Modelling the perception and composition of Western musical harmony

Peter Michael Combes Harrison

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

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

April 20, 2020

Declaration

I, Peter Harrison, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below. I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material. I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis. I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university. The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Date	Peter Harrison

Publications

Early versions of work described in this thesis appeared in the following conference proceedings:

Harrison, P. M. C., & Pearce, M. T. (2018). Dissociating sensory and cognitive theories of harmony perception through computational modeling. In Proceedings of ICMPC15/ESCOM10. Graz, Austria. `https://doi.org/10.31234/osf.io/wgjyv`

Harrison, P. M. C., & Pearce, M. T. (2018). An energy-based generative sequence model for testing sensory theories of Western harmony. In Proceedings of the 19th International Society for Music Information Retrieval Conference (pp. 160–167). Paris, France. `https://arxiv.org/abs/1807.00790`

Chapter 3 and 5 are respectively based on the following journal articles:

Harrison, P. M. C., & Pearce, M. T. (2020). Simultaneous consonance in music perception and composition. *Psychological Review, 127*(2), 216-244. `http://dx.doi.org/10.1037/rev0000169`

Harrison, P. M. C., & Pearce, M. T. (2020). A computational cognitive model for the analysis and generation of voice leadings. *Music Perception, 37*(3), 208-224. `https://doi.org/10.1525/mp.2020.37.3.208`

Early versions of Chapters 2, 3 and 5 were posted as the following preprints:

Harrison, P. M. C., & Pearce, M. T. (2019). Representing harmony in computational music cognition. *PsyArXiv.* `https://doi.org/10.31234/osf.io/xswp4`

Harrison, P. M. C., & Pearce, M. T. (2019). Instantaneous consonance in the perception and composition of Western music. *PsyArXiv*. https://doi.org/10.31234/osf.io/6jsug

Harrison, P. M. C., & Pearce, M. T. (2019). A computational model for the analysis and generation of chord voicings. *PsyArXiv*. https://doi.org/10.31234/osf.io/wrgj7

The following other articles were also produced during the doctoral study period:

Harrison, P. M. C., Musil, J. J., & Müllensiefen, D. (2016). Modelling melodic discrimination tests: Descriptive and explanatory approaches. *Journal of New Music Research, 45*(3), 265–280. https://doi.org/10.1080/09298215.2016.1197953

Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*(7). https://doi.org/10.1038/s41598-017-03586-z

Harrison, P. M. C. & Müllensiefen, D. (2018). Development and validation of the Computerised Adaptive Beat Alignment Test (CA-BAT). *Scientific Reports*(8). https://doi.org/10.1038/s41598-018-30318-8

Harrison, P. M. C. Statistics and Experimental Design for Psychologists: A model comparison approach by Rory Allen (book review). *PsyPAG Quarterly.*

Harrison, P. M. C., Larrouy-Maestri, P., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods, 51*, 663-675. https://doi.org/10.3758/s13428-019-01225-1

Gelding, R., Harrison, P. M. C., Silas, S., Johnson, B. W., Thompson, W. F., & Müllensiefen, D. (2019). Developing an efficient test of musical imagery ability: Applying modern psychometric techniques to the Pitch Imagery Arrow Task. *PsyArXiv*. https://doi.org/10.31234/osf.io/8gvhz

de Fleurian, R., Harrison, P. M. C., Pearce, M. T., & Quiroga-Martinez, D. R. (2019). Reward prediction tells us less than expected about musical pleasure. *Proceedings of the National Academy of Sciences, 116*(42), 20813-20814. https://doi.org/10.1073/pnas.1913244116

Cheung, V., Harrison, P. M. C., 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*(23), 4084-4092.e4. `https://doi.org/10.1016/j.cub.2019.09.067`

Zioga, I., Harrison, P. M. C., Pearce, M. T., Bhattacharya, J., Luft, C. D. B. (2020). From learning to creativity: Identifying the behavioural and neural correlates of learning to predict human judgements of musical creativity. *NeuroImage, 206*. `https://doi.org/10.1016/j.neuroimage.2019.116311`

Harrison, P. M. C., Bianco, R., Chait, M., Pearce, M. T. (2020). PPM-Decay: A computational model of auditory prediction with memory decay. *bioRxiv*. `https://doi.org/10.1101/2020.01.09.900266`

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. *bioRxiv*. `https://doi.org/10.1101/2020.02.14.949404`

Harrison, P. M. C. (2020). psychTestR: An R package for designing and conducting behavioural psychological experiments. *PsyArXiv*. `https://doi.org/10.31234/osf.io/dyar7`

Abstract

Harmony is a fundamental structuring principle in Western music, determining how simultaneously occurring musical notes combine to form chords, and how successions of chords combine to form chord progressions. Harmony is interesting to psychologists because it unites many core features of auditory perception and cognition, such as pitch perception, auditory scene analysis, and statistical learning. A current challenge is to formalise our psychological understanding of harmony through computational modelling. Here we detail computational studies of three core dimensions of harmony: consonance, harmonic expectation, and voice leading. These studies develop and evaluate computational models of the psychoacoustic and cognitive processes involved in harmony perception, and quantitatively model how these processes contribute to music composition. Through these studies we examine long-standing issues in music psychology, such as the relative contributions of roughness and harmonicity to consonance perception, the roles of low-level psychoacoustic and high-level cognitive processes in harmony perception, and the probabilistic nature of harmonic expectation. We also develop cognitively informed computational models that are capable of both analysing existing music and generating new music, with potential applications in computational creativity, music informatics, and music psychology. This thesis is accompanied by a collection of open-source software packages that implement the models developed and evaluated here, which we hope will support future research into the psychological foundations of musical harmony.

Contents

List of Figures

List of Tables

Acknowledgements

I could not have hoped for a more dedicated supervisor than Marcus Pearce, who was always generous with time and advice throughout my studies. I'm grateful to Emmanouil Benetos and Matthew Purver, the other two members of my progression panel, for taking the time to examine my interim reports and for providing useful advice at my progression meetings. I'm indebted to many others who contributed useful feedback about the work described here, including David Baker, Manuel Anglada-Tort, Bastiaan van der Weij, Vincent Cheung, Andrew Goldman, Tyreek Jackson, Stefan Koelsch, Daniel Müllensiefen, David Huron, and several anonymous reviewers. I also enjoyed the general support of various research groups at the university, in particular the Music Cognition Lab, the Centre for Digital Music, the Cognitive Science group, and the doctoral training centre for Media & Arts Technology. I'm very grateful to my external examiners, Alan Marsden and Tuomas Eerola, who gave much valuable advice about improving this work. Lastly, I would like to thank Maddy Seale for always being willing to listen to my ideas, and for supporting and encouraging me throughout this process.

This work was supported by the EPSRC/AHRC grant EP/L01632X/1.

Chapter 1

Introduction

Western music, broadly conceived as the musical traditions that originated in Europe and now permeate the globalised world, is rooted in the notion of harmony. Harmony describes how simultaneously occurring musical notes are combined to form chords, and how successions of chords combine to form chord progressions. It has long fascinated musicians, music theorists, mathematicians, and scientists for its combinatorial complexity, its ability to evoke complex emotions in the listener, and its apparent deep connections to mathematics, psychoacoustics, and linguistics.

The psychological study of harmony has its roots in the late 18th century. The polymath Hermann von Helmholtz may be considered the grandfather of the field, developing a 'beating' theory of consonance perception that remains influential to this day (Helmholtz, 1863). Soon after Helmholtz, Karl Stumpf presented his own 'fusion' theory of consonance, bringing an emphasis on perceptual integration that prefigured the subsequent 'Gestalt' school of psychology (Stumpf, 1890, 1898). Following Helmholtz, many subsequent researchers continued to analyse consonance in terms of interactions between neighboring spectral components, taking advantage of new advances in psychoacoustic measurement to build formal models predicting the consonance of musical sonorities from the dissonance profiles of pairs of pure tones (Dillon, 2013; Hutchinson & Knopoff, 1978; Kameoka & Kuriyagawa, 1969b; Sethares, 1993; Vassilakis, 2001). Following Stumpf, other researchers continued to analyse consonance in terms of perceptual integration. In particular, Ernst Terhardt studied how listeners integrate different parts of the auditory spectrum to derive the percept of pitch, and argued that consonance could be understood as a consequence of this pitch-finding process (Terhardt, 1984). Richard Parncutt adopted and broadened Terhardt's hypothesis, arguing that these pitch-finding mechanisms were fundamental not only to simultaneous consonance (a property of individual chords) but also to sequential consonance (a property of successive chords), and using this hypothesis to rationalise various tenets of Western music theory (Parncutt, 1989). Recent research by Andrew Milne and colleagues pursues similar principles, showing that psychoacoustic processes can provide a good account of both simultaneous and sequential relationships in tonal music (Milne & Holland, 2016; Milne, Laney, & Sharp, 2015; Milne et al., 2016).

A contrasting line of harmony research has emphasised the role of higher-level cognition in harmony perception. Much of the early work in this area depended on the so-called 'harmonic priming' paradigm, developed by Jamshed Bharucha and colleagues (Bharucha, 1987; Bharucha & Pryor, 1986; Tekman & Bharucha, 1992, 1998). In this paradigm, participants were instructed to make perceptual judgments concerning certain chords within chord sequences, and researchers analysed the extent to which reaction times were affected by manipulations of the preceding chords in the sequence. Bharucha and colleagues had a particular interest in isolating listeners' hierarchical conceptions of tonality, which they modelled with MUSACT, a connectionist network with three layers in ascending levels of abstraction: tones, chords, and keys (Bharucha, 1987). Nodes within MUSACT were connected by edges with weights prespecified by the researcher to reflect commonly held principles of Western harmonic tonality, such as the idea that any particular key may be represented by three core chords, namely the tonic, the dominant, and the subdominant. Given a particular musical input comprising a particular set of tones, activations would spread through the network in proportion to the weights of the edges connecting each node. In a series of empirical studies, Bharucha and colleagues showed that MUSACT's activations could successfully predict harmonic priming effects in Western listeners (Bharucha, 1987; Bharucha & Pryor, 1986; Tekman & Bharucha, 1992, 1998). As these studies progressed, the researchers controlled more carefully for low-level psychoacoustic cues that might explain the priming effects (e.g. common harmonics between the context and the target), but continued to find reliable priming effects consistent with MUSACT's predictions (Tekman & Bharucha, 1992, 1998). Emmanuel Bigand and colleagues extended the harmonic priming paradigm to longer harmonic sequences, and continued to find evidence for high-level cognitive contributions to harmonic priming (Bigand, Madurell, Tillmann, & Pineau, 1999; Bigand & Pineau, 1997; Bigand, Poulin, Tillmann, Madurell, & Adamo, 2003; Tillmann & Bigand, 2001).

The MUSACT model demonstrated that listeners' internal knowledge of Western harmonic tonality could be parsimoniously represented as a connectionist network, but it did not show how this knowledge could be acquired. Barbara Tillmann and colleagues addressed this question, showing that a self-organising network could acquire the same tonal principles as the MUSACT model simply through exposure to harmonic sequences from tonal music, and that the resulting model could explain a wide variety of psychological results from the literature (Tillmann, Bharucha, & Bigand, 2000). This work was complemented by contemporaneous studies of statistical learning in human listeners; various studies in the linguistic domain highlighted the capacity of listeners to learn artificial grammars simply through passive exposure to exemplars from these grammars (Aslin, Saffran, & Newport, 1998; Saffran et al., 1996a, 1996b), and it was soon shown that analogous learning processes took place for artificial musical grammars (Saffran, Johnson, Aslin, & Newport, 1999). Most early studies of musical statistical learning focused on melody (e.g. Creel, Newport, & Aslin, 2004; Dienes & Longuet-Higgins, 2004; Endress, 2010; Rohrmeier et al., 2011), but several studies also isolated statistical-learning effects for artificial harmonic grammars, including grammars in non-Western tuning systems and grammars incorporating context-free structure (Jonaitis & Saffran, 2009; Loui, Wu, Wessel, & Knight, 2009; Rohrmeier & Cross, 2009). These findings supported a growing belief that harmony might be less of a psychoacoustic phenomenon and more of a cultural phenomenon, driven by similar statistical-learning processes to language cognition.

These documented similarities between linguistic learning and musical learning were reinforced by further documented similarities between linguistic and musical domains. Several researchers presented music-theoretic accounts of the sequential structure of Western harmony – harmonic ‘syntax’ – with an emphasis on hierarchical structure, recursion, and non-adjacent dependencies, properties shared with the syntax of natural language (Rohrmeier, 2011; Steedman, 1984). These apparent similarities between harmonic and linguistic syntax motivated a popular hypothesis that both types of syntax are processed by similar cognitive and neural mechanisms. Some evidence for this hypothesis comes from empirical studies showing suggestive parallels between electrical signatures of harmonic and linguistic syntax processing (e.g. Patel, Gibson, Ratner, Besson, & Holcomb, 1998; Koelsch et al., 2001; Koelsch, Gunter, Friederici, & Schröger, 2000; Koelsch, Schroger, & Gunter, 2002), and interference effects between the processing of harmonic syntax and linguistic syntax (e.g. Koelsch, Gunter, Wittfoth, & Sammler, 2005; Hoch, Poulin-Charronnat, & Tillmann, 2011; Kunert, Willems, & Hagoort, 2016; Slevc, Rosenberg, & Patel, 2009). Analogous results with melodic stimuli suggest that these relationships between linguistic and musical syntax may generalise to musical dimensions beyond harmony (e.g. Fedorenko, Patel, Casasanto, Winawer, & Gibson, 2009; Mirand & Ullman, 2007).

These connections between harmonic and linguistic syntax are appealing in that they suggest an evolutionary explanation for harmony cognition. However, the scope of this relationship has been questioned by various studies. First, it has been argued that interference effects between linguistic and music processing are less specific than previously thought, potentially reflecting more general cognitive phenomena such as attention (Escoffier & Tillmann, 2008; Perruchet & Poulin-Charronnat, 2013; Poulin-Charronnat, Bigand, Madurell, & Peereman, 2005) and cognitive control (Slevc & Okada, 2015). Second, it has been shown that many empirical studies purporting to demonstrate listeners' sensitivity to musical syntax can in fact be explained by psychoacoustic models of pitch perception and auditory short-term memory (Bigand, Delbé, Poulin-Charronnat, Leman, & Tillmann, 2014). At the time of writing, therefore, the degree of similarity between harmonic and linguistic syntax processing is still a matter of debate.

Computational modelling may prove to be a powerful strategy for tackling these debates about harmony cognition. By formalising theories of harmony cognition as computational models, we can minimise the scope for ambiguity and force ourselves to address the assumptions inherent in these theories. The resulting models automate the process of generating testable predictions from theories, thereby improving the efficiency and objectivity of the scientific process. Furthermore, good models can generate predictions for relatively naturalistic and complex stimuli, allowing researchers to run more realistic experiments and thereby enhance the ecological validity of their work. The resulting models may also prove useful outside harmony cognition research, for example by providing assistance or inspiration to composers and performers (e.g. Gebhardt, Davies, & Seeber, 2016; Parncutt & Strasburger, 1994), and by supporting the development of music information retrieval systems (e.g. Haas, Magalhães, & Wiering, 2012).

Many computational models of harmony perception have been presented over the decades. However, the modelling literature is limited as it stands. First, systematic model comparisons are few and far between, making it difficult to understand which cognitive models provide the best accounts of harmony perception, and making it harder for music engineers to choose the most promising cognitive models for practical applications.[1] Second, the existing models mostly address individual perceptual features without addressing how these different features combine to drive higher-level psychological phenomena (though see Bigand, Parncutt, & Lerdahl, 1996; Johnson-Laird, Kang, & Leong, 2012). Third, there is a lack of publicly available and comprehensive software libraries for these harmony models. A few relevant audio-analysis toolboxes do exist, but these generally implement only a small number of harmony models, and often

[1] One exception is Stolzenburg (2015), who provides quantitative performance comparisons for many different consonance models on a variety of perceptual datasets. However, the small size of these datasets and the use of correlation rather than regression limits the scope of the findings (see Section 3.4.1).

lack systematic perceptual validation (e.g. IPEM toolbox, Leman, Lesaffre, & Tanghe, 2001; Janata Lab Music Toolbox, Collins, Tillmann, Barrett, Delbé, & Janata, 2014; MIRToolbox, Lartillot, Toiviainen, & Eerola, 2008; Essentia, Bogdanov et al., 2013).

This thesis aims to address these problems and thereby advance the state of the art in the cognitive modelling of harmony. Our approach is characterised by the following priorities:

1. **Perceptual modelling.** One goal of this work is to improve our understanding of the cognitive processes underlying harmony perception in Western listeners. We therefore model various empirical datasets of harmony perception, with some of these datasets compiled from the literature, and some generated from our own experiments.

2. **Corpus modelling.** A second goal of this work is to understand the cognitive processes underlying the composition of harmony in Western music. We therefore apply our computational models to the analysis of large musical corpora representing various traditions of Western music composition. There are many ways to conduct corpus analyses, but we take a perceptual perspective, seeking to understand how harmonic practice may have been shaped by perceptual principles such as harmonicity, spectral similarity, and auditory scene analysis.

3. **Model interpretability.** The interpretability of a model is defined as the sense in which a researcher can inspect a model and understand how it generates its output. Modellers are often faced with a trade-off between interpretability and expressivity: more expressive models can theoretically capture more complex phenomena at the expense of reduced interpretability. Here we prioritise interpretability, and therefore favour techniques such as linear models, log-linear models, and Markov models over connectionist techniques such as feedforward neural networks, mixture-of-expert models, and recurrent neural networks. One possibility for future work is to substitute connectionist models for these conventional approaches, with the goal of improving the expressive capacity of our models; this may be particularly relevant for applications where interpretability is less important than predictive power (e.g. Bayesian priors for automatic harmony transcription systems). Interestingly, however, recent work has found that modern connectionist approaches (in particular, long short-term memory recurrent neural networks) can struggle when applied to harmony modelling (Landsnes et al., 2019; Sears et al., 2018a).

4. **Model integration.** Many existing models of harmony perception ad-

dress isolated perceptual mechanisms such as roughness or spectral similarity. Comparatively few models address how these different perceptual mechanisms combine to determine higher-level cognitive phenomena. Here we emphasise this process of integration, reasoning that harmony perception is likely to be driven by a complex collection of different psychoacoustic and cognitive processes. As a result, our models are typically hierarchically structured, combining multiple submodels representing various psychological processes.

5. **Model comparison.** Model comparison is a crucial part of computational cognitive modelling: it is the primary method by which different psychological theories are evaluated against one another. Systematic model comparisons are surprisingly rare in the harmony cognition literature. Here we take model comparison seriously, providing the first systematic comparisons of many harmony models in the literature (Sections 3, 4), and using these models as benchmarks against which to evaluate our own models.

6. **Symbolic versus acoustic modelling.** Music modelling exists in two main traditions: symbolic modelling and acoustic modelling. Symbolic models represent the musical input using abstract human-readable representations, such as 'Cmaj7', {60, 64, 67}, or {0, 4, 7}; acoustic models engage more directly with the musical sound, using representations such as waveforms and frequency spectra. Our work combines these symbolic and acoustic approaches. Our three computational models – consonance, harmonic expectation, and voice leading – each take symbolic representations as their inputs, but then derive various acoustic representations from these inputs, typically by expanding each musical tone into its implied harmonics and in some cases blurring the resulting spectrum to account for perceptual uncertainty. The resulting representations are then used for modelling various psychoacoustic properties of the stimuli, such as interference between partials and spectral similarity. A particular contribution of the present work is to demonstrate how these continuous psychoacoustic features may be incorporated into probabilistic generative models of harmonic structure, which have been historically limited to discrete symbolic features (Chapter 4).

7. **Open-source implementations.** Computational modelling can facilitate cumulative research by summarising psychological theories as portable software that can be applied by subsequent researchers to new research problems. Currently, however, only a small portion of harmony

models in the literature have publicly available software implementations. We address this issue by developing open-source implementations of many of these models, alongside implementations of our own models, which we release alongside this thesis. By doing so we help to facilitate the cumulative testing and improvement of these models.

We focus on three core phenomena in harmony cognition: consonance, harmonic expectation, and voice leading. Consonance describes how certain combinations of tones sound 'well' when heard simultaneously; harmonic expectation describes how certain chord progressions create expectations that certain chords will come next; voice leading describes how chord notes are distributed across octaves, and how these notes connect to form simultaneous melodies.[2]

A common question that recurs throughout this work is as follows:

> "To what extent is the perception and composition of Western harmony determined by low-level psychoacoustic processes versus high-level cognitive processes?"

Consonance. The field of consonance research has traditionally emphasised the role of psychoacoustic processes in consonance perception. However, it has also been argued that consonance partly depends on higher-level learning processes, whereby listeners develop familiarity with prevalent chords in their musical culture. Here we construct a collection of computational models addressing both sides of consonance perception, and apply these models to large datasets of perceptual data to investigate their relative contributions to musical consonance. Furthermore, we decompose the psychoacoustic account of consonance into two processes – interference between partials and periodicity/harmonicity detection – and examine their relative contributions to consonance perception, as well as their ability to predict chord distributions in large corpora of Western music.

[2] At the outset of this project, our primary goal was specifically to develop a cognitively motivated probabilistic model of polyphonic music, inspired by Pearce's (2005) multiple-viewpoint model of melodic expectation. Given the foundational role of consonance in Western polyphony, we resolved that consonance ought to be incorporated into our probabilistic model. We looked to the consonance literature for an appropriate consonance model to include, and found many candidate models but no systematic evaluations of these models. This prompted us to examine the problem of consonance in greater detail, resulting in Chapter 3. Consonance values typically return continuous outputs, which cannot be directly modelled using traditional multiple-viewpoint techniques (e.g. Conklin & Witten, 1995; Pearce, 2005). This prompted us to develop the viewpoint regression technique presented in Chapter 4. At this point it seemed expedient to decompose the problem of polyphony modelling into two subtasks, namely harmony modelling and voice-leading modelling. This approach seemed useful for improving model tractability and interpretability, and it helped to align the research with traditional Western music theory and pedagogy. The harmonic model became Chapter 4, and the voice-leading model became Chapter 5.

