. (1980). Beyond Orpheus: Studies in musical structure. *Journal of Aesthetics and Art Criticism*, *38*, 480–482.

Farbood, M. M. (2016). Memory of a tonal center after modulation. *Music Perception*, *34*, 71–93. https://doi.org/10.1525/mp.2016.34.1.71

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., & Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Frontiers in Neuroscience*, *9*, 157. https://doi.org/10.3389/fnins.2015.00157

Frieler, K., & Müllensiefen, D. (2005). The Simile algorithm for melodic similarity. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

Frieler, K. (2006). Generalized n-gram measures for melodic similarity. In V. Batagelj, H. Bock, A. Ferligoj & A. Žiberna (Eds.). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-34416-0_31

Galand, J. (2014). Topics and tonal processes. In D. Mirka (Ed.). Oxford University Press.

Galeazzi, F. (1796/2012). *The theoretical-practical elements of music, parts III and IV* (D. Burton & G. W. Harwood, Eds.). University of Illinois Press.

Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G., & McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 594–609. https://doi.org/10.1037/xhp0000141

Gjerdingen, R. (1999). An experimental music theory? In N. Cook & M. Everist (Eds.). Oxford University Press.

Gold, B. P., Pearce, M. T., Mas-Herrero, E., Dagher, A., & Zatorre, R. J. (2019). Predictability and uncertainty in the pleasure of music: A reward for learning? *The Journal of Neuroscience*, *39*, 9397–9409. https://doi.org/10.1523/JNEUROSCI.0428-19.2019

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.). Cambridge University Press.

Gotlieb, H., & Konečni, V. J. (1985). The effects of instrumentation, playing style, and structure in the Goldberg Variations by Johann Sebastian Bach. *Music Perception*, *3*, 87–101. https://doi.org/10.2307/40285323

Grachten, M., Arcos, J. L., & de Mántaras, R. L. (2004). Melodic Similarity: Looking for a good abstraction level. *Proceedings of the 5th International Conference on Music Information Retrieval*.

Granot, R. Y., & Jacoby, N. (2011). Musically puzzling I: Sensitivity to overall structure in the sonata form? *Musicae Scientiae*, *15*, 365–386. https://doi.org/10.1177/1029864911409508

Granot, R. Y., & Jacoby, N. (2012). Musically puzzling II: Sensitivity to overall structure in a Haydn E-minor sonata. *Musicae Scientiae*, *16*, 67–80. https://doi.org/10.1177/1029864911423146

Grout, D. J., Burkholder, J. P., & Palisca, C. V. (2010). *A history of western music* (8th ed.). W. W. Norton.

Hall, E. T. R., & Pearce, M. T. (2021). A model of large-scale thematic structure. *Journal of New Music Research*, *50*, 220–241. https://doi.org/10.1080/09298215.2021.1930062

Halpern, A. R., Bartlett, J. C., & Dowling, W. J. (1998). Perception of mode, rhythm, and contour in unfamiliar melodies: Effects of age and experience. *Music Perception*, *15*, 335–355. https://doi.org/10.2307/40300862

Hansen, N. C., Kragness, H. E., Vuust, P., Trainor, L., & Pearce, M. T. (2021). Predictive uncertainty underlies auditory boundary perception. *Psychological Science*, *32*, 1416–1425. https://doi.org/10.1177/0956797621997349

Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, *5*, 1052. https://doi.org/10.3389/fpsyg.2014.01052

Hansen, N. C., Vuust, P., & Pearce, M. T. (2016). "If you have to ask, you'll never know": Effects of specialised stylistic expertise on predictive processing of music. *PLOS One*, *11*, e0163584. https://doi.org/10.1371/journal.pone.0163584

Harrison, P. M. C., Bianco, R., Chait, M., & Pearce, M. T. (2020). PPM-Decay: A computational model of auditory prediction with memory decay. *PLOS Computational Biology*, *16*, e1008304. https://doi.org/10.1371/journal.pcbi.1008304

Hatten, R. (2014). The troping of topics in mozart's instrumental works. In D. Mirka (Ed.). Oxford University Press.