Harmonic expectation. Various empirical studies have accumulated purporting to demonstrate that harmonic expectation is a high-level cognitive process similar to syntax parsing in natural language. However, a recent study by Bigand et al. (2014) undercuts much of this work, showing that many of the empirical results can be explained by a low-level psychoacoustic model, and illustrating the unexpected difficulty of constructing stimuli that effectively isolate high-level cognitive processing from low-level psychoacoustic cues. Here we take an alternative approach, conducting a behavioural study using a large dataset of chord sequences drawn from an authentic popular music corpus (Burgoyne, 2011), and using computational modelling to analyse how different kinds of psychological features contribute to harmonic expectation.

Voice leading. Huron (2001, 2016) has argued that voice-leading practice in Western music primarily reflects low-level psychoacoustic processes, in particular auditory scene analysis (Bregman, 1990). While different aspects of this argument have received specific empirical support, there is currently no integrative computational model quantifying how these processes combine in practice. Here we develop such a model, and apply it to the analysis of voice-leading practice in chorale harmonisations by J. S. Bach, as well as to the generation of voice leadings for unseen chord sequences.

These investigations of consonance, harmonic expectation, and voice leading constitute the main contributions of the present thesis. We precede these investigations with Chapter 2, which examines the problem of representing harmony for computational cognition research. The thesis continues with Chapters 3, 4, and 5, which examine consonance, harmonic expectation, and voice leading respectively. The thesis then concludes with Chapter 6, which discusses the outcomes of the present work.

Chapter 2

Representing harmony

2.1 Introduction

Cognitive science seeks to understand the human mind in terms of mental representations and computational operations upon these representations. This computational metaphor is formalised by creating computational models of these cognitive representations and operations, which can then be tested against human behaviour (Thagard, 2019).

Music cognition research applies the techniques of cognitive science to the domain of music. This research has generated many important cognitive insights into fundamental aspects of music, resulting in the development of sophisticated computational cognitive models of melody (e.g. Pearce, 2018; Temperley, 2008), rhythm (e.g. Large & Jones, 1999; Weij, Pearce, & Honing, 2017), and tonality (e.g. Bharucha, 1987; Collins et al., 2014; Krumhansl, 1990; Leman, 2000a; Tillmann et al., 2000). In comparison, the field of cognitive harmony modelling has historically seen less progress, partly due to the high combinatorial complexity of the harmonic domain, and partly due to a lack of large-scale harmonic corpora (though see Hutchinson & Knopoff, 1978; Pardo & Birmingham, 2002; Parncutt, 1988, 1989; Ponsford, Wiggins, & Mellish, 1999; Sethares, 1993; Temperley, 1997, 2009; Temperley & Sleator, 1999; Vassilakis, 2001 ; Winograd, 1968). However, in recent years, increases in computing resources and new large corpora of harmonic transcriptions have encouraged a new wave of research into the computational modelling of harmony cognition (e.g. Miles, Rosen, & Grzywacz, 2017; Di Giorgi, Dixon, Zanoni, & Sarti, 2017; Hedges & Wiggins, 2016b; Landsnes et al., 2019; Sears et al., 2018b).

This cognitive modelling has focused primarily on modelling computational operations underlying harmony cognition (e.g. predictive processing, Landsnes et al., 2019; reward generation, Miles et al., 2017; complexity, Di Giorgi et al., 2017). Little attention has been paid to representational issues; instead, different representations are adopted by different researchers, guided in part by the researchers' music-theoretic intuitions and in part by the encodings of the available music corpora.

We think that it is worth examining these representational issues more systematically. Relying excessively on music-theoretic intuition risks diluting the formal objectivity of the cognitive approach, and potentially excludes cognitive scientists without a musical background. Relying excessively on idiosyncrasies of encoded corpora provides a backdoor for questionable assumptions, such as the idea that untrained listeners infer functional harmony in the same way as the musicians who created the corpus.

This chapter addresses this question of how to represent harmony for music cognition research. Our goal is to provide a systematic account of the different representational possibilities available to music cognition researchers, making explicit the cognitive assumptions of each representation, and describing the computational methods for translating between representations. We enumerate full alphabets for several of these representations, thereby defining methods for encoding chords as integers, which should be particularly useful for constructing statistical models of harmonic style. We also describe how these representations may be derived from common types of musical corpora. The chapter is accompanied by an open-source software package, *hrep*, written for the programming language R, that implements these different representations and encodings in a generalisable object-oriented framework.[1]

We focus in particular on low-level representations. By 'low-level', we mean representations that correspond to early stages of cognitive processing, and that are linked relatively unambiguously to the structure of the musical score and the auditory signal. Our rationale is that these low-level representations provide the starting point for most cognitive studies: they determine how musical corpora are encoded, they determine the input to statistical-learning models (e.g. *n*-gram models, Hidden Markov Models), and they determine the input to psychological feature extractors (e.g. root-finding models, key-finding models). However, we also discuss how various important higher-level representations may be extracted from these low-level representations.

2.2 Related work

Several large corpora of harmonic transcriptions have been released in recent years, including the iRb corpus (Broze & Shanahan, 2013), the Billboard corpus

[1] `http://hrep.pmcharrison.com`

(Burgoyne, 2011), the Annotated Beethoven Corpus (Neuwirth, Harasim, Moss, & Rohrmeier, 2018), the rock corpus of Temperley & De Clercq (2013), and the Beatles corpus of Harte, Sandler, Abdallah, & Gómez (2005). Each of these corpora uses different representation schemes, with these representation schemes differing in terms of human-readability and analytical subjectivity.

The iRb corpus (Broze & Shanahan, 2013) and the Billboard corpus (Burgoyne, 2011) express chords using letter names for chord roots and textual symbols for chord qualities; for example, a chord might be written as 'D:min7' where 'D' denotes the chord root and 'min7' denotes a minor-seventh chord quality.[2] These symbols are easy for musicians to read, but the chord-quality component is imprecise, being subject to the interpretation of the performer. This ambiguity makes it non-trivial to translate such symbols to acoustic or sensory representations, which is problematic for cognitive modelling.

The rock corpus of Temperley & De Clercq (2013) and the Annotated Beethoven Corpus of Neuwirth et al. (2018) provide deeper levels of analysis: they express chords relative to the prevailing tonality, and provide functional interpretations of the resulting chord progressions, writing for example 'V/vi' to denote the secondary dominant of the submediant. This information is useful for many music-theoretic analyses. However, such representations are typically unsatisfactory for perceptual modelling because they assume that listeners share the same interpretations as music theorists, and because the representations do not generalise to non-tonal musical systems.

The Beatles corpus of Harte et al. (2005) avoids some of these problems. Instead of being denoted with textual labels, chord qualities are denoted as collections of pitch classes expressed relative to the chord root, with for example `D#:(b3,5,b7)` denoting a D# minor-seventh chord. Chord progressions are expressed independent of key context and without functional interpretation, eliminating much of the subjectivity of the previously described representations. However, the representation still relies on the notion of 'chord root', a concept from Western music theory that is not relevant to all musical styles, limiting the representation's suitability for general cognitive modelling.

The General Chord Type algorithm of Cambouropoulos (2016) generalises the notion of 'chord root' to non-Western musical styles (see also Cambouropoulos, Kaliakatsos-Papakostas, & Tsougras, 2014). The algorithm is parametrised by a style-dependent 'consonance vector' that quantifies the relative consonance of different intervals within the musical style. Using a consonance vector representing Western tonal conventions, the algorithm reliably reproduces chord roots as annotated by Western music theorists (Cambouropoulos, 2016). Applied to

[2] A *chord quality* defines a chord's pitch-class content relative to the chord root. See Section 2.4.2 for a definition of 'pitch class' and Section 2.8.2 for a definition of 'chord root'.

non-Western tonal systems, the algorithm produces harmonic representations that seem to work well for statistical modelling and music generation (Cambouropoulos et al., 2014). However, the algorithm is only loosely motivated by human cognition, and has yet to receive systematic perceptual validation.

Alternative representation schemes come from the music-theoretic tradition of pitch-class set theory, in particular the *pitch-class set* and the *prime form* pitch-class set (Forte, 1973; Rahn, 1980). These representations can be unambiguously computed from musical scores, and can be efficiently encoded as short lists of integers. The pitch-class set representation works particularly well as a characterisation of chord perception, and so we include it in our collection of representations (Section 2.4.2). However, the representation is not sufficient for all cognitive modelling, as it reveals little about sensory properties of the chord such as consonance (e.g. Hutchinson & Knopoff, 1978) or tonal distance (e.g. Milne, Sethares, Laney, & Sharp, 2011). Some insight can be gained by computing the pitch-class set's *interval vector*, which summarises the frequency of different intervals in the sonority, information that can be used to estimate the pitch-class set's consonance (Huron, 1994; Parncutt et al., 2018). However, this approach still only offers limited insight into the chord's perceptual qualities.

Other representations have been developed to characterise the intervallic relationships between successive chords, such as the voice leadings described by Tymoczko (2008), the voice-leading types of Quinn & Mavromatis (2011) and Sears, Arzt, Frostel, Sonnleitner, & Widmer (2017), the interval function of Lewin (1987), and the directed interval class vector of Cambouropoulos (2016). Such representations are indisputably valuable for music analysis, but they fall more in the realm of voice leading (connecting the notes of successive musical chords to form melodies) than of harmony, the topic of the present work.

In conclusion, many harmonic representation schemes exist in the literature, but many are not well-suited to be low-level representations for cognitive modelling. Ideally, such representations should be unambiguously computable from symbolic music corpora, well-motivated by cognitive theory, generalisable across musical styles, and able to capture both the discrete nature of chord categories and the sensory implications of a given chord's acoustic spectrum. The following section compiles a collection of representations intended to address this goal.

2.3 Low-level representations

The harmonic dimension of a music composition may be characterised as a sequence of musical chords, where a chord is defined as a collection of musical notes that are sounded in close temporal proximity and perceived as an integrated auditory object. Here we will describe different low-level representations for the chords within a chord sequence.

We organise our low-level harmonic representations into three categories: the *symbolic*, the *acoustic*, and the *sensory*.[3] Symbolic representations are succinct and human-readable descriptions of musical chords, such as are commonly used by performing musicians and music analysts; we are particularly interested in symbolic representations that reflect cognitive representations internal to the listener. Acoustic representations characterise musical sound, as created for example by a musician performing a symbolic score. Sensory representations then reflect a listener's perceptual images of the resulting sound. Generally speaking, symbolic representations tend to be discrete, corresponding to well-delineated perceptual categories, whereas acoustic and sensory representations tend to be continuous and high-dimensional.

Many of these representations express some kind of invariance. Invariance means that a chord retains some kind of musical identity under a given operation. For example, transposition invariance means that a chord retains its identity when all its pitches are transposed by the same pitch interval; octave invariance means that a chord retains its identity when individual pitches are transposed by octave intervals.

An invariance principle may also be formulated as an *equivalence relation*: saying that a chord is invariant under a given operation is the same as saying that a chord is equivalent to all other chords produced by applying the operation. Equivalence relations partition sets into *equivalence classes*, sets of objects that are all equal under some equivalence relation. If we label each object by its equivalence class, we produce a new representation that embodies the original invariance principle behind the equivalence relation. For example, if we begin with the seven triads of the C major scale (C major, D minor, E minor, …) and add transposition invariance, we get three equivalence classes: one of major triads (C major, F major, G major), one of minor triads (D minor, E minor, A minor), and one of diminished triads (B diminished). These equivalence classes therefore define a new representation with three transposition-invariant chord qualities: major, minor, and diminished. See Callender, Quinn, & Tymoczko (2008), Tymoczko (2011), and Lewin (1987) for further discussions of equivalence classes induced by musical invariances.

Experimental psychology delivers clear evidence for various invariance principles within music perception. Octave invariance reflects the fact that listeners perceive tones separated by octaves to share some essential quality termed *chroma* (Bachem, 1950). Transposition invariance reflects the phenomenon of

[3] These categories were inspired by Babbitt's (1965) graphemic, acoustic, and auditory representational categories.

relative pitch perception (Plantinga & Trainor, 2005). Tones with different spectra but identical fundamental frequencies share an invariant perceptual quality termed *pitch* (Stainsby & Cross, 2009). These invariance principles are reflected by representation schemes used by music theorists, such as pitch-class set notation (octave invariance, tone spectrum invariance) and Roman numeral notation (octave and tone spectrum invariance for chords, transposition invariance for chord sequences). These perceptual invariances should likewise be incorporated into cognitive models of harmony processing.

A second motivation for incorporating invariance into cognitive models is to improve the tractability of statistical learning. Listeners are thought to internalise the conventions of Western harmony through statistical learning (e.g. Jonaitis & Saffran, 2009; Loui et al., 2009; Rohrmeier & Cross, 2009), but when we try to simulate this process computationally we are faced by a serious computational complexity problem. Suppose that a chord is represented as a pitch set containing up to 12 notes drawn from an 80-note range. There are approximately $N = 7.3 \times 10^{13}$ such chords, and if we want to build a simple model of first-order transition probabilities between these chords, we need to construct an array containing $N \times N = 5.3 \times 10^{27}$ elements. This is impractical on modern computers, and implausible for human brains. Even if we were able to represent such an array, it would be difficult to estimate its parameters effectively without an impossible amount of training data. However, if we apply invariance principles to combine musically equivalent chords, then we can substantially reduce the scale of the problem. In this case, using octave invariance to represent each chord as a pitch-class set reduces the array to 1.7×10^7 values. Applying transpositional invariance further reduces the matrix to 1.4×10^6 values. This array is still large, but it is straightforward to represent it computationally and to estimate its important parameters with a reasonable amount of training data. These kinds of invariance principles can therefore be very useful for ensuring the tractability of harmonic statistical learning.

We organise our different low-level representations as a network (Figure 2.1). This network clarifies the sequence of computational operations required for translating between different representations, and shows how these representations may be organised according to different perceptual invariances. This structure is made more explicit in Table 2.1, which lists the invariance principles that apply to each representation, their alphabet sizes, and the corresponding class names in the *hrep* package.

Our presentation of these low-level representations is framed from the perspective of Western music. This is intentional: unlike melody and rhythm, harmony is relatively specific to Western musical traditions. We also assume equal-tempered tuning, which is common but not universal within Western mu-

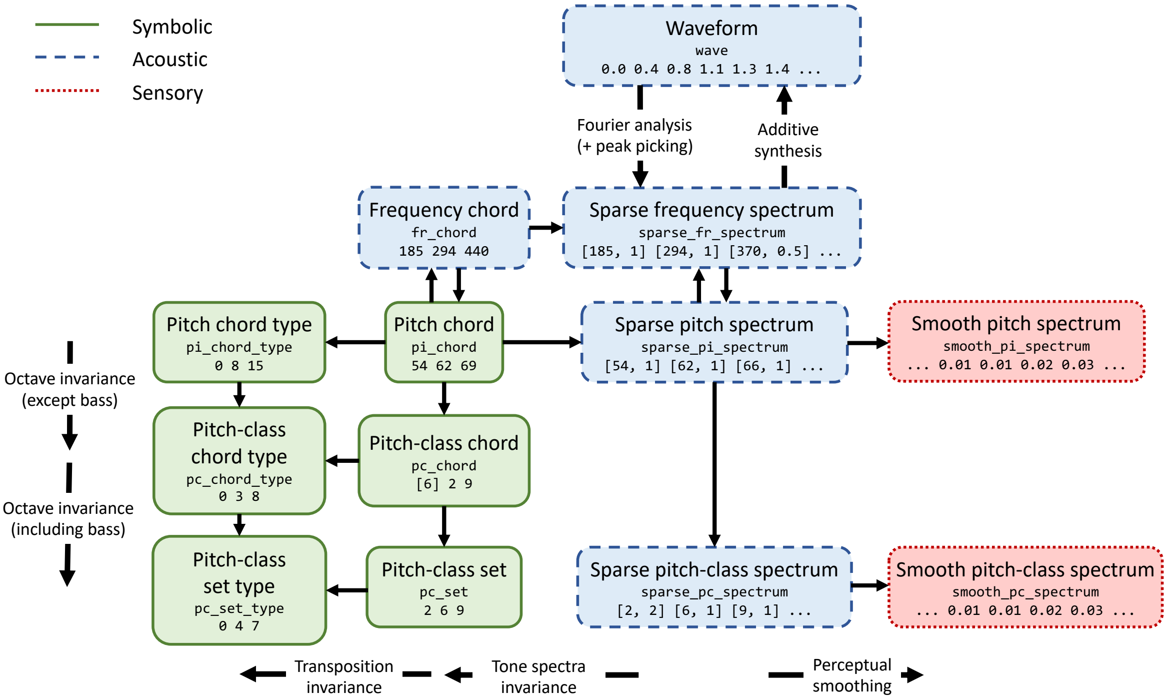

Figure 2.1: A network of low-level harmony representations, organised into three representational categories: symbolic, acoustic, and sensory. Arrows between representations indicate well-defined computational translations between representations.

sic. Nonetheless, much of the material presented here could easily generalise to other tuning systems.

Table 2.1: The 13 low-level harmony representations with their alphabet sizes and invariances.

Name			Invariance				
Textual	Computer	Alphabet size	Tone spectra	Tuning	Inversion	Octave	Transposition
Symbolic							
Pitch chord	`pi_chord`		✓	✓			
Pitch-class set	`pc_set`	4,095	✓	✓	✓	✓	
Pitch-class chord	`pc_chord`	24,576	✓	✓		(✓)	
Pitch chord type	`pi_chord_type`		✓	✓			✓
Pitch-class chord type	`pc_chord_type`	2,048	✓	✓		(✓)	✓
Pitch-class set type	`pc_set_type`	351	✓	✓	✓	✓	✓
Acoustic							
Frequency chord	`fr_chord`		✓				
Sparse pitch spectrum	`sparse_pi_spectrum`						
Sparse pitch-class spectrum	`sparse_pc_spectrum`					✓	
Sparse frequency spectrum	`sparse_fr_spectrum`						
Waveform	`wave`						
Sensory							
Smooth pitch spectrum	`smooth_pi_spectrum`						
Smooth pitch-class spectrum	`smooth_pc_spectrum`					✓	

Note. Here 'inversion' refers to changing the chord's bass pitch class while maintaining its pitch-class set.

2.4 Symbolic representations

We begin by defining six symbolic representations, using the naming conventions below:

1. 'Pitch-class' representations express octave invariance;

2. 'Chord' representations identify which pitch class corresponds to the lowest pitch in the chord, termed the *bass note*;

3. 'Type' representations express transposition invariance.

These six representations are termed pitch chords, pitch-class sets, pitch-class chords, pitch chord types, pitch-class chord types, and pitch-class set types respectively. The term 'pitch-class set' is common in previous research, but the other terms are mostly new.[4] Table 2.2 displays these six representations as derived for several example chords.

In theories of Western tonal harmony, the term 'chord' usually refers to an established harmonic category such as 'major chord', 'diminished chord', 'Neapolitan chord', etcetera. We wish to avoid limiting our representation scheme to Western tonal harmony, and so we adopt a more inclusive definition of 'chord', namely 'a collection of pitches or pitch classes, one of which is labelled as the bass pitch class'. Such structures are sometimes termed 'sonorities' in music theory.

Some of these symbolic representations are candidates for what has been termed the *musical surface*, defined by Jackendoff (1987) as the 'lowest level of representation that has musical significance' (p. 219) and by Cambouropoulos (2016) as a 'minimal discrete representation of the musical sound continuum in terms of note-like events (each note described by pitch, onset, duration, and possibly dynamic markings and timbre/instrumentation)' (p. 31). Certainly these symbolic representations are minimal in the sense that they discard much of the information present in acoustic representations, while retaining sufficient information to support meaningful musical analyses (e.g. Huron & Sellmer, 1992; Parncutt, Sattmann, Gaich, & Seither-Preisler, 2019; Rohrmeier & Cross, 2008). However, this is not to say that still lower levels of representation do not also have musical significance: important musical phenomena such as consonance and harmonic distance are often best explained by appealing to lower-level acoustic and sensory representations (see Sections 2.5 and 2.6). It is also true that, in some contexts, it might prove useful to define the musical surface in terms of higher-level representations such as chord roots

[4] Note however that the definition of 'pitch-class set' varies in the literature; for example, some music theorists take it to imply octave invariance, inversion (reflection) invariance, and transposition invariance, whereas we solely take it to imply octave invariance.

Table 2.2: Symbolic representations for three example chords: (**A**) C major triad, 1st inversion; (**B**) C major triad, second inversion; (**C**) E dominant seventh, third inversion.

	Chord		
Representation	**A**	**B**	**C**
`pi_chord`	{52, 60, 67}	{43, 52, 60}	{50, 59, 64, 68}
`pc_set`	{0, 4, 7}	{0, 4, 7}	{2, 4, 8, 11}
`pc_chord`	(4, {0, 4, 7})	(7, {0, 4, 7})	(2, {2, 4, 8, 11})
`pi_chord_type`	{0, 8, 15}	{0, 9, 17}	{0, 9, 14, 18}
`pc_chord_type`	{0, 3, 8}	{0, 5, 9}	{0, 2, 6, 9}
`pc_set_type`	{0, 4, 7}	{0, 4, 7}	{0, 3, 6, 8}

(Cambouropoulos, 2010, 2016). Instead of committing to a particular definition of the musical surface, we therefore prefer to present a collection of symbolic, acoustic, and sensory representations that can be selected from according to the task at hand.

2.4.1 Pitch chord

The *pitch chord* representation is the most granular of the symbolic representations considered here. The other five symbolic representations are defined by partitioning this representation into equivalence classes.

The pitch chord representation expresses each chord as a finite and non-empty set of pitches. Each pitch is represented as a MIDI note number, where 60 corresponds to middle C (C4, c. 262 Hz), and integers correspond to semitones. Correspondingly, a chord containing n unique notes is written as a set of n integers; for example, the set {54, 62, 69} represents a D major triad in first inversion comprising the notes F#3, D4, A4. Transposition then corresponds to simple integer addition. Note that duplicate pitches are not encoded, reflecting how such pitches tend to be perceptually fused in the mind of the listener.

Pitch is a continuous phenomenon, and real musical instruments rarely play in exact 12-tone equal temperament (see e.g. Parncutt & Hair, 2018). It is possible to embed this continuity in the pitch chord representation, writing sets such as {59.9, 64.2, 67.3} to express deviations from equal temperament (in this case -10 cents, $+20$ cents, and $+30$ cents respectively). However, listeners tend to represent pitch categorically (Stainsby & Cross, 2009), and it is generally useful to follow this principle in the symbolic representations, while allowing for continuous pitch in the acoustic and sensory representations. In what follows we will assume that all symbolic representations are treated categorically and hence limited to integer values.

Even assuming categorical representation, many pitch chords are possible: there are 7.3×10^{13} possible pitch chords that can be created by drawing 1–10 pitches from a candidate set of 80 pitches. In the *hrep* package, pitch chords are represented as `pi_chord` objects, which are themselves internally represented as numeric vectors sorted in ascending order.

2.4.2 Pitch-class set

Pitch-class sets express the principle of octave invariance. The representation is defined by partitioning the set of pitch chords under the following equivalence relation: two chords are equivalent if they comprise the same pitch classes. Pitch classes are defined as equivalence classes over pitches: two pitches correspond to the same pitch class if they are separated by an integer number of octaves.

A pitch-class set may be represented as a set of integers, where each integer identifies a pitch class. These pitch classes are computed by applying the modulo 12 operator to MIDI note numbers; for example, the pitches 48 and 60 both correspond to a pitch class of 0, whereas the pitches 49 and 61 correspond to a pitch class of 1. Computing the pitch-class set for a given chord involves applying the modulo 12 operator to each chord pitch and then discarding any duplicate pitch classes: for example, the pitch chord $\{54, 62, 69, 74\}$ produces the pitch-class set $\{2, 6, 9\}$. A pitch-class set may be transposed by x semitones by adding x to each integer and then applying the modulo 12 operator.

There are 4,095 possible non-empty pitch-class sets in the Western 12-tone scale. In the *hrep* package, pitch chords are represented as `pc_set` objects, which are themselves internally represented as numeric vectors sorted in ascending order.

2.4.3 Pitch-class chord

Pitch-class chords express a limited form of octave invariance that differentiates chords with different bass pitch classes. Musicians use the term *inversion* to describe the process of changing a chord's bass pitch class while keeping the pitch-class set the same; we might therefore say that pitch-class chords lack inversion invariance. The identity of the bass pitch class is considered important in music theory, and is encoded in both figured-bass and Roman numeral notation schemes. Correspondingly, we might hypothesise that the identity of the bass pitch class contributes significantly to a chord's perceptual identity.

The pitch-class chord representation may be formally defined by the following equivalence relation: two chords are equivalent if a) they have the same pitch-class set and b) they have the same bass pitch class, with the bass pitch class being defined as the pitch class of the lowest chord pitch. Correspondingly, a pitch-class chord may be represented as a tuple of its bass pitch class and its pitch-class set; for example, the pitch chord {54, 62, 69, 74} can be represented as the pitch-class chord (6, {2, 6, 9}). Like pitch-class sets, pitch-class chords may be transposed by x semitones by adding x to each integer modulo 12.

There are 24,576 possible non-empty pitch-class chords in the Western 12-tone scale. In the *hrep* package, pitch-class chords are represented as `pc_chord` objects, which are themselves internally represented as numeric vectors, with the first element corresponding to the bass pitch class and the remaining elements corresponding to the non-bass pitch classes sorted in ascending order.

2.4.4 Pitch chord type

Pitch chord types express transposition invariance, and are defined by the following equivalence relation: chords are equivalent if they can be transposed to the same pitch chord representation. Pitch chord types are represented as sets of integers, where each chord note is represented as its pitch interval from the bass note. Given a pitch chord, the pitch chord type may be computed by simply subtracting the bass note from all pitches: for example, the pitch chord type of the pitch chord {54, 62, 69} is {0, 8, 15}. In the *hrep* package, pitch chord types are represented as `pi_chord_type` objects, which are themselves internally represented as numeric vectors sorted in ascending order.

2.4.5 Pitch-class chord type

The *pitch-class chord type* representation combines the invariances of the pitch chord type representation and the pitch-class chord representation: it expresses both transposition invariance and a limited form of octave invariance that differentiates chords with different bass pitch classes. The representation may also be defined by the following equivalence relation: chords are equivalent if they can be transposed to the same pitch-class chord representation. Pitch-class chord types are represented as sets of integers, where each chord note is represented as its pitch-class interval from the bass pitch class.[5] For example, the pitch-class chord type of the pitch-class chord (6, {2, 9}) is {0, 3, 8}. In the *hrep* package, pitch-class chord types are represented as `pc_chord_type` objects, which are themselves internally represented as numeric vectors sorted in ascending order. There are 2,048 possible non-empty pitch-class chord types in the Western 12-tone scale.

[5] The pitch-class interval from x to y is defined as $(y - x) \bmod 12$.

2.4.6 Pitch-class set type

The *pitch-class set type* representation combines the invariances of the pitch chord type representation and the pitch-class set representation, namely transposition invariance and octave invariance. The representation may be formally defined by the following equivalence relation: chords are equivalent if they can be transposed to the same pitch-class set representation.

Computing pitch-class set types is less straightforward than computing other chord types, because there is no longer a bass pitch class to anchor the representation. Fortunately, the music-theoretic discipline of pitch-class set theory provides a solution to this problem.

We begin by finding the 'normal order' (also known as the 'normal form') of the chord's pitch-class set using Rahn's (1980) algorithm, which comprises the following steps:[6]

1. Write the pitch-class set as an ascending list of integers.
2. Enumerate all possible 'cycles' of this list by repeatedly moving the first element to the end of the list. For example, the pitch-class set {0, 7, 9} has three cycles: (0, 7, 9), (7, 9, 0), and (9, 0, 7).
3. Look for the most 'compact' cycle, that is, the cycle with the smallest ascending pitch-class interval between its first and last elements. In the case of a tie, look for the cycle with the smallest ascending pitch-class interval between the first and the second-to-last element. In the case of a further tie, look at the third-to-last-element, and so on. If there is a still a tie, choose the cycle with the smallest initial pitch-class number. In the example above, the most compact cycle is (7, 9, 0).

Having identified the pitch-class set's normal order, the pitch-class set type is computed by transposing the normal-order pitch class set so that the first element is 0; this simply means subtracting the first pitch class from all pitch classes, and expressing the result modulo 12. The pitch-class set type of {0, 7, 9} is therefore {0, 2, 5}.

There are 351 possible pitch-class set types in the Western 12-tone scale. In the *hrep* package, pitch-class chord types are represented as `pc_set_type` objects, which are themselves internally represented as numeric vectors sorted in ascending order.

[6] See Forte (1973) for a similar algorithm, and see Straus (1991) for an alternative presentation of Rahn's algorithm.

2.4.7 Further issues

Translating to pitch chords

A common task in harmony cognition research is to play a participant a chord sequence from a musical corpus that only provides pitch classes, not pitch heights (e.g. the pitch-class chord representation). If the researcher wishes to play these chords using standard musical instruments, they must assign pitch heights to these chord tones. It is possible to address this problem using simple heuristics, such as assigning the bass pitch class to the octave below middle C and the remaining pitch classes to the octave above middle C (e.g. Chapter 4), but the outcome is typically aesthetically unsatisfying and musically unrealistic. To address this problem, Chapter 5 presents a data-driven algorithm that finds optimised voicings for chord sequences on the basis of a variety of psychoacoustic features. This algorithm is implemented in a publicly available R package called *voicer*.[7]

Potential extensions

Each of these representations conveys a binary notion of presence: a chord component (e.g. a pitch or pitch class) may be absent or present, but nothing in between. It could be useful to incorporate a more graduated notion of presence, to represent the fact that given chord components might be weighted more than others, perhaps by doubling pitch classes, playing notes more loudly, or situating notes in more perceptually salient registers. It could also be useful to differentiate newly sounded notes from notes that are sustained from a previous chord. Both of these types of metadata could be incorporated as numeric or Boolean vectors with the same length as the original note vector.

Omissions

We have purposefully omitted the 'prime form' representation from pitch-class set theory, which expresses octave invariance, transposition invariance, and reflection invariance (Forte, 1973; Rahn, 1980).[8] Our reasoning is that there is little evidence for reflection invariance in harmony perception: for example, the major and minor triads are reflections of each other, yet they imply opposite emotional valences to Western listeners ('happy' versus 'sad' respectively).[9]

[7] `https://github.com/pmcharrison/voicer`

[8] Pitch-class set theorists typically use the word 'inversion' instead of 'reflection', but we use the latter to avoid confusion with the unrelated concept of inversion in tonal harmony. For example, suppose that we wish to reflect the major-triad pitch-class set $\{0, 4, 7\}$ about the pitch-class 4: we have $0 \rightarrow 8$, $4 \rightarrow 4$, and $7 \rightarrow 1$, producing the minor-triad pitch-class set $\{1, 4, 8\}$.

[9] However, see Dienes & Longuet-Higgins (2004) and Krumhansl, Sandell, & Sergeant (1987) for studies of inversion salience in the perception of serially presented tone rows.

We have also purposefully omitted several representations commonly used to study voice leading. A *multiset* generalises the notion of pitch chords and pitch-class sets to differentiate between chords containing different numbers of the same element, for example {C, C, E, G} (e.g. Tymoczko, 2008). A *voice leading* describes how each element of one chord moves to an element in the next chord, for example {F → E, A → G, C → C} (Tymoczko, 2008), and a *voice-leading type* is a transpositionally equivalent set of voice leadings (Quinn & Mavromatis, 2011; see Sears et al., 2017 for a related definition).[10] By omitting these voice-leading representations, we acknowledge the conventional distinction between harmony (the construction and arrangement of chords) and voice leading (connecting the constituent notes of successive musical chords to form simultaneous melodies).

2.5 Acoustic representations

Acoustic representations describe how chords manifest as sound. If the listener is modelled as an information-processing system, then acoustic representations correspond to the listener's input data formats.

As previously noted, our symbolic representations assume categorical pitch perception, where each chord note is subsumed into semitone-width categories. The acoustic representations relax this assumption, allowing chord tones to take continuous pitch and frequency values. This reflects the non-cognitive nature of the sound signal.

2.5.1 Frequency chord

Each *frequency chord* is represented as a set of positive real numbers, corresponding to the fundamental frequencies of the chord tones as realised by the performer. The mapping between chord pitches and chord frequencies is specified by a tuning system. One possible tuning system is 12-tone equal temperament, where the octave is defined as a 2:1 frequency ratio and the semitones equally divide the octave. This tuning system is formalised as follows:

$$f = f_{ref} \times 2^{(p-69)/12} \tag{2.1}$$

where f is the frequency (Hz), f_{ref} is the reference frequency (typically 440 Hz), and p is the pitch, expressed as a MIDI note number. In practice, stretched tunings are sometimes adopted, such that the octave corresponds to slightly

[10] Sears et al.'s (2017) representation only captures the voice leading between successive bass notes, and expresses each successive chord as a set of pitch-class intervals above the bass note; it may be considered a hybrid between harmonic and voice-leading approaches.

more than a 2:1 ratio. One such stretched tuning can be produced by replacing 12 with 11.9 in Equation 2.1, producing a stretch of approximately 10 cents per octave (Parncutt & Strasburger, 1994).

Many instruments do not force a tuning system upon the performer, but instead support dynamic pitch adjustments that can reflect the particular harmonic and melodic functions of the notes being played. It is difficult to simulate these dynamic adjustments from the musical score alone; however, these adjustments can theoretically be recovered from audio signals using polyphonic signal transcription (see Section 2.9.3).

2.5.2 Sparse frequency spectrum

The *sparse frequency spectrum* is generated from the frequency chord by incorporating knowledge of the spectral content of the tones used to play the chord (Figure 2.2A). Each chord tone is modelled as a *complex tone*, defined as a superposition of pure tones, each of known frequency and amplitude. This definition ignores phase as well as any temporal evolution of the chord tones. Pitched instruments are typically modelled as *harmonic complex tones*, where the ith partial has a frequency i times that of the fundamental frequency and an amplitude that decreases steadily with increasing i. The ith partial is then termed the ith harmonic. Similar to previous literature (e.g. Hutchinson & Knopoff, 1978; Milne & Holland, 2016; Parncutt & Strasburger, 1994), we default to modelling each complex tone with 11 partials, and give the ith partial an amplitude of $1/i$, with the amplitudes being provided in arbitrary units. These defaults are somewhat arbitrary, and the researcher is encouraged to adjust the number of harmonics and their amplitudes to reflect the particular sound being modelled.

Once each chord tone is decomposed into its partials, these partials must be combined together into one spectrum. For this combination process, it is common to define a procedure for combining partials from different tones with (approximately) the same frequency. Such a procedure requires two decisions: a) how close do two partials need to be before they will be combined, and b) what is the amplitude of the resulting partial. The latter question is easy to answer: in most musical instruments, different chord tones will be produced by different oscillators, so they will have incoherent phases, and it is a standard result that the incoherent superposition of two waves with (approximately) identical frequencies and amplitudes x, y yields a resultant amplitude of

$$(x^2 + y^2)^{0.5}. \tag{2.2}$$

The former question has a less consistent answer in the literature. Here we adopt the heuristic of expressing each frequency as a (possibly non-integer) MIDI

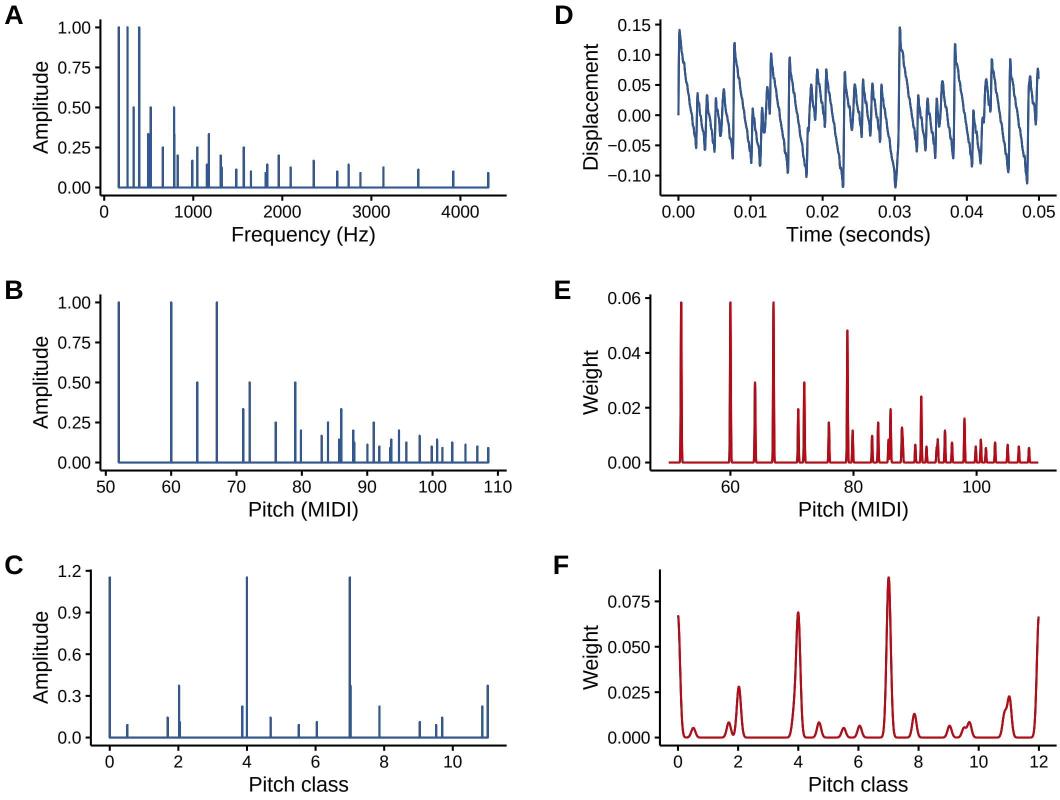

Figure 2.2: Acoustic and sensory representations for a C major chord in first inversion comprising the MIDI pitches {52, 60, 67}, synthesised with 11 harmonics with amplitude roll-off $1/n$. (**A**) Sparse frequency spectrum; (**B**) sparse pitch spectrum; (**C**) sparse pitch-class spectrum; (**D**) waveform; (**E**) smooth pitch spectrum; (**F**) smooth pitch-class spectrum.

note number, rounding the MIDI note number to six decimal places, combining any partials with identical results, and translating the MIDI note numbers back to frequencies. Six decimal places should be sufficiently precise to differentiate non-identical partials in most practical contexts, but sufficiently lax to be robust to floating-point inaccuracies in software implementations. Alternatively, one could adopt a perceptually motivated criterion related to perceptual discrimination thresholds.

Ideally, computational models taking sparse frequency spectra as input should be invariant, in the limit, to the arbitrary decision of whether or not to combine to partials with almost-identical frequencies; unfortunately, this is seldom true in practice. For example, Dillon (2013) explains how several historic consonance models (e.g. Hutchinson & Knopoff, 1978; Sethares, 1993, 2005) fail to express invariance to partial combination, and presents an alternative consonance model that captures this desired invariance. However, this model has yet to be systematically tested, and it is therefore unclear how much the invariance principle matters in practice.

2.5.3 Sparse pitch spectrum

The *sparse pitch spectrum* corresponds to a version of the sparse frequency spectrum where partial frequencies are expressed as MIDI note numbers (Figure 2.2B). This logarithmic transformation provides a more intuitive way of visualising how spectral content maps to the Western scale, and provides a better account of the subjective notion of pitch distance. The mapping between frequency and MIDI note number is defined by Equation 2.1; note that the resulting MIDI note numbers can take non-integer values.

2.5.4 Sparse pitch-class spectrum

The *sparse pitch-class spectrum* corresponds to an octave-invariant version of the sparse pitch spectrum (Figure 2.2C). It can be computed from the sparse pitch spectrum by expressing each pitch as a pitch class (using the modulo 12 operation) and combining partials with identical pitch classes. As before, pitch classes are rounded to six decimal places before testing for equality, and amplitudes are combined assuming incoherent summation (Equation 2.1).

It is also possible to compute pitch-class spectra from pitch-class sets, because both representations are invariant to octave transpositions of chord tones. This may be achieved by constructing a pitch chord containing one instantiation of every pitch class in the pitch-class set, and then computing the sparse pitch spectrum and hence the sparse pitch-class spectrum as described above.

The sparse pitch-class spectrum is closely related to the *chroma vector* representation commonly used in music audio analysis (e.g. Collins et al., 2014; Lartillot et al., 2008); both are octave-invariant representations that capture the perceptual notion of chroma. The primary difference between the two is that the former represents the pitch-class domain as a continuous space, whereas the latter partitions the pitch-class domain into 12 categories. The former approach is preferable for an acoustic representation because it avoids imposing the cultural assumption of a specific scale system.

2.5.5 Waveform

The *waveform* characterises the chord as the fluctuation of sound pressure over time (Figure 2.2D). It may be computed from the sparse frequency spectrum using additive synthesis, where separate sine waves are computed for each partial in the spectrum, and then additively combined. This operation may be approximately reversed by conducting a Fourier transform and then applying a spectral peak-picking algorithm (see e.g. Essentia toolbox, Bogdanov et al., 2013; Mirtoolbox, Lartillot et al., 2008).

2.6 Sensory representations

Sensory representations describe perceptual internalisations of acoustic signals. They may be seen as indirect consequences of symbolic representations, in that explicit symbolic representations in musical scores are translated by musicians into performances, which are then heard by listeners. They may also be seen as psychological predecessors to listeners' internal and implicit symbolic representations, automatically derived from sensory representations through processes of categorical perception and template recognition (Figure 2.3).