Hepokoski, J., & Darcy, W. (2006). *Elements of Sonata Theory: Norms, types, and deformations in the Late-Eighteenth-Century Sonata*. Oxford University Press.

Hsu, J.-L., Liu, C.-C., & Chen, A. L. P. (2001). Discovering nontrivial repeating patterns in music data. *IEEE Transactions on Multimedia*, *3*, 311–325. https://doi.org/10.1109/6046.944475

Huron, D. (2001). What is a musical feature? Forte's analysis of Brahms's Opus 51, No. 1, revisited. *Music Theory Online*, *7*, 69.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT Press.

Huron, D. (2013). A psychological approach to musical form: The habituation–fluency theory of repetition. *Current Musicology*, 7–35. https://doi.org/https://doi.org/10.7916/cm.v0i96.5312

Irving, J. (2010). *Understanding Mozart's Piano Sonatas*. Ashgate Publishing, Ltd.

Janssen, B., Burgoyne, J. A., & Honing, H. (2017). Predicting variation of folk songs: A corpus analysis study on the memorability of melodies. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00621

Janssen, B., de Haas, W. B., Volk, A., & van Kranenburg, P. (2014). Finding repeated patterns in music: State of knowledge, challenges, perspectives. In M. Aramaki, O. Derrien, R. Kronland-Martinet & S. Ystad (Eds.). Springer International Publishing. https://doi.org/10.1007/978-3-319-12976-1_18

Janssen, B., van Kranenburg, P., & Volk, A. (2017). Finding occurrences of melodic segments in folk songs employing symbolic similarity measures. *Journal of New Music Research*, *46*, 118–134. https://doi.org/10.1080/09298215.2017.1316292

Juhász, Z. (2004). Segmentation of Hungarian folk songs using an entropy-based learning system. *Journal of New Music Research*, *33*(1), 5–15. https://doi.org/10.1076/jnmr.33.1.5.35395

Karno, M., & Konečni, V. J. (1992). The effects of structural interventions in the first movement of Mozart's Symphony in G Minor K. 550 on aesthetic preference. *Music Perception*, *10*, 63–72. https://doi.org/10.2307/40285538

Karydis, I., Nanopoulos, A., & Manolopoulos, Y. (2007). Finding maximum-length repeating patterns in music databases. *Multimedia Tools and Applications*, *32*, 49–71. https://doi.org/10.1007/s11042-006-0068-5

Kinderman, W. (2006). *Mozart's Piano Music*. Oxford University Press.

Kivy, P. (1993). *The fine art of repetition: Essays in the philosophy of music*. Cambridge University Press.

Kivy, P. (2017). On the recent remarriage of music to philosophy. *The Journal of Aesthetics and Art Criticism*, *75*, 429–438. https://doi.org/10.1111/jaac.12402

Koelsch, S., Rohrmeier, M., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences*, *110*, 15443–15448. https://doi.org/10.1073/pnas.1300272110

Konečni, V. J. (1984). Elusive effects of artists' 'messages'. In W. R. Crozier & A. J. Chapman (Eds.). North-Holland. https://doi.org/10.1016/S0166-4115(08)62346-8

Koniari, D., Predazzer, S., & Mélen, M. (2001). Categorization and schematization processes used in music perception by 10- to 11-year-old children. *Music Perception*, *18*, 297–324. https://doi.org/10.1525/mp.2001.18.3.297

Kramer, J. D. (2004). The concept of disunity and musical analysis. *Music Analysis*, *23*, 361–372. https://doi.org/10.1111/j.0262-5245.2004.00210.x

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. Oxford University Press.