2.6.1 Smooth pitch/pitch-class spectra

In order to simulate how perceptual inaccuracies reduce the effective resolution of pitch perception, computational implementations of sensory representations can apply pitch-domain *smoothing.* Milne and colleagues have shown that a simple model of this smoothing effect is remarkably effective for reproducing a variety of musical phenomena (e.g. diatonic tonality, Milne et al., 2015; harmonic distance, Milne & Holland, 2016; perceived change, Dean, Milne, & Bailes, 2019). Our smooth spectra representations are computed from sparse spectra using this smoothing technique. The sparse spectrum is first expressed as a *dense spectrum*, a numeric vector where each bin corresponds to a pitch

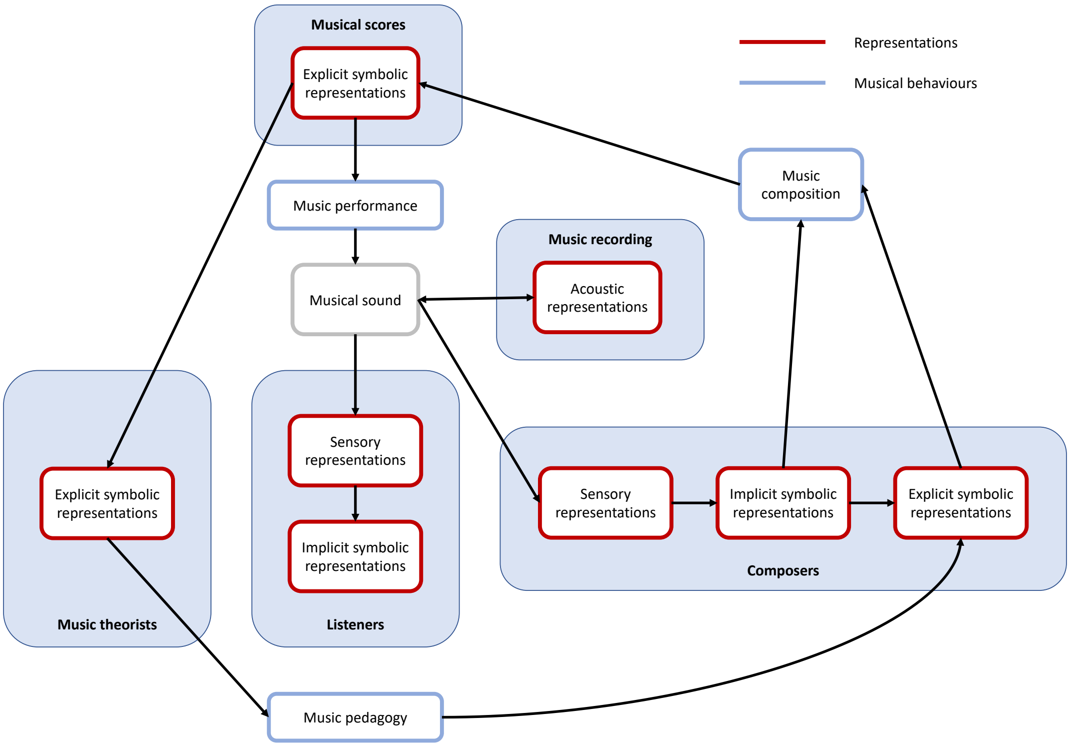

Figure 2.3: Schematic diagram illustrating the role of symbolic, acoustic, and sensory representations in different musical activities. Explicit symbolic representations are used in musical scores, which are analysed by music theorists to yield insights into musical practice, insights which are then conveyed to composers through music pedagogy. These explicit symbolic representations reflect implicit symbolic representations held by music listeners, which are themselves derived from sensory representations that listeners construct from musical sound, which can be recorded using acoustic representations.

window of one cent (i.e. a hundredth of a semitone) and each bin value corresponds to the spectral amplitude in that pitch window. This is achieved by initialising a 0-indexed vector of zeros with 1200 elements per octave, where the 0th element corresponds to 0 cents, the 1st element corresponds to 1 cent, and so on, and then iterating through every partial in the sparse spectrum, rounding its frequency to the nearest cent, and incrementing the corresponding element in the vector using the incoherent amplitude summation rule (Equation 2.1). To produce the smooth spectrum, we then convolve this dense spectrum with a Gaussian function with unit mass and standard deviation of 10 cents, after Dean et al. (2019). In the case of pitch-class spectra, this should be a circular convolution, so that the smoothing wraps round the extremes of the pitch-class vector. Perceptual similarity between smooth spectra can then be simulated using geometric similarity measures such as cosine similarity:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_i x_i y_i}{\sqrt{\sum_j x_j^2 \sum_k y_k^2}} \tag{2.3}$$

where $s(\mathbf{x}, \mathbf{y})$ is the cosine similarity between vectors $\mathbf{x}$ and $\mathbf{y}$, and x_i, y_i denote the ith elements of $\mathbf{x}$ and $\mathbf{y}$ respectively. This method can be used to produce two types of smooth spectrum: the *smooth pitch spectrum* (Figure 2.2E), and the *smooth pitch-class spectrum* (Figure 2.2F). Both pitch (Dean et al., 2019; Milne et al., 2016) and pitch-class versions (Milne & Holland, 2016; Milne et al., 2015) have proved useful in perceptual modelling.

Several subtleties should be acknowledged when implementing these spectral smoothing techniques. First, the original presentation of the technique treated the smoothed spectrum as an 'expectation vector', with each element being interpreted as the expected number of tones that the listener should hear at that pitch or pitch class (Milne et al., 2011). However, this probabilistic interpretation has limited empirical basis, and subsequent work often prefers a less specific 'perceptual weight' interpretation. Second, Milne assumed that the ith harmonic of a harmonic complex tone should additively contribute a value of $1/i^\rho$ to this perceptual weight spectrum, where ρ is a numeric parameter optimised to the data. This assumption of additive combination in perceptual space is difficult to reconcile with the formula for the acoustic combination of incoherent partials (Equation 2.1). However, in later work Milne and colleagues remove this inconsistency by applying the smoothing directly to the acoustic amplitude spectrum (Dean et al., 2019). We prefer this latter approach for its applicability to arbitrary acoustic spectra, but implement both approaches in the *hrep* package.

As is apparent from Figure 2.1, the computational translation from pitch-class sets to smooth pitch-class spectra is not well-defined. This is because pitch-class sets discard information about the number of chord pitches representing each pitch class, information which is retained implicitly in the smooth pitch-class spectrum. However, if necessary a representative spectrum can still be computed by assuming that each pitch class is assigned exactly one chord pitch.

2.6.2 Potential alternatives

Several other researchers have formulated perceptual representation schemes that can be applied to harmony modelling. Parncutt & Strasburger (1994) present a psychoacoustic model of music perception, based on Terhardt's (1982a) pitch perception model, that incorporates features such as auditory masking and spectral-domain pitch detection. We have implemented this model in the freely available R package *parn94*.[11] This model can be used as an alternative to the smooth pitch spectrum representation described here. Leman (2000a) provides an alternative psychoacoustic model of pitch perception, which combines Van Immerseel & Martens' (1992) simulation of the auditory periphery with a time-domain pitch detection algorithm. The model takes an audio signal as input, and returns a time-varying vector of activations for different pitch windows; this representation can then be used as an alternative for the smooth pitch spectrum representation.[12] Collins et al. (2014) implement two further representations that derive from Leman's (2000a) representation: a chroma vector representation and a tonal space representation. The chroma vector representation adds octave invariance, whereas the tonal space representation projects Leman's representation onto a self-organising map in the shape of a torus, intended to capture the topological structure of Western tonality.[13] Each of these models (Collins et al., 2014; Leman, 2000a; Parncutt, 1989) add significant complexity to the process of deriving sensory representations; future work needs to examine the extent to which this added complexity is useful for characterising harmony cognition.

2.7 Alphabets

When analysing the statistics of chord progressions, it is useful to define an 'alphabet' that enumerates each of the possible values for the chord representation. For example, this alphabet could define the indices of an n-gram transition

[11] `https://github.com/pmcharrison/parn94`; `https://doi.org/10.5281/zenodo.2545759`

[12] At the time of writing, Leman's (2000a) model was available in the IPEM toolbox, hosted at `https://github.com/IPEM/IPEMToolbox`.

[13] At the time of writing, the implementation of Collins et al. (2014) was available in the Janata Lab Music Toolbox, hosted at `http://atonal.ucdavis.edu/resources/software/jlmt/`.

matrix, or the coding of a one-hot input vector for a neural network. The classic harmonic alphabet from the literature is the 'Forte number', which indexes the prime forms of pitch-class sets;[14] however, this representation is not part of our representational network for reasons discussed previously. We instead define new alphabets for four representations in our network: pitch-class sets, pitch-class set types, pitch-class chord types, and pitch-class chords.

2.7.1 Pitch-class set

As defined above, a pitch-class set is a non-empty unordered set of integers between 0 and 11 inclusive (e.g. {0, 4, 7}). A pitch-class set may be encoded as follows:

1. Represent the pitch-class set as a 0-indexed binary vector of length 12, where the ith element is 1 if the pitch-class set contains the pitch class i and 0 otherwise. The pitch-class set {0, 4, 7} would therefore receive the binary vector representation '100010010000'.
2. Reinterpret this binary vector as a base-2 number.
3. Convert this base-2 number to a base-10 integer.[15]

The result is an integer encoding that ranges between 1 and 4,095; for example, {11} is coded as 1, {10} is coded as 2, {10, 11} is coded as 3, and {0, 4, 7} is coded as 2,192. The full alphabet can then be enumerated by iterating through the integers from 1 to 4,095.

2.7.2 Pitch-class set type

As defined above, pitch-class set types define transpositionally invariant equivalence classes over pitch-class sets; for example, the pitch-class set type {0, 4, 8} contains the pitch-class sets {0, 4, 8}, {1, 5, 9}, {2, 6, 10}, and {3, 7, 11}. The alphabet of pitch-class set types is constructed as follows:

1. Initialise an empty list.
2. Iterate sequentially through the pitch-class set alphabet, and:
   (a) Compute the pitch-class set type of the pitch-class set;
   (b) If this is the first occurrence of that pitch-class set type, append it to the list.

[14] See Carter (2002) and Martino (1961) for related efforts.

[15] A binary vector of length 12 can be converted to a base-10 integer by multiplying each ith element ($1 \leq i \leq 12$) by 2^{12-i} and summing the result.

The resulting alphabet maps pitch-class set types to integers between 1 and 351; for example, {0} maps to 1, {0, 1} maps to 2, {0, 2} maps to 3, and {0, 1, 2} maps to 4.

2.7.3 Pitch-class chord type

As defined above, pitch-class chord types define transpositionally invariant equivalence classes over pitch-class chords; for example, the pitch-class chords (0, {4, 7}) and (1, {5, 8}) both belong to the pitch-class chord type {0, 4, 7}. By definition, the integer-vector representation of the pitch-class chord type must begin with 0; the remaining elements may then be drawn arbitrarily from the integers 1 to 11. Correspondingly, pitch-class chord types are encoded as follows:

1. Construct a 1-indexed binary vector of length 11, where the ith element is 1 if the integer vector contains i and 0 otherwise.
2. Convert the resulting binary vector to base 10.
3. Add 1.

The resulting alphabet maps pitch-class chord types to integers between 1 and 2,048; for example, {0} maps to 1, {0, 1} maps to 1,025, and {0, 4, 7} maps to 145.

2.7.4 Pitch-class chord

Each of the 2,048 pitch-class chord types is an equivalence class containing exactly 12 pitch-class chords with 12 different bass pitch classes. Correspondingly, pitch-class chords are encoded as follows:

1. Take the bass pitch class of the pitch-class chord, expressed as an integer.
2. Multiply it by 2,048.
3. Add the integer encoding of the pitch-class chord type.

The resulting alphabet maps pitch-class chords to integers between 1 and 24,576; for example, (0, {3, 6}) maps to 289, (0, {4, 7}) maps to 145, and (4, {0, 7}) maps to 8,457.

2.8 Deriving higher-level representations

The low-level representations described above are intended to reflect relatively early stages of perceptual and cognitive processing. We will now discuss several higher-level cognitive representations for harmonic structure, alongside ways in which these representations may be simulated.

2.8.1 Voice leading

Chords played in succession can imply latent melodic lines, even if these melodic lines are not made explicit by performance characteristics such as timbre and loudness. Under Bregman's (1990) terminology of auditory scene analysis, we can say that successive notes in chord progressions are grouped into parallel auditory streams, where individual streams cohere on the basis of pitch proximity between adjacent notes. This organization into auditory streams, or melodic lines, may be termed *voice leading*.

We may wish to simulate this process of inferring voice leading from a chord sequence. Tymoczko (2006) has presented an efficient algorithm for computing voice leadings within pairs of chords; this algorithm deterministically finds a voice leading that minimises the aggregate distance moved by the individual melodic lines. In its original presentation, the algorithm finds minimal voice leadings between pitch-class sets, but it can trivially be modified to find minimal voice leadings between pitch chords. We have implemented this algorithm in the freely available R package *minVL*.[16]

2.8.2 Chord roots

Chord roots have been considered central to harmonic analysis since Rameau (1722). A chord root may be defined as a single pitch class that summarises the tonal content of a given chord. Chords are often notated with reference to these roots: for example, lead sheets typically represent chords by combining the root pitch class (written in letter-name notation, e.g. 'D') with a textual label expressing the other pitch classes relative to the root (e.g. 'maj7' implies that the chord contains pitch classes 0, 4, 7, and 11 semitones above the root). Roman numeral notation also uses chord roots, which are typically expressed relative to the local tonic (see Section 2.8.3).

It has been proposed that chord roots have a psychological basis in pitch perception (Parncutt, 1988; Terhardt, 1974, 1982). Terhardt and Parncutt argue that listeners infer fundamental frequencies of acoustic spectra by comparing these spectra to harmonic templates with different fundamental frequencies.

[16] `https://github.com/pmcharrison/minVL`

Candidate fundamental frequencies are perceived as *virtual pitches*, which Terhardt and Parncutt equate with chord roots. Recent empirical work indeed supports the notion that chord roots correspond to chord tones with particularly high perceptual salience, and that this perceptual salience can be predicted by computational models of pitch detection (Parncutt et al., 2019).

These empirical results motivate the use of root-finding models for simulating higher-level cognitive representations of harmonic structure. Various root-finding models exist, some deriving from psychoacoustics, some deriving from music theory, and some deriving from statistical modelling. A key example of the psychoacoustic approach is Parncutt's (1988) root-finding model, which was derived from Terhardt's (1982) root-finding model and subsequently extended in Parncutt (1997). The model provides a straightforward hand-computable approach to estimating the chord root for a given pitch-class set; we have implemented it in the freely available R package *parn88*.[17] Parncutt's (1989) model relaxes the octave equivalence assumption and provides a much more detailed account of peripheral psychoacoustic processes such as masking (see also Parncutt & Strasburger, 1994); we have implemented this model in the R package *parn94*.[18] Examples of the music-theoretic approach include Temperley's (1997) algorithm, Pardo & Birmingham's (2002) template-matching algorithm, Sapp's (2007) stack-of-thirds algorithm, and Cambouropoulos et al.'s (2014) general chord type algorithm. Raphael & Stoddard's (2004) algorithm follows the statistical approach, performing harmonic analysis using Hidden Markov Models. These root-finding algorithms provide several options for simulating root-based cognitive representations of harmonic structure.

2.8.3 Tonality

The notion of tonality is fundamental to Western music theory. Common-practice tonality is said to be organised around 24 possible keys, where each key is defined by a) choosing one of the 12 pitch classes to be the tonic and b) choosing one of the two diatonic modes: major and minor. The chosen key is then reflected in the distribution of notes selected by the composer. The seven most common pitch classes in a given key constitute the key's scale, from which most of the pitch material is drawn.

Music theorists often represent chord progressions relative to the tonic. Such representations may be termed *scale-degree representations*, because each pitch class is represented as a degree of the key's underlying scale. It is thought that Western listeners also form implicit scale-degree representations when listening

[17] `https://github.com/pmcharrison/parn88`

[18] `https://github.com/pmcharrison/parn94`

to Western tonal music (e.g. Arthur, 2018). The key challenge in constructing such representations automatically is to simulate the process of *key finding*; that is, given a musical extract, to estimate the underlying key at each point in the extract.

Various key-finding models have been introduced over the years. Arguably most influential is the Krumhansl-Schmuckler algorithm (Krumhansl, 1990), which estimates the key of a given musical passage by correlating pitch-class prevalences in a given musical extract (termed *key profiles*) with experimentally determined key profiles for major and minor keys. Several variants on this approach have subsequently been presented, with these variants modifying the key profiles (Aarden, 2003; Albrecht & Shanahan, 2013; Bellmann, 2005; Sapp, 2011; Temperley, 1999), replacing key profiles with interval profiles (Madsen & Widmer, 2007), adding psychoacoustic modelling (Huron & Parncutt, 1993), implementing sliding windows (Shmulevich & Yli-Harja, 2000), and incorporating self-organising maps (Schmuckler & Tomovski, 2005). Outside the Krumhansl-Schmuckler tradition, several algorithms instead take statistical approaches to key finding, including Temperley's (2007) Bayesian model, Hu & Saul's (2009) probabilistic topic model, Quinn's (2010) chord-progression model, and White's (2018) feedforward model. Chew (2002) has also presented an approach for identifying key regions within a musical piece based on the geometric Spiral Array model (Chew, 2000). These models provide many options for simulating key-based representations of harmonic progressions.

For music theorists, tonality typically implies more than key finding: it also describes high-level principles such as how certain chords are more stable than others, how certain chords tend to resolve to other chords, and how harmonic progressions can be organised into hierarchies that potentially span the entire musical piece. These principles are typically characterised as both musical and perceptual: they are intended to describe both how tonal music is created by composers and how it is perceived by listeners. Various models have been developed to summarise these principles of tonal harmony, which typically embed chords within some kind of 'tonal space' within which the notion of distance between chords is well-defined. One prominent example is Lerdahl's (1988) Tonal Pitch Space model, which operates on a symbolic level and is based on a hierarchical conceptualisation of the Western chromatic scale. A contrasting example is the Tonal Space model of Janata et al. (2002; see also Collins et al., 2014), which projects audio input onto a toroidal ('doughnut-shaped') self-organising map. Pretrained on an artificial melody that modulates through all 24 diatonic keys, the Tonal Space model is supposed to learn a topological representation of Western tonality (see also Toiviainen & Krumhansl, 2003). These kinds of models may be used to simulate tonality-based cognitive representations of harmonic progressions.

2.8.4 Functional harmony

Music theorists often assign chords to functional categories such as 'tonic', 'dominant', and 'subdominant'. These categories reflect the chord's implications for surrounding chords: for example, dominant chords are particularly likely to be followed by tonic chords. The categories are typically associated with certain chord roots, expressed relative to the prevailing tonality: for example, the dominant chord category is particularly associated with chord roots situated on the fifth degree of the prevailing diatonic scale. It is possible that untrained Western listeners also possess implicit knowledge of these functional categories, acquired through incidental exposure to Western music. Recent work has demonstrated that the acquisition of these functional categories can be simulated using computational techniques such as hidden Markov models (White & Quinn, 2018) and the information bottleneck algorithm (Jacoby, Tishby, & Tymoczko, 2015). Such algorithms could be incorporated within cognitive models of harmony perception to simulate implicit knowledge of functional chord categories.

2.9 Corpus translation

Music cognition research often takes advantage of corpora of music compositions, typically either to analyse the cognitive principles underlying the creative process, or to simulate processes of statistical learning from lifetime musical exposure. Various corpora of chord sequences exist, but these corpora are typically not expressed in the representation schemes described above. We will now describe various solutions to this problem.

2.9.1 Textual symbols

Many digitally encoded music corpora notate chords using textual symbols such as 'min9' and 'sus4' (e.g. the Billboard Corpus, Burgoyne, 2011; the iRb corpus, Broze & Shanahan, 2013). Translating these corpora into our representational network requires decoding these symbols. Unfortunately this is a poorly defined task, as there is no canonic mapping between these symbols and pitch-class content; instead, musicians use their experience to interpret these symbols, and can exert freedom to choose different chord realisations on the basis of musical context and personal preference. Ideally, a cognitive musicologist would formalise this process by analysing many performances and quantifying the frequency with which different textual symbols are mapped to different chord pitches in different contexts. However, in the absence of such prior work, a useful stopgap is to construct a mapping between textual symbols and chord pitches on the basis of music theory, and use this mapping to translate corpora such as the Billboard Corpus and the iRb corpus into our representational network. While such a mapping will not capture all the nuances of real-life performance, it should nonetheless provide a useful approximation for supporting downstream psychoacoustic and cognitive analyses.

We have compiled such a mapping between textual symbols and chord pitches. Reflecting the typical construction of lead sheets, these textual symbols correspond to *chord qualities*, expressing the pitch-class content of a chord relative to the chord's root. This mapping can be accessed through the function `decode_chord_quality` in the *hrep* R package. Assuming that the chord's root (and possibly its bass note) can be extracted by an automatic parser, this mapping is theoretically sufficient to convert the chord symbols of a lead sheet into pitch-class chord representations. In practice, new corpora and notation schemes are likely to contain occasional textual symbols not present in our mapping; however, these cases should be rare enough for the researcher to update the mapping manually. We would like to encourage future researchers to submit such updates to the package's online repository,[19] so that the mapping can become more comprehensive with time. Using this mapping, we have translated the Billboard popular music corpus and the iRb corpus into pitch-class chord notation. The resulting corpora are available in our *hcorp* package.[20]

It may sometimes be useful to perform the reverse operation, mapping integer-based representations of chords to textual symbols. To this end, Hedges & Wiggins (2016b) provide a useful algorithm that takes as input a pitch-class set, which should be expressed relative to the chord root, and returns a textual label corresponding to the chord quality, for example 'min7', 'dim', or 'sus'. Combined with a root-finding algorithm (e.g. Cambouropoulos et al., 2014; Parncutt, 1988, 1997; Sapp, 2007), this algorithm could be used to convert the symbolic representations described here into a text-based format readable by musicians.

2.9.2 Polyphonic scores

Music corpora from the tradition of Western art music typically do not notate chord sequences explicitly, but instead notate the full polyphonic texture of the original musical scores. An important challenge for harmonic analysis is then to derive chord sequences from these polyphonic textures.

[19] `http://hrep.pmcharrison.com`

[20] `https://github.com/pmcharrison/hcorp`

One approach is *full expansion* (Conklin, 2002), also known as 'salami slicing' or 'chordifying'. Here the composition is partitioned into chords at each point where a new onset occurs in any voice.[21] Each chord is then defined as the full set of pitches that sound within the corresponding partition. An advantage of this approach is that it is straightforward to automate, and it captures much of the detail present in the original musical score. We have translated a corpus of 370 chorale harmonisations by J. S. Bach into pitch chords using this approach, and have made the resulting corpus available in the *hcorp* package.[22]

A limitation of full expansion is that it fails to differentiate the function of different tones within a sonority. Music theorists often differentiate tones into 'chord tones' and 'non-chord tones', where the chord tones represent a prototypical sonority such as a major or minor triad, and the non-chord tones embellish this sonority through techniques such as suspensions and passing notes. Depending on the application, a researcher may wish to exclude chord tones from their chord sequences, and potentially even add implied but missing chord tones (e.g. the fifth in a seventh chord). This differentiation of chord tones from non-chord tones is vital for chord labelling (deriving labels such as 'maj9', 'sus4') and for functional harmonic analysis. Unfortunately, the process is rather subjective, as epitomised by the famous debates over the analysis of the Tristan chord (Martin, 2008). Nonetheless, computational approximations to this process are possible. For example, Rohrmeier & Cross (2008) provide a simple approach for chorale harmonisations by J. S. Bach, where pitch-class sets are generated by full expansion, and then one pitch-class set is chosen for each quarter-note segment of the chorale, with this pitch-class set being chosen to minimise dissonance. Pardo & Birmingham (2002) present a rule-based system where chord qualities are inferred using template matching, and partition points are found using dynamic programming. Alternative approaches to chord-tone detection and chord labelling can be found in Barthélemy & Bonardi (2001), Chen & Su (2018), Chuan & Chew (2011), Ju, Condit-Schultz, Arthur, & Fujinaga (2017), Kröger, Passos, Sampaio, & Cidra (2008), Masada & Bunescu (2017), Mearns (2013), Raphael & Stoddard (2004), Temperley & Sleator (1999), and Winograd (1968). These works vary in the extent to which they apply explicit music-theoretic rules; recent work minimises the role of explicit knowledge, and instead uses algorithms that learn these principles from musical data. In almost all cases, these approaches are targeted towards replicating the annotations of expert music theorists; we are unaware of any such work actively engaging with empirical studies of music perception. This strategy is perfectly acceptable for

[21] Alternatively, one could partition at both onset and offset points.

[22] The original corpus was sourced from KernScores (`http://kern.humdrum.org`; Sapp, 2005).

many applications in music informatics, but it is dissatisfying for applications in music cognition research. Future work should ideally examine the psychological foundations of chord-tone identification and chord labelling and use this knowledge to improve the cognitive validity of these algorithms.

2.9.3 Audio

Chord sequences can also be derived from audio files. This is generally a harder task than deriving chord sequences from symbolic music representations, but one that has received much attention in recent years. Effective chord transcription algorithms could vastly widen the amount of data available for music corpus analyses.

Many systems have been proposed for deriving sequences of chord labels from audio files (e.g. Boulanger-Lewandowski, Bengio, & Vincent, 2013; Haas et al., 2012; Lee & Slaney, 2008; Mauch, 2010; Mauch, Noland, & Dixon, 2009). These systems are typically data-driven, training either Hidden Markov Models or deep neural networks on datasets pairing audio frames with chord labels, and producing algorithms capable of generating chord labels for unseen audio. These chord labels could then be translated to our representational network using the methods described above.

One could alternatively begin by automatically transcribing the full polyphonic score. Automatic polyphonic transcription remains a very difficult task, but substantial progress has been made in recent years (e.g. Benetos & Dixon, 2013; Boulanger-Lewandowski, Vincent, & Bengio, 2012; Sigtia, Benetos, & Dixon, 2016). Having transcribed a full polyphonic score, chord sequences could then be derived using full expansion, potentially also applying a symbolic chord-tone detection algorithm.

2.10 Conclusion

This chapter's primary contribution is the compilation of low-level harmonic representations for music cognition research. Our description clarifies the implicit cognitive assumptions of each representation, and provides computational methods for translating between representations. We hope that this explicit account will assist cognitive scientists by providing a terminological standard for these representations, providing guidance on choosing representations, and helping researchers to reconcile music corpora from heterogeneous data sources.

A second contribution is the enumeration of alphabets for four harmonic representations: pitch-class chords, pitch-class sets, pitch-class chord types, and pitch-class set types. These enumerated alphabets provide bijective mappings between set-based chord representations (e.g. '(0, {3, 6})') and integer-based representations (e.g. '289'). This enumeration is particularly useful for statistical modelling, because standard techniques (e.g. *n*-gram modelling, recurrent neural networks) often require the input data to be encoded in this integer-based format. This integer-based encoding is also useful for representing large corpora efficiently on computer systems. Defining these alphabets here will hopefully save future researchers from duplicating this work with future projects, and provide a useful standard for communicating music corpora that use these representations.

A third contribution is the creation of the open-source R package *hrep*, which implements these different representations and encodings in an easy-to-use object-oriented framework. In our experience, a large part of harmony modelling projects tends to be occupied with the scientifically uninteresting task of taking musical corpora and translating them to the appropriate representation for a given statistical model or feature extractor. By centralising these processes in a single R package, we expect that we can save researchers significant time in this research pipeline, and facilitate rapid prototyping of different analysis approaches. Of course, many music researchers use other programming languages such as Python and Lisp, and the *hrep* package will be less useful here. However, the *hrep* package can still be used for preprocessing musical inputs to programs written in these languages, and it provides a template according to which analogous harmonic representation packages could be developed.

We have discussed how higher-level cognitive representations, such as voice leadings, chord roots, tonality, and functional harmony, can be derived from the representations included here. These representations are typically more computationally complex to implement, and are still active topics of research, so we have not implemented them in our *hrep* package. However, we have reviewed the various computational methods for deriving these higher-level representations from the low-level representations described here, and have made some of these methods available in standalone R packages.

We also discussed how to translate different musical corpora into these low-level representations. Various approaches exist, depending on whether the corpus is represented as textual chord symbols, polyphonic scores, or audio. One contribution of this chapter is to provide a systematic mapping between common chord symbols and the pitch-class sets implied by these chord symbols. This should help future researchers translate music corpora into the representations described here.

In our eyes, the main limitation of this harmonic analysis approach is the coercion of Western music to sequences of discrete chords. This is a common approach in music theory and music psychology, where the simplification helps the tractability of music analysis and psychological experimentation. However, real Western music rarely has such a simple structure. Chords are often formed implicitly by the superposition of multiple melodic or contrapuntal parts, and elaborated by structures such as passing notes and appoggiaturas. An important goal for future cognitive research is to examine the way in which listeners extract chord-like representations from these complex musical textures.

Chapter 3

Simultaneous consonance

3.1 Introduction

Simultaneous consonance is a salient perceptual phenomenon that arises from simultaneously sounding musical tones. Consonant tone combinations tend to be perceived as pleasant, stable, and positively valenced; dissonant combinations tend conversely to be perceived as unpleasant, unstable, and negatively valenced. The opposition between consonance and dissonance underlies much of Western music (e.g. Dahlhaus, 1990; Hindemith, 1945; Parncutt & Hair, 2011; Rameau, 1722; Schoenberg, 1978).[1]

Many psychological explanations for simultaneous consonance have been proposed over the centuries, including amplitude fluctuation (Vassilakis, 2001), masking of neighboring partials (Huron, 2001), cultural familiarity (Johnson-Laird et al., 2012), vocal similarity (Bowling, Purves, & Gill, 2018), fusion of chord tones (Stumpf, 1890), combination tones (Hindemith, 1945), and spectral evenness (Cook, 2009). Recently, however, a consensus is developing that consonance primarily derives from a chord's harmonicity (Bidelman & Krishnan, 2009; Bowling & Purves, 2015; Cousineau, McDermott, & Peretz, 2012; Lots & Stone, 2008; McDermott, Lehr, & Oxenham, 2010; Stolzenburg, 2015), with this effect potentially being moderated by musical exposure (McDermott et al., 2010; McDermott, Schultz, Undurraga, & Godoy, 2016).

In this chapter we question whether harmonicity is truly sufficient to explain simultaneous consonance perception. First, we critically review historic consonance research from a broad variety of disciplines, including psychoacoustics, cognitive psychology, animal behaviour, computational musicology, and ethnomusicology. Second, we reanalyse consonance perception data from

[1] By 'Western music' we refer broadly to the musical traditions of Europe and music derived from these traditions; by 'Western listeners' we refer to listeners from these musical traditions.

four previous studies representing more than 500 participants (Bowling et al., 2018; Johnson-Laird et al., 2012; Lahdelma & Eerola, 2016; Schwartz, Howe, & Purves, 2003). Third, we model chord prevalences in three large musical corpora representing more than 100,000 compositions (Broze & Shanahan, 2013; Burgoyne, 2011; Viro, 2011). On the basis of these analyses, we estimate the degree to which different psychological mechanisms contribute to consonance perception in Western listeners.

Computational modelling is a critical part of the approach. We review the state of the art in consonance modelling, empirically evaluate 20 of these models, and use these models to test competing theories of consonance. Our work results in three new consonance models: a harmonicity model based on smooth pitch-class spectra, a corpus-based cultural familiarity model, and a composite model of consonance perception that captures interference between partials, harmonicity, and cultural familiarity. We release these new models in an accompanying R package, *incon*, alongside new implementations of 14 other models from the literature (see Section 3.9.2 for details). In doing so, we hope to facilitate future consonance research in both psychology and empirical musicology.

3.2 Terminology

A collection of notes is said to be *consonant* if the notes 'sound well together', and conversely *dissonant* if the notes 'sound poorly together'. In its broadest definitions, consonance is associated with many different musical concepts, including diatonicism, centricism, stability, tension, similarity, and distance (Parncutt & Hair, 2011). For psychological studies, however, it is often useful to provide a stricter operationalisation of consonance, and so researchers commonly define consonance to their participants as the *pleasantness*, *beauty*, or *attractiveness* of a chord (e.g. Bowling & Purves, 2015; Bowling et al., 2018; Cousineau et al., 2012; McDermott et al., 2010, 2016).

In this chapter we use the term 'simultaneous' to restrict consideration to the notes within the chord, as opposed to sequential relationships between the chord and its musical context. Simultaneous and sequential consonance are sometimes termed *vertical* and *horizontal* consonance respectively, by analogy with the physical layout of the Western musical score (Parncutt & Hair, 2011). These kinds of chordal consonance may also be distinguished from 'melodic' consonance, which refers to the intervals of a melody. For the remainder of this chapter, the term 'consonance' will be taken to imply 'simultaneous consonance' unless specified otherwise.

Consonance and dissonance are often treated as two ends of a continuous scale, but some researchers treat the two as distinct phenomena (e.g. Parncutt

& Hair, 2011). Under such formulations, consonance is typically treated as the perceptual correlate of harmonicity, and dissonance as the perceptual correlate of roughness (see Section 3.3). Here we avoid this approach, and instead treat consonance and dissonance as antonyms.

3.3 Consonance theories

Here we review current theories of consonance perception. We pay particular attention to three classes of theories – periodicity/harmonicity, interference between partials, and culture – that we consider to be particularly well-supported by the empirical literature. We also discuss several related theories, including vocal similarity, fusion, and combination tones.

3.3.1 Periodicity/harmonicity

Human vocalisations are characterised by repetitive structure termed *periodicity.* This periodicity has several perceptual correlates, of which the most prominent is *pitch.* Broadly speaking, pitch corresponds to the waveform's repetition rate, or *fundamental frequency*: Faster repetition corresponds to higher pitch.

Sound can be represented either in the time domain or in the frequency domain. In the time domain, periodicity manifests as repetitive waveform structure. In the frequency domain, periodicity manifests as *harmonicity*, a phenomenon where the sound's frequency components are all integer multiples of the fundamental frequency.[2] These integer-multiple frequencies are termed *harmonics*; a sound comprising a full set of integer multiples is termed a *harmonic series.* Each periodic sound constitutes a (possibly incomplete) harmonic series rooted on its fundamental frequency; conversely, every harmonic series (incomplete or complete) is periodic in its fundamental frequency. Harmonicity and periodicity are therefore essentially equivalent phenomena, and we will denote both by writing 'periodicity/harmonicity'.

Humans rely on periodicity/harmonicity analysis to understand the natural environment and to communicate with others (e.g. Oxenham, 2018), but the precise mechanisms of this analysis remain unclear. The primary extant theories are time-domain *autocorrelation* theories and frequency-domain *pattern-matching* theories (de Cheveigné, 2005). Autocorrelation theories state that listeners detect periodicity by computing the signal's correlation with a delayed version of itself as a function of delay time; peaks in the autocorrelation function correspond to potential fundamental frequencies (Balaguer-Ballester, Denham, &

[2] In particular, the fundamental frequency is equal to the greatest common divisor of the frequency components.

Meddis, 2008; Bernstein & Oxenham, 2005; Cariani, 1999; Cariani & Delgutte, 1996; de Cheveigné, 1998; Ebeling, 2008; Langner, 1997; Licklider, 1951; Meddis & Hewitt, 1991b, 1991a; Meddis & O'Mard, 1997; Slaney & Lyon, 1990; Wightman, 1973). Pattern-matching theories instead state that listeners infer fundamental frequencies by detecting harmonic patterns in the frequency domain (Bilsen, 1977; Cohen, Grossberg, & Wyse, 1995; Duifhuis, Willems, & Sluyter, 1982; Goldstein, 1973; Shamma & Klein, 2000; Terhardt, 1974; Terhardt et al., 1982b). Both of these explanations have resisted definitive falsification, and it is possible that both mechanisms contribute to periodicity/harmonicity detection (de Cheveigné, 2005).

The prototypically consonant intervals of Western music tend to exhibit high periodicity/harmonicity (Figure 3.1). For example, perfect fifths are typically performed as complex tones that approximate 3:2 frequency ratios, where every second cycle of the lower-frequency waveform approximately coincides with every third cycle of the higher-frequency waveform. The resulting waveform therefore has a relatively high repetition rate, or periodicity (Figure 3.1B). In contrast, the dissonant tritone cannot be easily approximated by a simple frequency ratio, and so its fundamental frequency (approximate or otherwise) must be much lower than that of the lowest tone. We therefore say that the tritone has relatively low periodicity (Figure 3.1C).

It has correspondingly been proposed that periodicity/harmonicity determines consonance perception (Bidelman & Heinz, 2011; Boomsliter & Creel, 1961; Bowling & Purves, 2015; Bowling et al., 2018; Cousineau et al., 2012; Ebeling, 2008; Heffernan & Longtin, 2009; Lee, Skoe, Kraus, & Ashley, 2015; Lots & Stone, 2008; McDermott et al., 2010; Milne et al., 2016; Nordmark & Fahlén, 1988; Patterson, 1986; Spagnolo, Ushakov, & Dubkov, 2013; Stolzenburg, 2015; Terhardt, 1974; Ushakov, Dubkov, & Spagnolo, 2010).[3] The nature of this potential relationship depends in large part on the unresolved issue of whether listeners detect periodicity/harmonicity using autocorrelation or pattern-matching (de Cheveigné, 2005), as well as other subtleties of auditory processing such as masking (Parncutt, 1989; Parncutt & Strasburger, 1994), octave invariance (Milne et al., 2016; Parncutt, 1988; Parncutt et al., 2018), and nonlinear signal transformation (Lee et al., 2015; Stolzenburg, 2017). It is also unclear precisely how consonance develops from the results of periodicity/harmonicity detection; competing theories suggest that consonance is determined by the inferred fundamental frequency (Boomsliter & Creel, 1961; Stolzenburg, 2015), the absolute degree of harmonic template fit at the fundamental frequency (Bowling et al.,

[3] Periodicity theories of consonance predating the 20th century can be found in the work of Galileo Galilei, Gottfried Wilhelm Liebniz, Leonhard Euler, Theodor Lipps, and A. J. Polak (Plomp & Levelt, 1965).

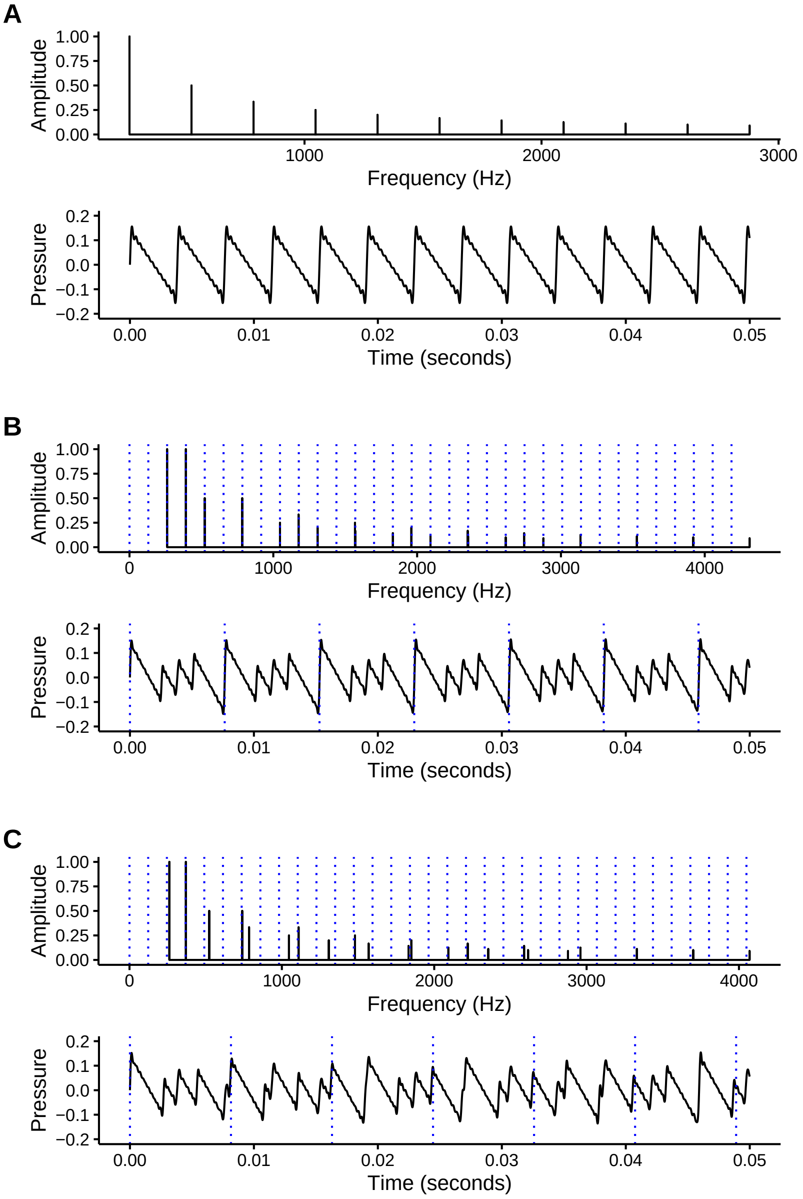

Figure 3.1: Acoustic spectra and waveforms for (**A**) a single harmonic complex tone, (**B**) two harmonic complex tones separated by an equal-tempered perfect fifth, and (**C**) two harmonic complex tones separated by an equal-tempered tritone. Each complex tone has 11 harmonics, with the ith harmonic having an amplitude of $1/i$. The blue dotted lines mark harmonic/periodic grids, which are aligned for the harmonic/periodic perfect fifth but misaligned for the inharmonic/aperiodic tritone.

2018; Gill & Purves, 2009; Milne et al., 2016; Parncutt, 1989; Parncutt & Strasburger, 1994), or the degree of template fit at the fundamental frequency relative to that at other candidate fundamental frequencies (Parncutt, 1988; Parncutt et al., 2018). This variety of hypotheses is reflected in a diversity of computational models of musical periodicity/harmonicity perception (Ebeling, 2008; Gill & Purves, 2009; Lartillot et al., 2008; Milne et al., 2016; Parncutt, 1988, 1989; Parncutt & Strasburger, 1994; Spagnolo et al., 2013; Stolzenburg, 2015). So far these models have only received limited empirical comparison (e.g. Stolzenburg, 2015).

It is clear why periodicity/harmonicity should be salient to human listeners: Periodicity/harmonicity detection is crucial for auditory scene analysis and for natural speech understanding (e.g. Oxenham, 2018). It is less clear why periodicity/harmonicity should be positively valenced, and hence associated with consonance. One possibility is that long-term exposure to vocal sounds (Schwartz et al., 2003) or Western music (McDermott et al., 2016) induces familiarity with periodicity/harmonicity, in turn engendering liking through the mere exposure effect (Zajonc, 2001). A second possibility is that the ecological importance of interpreting human vocalisations creates a selective pressure to perceive these vocalisations as attractive (Bowling et al., 2018).

3.3.2 Interference between partials

Musical chords can typically be modelled as *complex tones*, superpositions of finite numbers of sinusoidal *pure tones* termed *partials*. Each partial is characterised by a frequency and an amplitude. It is argued that neighboring partials can interact to produce *interference* effects, with these interference effects subsequently being perceived as dissonance (Dillon, 2013; Helmholtz, 1863; Hutchinson & Knopoff, 1978; Kameoka & Kuriyagawa, 1969a, 1969b; Mashinter, 2006; Plomp & Levelt, 1965; Sethares, 1993; Vassilakis, 2001).

Pure-tone interference has two potential sources: *beating* and *masking*. Beating develops from the following mathematical identity for the addition of two equal-amplitude sinusoids:

$$\cos(2\pi f_1 t) + \cos(2\pi f_2 t) = 2\cos\left(2\pi \bar{f} t\right)\cos\left(\pi \delta t\right) \tag{3.1}$$

where f_1, f_2 are the frequencies of the original sinusoids ($f_1 > f_2$), $\bar{f} = (f_1 + f_2)/2$, $\delta = f_1 - f_2$, and t denotes time. For sufficiently large frequency differences, listeners perceive the left hand side of Equation 3.1, corresponding to two separate pure tones at frequencies f_1, f_2. For sufficiently small frequency differences, listeners perceive the right hand side of Equation 3.1, corresponding to a tone of intermediate frequency $\bar{f} = (f_1 + f_2)/2$ modulated by a sinusoid of frequency $\delta/2 = (f_1 - f_2)/2$. This modulation is perceived as amplitude fluctuation with frequency equal to the modulating sinusoid's zero-crossing rate, $f_1 - f_2$. Slow amplitude fluctuation (c. 0.1–5 Hz) is perceived as a not unpleasant oscillation in loudness, but fast amplitude fluctuation (c. 20–30 Hz) takes on a harsh quality described as *roughness*. This roughness is thought to contribute to dissonance perception.

Masking describes situations where one sound obstructs the perception of another sound (e.g. Patterson & Green, 2012; Scharf, 1971). Masking in general is a complex phenomenon, but the mutual masking of pairs of pure tones can be approximated by straightforward mathematical models (Parncutt, 1989; Parncutt & Strasburger, 1994; Terhardt et al., 1982a; Wang, Shen, Guo, Tang, & Hamade, 2013). These models embody long-established principles that masking increases with smaller frequency differences and with higher sound pressure level.