Krumhansl, C. L. (1991). Memory for musical surface. *Memory & Cognition*, *19*, 401–411. https://doi.org/10.3758/BF03197145

Laaksonen, A., & Lemström, K. (2019). Transposition and time-warp invariant algorithm for detecting repeated patterns in polyphonic music. *6th International Conference on Digital Libraries for Musicology*, 38–42. https://doi.org/10.1145/3358664.3358670

Laaksonen, A., & Lemström, K. (2021). Discovering distorted repeating patterns in polyphonic music through longest increasing subsequences. *Journal of Mathematics and Music*, *15*, 99–111. https://doi.org/10.1080/17459737.2021.1896811

Laitinen, M., & Lemström, K. (2010). Geometric algorithms for melodic similarity. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

Lalitte, P., & Bigand, E. (2006). Music in the moment? revisiting the effect of large scale structures. *Perceptual and Motor Skills*, *103*, 811–828. https://doi.org/10.2466/PMS.103.3.811-828

Lamont, A., & Dibben, N. (2001). Motivic structure and the perception of similarity. *Music Perception*, *18*, 245–274. https://doi.org/10.1525/mp.2001.18.3.245

Laney, R., Samuels, R., & Capulet, E. (2015). Cross entropy as a measure of musical contrast. In T. Collins, D. Meredith & A. Volk (Eds.). Springer, Cham. https://doi.org/10.1007/978-3-319-20603-5_20

Lartillot, O. (2014). In-depth motivic analysis based on multiparametric closed pattern and cyclic sequence mining. *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 361–366.

Lattner, S., Grachten, M., Agres, K., & Cancino Chacón, C. E. (2015). Probabilistic segmentation of musical sequences using restricted boltzmann machines. In T. Collins, D. Meredith & A. Volk (Eds.), *Mathematics and computation in music* (pp. 323–334). Springer International.

Lemström, K. (2010). Towards more robust geometric content-based music retrieval. *Proceedings of the Conference of the International Society for Music Information Retrieval*.

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

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, *10*, 707–710.

Li, M., Chen, X., Li, X., Ma, B., & Vitányi, P. (2003). The similarity metric. *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 863–872.

Lu, L., & Zhang, H.-J. (2003). Automated extraction of music snippets. *Proceedings of the Eleventh ACM International Conference on Multimedia*, 140–147. https://doi.org/10.1145/957013.957043

Margulis, E. H. (2012). Musical repetition detection across multiple exposures. *Music Perception*, *29*, 377–385. https://doi.org/10.1525/mp.2012.29.4.377

Margulis, E. H. (2013). Aesthetic responses to repetition in unfamiliar music. *Empirical Studies of the Arts*, *31*, 45–57. https://doi.org/10.2190/EM.31.1.c

Margulis, E. H. (2014). *On repeat: How music plays the mind*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199990825.001.0001

Margulis, E. H. (2020). Repetition. In A. Rehding & S. Rings (Eds.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190454746.013.22

Marsden, A. (2010). Recognition of variations using automatic Schenkerian Reduction. *Proceedings of the International Society of Music Information Retrieval*.

Marvin, E. W., & Brinkman, A. (1999). The effect of modulation and formal manipulation on perception of tonic closure by expert listeners. *Music Perception*, *16*, 389–408. https://doi.org/10.2307/40285801

McAdams, S., Vines, B. W., Vieillard, S., Smith, B. K., & Reynolds, R. (2004). Influences of large-scale form on continuous ratings in response to a contemporary piece in a live concert setting. *Music Perception*, *22*, 297–350. https://doi.org/10.1525/mp.2004.22.2.297

Meek, C., & Birmingham, W. P. (2001). Thematic extractor. *Proceedings of the International Society of Music Information Retrieval Conference 2001*, 119–128.