Beating and masking are both closely linked with the notion of *critical bands*. The notion of critical bands comes from modelling the cochlea as a series of overlapping *band-pass filters*, areas that are preferentially excited by spectral components within a certain frequency range (Zwicker, Flottorp, & Stevens, 1957). Beating typically only arises from spectral components localised to the same critical band (Daniel & Weber, 1997). The mutual masking of pure tones approximates a linear function of the number of critical bands separating them (termed *critical-band distance*), with additional masking occurring from pure tones within the same critical band that are unresolved by the auditory system (Terhardt et al., 1982a).

Beating and masking effects are both considerably stronger when two tones are presented diotically (to the same ear) rather than dichotically (to different ears) (Buus, 1997; Grose, Buss, & Hall III, 2012). This indicates that these phenomena depend, in large part, on physical interactions in the inner ear.

There is a long tradition of research relating beating to consonance, mostly founded on the work of Helmholtz (1863; Aures, 1985a, cited in Daniel & Weber, 1997; Hutchinson & Knopoff, 1978; Kameoka & Kuriyagawa, 1969a, 1969b; Mashinter, 2006; Parncutt et al., 2018; Plomp & Levelt, 1965; Sethares, 1993; Vassilakis, 2001).[4] The general principle shared by this work is that dissonance develops from the accumulation of roughness deriving from the beating of neighboring partials.

In contrast, the literature linking masking to consonance is relatively sparse. Huron (2001) suggests that masking induces dissonance because it reflects a compromised sensitivity to the auditory environment, with analogies in visual

[4] Earlier work in a similar line can be found in Sorge (1747), cited in Plomp & Levelt (1965) and Sethares (2005).

processing such as occlusion or glare. Aures (1984; cited in Parncutt, 1989) and Parncutt (1989; Parncutt & Strasburger, 1994) also state that consonance reduces as a function of masking. Unfortunately, these ideas have yet to receive much empirical validation; a difficulty is that beating and masking tend to happen in similar situations, making them difficult to disambiguate (Huron, 2001).

The kind of beating that elicits dissonance is achieved by small, but not too small, frequency differences between partials (Figure 3.2A). With very small frequency differences, the beating becomes too slow to elicit dissonance (Hutchinson & Knopoff, 1978; Kameoka & Kuriyagawa, 1969a; Plomp & Levelt, 1965). The kind of masking that elicits dissonance is presumably also maximised by small, but not too small, frequency differences between partials. For moderately small frequency differences, the auditory system tries to resolve two partials, but finds it difficult on account of mutual masking, with this difficulty eliciting negative valence (Huron, 2001). For very small frequency differences, the auditory system only perceives one partial, which becomes purer as the two acoustic partials converge on the same frequency.

Musical sonorities can often be treated as combinations of *harmonic complex tones*, complex tones whose spectral frequencies follow a harmonic series. The interference experienced by a combination of harmonic complex tones depends on the fundamental frequencies of the complex tones. A particularly important factor is the ratio of these fundamental frequencies. Certain ratios, in particular the simple-integer ratios approximated by prototypically consonant musical chords, tend to produce partials that either completely coincide or are widely spaced, hence minimising interference (Figure 3.2B).

Interference between partials also depends on pitch height. A given frequency ratio occupies less critical-band distance as absolute frequency decreases, typically resulting in increased interference. This mechanism potentially explains why the same musical interval (e.g. the major third, 5:4) can sound consonant in high registers and dissonant in low registers.

It is currently unusual to distinguish beating and masking theories of consonance, as we have done above. Most previous work solely discusses beating and its psychological correlate, roughness (e.g. Cousineau et al., 2012; McDermott et al., 2010, 2016; Parncutt & Hair, 2011; Parncutt et al., 2018; Terhardt, 1984). However, we contend that the existing evidence does little to differentiate beating and masking theories, and that it would be premature to discard the latter in favour of the former. Moreover, we show later in this chapter that computational models that address beating explicitly (e.g. Wang et al., 2013) seem to predict consonance worse than generic models of interference between partials (e.g. Hutchinson & Knopoff, 1978; Sethares, 1993; Vassilakis, 2001).

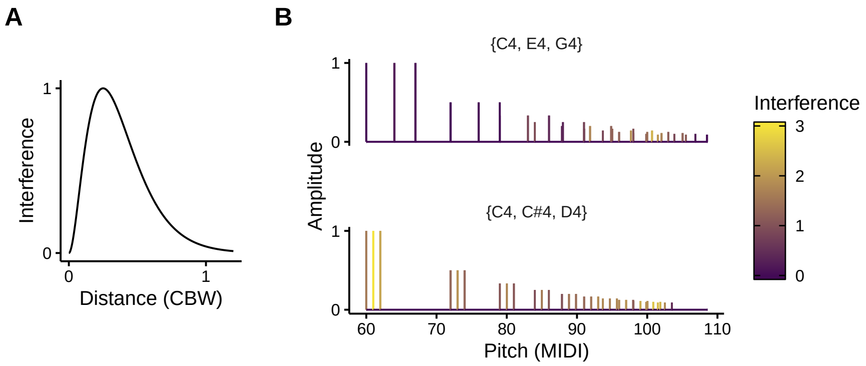

Figure 3.2: **A**: Interference between partials as a function of distance in critical bandwidths, after Hutchinson & Knopoff (1978). **B**: Interference patterns in two musical sonorities, after Hutchinson & Knopoff (1978). {C4, E4, G4} is a consonant major triad, and experiences low interference. {C4, C♯4, D4} is a dissonant cluster chord, and experiences high interference.

For now, therefore, it seems wise to contemplate both beating and masking as potential contributors to consonance.

3.3.3 Culture

Consonance may also be determined by a listener's cultural background (Arthurs, Beeston, & Timmers, 2018; Guernsey, 1928; Johnson-Laird et al., 2012; Lundin, 1947; McDermott et al., 2016; McLachlan, Marco, Light, & Wilson, 2013; Omigie, Dellacherie, & Samson, 2017; Parncutt, 2006b; Parncutt & Hair, 2011). Several mechanisms for this effect are possible. Through the mere exposure effect (Zajonc, 2001), exposure to common chords in a musical style might induce familiarity and hence liking. Through classical conditioning, the co-occurrence of certain musical features (e.g. interference) with external features (e.g. the violent lyrics in death metal music, Olsen, Thompson, & Giblin, 2018) might also induce aesthetic responses to these musical features.

It remains unclear which musical features might become consonant through familiarity. One possibility is that listeners become familiar with acoustic phenomena such as periodicity/harmonicity (McDermott et al., 2016). A second possibility is that listeners internalise Western tonal structures such as diatonic scales (Johnson-Laird et al., 2012). Alternatively, listeners might develop a granular familiarity with specific musical chords (McLachlan et al., 2013).

3.3.4 Other theories

Vocal similarity

Vocal similarity theories hold that consonance derives from acoustic similarity to human vocalisations (e.g. Bowling & Purves, 2015; Bowling et al., 2018; Schwartz et al., 2003). A key feature of human vocalisations is periodicity/harmonicity, leading some researchers to operationalise vocal similarity as the latter (Gill & Purves, 2009). In such cases, vocal similarity theories may be considered a subset of periodicity/harmonicity theories. However, Bowling et al. (2018) additionally operationalise vocal similarity as the absence of fundamental frequency intervals smaller than 50 Hz, arguing that such intervals are rarely found in human vocalisations. Indeed, such intervals are negatively associated with consonance; however, this phenomenon can also be explained by interference minimisation. To our knowledge, no studies have shown that vocal similarity contributes to consonance through paths other than periodicity/harmonicity and interference. We therefore do not evaluate vocal similarity separately from interference and periodicity/harmonicity.

Fusion

Stumpf (1890, 1898) proposed that consonance derives from *fusion*, the perceptual merging of multiple harmonic complex tones. The substance of this hypothesis depends on the precise definition of fusion. Some researchers have operationalised fusion as *perceptual indiscriminability*, that is, an inability to identify the constituent tones of a sonority (DeWitt & Crowder, 1987; McLachlan et al., 2013). This was encouraged by Stumpf's early experiments investigating how often listeners erroneously judged tone pairs as single tones (DeWitt & Crowder, 1987; Schneider, 1997). Subsequently, however, Stumpf wrote that fusion should not be interpreted as indiscriminability but rather as the formation of a coherent whole, with the sophisticated listener being able to attend to individual chord components at will (Schneider, 1997). Stumpf later wrote that he was unsure whether fusion truly caused consonance; instead, he suggested that fusion and consonance might both stem from harmonicity recognition (Plomp & Levelt, 1965; Schneider, 1997).

Following Stumpf, several subsequent studies have investigated the relationship between fusion and consonance, but with mixed findings. Guernsey (1928) and DeWitt & Crowder (1987) tested fusion by playing participants different dyads and asking how many tones these chords contained. In both studies, prototypically consonant musical intervals (octaves, perfect fifths) were most likely to be confused for single tones, supporting a link between consonance and fusion. McLachlan et al. (2013) instead tested fusion with a pitch-matching task, where each trial cycled between a target chord and a probe tone, and participants were instructed to manipulate the probe tone until it matched a specified chord tone (lowest, middle, or highest). Pitch-matching accuracy increased for prototypically consonant chords, suggesting (contrary to Stumpf's claims) that consonance was *inversely* related to fusion. It is difficult to conclude much about Stumpf's claims from these studies, partly because different studies have yielded contradictory results, and partly because none of these studies tested for *causal* effects of fusion on consonance, as opposed to consonance and fusion both being driven by a common factor of periodicity/harmonicity.

Combination tones

Combination tones are additional spectral components introduced by nonlinear sound transmission in the ear's physical apparatus (e.g. Parncutt, 1989; Smoorenburg, 1972; Wever, Bray, & Lawrence, 1940). For example, two pure tones of frequencies $f_1, f_2 : f_1 < f_2$ can elicit combination tones including the *simple difference tone* ($f = f_2 - f_1$) and the *cubic difference tone* ($f = 2f_1 - f_2$) (Parncutt, 1989; Smoorenburg, 1972).

Combination tones were once argued to be an important mechanism for pitch perception, reinforcing a complex tone's fundamental frequency and causing it to be perceived even when not acoustically present (e.g. Fletcher, 1924; see Parncutt, 1989). Combination tones were also argued to have important implications for music perception, explaining phenomena such as chord roots and perceptual consonance (Hindemith, 1945; Krueger, 1910; Tartini, 1754, cited in Parncutt, 1989). However, subsequent research showed that the missing fundamental persisted even when the difference tone was removed by acoustic cancellation (Schouten, 1938, described in Plomp, 1967), and that, in any case, difference tones are usually too quiet to be audible for typical speech and music listening (Plomp, 1965). We therefore do not consider combination tones further.

Loudness and sharpness

Aures (1985a, 1985b) describes four aspects of sensory consonance: *tonalness*, *roughness*, *loudness*, and *sharpness*. Tonalness is a synonym for periodicity/harmonicity, already discussed as an important potential contributor to consonance. Roughness is an aspect of interference, also an important potential contributor to consonance. Loudness is the perceptual correlate of a sound's energy content; sharpness describes the energy content of high spectral frequencies. Historically, loudness and sharpness have received little attention in the study of musical consonance, perhaps because music theorists and psychologists have primarily been interested in the consonance of transposition-invariant and loudness-invariant structures such as pitch-class sets, for which loudness and sharpness are undefined. We do not consider these phenomena further in this chapter, but they may ultimately prove necessary for achieving a complete perceptual account of consonance.

Evenness

The constituent notes of a musical chord can be represented as points on a *pitch line* or a *pitch-class circle* (e.g. Tymoczko, 2016). The *evenness* of the resulting distribution can be characterised in various ways, including the difference in successive interval sizes (Cook, 2009, 2017; Cook & Fujisawa, 2006), the difference between the largest and smallest interval sizes (Parncutt et al., 2018), and the standard deviation of interval sizes (Parncutt et al., 2018). In the case of Cook's (2009, 2017; 2006) models, each chord note is expanded into a harmonic complex tone, and pitch distances are computed between the resulting partials; in the other cases, pitch distances are computed between fundamental frequencies, presumably as inferred through periodicity/harmonicity detection.

Evenness may contribute negatively to consonance. When a chord contains multiple intervals of the same size, these intervals may become confusable and impede perceptual organisation, hence decreasing consonance (Cook, 2009, 2017; Cook & Fujisawa, 2006; Meyer, 1956). For example, a major triad in pitch-class space contains the intervals of a major third, a minor third, and a perfect fourth, and each note of the triad participates in a unique pair of these intervals, one connecting it to the note above, and one connecting it to the note below. In contrast, an augmented triad contains only intervals of a major third, and so each note participates in an identical pair of intervals. Correspondingly, the individual notes of the augmented triad may be considered less distinctive than those of the major triad.

Evenness may also contribute positively, but indirectly, to consonance. Spacing harmonics evenly on a critical-band scale typically reduces interference, thereby increasing consonance (see e.g. Huron & Sellmer, 1992; Plomp & Levelt, 1965). Evenness also facilitates efficient voice leading, and therefore may contribute positively to sequential consonance (Parncutt et al., 2018; Tymoczko, 2011).

Evenness is an interesting potential contributor to consonance, but so far it has received little empirical testing. We do not consider it to be sufficiently well-supported to include in this chapter's analyses, but we encourage future empirical research on the topic.

Table 3.1: Summarised evidence for the mechanisms underlying Western consonance perception.

Evidence	Interference	Periodicity	Culture
Stimulus effects			
Tone spectra	✓		
Pitch height	✓		
Dichotic presentation	✠		
Familiarity			(✓)
Chord structure	(✓)	(✓)	(✓)
→ *Section 3.6*	✓	✓	(✓)
Listener effects			
Western listeners	(✗)	✓	
Congenital amusia		✓	
Non-Western listeners			✓
Infants			(✠)
Animals			(✠)
Composition effects			
Musical scales	✓		
Manipulation of interference	✓		✓
Chord spacing (Western music)	✓		
Chord prevalences (Western music)	(✓)	(✓)	
→ *Section 3.7*	✓	✓	

Note. Each row identifies a section in Section 3.4. '✓' denotes evidence that a mechanism contributes to Western consonance perception. '✗' denotes evidence that a mechanism is *not* relevant to Western consonance perception. '✠' denotes evidence that a mechanism is insufficient to explain Western consonance perception. Parentheses indicate tentative evidence; blank spaces indicate a lack of evidence.

3.4 Current evidence

Evidence for disambiguating different theories of consonance perception can be organised into three broad categories: *stimulus effects*, *listener effects*, and *composition effects*. We review each of these categories in turn, and summarise our conclusions in Table 3.1.

3.4.1 Stimulus effects

We begin by discussing *stimulus effects*, ways in which consonance perception varies as a function of the stimulus.

Tone spectra

A chord's consonance depends on the spectral content of its tones. With harmonic tone spectra, peak consonance is observed when the fundamental frequencies are related by simple frequency ratios (e.g. Stolzenburg, 2015). With pure tone spectra, these peaks at integer ratios disappear, at least for musically untrained listeners (Kaestner, 1909; Plomp & Levelt, 1965). With inharmonic tone spectra, the peaks at integer ratios are replaced by peaks at ratios determined by the inharmonic spectra (Geary, 1980; Pierce, 1966; Sethares, 2005).[5] The consonance of harmonic tone combinations can also be increased by selectively deleting harmonics responsible for interference (Vos, 1986), though Nordmark & Fahlén (1988) report limited success with this technique.

Interference theories clearly predict these effects of tone spectra on consonance (for harmonic and pure tones, see Plomp & Levelt, 1965; for inharmonic tones, see Sethares, 1993, 2005). In contrast, neither periodicity/harmonicity nor cultural theories clearly predict these phenomena. This suggests that interference does indeed contribute towards consonance perception.

Pitch height

A given interval ratio typically appears less consonant if it appears at low frequencies (Plomp & Levelt, 1965). Interference theories predict this phenomenon by relating consonance to pitch distance on a critical-bandwidth scale; a given ratio corresponds to a smaller critical-bandwidth distance if it appears at lower frequencies (Plomp & Levelt, 1965). In contrast, neither periodicity/harmonicity nor cultural theories predict this sensitivity to pitch height.

Dichotic presentation

Interference between partials is thought to take place primarily within the inner ear (Plomp & Levelt, 1965). Correspondingly, the interference of a given pair of pure tones can be essentially eliminated by dichotic presentation, where each tone is presented to a separate ear. Periodicity/harmonicity detection, meanwhile, is thought to be a central process that combines information from both ears (Cramer & Huggins, 1958; Houtsma & Goldstein, 1972). Correspondingly, the contribution of periodicity/harmonicity detection to consonance perception should be unaffected by dichotic presentation.

Bidelman & Krishnan (2009) report consonance judgments for dichotically presented pairs of complex tones. Broadly speaking, participants continued to differentiate prototypically consonant and dissonant intervals, suggesting that interference is insufficient to explain consonance. Unexpectedly, however, the tritone and perfect fourth received fairly similar consonance ratings. This finding needs to be explored further.

[5] Audio examples from Sethares (2005) are available at `http://sethares.engr.wisc.edu/html/soundexamples.html`.

Subsequent studies have investigated the effect of dichotic presentation on consonance judgments for pairs of pure tones (Cousineau et al., 2012; McDermott et al., 2010, 2016). These studies show that dichotic presentation reliably increases the consonance of small pitch intervals, in particular major and minor seconds, as predicted by interference theories. This would appear to support interference theories of consonance, though it is unclear whether these effects generalise to the complex tone spectra of real musical instruments.

Familiarity

McLachlan et al. (2013, Experiment 2) trained nonmusicians to perform a pitch-matching task on two-note chords. After training, participants judged chords from the training set as more consonant than novel chords. These results could be interpreted as evidence that consonance is positively influenced by exposure, consistent with the mere exposure effect, and supporting a cultural theory of consonance. However, the generalisability of this effect has yet to be confirmed.

Chord structure

Western listeners consider certain chords (e.g. the major triad) to be more consonant than others (e.g. the augmented triad). It is possible to test competing theories of consonance by operationalising the theories as computational models and testing their ability to predict consonance judgments.

Unfortunately, studies using this approach have identified conflicting explanations for consonance:

a) Interference (Hutchinson & Knopoff, 1978);

b) Interference and additional unknown factors (Vassilakis, 2001);

c) Interference and cultural knowledge (Johnson-Laird et al., 2012);

d) Periodicity/harmonicity (Stolzenburg, 2015);

e) Periodicity/harmonicity and interference (Marin, Forde, Gingras, & Stewart, 2015);

f) Interference and sharpness (Lahdelma & Eerola, 2016);

g) Vocal similarity (Bowling et al., 2018).

These contradictions may often be attributed to methodological problems:

a) Different studies test different theories, and rarely test more than two theories simultaneously.

b) Stimulus sets are often too small to support reliable inferences.[6]

c) Stolzenburg (2015) evaluates models using pairwise correlations, implicitly assuming that only one mechanism (e.g. periodicity/harmonicity, interference) determines consonance. Multiple regression would be necessary to capture multiple simultaneous mechanisms.

d) The stimulus set of Marin et al. (2015) constitutes 12 dyads each transposed four times; the conditional dependencies between transpositions are not accounted for in the linear regressions, inflating Type I error.

e) Johnson-Laird et al. (2012) do not report coefficients or p-values for their fitted regression models; they do report hierarchical regression statistics, but these statistics do not test their primary research question, namely whether interference and cultural knowledge *simultaneously* contribute to consonance.

f) The audio-based periodicity/harmonicity model used by Lahdelma & Eerola (2016) fails when applied to complex stimuli such as chords (see Section 3.6).

These methodological problems and contradictory findings make it difficult to generalise from this literature.

3.4.2 Listener effects

We now discuss *listener effects*, ways in which consonance perception varies as a function of the listener.

Western listeners

McDermott et al. (2010) tested competing theories of consonance perception using an individual-differences approach. They constructed three psychometric measures, testing:

a) *Interference preferences*, operationalised by playing listeners pure-tone dyads and subtracting preference ratings for dichotic presentation (one tone in each ear) from ratings for diotic presentation (both tones in both ears);

b) *Periodicity/harmonicity preferences*, operationalised by playing listeners subsets of a harmonic complex tone and subtracting preference ratings for the original version from ratings for a version with perturbed harmonics;

[6] For example, Stolzenburg (2015, Table 4) tabulates correlation coefficients for 15 consonance models as evaluated on 12 dyads; the median correlation of .939 has a 95% confidence interval spanning from .79 to .98, encompassing all but one of the reported coefficients.

c) *Consonance preferences*, operationalised by playing listeners 14 musical chords, and subtracting preference ratings for the globally least-preferred chords from the globally most-preferred chords.

Consonance preferences correlated with periodicity/harmonicity preferences but not with interference preferences. This suggests that consonance may be driven by periodicity/harmonicity, not interference. However, these findings must be considered preliminary given the limited construct validation of the three psychometric measures. Future work must examine whether these measures generalise to a wider range of stimulus manipulations and response paradigms.

Congenital amusia

Congenital amusia is a lifelong cognitive disorder characterised by difficulties in performing simple musical tasks (Ayotte, Peretz, & Hyde, 2002; Stewart, 2011). Using the individual-differences tests of McDermott et al. (2010) (see Section 3.4.2), Cousineau et al. (2012) found that amusics exhibited no aversion to traditionally dissonant chords, normal aversion to interference, and an inability to detect periodicity/harmonicity. Since the aversion to interference did not transfer to dissonant chords, Cousineau et al. (2012) concluded that interference is irrelevant to consonance perception. However, Marin et al. (2015) subsequently identified small but reliable preferences for consonance in amusics, and showed with regression analyses that these preferences were driven by interference, whereas non-amusic preferences were driven by both interference and periodicity/harmonicity. This discrepancy between Cousineau et al. (2012) and Marin et al. (2015) needs further investigation.