Mélen, M., & Wachsmann, J. (2001). Categorization of musical motifs in infancy. *Music Perception*, *18*, 325–346. https://doi.org/10.1525/mp.2001.18.3.325

Melkonian, O., Ren, I. Y., Swierstra, W., & Volk, A. (2019). What constitutes a musical pattern? *Proceedings of the 7th ACM SIGPLAN International Workshop on Functional Art, Music, Modeling, and Design*, 95–105. https://doi.org/10.1145/3331543.3342587

Meredith, D. (2015). Music analysis and point-set compression. *Journal of New Music Research*, *44*(3), 245–270. https://doi.org/10.1080/09298215.2015.1045003

Meredith, D., Lemström, K., & Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, *31*, 321–345. https://doi.org/10.1076/jnmr.31.4.321.14162

Meyer, L. B. (1956). *Emotion and meaning in music*. University of Chicago Press.

Meyer, L. B. (1989). *Style and music: Theory, history, and ideology*. University of Chicago Press.

Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Transactions on Communications*, *38*, 1917–1921. https://doi.org/10.1109/26.61469

Müllensiefen, D., & Frieler, K. (2004). Optimizing measures of melodic similarity for the exploration of a large folk song database. *Proceedings of the International Conference on Music Information Retrieval*.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLOS One*, *9*, e89642. https://doi.org/10.1371/journal.pone.0089642

Müller, M. (2015). *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer. https://doi.org/10.1007/978-3-319-21945-5

Narmour, E. (1990). *The analysis and cognition of basic melodic structures: The implication-realization model.* University of Chicago Press.

Narmour, E. (1992). *The analysis and cognition of melodic complexity: The Implication-realisation Model.* University of Chicago Press.

Nettl, B. (2010). *The study of ethnomusicology: Thirty-one issues and concepts*. University of Illinois Press.

Nieto, O., & Farbood, M. (2014). Identifying polyphonic patterns from audio recordings using music segmentation techniques. *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 411–416.

Nieto, O., & Bello, J. P. (2016). Systematic exploration of computational music structure research. *Proceedings of the 17th International Society for Music Information Retrieval Conference*, 547–553.

Nieto, O., Mysore, G. J., Wang, C.-i., Smith, J. B., Schlüter, J., Grill, T., & McFee, B. (2020). Audio-based music structure analysis: Current trends, open challenges, and applications. *Transactions of the International Society for Music Information Retrieval*, *3*(1). https://doi.org/10.5334/tismir.54

Ockelford, A. (1991). The role of repetition in perceived musical structures. *Representing musical structure*, 129–160.

Ockelford, A. (2004). On similarity, derivation and the cognition of musical structure. *Psychology of Music*, *32*, 23–74. https://doi.org/10.1177/0305735604039282

Ockelford, A. (2005). *Repetition in music*. Routledge. https://doi.org/10.4324/9781315088884

Ockelford, A. (2006). Implication and expectation in music: A zygonic model. *Psychology of Music*, *34*, 81–142. https://doi.org/10.1177/0305735606059106

Ockelford, A. (2009). Similarity relations between groups of notes: Music-theoretical and music-psychological perspectives. *Musicae Scientiae*, *13*, 47–98. https://doi.org/10.1177/102986490901300104

Ockelford, A. (2010). Exploring the structural principles underlying the capacity of groups of notes to function concurrently in music. *Musicae Scientiae*, *14*, 149–185. https://doi.org/10.1177/10298649100140S210

Omigie, D., Pearce, M. T., & Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, *50*, 1483–1493. https://doi.org/10.1016/j.neuropsychologia.2012.02.034

Omigie, D., Pearce, M. T., Williamson, V. J., & Stewart, L. (2013). Electrophysiological correlates of melodic processing in congenital amusia. *Neuropsychologia*, *51*, 1749–1762. https://doi.org/10.1016/j.neuropsychologia.2013.05.010

Orio, N., & Rodà, A. (2009). A measure of melodic similarity based on a graph representation of the music structure. *Proceedings of the 10th International Society for Music Information Retrieval Conference*.