Non-Western listeners

Cross-cultural research into consonance perception has identified high similarity between the consonance judgments of Western and Japanese listeners (Butler & Daston, 1968), but low similarity between Western and Indian listeners (Maher, 1976), and between Westerners and native Amazonians from the Tsimane' society (McDermott et al., 2016). Exploring these differences further, McDermott et al. (2016) found that Tsimane' and Western listeners shared an aversion to interference and an ability to perceive periodicity/harmonicity, but, unlike Western listeners, the Tsimane' had no *preference* for periodicity/harmonicity.

These results suggest that cultural exposure significantly affects consonance perception. The results of McDermott et al. (2016) additionally suggest that this effect of cultural exposure may be mediated by changes in preference for periodicity/harmonicity.

Infants

Consonance perception has been demonstrated in toddlers (Di Stefano et al., 2017), 6-month-old infants (Crowder, Reznick, & Rosenkrantz, 1991; Trainor & Heinmiller, 1998), 4-month-old infants (Trainor, Tsang, & Cheung, 2002; Zentner & Kagan, 1998), 2-month-old infants (Trainor et al., 2002), and newborn infants (Masataka, 2006; Perani et al., 2010; Virtala, Huotilainen, Partanen, Fellman, & Tervaniemi, 2013). Masataka (2006) additionally found preserved consonance perception in newborn infants with deaf parents. These results suggest that consonance perception does not solely depend on cultural exposure.

A related question is whether infants *prefer* consonance to dissonance. Looking-time paradigms address this question, testing whether infants preferentially look at consonant or dissonant sound sources (Crowder et al., 1991; Masataka, 2006; Plantinga & Trehub, 2014; Trainor & Heinmiller, 1998; Trainor et al., 2002; Zentner & Kagan, 1998). With the exception of Plantinga & Trehub (2014), these studies each report detecting consonance preferences in infants. However, Plantinga & Trehub (2014) failed to replicate several of these results, and additionally question the validity of looking-time paradigms, noting that looking times may be confounded by features such as familiarity and comprehensibility. These problems may partly be overcome by physical play-based paradigms (e.g. Di Stefano et al., 2017), but such paradigms are unfortunately only applicable to older infants.

In conclusion, therefore, it seems that young infants perceive some aspects of consonance, but it is unclear whether they prefer consonance to dissonance. These conclusions provide tentative evidence that consonance perception is not solely cultural.

Animals

Animal studies could theoretically provide compelling evidence for non-cultural theories of consonance. If animals were to display sensitivity or preference for consonance despite zero prior musical exposure, this would indicate that consonance could not be fully explained by cultural learning.

Most studies of consonance perception in animals fall into two categories: *discrimination* studies and *preference* studies (see Toro & Crespo-Bojorque, 2017 for a review). Discrimination studies investigate whether animals can be taught to discriminate consonance from dissonance in unfamiliar sounds. Preference studies investigate whether animals prefer consonance to dissonance.

Discrimination studies have identified consonance discrimination in several non-human species, but methodological issues limit interpretation of their findings. Experiment 5 of Hulse, Bernard, & Braaten (1995) suggests that starlings may be able to discriminate consonance from dissonance, but their stimulus set contains just four chords. Experiment 2 of Izumi (2000) suggests that Japanese monkeys may be able to discriminate consonance from dissonance, but this study likewise relies on just four chords at different transpositions. Watanabe, Uozumi, & Tanaka (2005) claim to show consonance discrimination in Java sparrows, but the sparrows' discriminations can also be explained by interval-size judgments.[7] Conversely, studies of pigeons (Brooks & Cook, 2010) and rats (Crespo-Bojorque & Toro, 2015) have failed to show evidence of consonance discrimination (but see also Borchgrevink, 1975).[8]

Preference studies have identified consonance preferences in several nonhuman animals. Using stimuli from a previous infant consonance study (Zentner & Kagan, 1998), Chiandetti & Vallortigara (2011) found that newly hatched domestic chicks spent more time near consonant sound sources than dissonant sound sources. Sugimoto et al. (2010) gave an infant chimpanzee the ability to select between consonant and dissonant two-part melodies, and found that the chimpanzee preferentially selected consonant melodies. However, these studies have yet to be replicated, and both rely on borderline p-values ($p = .03$). Other studies have failed to demonstrate consonance preferences in Campbell's monkeys (Koda et al., 2013) or cotton-top tamarins (McDermott & Hauser, 2004).

These animal studies provide an important alternative perspective on consonance perception. However, recurring problems with these studies include small stimulus sets, small sample sizes, and a lack of replication studies. Future work should address these problems.

3.4.3 Composition effects

Here we consider how compositional practice may provide evidence for the psychological mechanisms underlying consonance perception.

Musical scales

A *scale* divides an octave into a set of pitch classes that can subsequently be used to generate musical material. Scales vary cross-culturally, but certain cross-cultural similarities between scales suggest common perceptual biases.

Gill & Purves (2009) argue that scale construction is biased towards harmonicity maximisation, and explain harmonicity maximisation as a preference

[7] 0/12 of their consonant chords contain intervals smaller than a minor third, whereas 15/16 of their dissonant chords contain such intervals.

[8] Toro & Crespo-Bojorque (2017) also claim that consonance discrimination has been demonstrated in black-capped chickadees, but we disagree in their interpretation of the cited evidence (Hoeschele, Cook, Guillette, Brooks, & Sturdy, 2012).

for vocal-like sounds. They introduce a computational model of harmonicity, which successfully recovers several important scales in Arabic, Chinese, Indian, and Western music. However, they do not test competing consonance models, and admit that their results may also be explained by interference minimisation.

Gamelan music and Thai classical music may help distinguish periodicity/harmonicity from interference. Both traditions use inharmonic scales whose structures seemingly reflect the inharmonic spectra of their percussion instruments (Sethares, 2005). Sethares provides computational analyses relating these scales to interference minimisation; periodicity/harmonicity, meanwhile, offers no obvious explanation for these scales.[9] These findings suggest that interference contributes cross-culturally to consonance perception.

Manipulation of interference

Western listeners typically perceive interference as unpleasant, but various other musical cultures actively promote it. Interference is a key feature of the Middle Eastern *mijwiz*, an instrument comprising two blown pipes whose relative tunings are manipulated to induce varying levels of interference (Vassilakis, 2005). Interference is also promoted in the vocal practice of *beat diaphony*, or *Schwebungsdiaphonie*, where two simultaneous voice parts sing in close intervals such as seconds. Beat diaphony can be found in various musical traditions, including music from Lithuania (Ambrazevičius, 2017; Vyčinienė, 2002), Papua New Guinea (Florian, 1981), and Bosnia (Vassilakis, 2005). In contrast to Western listeners, individuals from these traditions seem to perceive the resulting sonorities as consonant (Florian, 1981). These cross-cultural differences indicate that the aesthetic valence of interference is, at least in part, culturally determined.

Chord spacing (Western music)

In Western music, chords seem to be spaced to minimise interference, most noticeably by avoiding small intervals in lower registers but permitting them in higher registers (Huron & Sellmer, 1992; McGowan, 2011; Plomp & Levelt, 1965). Periodicity theories of consonance provide no clear explanation for this phenomenon.

Chord prevalences (Western music)

Many theorists have argued that consonance played an integral role in determining Western compositional practice (e.g. Dahlhaus, 1990; Hindemith, 1945;

[9] It would be worth testing this formally, applying periodicity/harmonicity consonance models to the inharmonic tone spectra of Gamelan and Thai classical music, and relating the results to scale structure.

Rameau, 1722). If so, it should be possible to test competing theories of consonance by examining their ability to predict compositional practice.

Huron (1991) analysed prevalences of different intervals within 30 polyphonic keyboard works by J. S. Bach, and concluded that they reflected dual concerns of minimising interference and minimising tonal fusion. Huron argued that interference was minimised on account of its negative aesthetic valence, whereas tonal fusion was minimised to maintain perceptual independence of the different voices.

Parncutt et al. (2018) tabulated chord types in seven centuries of vocal polyphony, and related their occurrence rates to several formal models of diatonicity, interference, periodicity/harmonicity, and evenness. Most models correlated significantly with chord occurrence rates, with fairly stable coefficient estimates across centuries. These results suggest that multiple psychological mechanisms contribute to consonance.

However, these findings must be treated as tentative, for the following reasons:

a) The parameter estimates have low precision due to the small sample sizes (12 dyads in Huron, 1991; 19 triads in Parncutt et al., 2018);[10]

b) The pairwise correlations reported in Parncutt et al. (2018) cannot capture effects of multiple concurrent mechanisms (e.g. periodicity/harmonicity and interference).

3.4.4 Discussion

Table 3.1 summarises the evidence contributed by these diverse studies. We now use this evidence to re-evaluate some claims in the recent literature.

Role of periodicity/harmonicity

Recent work has claimed that consonance is primarily determined by periodicity/harmonicity, with the role of periodicity/harmonicity potentially moderated by musical background (Cousineau et al., 2012; McDermott et al., 2010, 2016). In our view, a significant contribution of periodicity/harmonicity to consonance is indeed supported by the present literature, in particular by individual-differences research and congenital amusia research (Table 3.1). A moderating effect of musical background also seems likely, on the basis of cross-cultural variation in music perception and composition. However, quantitative descriptions of these effects are missing: It is unclear what proportion of consonance

[10] For example, a correlation coefficient of $r = 0.5$ with 19 triads has a 95% confidence interval of [0.06, 0.78].

may be explained by periodicity/harmonicity, and it is unclear how sensitive consonance is to cultural exposure.

Role of interference

Recent work has also claimed that consonance is independent of interference (Bowling & Purves, 2015; Bowling et al., 2018; Cousineau et al., 2012; McDermott et al., 2010, 2016). In our view, the wider literature is inconsistent with this claim (Table 3.1). The main evidence against interference comes from the individual-differences study of McDermott et al. (2010), but this evidence is counterbalanced by several positive arguments for interference, including studies of tone spectra, pitch height, chord voicing in Western music, scale tunings in Gamelan music and Thai classical music, and cross-cultural manipulation of interference for expressive effect.

Role of culture

Cross-cultural studies of music perception and composition make it clear that culture contributes to consonance perception (Table 3.1). The mechanisms of this effect remain unclear, however: Some argue that Western listeners internalise codified conventions of Western harmony (Johnson-Laird et al., 2012), whereas others argue that Westerners simply learn aesthetic preferences for periodicity/harmonicity (McDermott et al., 2016). These competing explanations have yet to be tested.

Conclusions

We conclude that consonance perception in Western listeners is likely to be driven by multiple psychological mechanisms, including interference, periodicity/harmonicity, and cultural background (Table 3.1). This conclusion is at odds with recent claims that interference does not contribute to consonance perception (Cousineau et al., 2012; McDermott et al., 2010, 2016). In the rest of this chapter, we therefore examine our proposition empirically, computationally modelling large datasets of consonance judgments and music compositions.

3.5 Computational models

We begin by reviewing prominent computational models of consonance from the literature, organising them by psychological theory and by modelling approach (Figure 3.3).

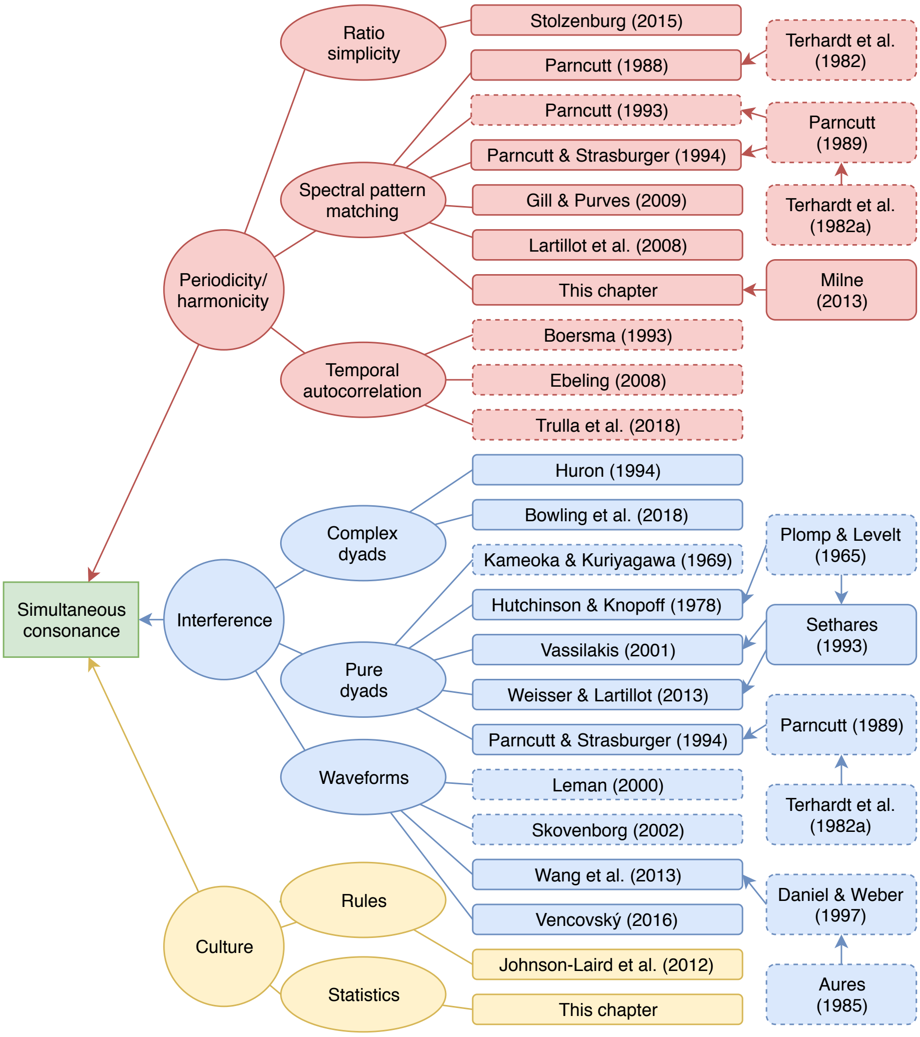

Figure 3.3: Consonance models organised by psychological theory and modelling approach. Dashed borders indicate models not evaluated in our empirical analyses. Arrows denote model revisions.

3.5.1 Periodicity/harmonicity: Ratio simplicity

Chords tend to be more periodic when their constituent tones are related by simple frequency ratios. Ratio simplicity can therefore provide a proxy for periodicity/harmonicity. Previous research has formalised ratio simplicity in various ways, with the resulting measures predicting the consonance of just-tuned chords fairly well (e.g. Euler, 1739; Geer, Levelt, & Plomp, 1962; Levelt, Geer, & Plomp, 1966; Schellenberg & Trehub, 1994).[11] Unfortunately, these measures generally fail to predict consonance for chords that are not just-tuned. A particular problem is disproportionate sensitivity to small tuning deviations: For example, an octave stretched by 0.001% still sounds consonant, despite corresponding to a very complex frequency ratio (200,002:100,000). However, Stolzenburg (2015) provides an effective solution to this problem, described below.

Stolzenburg (2015)

Stolzenburg's (2015) model avoids sensitivity to small tuning deviations by introducing a preprocessing step where each note is adjusted to maximise ratio simplicity with respect to the bass note. These adjustments are not permitted to change the interval size by more than 1.1%. Stolzenburg argues that such adjustments are reasonable given human perceptual inaccuracies in pitch discrimination. Having expressed each chord frequency as a fractional multiple of the bass frequency, ratio simplicity is then computed as the lowest common multiple of the fractions' denominators. Stolzenburg terms this expression *relative periodicity*, and notes that, assuming harmonic tones, relative periodicity corresponds to the chord's overall period length divided by the bass tone's period length. Relative periodicity values are then postprocessed with logarithmic transformation and smoothing to produce the final model output (see Stolzenburg, 2015 for details).

3.5.2 Periodicity/harmonicity: Spectral pattern matching

Spectral pattern-matching models of consonance follow directly from spectral pattern-matching theories of pitch perception (see Section 3.3). These models operate in the frequency domain, searching for spectral patterns characteristic of periodic sounds.

[11] A chord is just-tuned when its pitches are drawn from a just-tuned scale. A just-tuned scale is a scale tuned to maximise ratio simplicity.

Terhardt (1982); Parncutt (1988)

Terhardt (1982) and Parncutt (1988) both frame consonance in terms of chord-root perception. In Western music theory, the chord root is a pitch class summarising a chord's tonal content, which (according to Terhardt and Parncutt) arises through pattern-matching processes of pitch perception. Consonance arises when a chord has a clear root; dissonance arises from root ambiguity.

Both Terhardt's (1982) and Parncutt's (1988) models use harmonic templates quantised to the Western twelve-tone scale, with the templates represented as octave-invariant pitch class sets. Each pitch class receives a numeric weight, quantifying how well the chord's pitch classes align with a harmonic template rooted on that pitch class. These weights preferentially reward coincidence with primary harmonics such as the octave, perfect fifth, and major third.[12] The chord root is estimated as the pitch class with the greatest weight; root ambiguity is then operationalised by dividing the total weight by the maximum weight. According to Terhardt and Parncutt, root ambiguity should then negatively predict consonance.

Parncutt (1989); Parncutt & Strasburger (1994)

Parncutt's (1989) model constitutes a musical revision of Terhardt et al.'s (1982a) pitch perception algorithm. Parncutt & Strasburger's (1994) model, in turn, represents a slightly updated version of Parncutt's (1989) model.

Like Parncutt's (1988) model, Parncutt's (1989) model formulates consonance in terms of pattern-matching pitch perception. As in Parncutt (1988), the algorithm works by sweeping a harmonic template across an acoustic spectrum, seeking locations where the template coincides well with the acoustic input; consonance is elicited when the location of best fit is unambiguous. However, Parncutt's (1989) algorithm differs from Parncutt (1988) in several important ways:

a) Chord notes are expanded into their implied harmonics;

b) Psychoacoustic phenomena such as hearing thresholds, masking, and audibility saturation are explicitly modelled;

c) The pattern-matching process is no longer octave-invariant.

Parncutt (1989) proposes two derived measures for predicting consonance: *pure tonalness* and *complex tonalness*.[13] Pure tonalness describes the extent

[12] The weights assigned to each harmonic differ between studies; Terhardt (1982) used binary weights, but Parncutt (1988) introduced graduated weights, which he updated in later work (see Parncutt, 2006a).

[13] These measures were later termed pure and complex *sonorousness* by Parncutt & Strasburger (1994).

to which the input spectral components are audible, after accounting for hearing thresholds and masking. Complex tonalness describes the audibility of the strongest virtual pitch percept. The former may be considered a interference model, the latter a periodicity/harmonicity model.

Parncutt & Strasburger (1994) describe an updated version of Parncutt's (1989) algorithm. The underlying principles are the same, but certain psychoacoustic details differ, such as the calculation of pure-tone audibility thresholds and the calculation of pure-tone height. We evaluate this updated version here.

Parncutt (1993) presents a related algorithm for modelling the perception of octave-spaced tones (also known as Shepard tones). Since octave-spaced tones are uncommon in Western music, we do not evaluate the model here.

Gill & Purves (2009)

Gill & Purves (2009) present a pattern-matching periodicity/harmonicity model which they apply to various two-note chords. They assume just tuning, which allows them to compute each chord's fundamental frequency as the greatest common divisor of the two tones' frequencies. They then construct a hypothetical harmonic complex tone rooted on this fundamental frequency, and calculate what proportion of this tone's harmonics are contained within the spectrum of the original chord. This proportion forms their periodicity/harmonicity measure. This approach has been shown to generalise well to three- and four-note chords (Bowling et al., 2018). However, the model's cognitive validity is limited by the fact that, unlike human listeners, it is very sensitive to small deviations from just tuning or harmonic tone spectra.

Peeters et al. (2011); Bogdanov et al. (2013); Lartillot et al. (2008)

Several prominent audio analysis toolboxes – the Timbre Toolbox (Peeters et al., 2011), Essentia (Bogdanov et al., 2013), and MIRtoolbox (Lartillot et al., 2008) – contain inharmonicity measures. Here we examine their relevance for consonance modelling.

The inharmonicity measure in the Timbre Toolbox (Peeters et al., 2011) initially seems relevant for consonance modelling, being calculated by summing each partial's deviation from harmonicity. However, the algorithm's preprocessing stages are clearly designed for single tones rather than tone combinations. Each input spectrum is preprocessed to a harmonic spectrum, slightly deformed by optional stretching; this may be a reasonable approximation for single tones, but it is inappropriate for tone combinations. We therefore do not consider this model further.

Essentia (Bogdanov et al., 2013) contains an inharmonicity measure defined similarly to the Timbre Toolbox (Peeters et al., 2011). As with the Timbre Toolbox, this feature is clearly intended for single tones rather than tone combinations, and so we do not consider it further.

MIRtoolbox (Lartillot et al., 2008) contains a more flexible inharmonicity measure. First, the fundamental frequency is estimated using autocorrelation and peak-picking; inharmonicity is then estimated by applying a sawtooth filter to the spectrum, with troughs corresponding to integer multiples of the fundamental frequency, and then integrating the result. This measure seems more likely to capture inharmonicity in musical chords, and indeed it has been recently used in consonance perception research (Lahdelma & Eerola, 2016). However, systematic validations of this measure are lacking.

Milne (2013); This chapter

Milne (2013) presents a periodicity/harmonicity model that operates on smooth pitch-class spectra (see also Milne et al., 2016). The model takes a pitch-class set as input, and expands all tones to idealised harmonic spectra. These spectra are superposed additively, and then blurred by convolution with a Gaussian distribution, mimicking perceptual uncertainty in pitch processing. The algorithm then sweeps a harmonic template over the combined spectrum, calculating the cosine similarity between the template and the combined spectrum as a function of the template's fundamental frequency. The resulting cosine similarity profile is termed the *virtual pitch-class spectrum*, and corresponds hypothetically to the perceptual salience of different candidate fundamental frequencies. The pitch class eliciting the maximal cosine similarity is identified as the fundamental pitch class, and the resulting cosine similarity is taken as the periodicity/harmonicity estimate. High similarity means that the chord can be well-approximated by a single harmonic complex tone.