Parks, R. (2003). Music's inner dance: Form, pacing and complexity in Debussy's music. In S. Trezise (Ed.). Cambridge University Press. https://doi.org/10.1017/CCOL9780521652438.013

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

Pearce, M. T. (2018). Statistical learning and probabilistic prediction in music cognition: Mechanisms of stylistic enculturation. *Annals of the New York Academy of Sciences*, *1423*, 378–395. https://doi.org/10.1111/nyas.13654

Pearce, M. T., & Müllensiefen, D. (2017). Compression-based modelling of musical similarity perception. *Journal of New Music Research*, *46*, 135–155. https://doi.org/10.1080/09298215.2017.1305419

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, *39*, 1367–1391. https://doi.org/10.1068/p6507

Pearce, M. T., & Rohrmeier, M. (2018). Musical syntax II: Empirical perspectives. In R. Bader (Ed.). Springer-Verlag. https://doi.org/10.1007/978-3-662-55004-5_26

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*, 302–313. https://doi.org/10.1016/j.neuroimage.2009.12.019

Pesek, M., Leonardis, A., & Marolt, M. (2017). SymCHM—an unsupervised approach for pattern discovery in symbolic music with a compositional hierarchical model. *Applied sciences*, *7*(11), 1135. https://doi.org/10.3390/app7111135

Pollard-Gott, L. (1983). Emergence of thematic concepts in repeated listening to music. *Cognitive Psychology*, *15*, 66–94. https://doi.org/10.1016/0010-0285(83)90004-X

Potter, C. (2003). Debussy and nature. In S. Trezise (Ed.). Cambridge University Press. https://doi.org/10.1017/CCOL9780521652438.010

Prince, J. B. (2014). Contributions of pitch contour, tonality, rhythm, and meter to melodic similarity. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 2319–2337. https://doi.org/10.1037/a0038010

Prince, J. B., Thompson, W. F., & Schmuckler, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, *35*, 1598–1617. https://doi.org/10.1037/a0016456

Rafael, B., Oertl, S., Affenzeller, M., & Wagner, S. (2009). Using heuristic optimization for segmentation of symbolic music. *International Conference on Computer Aided Systems Theory*, 641–648.

Rafael, B., & Oertl, S. M. (2010). MTSSM—a framework for multi-track segmentation of symbolic music. *International Journal of Computer and Information Engineering*, *4*(1), 7–13.

R-Core-Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ren, I. Y., Koops, H. V., Volk, A., & Swierstra, W. (2017). In search of the consensus among musical pattern discovery algorithms. *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 671–679.

Reti, R. (1978). *The thematic process in music*. Greenwood Press.

Reynolds, D. (2009). Gaussian mixture models (S. Z. Li & A. Jain, Eds.). *Encyclopedia of Biometrics*, 659–663. https://doi.org/10.1007/978-0-387-73003-5_196

Reynolds, R. (2002). *Form and method: Composing music* (S. McAdams, Ed.). Routledge.

Rizo, D., & Inesta, J. M. (2010). Trees and combined methods for monophonic music similarity evaluation. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

Robert, C. P., & Casella, G. (2004). The Metropolis—Hastings algorithm. https://doi.org/10.1007/978-1-4757-4145-2_7

Rodríguez-López, M. E., & Volk, A. (2015). Location constraints for repetition-based segmentation of melodies. In T. Collins, D. Meredith & A. Volk (Eds.), *Mathematics and computation in music* (pp. 73–84). Springer International. https://doi.org/10.1007/978-3-319-20603-5_7

Rohrmeier, M., & Pearce, M. (2018). Musical syntax I: Theoretical perspectives. In R. Bader (Ed.). Springer-Verlag. https://doi.org/10.1007/978-3-662-55004-5_25

Roig, C., Tardón, L. J., Barbancho, A. M., & Barbancho, I. (2013). Submission to MIREX 2013 symbolic melodic similarity. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

Rolison, J. J., & Edworthy, J. (2012). The role of formal structure in liking for popular music. *Music Perception*, *29*, 269–284. https://doi.org/10.1525/mp.2012.29.3.269

Rumph, S. (2014). Topical figurae: The double articulation of topics. In D. Mirka (Ed.). Oxford University Press.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52. https://doi.org/10.1016/S0010-0277(98)00075-4

Salzer, F. (1962). *Structural hearing: Tonal coherence in music*. Dover Publications.

Sauvé, S. A. (2018). *Prediction in polyphony: Modelling musical auditory scene analysis*. Queen Mary University of London.

Schenker, H. (1994). *The masterwork in music: A yearbook. vol. 1, (1925)* (W. Drabkin & I. Bent, Eds.). Cambridge University Press.

Schoenberg, A. (1967). *Fundamentals of musical composition* (G. Strang & L. Stein, Eds.; 4th ed.). Faber & Faber.

Serra, J., Müller, M., Grosche, P., & Arcos, J. L. (2012). Unsupervised detection of music boundaries by time series structure features. *Proceedings of the AAAI Conference on Artificial Intelligence*, *26*(1), 1613–1619.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. https://doi.org/10.1126/science.3629243

Sleator, D., & Temperley, D. (2003). *The Melisma Music Analyzer*.

Smyth, D. (1993). "balanced interruption" and the formal repeat. *Music Theory Spectrum*, *15*, 76–88. https://doi.org/10.2307/745910

Spyra, J., Stodolak, M., & Woolhouse, M. (2021). Events versus time in the perception of nonadjacent key relationships. *Musicae Scientiae*, *25*, 212–225. https://doi.org/10.1177/1029864919867463

Suyoto, I. S. H., & Uitdenbogerd, A. L. (2010). Simple orthogonal pitch with IOI symbolic music matching. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

Takasu, A., Yanase, T., Kanazawa, T., & Adachi, J. (1999). Music structure analysis and its application to theme phrase extraction. In S. Abiteboul & A.-M. Vercoustre (Eds.). Springer Berlin Heidelberg.

Tan, S.-L., & Spackman, M. P. (2005). Listeners' judgments of the musical unity of structurally altered and intact musical compositions. *Psychology of Music*, *33*, 133–153. https://doi.org/10.1177/0305735605050648

Tan, S.-L., Spackman, M. P., & Peaslee, C. L. (2006). The effects of repeated exposure on liking and judgments of musical unity of intact and patchwork compositions. *Music Perception*, *23*, 407–421. https://doi.org/10.1525/mp.2006.23.5.407

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

Temperley, D. (2007). *Music and probability*. MIT Press.

Temperley, D. (2009). A unified probabilistic model for polyphonic music analysis. *Journal of New Music Research*, *38*, 3–18. https://doi.org/10.1080/09298210902928495

Temperley, D. (2014). Information flow and repetition in music. *Journal of Music Theory*, *58*, 155–178. https://doi.org/10.1215/00222909-2781759

Temperley, D. (2019). Uniform information density in music. *Music Theory Online*, *25*. https://doi.org/10.30535/mto.25.2.5

Tharwat, A. (2018). Independent component analysis: An introduction. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2018.08.006

Tillmann, B., & Bigand, E. (1996). Does formal musical structure affect perception of musical expressiveness? *Psychology of Music*, *24*, 3–17. https://doi.org/10.1177/0305735696241002

Tillmann, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *The Journal of Aesthetics and Art Criticism*, *62*. https://doi.org/10.1111/j.1540-594X.2004.00153.x

Tillmann, B., Bigand, E., & Madurell, F. (1998). Local versus global processing of harmonic cadences in the solution of musical puzzles. *Psychological Research*, *61*, 157–174. https://doi.org/10.1007/s004260050022

Toiviainen, P. (2007). Similarity perception in listening to music. *Musicae Scientiae*, *11*, 3–6. https://doi.org/10.1177/1029864907011001011

Typke, R., den Hoed, M., de Nooijer, J., Wiering, F., & Veltkam, R. C. (2005). A ground truth for half a million musical incipits. *Journal of Digital Information Management*, *3*, 34–39.

Uitdenbogerd, A., & Zobel, J. (1998). Manipulation of music for melody matching. *Proceedings of the sixth ACM international conference on Multimedia - MULTIMEDIA '98*, 235–240. https://doi.org/10.1145/290747.290776

Uitdenbogerd, A., & Zobel, J. (1999). Melodic matching techniques for large music databases. *Proceedings of the seventh ACM international conference on Multimedia (Part 1) - MULTIMEDIA '99*, 57–66. https://doi.org/10.1145/319463.319470

Ullrich, K., Schlüter, J., & Grill, T. (2014). Boundary detection in music structure analysis using convolutional neural networks. *Proceedings of the 15th International Society for Music Information Retrieval Conference*, 417–422.

Urbano, J. (2013). A geometric model supported with hybrid sequence alignment. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

van der Weij, B., Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, *8*, 824. https://doi.org/10.3389/fpsyg.2017.00824

Velarde, G., Meredith, D., & Weyde, T. (2016). A wavelet-based approach to pattern discovery in melodies. In D. Meredith (Ed.), *Computational music analysis* (pp. 303–333). Springer. https://doi.org/10.1007/978-3-319-25931-4_12

Velardo, V., Vallati, M., & Jan, S. (2016). Symbolic melodic similarity: State of the art and future challenges. *Computer Music Journal*, *40*, 70–83. https://doi.org/10.1162/COMJ_a_00359

Vempala, N. N., & Russo, F. A. (2015). An empirically derived measure of melodic similarity. *Journal of New Music Research*, *44*, 391–404. https://doi.org/10.1080/09298215.2015.1080284

Volk, A., de Haas, W. B., & van Kranenburg, P. (2012). Towards modelling variation in music as foundation for similarity. In E. Cambouropoulos, C. C. Tsougras, P. P. Mavromatis & K. Pastiadis (Eds.).

Volk, A., Chew, E., Margulis, E. H., & Anagnostopoulou, C. (2016). Music similarity: Concepts, cognition and computation. *Journal of New Music Research*, *45*, 207–209. https://doi.org/10.1080/09298215.2016.1232412

Volk, A., & van Kranenburg, P. (2012). Melodic similarity among folk songs: An annotation study on similarity-based categorization in music. *Musicae Scientiae*, *16*, 317–339. https://doi.org/10.1177/1029864912448329

Wang, C., Li, J., & Shi, S. (2006). N-gram inverted index structures on music data for theme mining and content-based information retrieval. *Pattern Recognition Letters*, *27*, 492–503. https://doi.org/10.1016/j.patrec.2005.09.012

Welker, R. L. (1982). Abstraction of themes from melodic variations. *Journal of Experimental Psychology: Human Perception and Performance*, *8*, 435–447. https://doi.org/10.1037/0096-1523.8.3.435

Whittall, A. (2001). Form. *Grove Music Online*, 493–513.

Wolkowicz, J., & Kešelj, V. (2011). Text information retrieval approach to music information retrieval. *Proceedings of the Annual Music Information Retrieval Evaluation Exchange*.

Woolhouse, M., Cross, I., & Horton, T. (2016). Perception of nonadjacent tonic-key relationships. *Psychology of Music*, *44*, 802–815. https://doi.org/10.1177/0305735615593409

Yazawa, S., Hasegawa, Y., Kanamori, K., & Hamanaka, M. (2013). Melodic similarity based on extension implication-realization model. *MIREX Symbolic Melodic Similarity Results*.

Ziv, N., & Eitan, Z. (2007). Themes as prototypes: Similarity judgments and categorization tasks in musical contexts. *Musicae Scientiae*, *11*, 99–133. https://doi.org/10.1177/1029864907011001051