Milne's (2013) model is appealing for its ability to deal with arbitrary tunings, unlike many other pattern-matching harmonicity models. However, when experimenting with the model, we identified a potential limitation: the model assumes that the listener only identifies one fundamental frequency in the chord, but with larger chords it seems likely that the listener hears multiple fundamental frequencies. Correspondingly, instead of hypothesising that consonance depends on how well the chord can be approximated by a single harmonic complex tone, we might hypothesise that consonance should depend on how well the chord can be approximated by a small number of harmonic complex tones. There are many ways that this principle might be operationalised; we decided to work with the 'peakiness' of the virtual pitch-class spectrum. A 'peaky' virtual pitch-class spectrum resembles a small collection of harmonic complex tones; in contrast, a flat spectrum does not allude to any harmonic complex tones. We operationalised peakiness by treating the cosine-similarity profile as a probability distribution, and computing the Kullback-Leibler divergence to this distribution from a uniform distribution. This information-theoretic quantity is equivalent to the negative continuous entropy of the virtual pitch-class spectrum, as formulated by Jaynes (1968). Future work could profitably evaluate alternative peakiness measures.

We now provide a formal definition of this model. This definition uses Milne et al.'s (2011) original definition of the smooth pitch-class spectrum, but similar results could be achieved by using the updated definition in Section 2.6.1 of this thesis.

We begin by formalising a pitch-class spectrum as a continuous function that describes perceptual weight as a function of pitch class (p_c). Perceptual weight is interpreted as the strength of perceptual evidence for a given pitch class. Here pitch classes (p_c) take continuous values in the interval $[0, 12)$ and relate to frequency (f, Hz scale) as follows:

$$p_c = \left[9 + 12 \log_2 \left(\frac{f}{440}\right)\right] \bmod 12. \tag{3.2}$$

Pitch-class sets are transformed to pitch-class spectra by expanding each pitch class into its implied harmonics. Pitch classes are modelled as harmonic complex tones with 12 harmonics, after Milne & Holland (2016). The jth harmonic in a pitch class has level $j^{-\rho}$, where ρ is the roll-off parameter ($\rho > 0$). Partials are represented by Gaussians with mass equal to partial level, mean equal to partial pitch class, and standard deviation σ. Perceptual weights combine additively.

We express a pitch-class spectrum mathematically as the function $W(p_c, X)$, which returns the perceptual weight at pitch-class p_c for an input pitch-class set $X = \{x_1, x_2, \ldots, x_m\}$:

$$W(p_c, X) = \sum_{i=1}^{m} T(p_c, x_i), \tag{3.3}$$

where i indexes the pitch classes. The term $T(p_c, x)$ corresponds to the contribution of a harmonic complex tone with fundamental pitch class x to an observation at pitch class p_c:

$$T(p_c, x) = \sum_{j=1}^{12} g\left(p_c, j^{-\rho}, h(x, j)\right), \tag{3.4}$$

where j indexes the harmonics in the complex tone. The term $g(p_c, l, p_x)$ corresponds to the contribution from a harmonic with level l and pitch-class p_x to an observation at pitch-class p_c:

$$g(p_c, l, p_x) = \frac{l}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{d(p_c, p_x)}{\sigma}\right)^2\right). \tag{3.5}$$

The term $d(p_x, p_y)$ corresponds to the distance between two pitch classes p_x and p_y:

$$d(p_x, p_y) = \min\left(|p_x - p_y|, 12 - |p_x - p_y|\right). \tag{3.6}$$

Lastly, $h(x, j)$ is the pitch class of the jth partial of a harmonic complex tone with fundamental pitch class x:

$$h(x, j) = (x + 12\log_2 j) \bmod 12. \tag{3.7}$$

ρ and σ are set to 0.75 and 0.0683 after Milne & Holland (2016).

The cosine similarity of two pitch-class spectra X, Y is defined as follows:

$$S(X, Y) = \frac{\int_0^{12} W(z, X)W(z, Y)\,dz}{\sqrt{\int_0^{12} W(z, X)^2\,dz}\sqrt{\int_0^{12} W(z, Y)^2\,dz}} \tag{3.8}$$

with W as defined in Equation 3.3. The measure takes values in the interval $[0, 1]$, where 1 indicates maximal similarity. The *virtual pitch-class spectrum Q* is then defined as the spectral similarity of the pitch-class set X to a harmonic complex tone with pitch class p_c:

$$Q(p_c, X) = S(p_c, X) \tag{3.9}$$

with S as defined in Equation 3.8. Normalising Q to unit mass produces Q':

$$Q'(p_c, X) = \frac{Q(p_c, X)}{\int_0^{12} Q(y, X)\,dy}. \tag{3.10}$$

$H(X)$, the harmonicity of a pitch-class set X, is finally computed as the Kullback-Leibler divergence to the virtual pitch-class spectrum from a uniform distribution:

$$H(X) = \int_0^{12} Q'(y, X)\log_2\left(12\,Q'(y, X)\right) dy. \tag{3.11}$$

Following Milne et al. (2011), we do not compute these integrations exactly, but instead approximate them using the rectangle rule with 1,200 subintervals, each corresponding to a 1-cent bin. A reference implementation of the resulting model is available in the R package *har18*.[14]

3.5.3 Periodicity/harmonicity: Temporal autocorrelation

Temporal autocorrelation models of consonance follow directly from autocorrelation theories of pitch perception (see Section 3.3). These models operate in the time domain, looking for time lags at which the signal correlates with itself: High autocorrelation implies periodicity and hence consonance.

Boersma (1993)

Boersma's (1993) autocorrelation algorithm can be found in the popular phonetics software Praat. The algorithm tracks the fundamental frequency of an acoustic input over time, and operationalises periodicity as the *harmonics-to-noise ratio*, the proportion of power contained within the signal's periodic component. Marin et al. (2015) found that this algorithm had some power to predict the relative consonance of different dyads. However, the details of the algorithm lack psychological realism, having been designed to solve an engineering problem rather than to simulate human perception. This limits the algorithm's appeal as a consonance model.

Ebeling (2008)

Ebeling's (2008) autocorrelation model estimates the consonance of pure-tone intervals. Incoming pure tones are represented as sequences of discrete pulses, reflecting the neuronal rate coding of the peripheral auditory system. These pulse sequences are additively superposed to form a composite pulse sequence, for which the autocorrelation function is computed. The *generalised coincidence function* is then computed by integrating the squared autocorrelation function over a finite positive range of time lags. Applied to pure tones, the generalised coincidence function recovers the traditional hierarchy of intervallic consonance, and mimics listeners in being tolerant to slight mistunings. Ebeling presents this as a positive result, but it is inconsistent with Plomp & Levelt's (1965) observation that, after accounting for musical training, pure tones do not exhibit the traditional hierarchy of intervallic consonance. It remains unclear whether the model would successfully generalise to larger chords or to complex tones.

[14] `https://github.com/pmcharrison/har18`

Trulla et al. (2018)

Trulla et al.'s (2018) model uses *recurrence quantification analysis* to model the consonance of pure-tone intervals. Recurrence quantification analysis performs a similar function to autocorrelation analysis, identifying time lags at which waveform segments repeat themselves. Trulla et al. (2018) use this technique to quantify the amount of repetition within a waveform, and show that repetition is maximised by traditionally consonant frequency ratios, such as the just-tuned perfect fifth (3:2). The algorithm constitutes an interesting new approach to periodicity/harmonicity detection, but one that lacks much cognitive or neuroscientific backing. As with Ebeling (2008), it is also unclear how well the algorithm generalises to larger chords or to different tone spectra, and the validation suffers from the same problems described above for Ebeling's model.

Summary

Autocorrelation is an important candidate mechanism for consonance perception. However, autocorrelation consonance models have yet to be successfully generalised outside simple tone spectra and two-note intervals. We therefore do not evaluate these models in the present work, but we look forward to future research in this area (see e.g. Tabas et al., 2017).

3.5.4 Interference: Complex dyads

Complex-dyad models of interference search chords for complex dyads known to elicit interference. These models are typically hand-computable, making them well-suited to quick consonance estimation.

Huron (1994)

Huron (1994) presents a measure termed *aggregate dyadic consonance*, which characterises the consonance of a pitch-class set by summing consonance ratings for each pitch-class interval present in the set. These consonance ratings are derived by aggregating perceptual data from previous literature.

Huron (1994) originally used aggregate dyadic consonance to quantify a scale's ability to generate consonant intervals. Parncutt et al. (2018) subsequently applied the model to musical chords, and interpreted the output as an interference measure. The validity of this approach rests on the assumption that interference is additively generated by pairwise interactions between spectral components; a similar assumption is made by pure-dyad interference models (see Section 3.5.5). A further assumption is that Huron's dyadic consonance ratings solely reflect interference, not (for example) periodicity/harmonicity; this assumption is arguably problematic, especially given recent claims that dyadic consonance is driven by periodicity/harmonicity, not interference (McDermott et al., 2010; Stolzenburg, 2015).

Bowling et al. (2018)

Bowling et al. (2018) primarily explain consonance in terms of periodicity/harmonicity, but also identify dissonance with chords containing pitches separated by less than 50 Hz. They argue that such intervals are uncommon in human vocalisations, and therefore elicit dissonance. We categorise this proposed effect under interference, in line with Parncutt et al.'s (2018) argument that these small intervals (in particular minor and major seconds) are strongly associated with interference.

3.5.5 Interference: Pure dyads

Pure-dyad interference models work by decomposing chords into their pure-tone components, and accumulating interference contributions from each pair of pure tones.

Plomp & Levelt (1965); Kameoka & Kuriyagawa (1969b)

Plomp & Levelt (1965) and Kameoka & Kuriyagawa (1969b) concurrently established an influential methodology for consonance modelling: use perceptual experiments to characterise the consonance of pure-tone dyads, and estimate the dissonance of complex sonorities by summing contributions from each pure dyad. However, their original models are rarely used today, having been supplanted by later work.

Hutchinson & Knopoff (1978)

Hutchinson & Knopoff (1978) describe a pure-dyad interference model in the line of Plomp & Levelt (1965). Unlike Plomp & Levelt, Hutchinson & Knopoff sum dissonance contributions over all harmonics, rather than just neighboring harmonics. The original model is not fully algebraic, relying on a graphically depicted mapping between interval size and pure-dyad dissonance; a useful modification is the algebraic approximation introduced by Bigand et al. (1996), which we adopt here (see also Mashinter, 2006).

Hutchinson & Knopoff (1978) only applied their model to complex-tone dyads. They later applied their model to complex-tone triads (Hutchinson & Knopoff, 1979), and for computational efficiency introduced an approximation decomposing the interference of a triad into the contributions of its constituent complex-tone dyads (see previous discussion of Huron, 1994). With modern computers, this approximation is unnecessary and hence rarely used.

Sethares (1993); Vassilakis (2001); Weisser & Lartillot (2013)

Several subsequent studies have preserved the general methodology of Hutchinson & Knopoff (1978) while introducing various technical changes. Sethares (1993) reformulated the equations linking pure-dyad consonance to interval size and pitch height. Vassilakis (2001) and Weisser & Lartillot (2013) subsequently modified Sethares's (1993) model, reformulating the relationship between pure-dyad consonance and pure-tone amplitude. These modifications generally seem principled, but the resulting models have received little systematic validation.

Parncutt (1989); Parncutt & Strasburger (1994)

As discussed above (see Section 3.5.2), the pure tonalness measure of Parncutt (1989) and the pure sonorousness measure of Parncutt & Strasburger (1994) may be categorised as interference models. Unlike other pure-dyad interference models, these models address masking, not beating.

3.5.6 Interference: Waveforms

Dyadic models present a rather simplified account of interference, and struggle to capture certain psychoacoustic phenomena such as effects of phase (e.g. Pressnitzer & McAdams, 1999) and waveform envelope shape (e.g. Vencovský, 2016) on roughness. The following models achieve a more detailed account of interference by modelling the waveform directly.

Leman (2000b)

Leman's (2000b) synchronisation index model measures beating energy within roughness-eliciting frequency ranges. The analysis begins with Immerseel & Martens's (1992) model of the peripheral auditory system, which simulates the frequency response of the outer and middle ear, the frequency analysis of the cochlea, hair-cell transduction from mechanical vibrations to neural impulses, and transmission by the auditory nerve. Particularly important is the half-wave rectification that takes place in hair-cell transduction, which physically instantiates beating frequencies within the Fourier spectrum. Leman's model then filters the neural transmissions according to their propensity to elicit roughness, and calculates the energy of the resulting spectrum as a roughness estimate. Leman illustrates model outputs for several amplitude-modulated tones, and for two-note chords synthesised with harmonic complex tones. The initial results seem promising, but we are unaware of any studies systematically fine-tuning or validating the model.

Skovenborg & Nielsen (2002)

Skovenborg & Nielsen's (2002) model is conceptually similar to Leman's (2000b) model. The key differences are simulating the peripheral auditory system using the HUTear MATLAB toolbox (Härmä & Palomäki, 1999), rather than Immerseel & Martens's (1992) model, and adopting different definitions of roughness-eliciting frequency ranges. The authors provide some illustrations of the model's application to two-tone intervals of pure and complex tones. The model recovers some established perceptual phenomena, such as the dissonance elicited by small intervals, but also exhibits some undesirable behaviour, such as multiple consonance peaks for pure-tone intervals, and oversensitivity to slight mistunings for complex-tone intervals. We are unaware of further work developing this model.

Aures (1985c); Daniel & Weber (1997); Wang et al. (2013)

Aures (1985c) describes a roughness model that has been successively developed by Daniel & Weber (1997) and Wang et al. (2013). Here we describe the model as implemented in Wang et al. (2013). Like Leman (2000b) and Skovenborg & Nielsen (2002), the model begins by simulating the frequency response of the outer and middle ear, and the frequency analysis of the cochlea. Unlike Leman (2000b) and Skovenborg & Nielsen (2002), the model does not simulate hair-cell transduction or transmission by the auditory nerve. Instead, the model comprises the following steps:

a) Extract the waveform envelope at each cochlear filter;

b) Filter the waveform envelopes to retain the beating frequencies most associated with roughness;

c) For each filter, compute the *modulation index*, summarising beating magnitude as a proportion of the total signal.

d) Multiply each filter's modulation index by a *phase impact factor*, capturing signal correlations between adjacent filters; high correlations yield higher roughness;

e) Multiply by a weighting factor identifying how different cochlear filters contribute more to the perception of roughness;

f) Square the result and sum over cochlear filters.

Unlike the models of Leman (2000b) and Skovenborg & Nielsen (2002), these three models are presented alongside objective perceptual validations. However, these validations are generally restricted to relatively artificial and non-musical stimuli.

Vencovský (2016)

Like Leman (2000b), Skovenborg & Nielsen (2002), and Wang et al. (2013), Vencovský's (2016) model begins with a sophisticated model of the peripheral auditory system. The model of Meddis (2011) is used for the outer ear, middle ear, inner hair cells, and auditory nerve; the model of Nobili, Vetešník, Turicchia, & Mammano (2003) is used for the basilar membrane and cochlear fluid. The output is a neuronal signal for each cochlear filter.

Roughness is then estimated from the neuronal signal's envelope, or beating pattern. Previous models estimate roughness from the amplitude of the beating pattern; Vencovský's (2016) model additionally accounts for the beating pattern's shape. Consider a single oscillation of the beating pattern; according to Vencovský's (2016) model, highest roughness is achieved when the difference between minimal and maximal amplitudes is large, and when the progression from minimal to maximal amplitudes (but not necessarily vice versa) is fast. Similar to previous models (Daniel & Weber, 1997; Wang et al., 2013), Vencovský's (2016) model also normalises roughness contributions by overall signal amplitudes, and decreases roughness when signals from adjacent cochlear channels are uncorrelated.

Vencovský (2016) validates the model on perceptual data from various types of artificial stimuli, including two-tone intervals of harmonic complex tones, and finds that the model performs fairly well. It is unclear how well the model generalises to more complex musical stimuli.

3.5.7 Culture

Cultural aspects of consonance perception have been emphasised by many researchers (see Section 3.3), but we are only aware of one pre-existing computational model instantiating these ideas: that of Johnson-Laird et al. (2012).

Johnson-Laird et al. (2012)

Johnson-Laird et al. (2012) provide a rule-based model of consonance perception in Western listeners. The model comprises three rules, organised in decreasing order of importance:

a) Chords consistent with a major scale are more consonant than chords only consistent with a minor scale, which are in turn more consonant than chords not consistent with either;

b) Chords are more consonant if they i) contain a major triad and ii) all chord notes are consistent with a major scale containing that triad;

c) Chords are more consonant if they can be represented as a series of pitch classes each separated by intervals of a third, optionally including one interval of a fifth.

Unlike most other consonance models, this model does not return numeric scores, but instead ranks chords in order of their consonance. Ranking is achieved as follows: Apply the rules one at a time, in decreasing order of importance, and stop when a rule identifies one chord as more consonant than the other. This provides an estimate of cultural consonance.

Johnson-Laird et al. (2012) suggest that Western consonance perception depends both on culture and on roughness. They capture this idea with their *dual-process model*, which adds an extra rule to the cultural consonance algorithm, applied only when chords cannot be distinguished on the cultural consonance criteria. This rule predicts that chords are more consonant if they exhibit lower roughness. The authors operationalise roughness using the model of Hutchinson & Knopoff (1978).

The resulting model predicts chordal consonance rather effectively (Johnson-Laird et al., 2012; Stolzenburg, 2015). However, a problem with this model is that the rules are hand-coded on the basis of expert knowledge. The rules could represent cultural knowledge learned through exposure, but they could also explain post-hoc rationalisations of perceptual phenomena. This motivates us to introduce an alternative corpus-based model, described below.

A corpus-based model of cultural familiarity

Here we introduce a simple corpus-based model of cultural familiarity, representing the hypothesis that listeners become familiar with chords in proportion to their frequency of occurrence in the listener's musical culture, and that this familiarity positively influences consonance through the mere exposure effect (Zajonc, 2001). We simulate a Western listener's musical exposure by tabulating the occurrences of different chord types in the Billboard dataset (Burgoyne, 2011), a large dataset of music from the US charts. We reason that this dataset should provide a reasonable first approximation to the musical exposure of the average Western listener, but note that this approach could easily be tailored to the specific musical backgrounds of individual listeners. See Section 3.9 for further details.

3.6 Perceptual analyses

We now reanalyse consonance perception data from four previous studies (Bowling et al., 2018; Johnson-Laird et al., 2012; Lahdelma & Eerola, 2016; Schwartz et al., 2003). These datasets correspond to consonance judgments for Western musical chords as made by listeners from Western musical cultures. We focus in particular on the dataset from Bowling et al. (2018), as it contains considerably more chord types than previous datasets (see Section 3.9 for details). We make all these datasets available in an accompanying R package, *inconData*.

Previous analyses of these datasets suffer from important limitations. Several studies show that a dataset is consistent with their proposed theory, but fail to test competing theories (Bowling et al., 2018; Schwartz et al., 2003). When competing theories are tested, each theory is typically operationalised using just one computational model (Johnson-Laird et al., 2012; Lahdelma & Eerola, 2016), and the choice of model is fairly arbitrary, because few comparative model evaluations are available in the literature. However, as we later show, models representing the same consonance theory can vary widely in performance. Furthermore, when multiple models are evaluated, parameter reliability is rarely considered, encouraging inferences to be made from statistically insignificant differences (Stolzenburg, 2015). Lastly, no studies simultaneously model contributions from periodicity/harmonicity, interference, and cultural familiarity, despite the implication from the empirical literature that all three phenomena may contribute to consonance perception.

Here we address these problems. Our primary goal is to re-evaluate competing theories of consonance perception; our secondary goal is to facilitate future consonance research. Towards these goals, we compile 20 consonance models, 15 of which we implement in this chapter's accompanying R package, and 5 of which are available in publicly available audio analysis toolboxes (Table 3.2).[15] We systematically evaluate these 20 models on our perceptual data, providing future researchers an objective basis for model selection. We then assess the evidence for a composite theory of consonance perception, evaluating the extent to which periodicity/harmonicity, interference, and cultural familiarity simultaneously contribute to consonance judgments. We include the resulting composite consonance model in the *incon* package.

For practical reasons, we do not try to evaluate every model in the literature.

[15] See Section 3.9.1 for more information on the model implementations.

In most cases, we only evaluate the latest published version of a given model, and avoid models with limited or discouraging perceptual validations (e.g. Leman, 2000b; Skovenborg & Nielsen, 2002). We also omit one model on the grounds of its complexity (Vencovský, 2016). See Section 3.9 for further details.

Table 3.2: Consonance models evaluated in the present work.

Reference	Original name	Input	Implementation
Periodicity/harmonicity			
Gill & Purves (2009)	Percentage similarity	Symbolic	incon (bowl18)
This chapter	Harmonicity	Symbolic	incon (har18)
Milne (2013)	Harmonicity	Symbolic	incon (har18)
Parncutt (1988)	Root ambiguity	Symbolic	incon (parn88)
Parncutt & Strasburger (1994)	Complex sonorousness	Symbolic	incon (parn94)
Stolzenburg (2015)	Smoothed relative periodicity	Symbolic	incon (stolz15)
Lartillot et al. (2008)	Inharmonicity	Audio	MIRtoolbox
Interference			
Bowling et al. (2018)	Absolute frequency intervals	Symbolic	incon (bowl18)
Huron (1994)	Aggregate dyadic consonance	Symbolic	incon
Hutchinson & Knopoff (1978)	Dissonance	Symbolic	incon (dycon)
Parncutt & Strasburger (1994)	Pure sonorousness	Symbolic	incon (parn94)
Sethares (1993)	Dissonance	Symbolic	incon (dycon)
Vassilakis (2001)	Roughness	Symbolic	incon (dycon)
Wang et al. (2013)	Roughness	Symbolic	incon (wang13)
Bogdanov et al. (2013)	Dissonance	Audio	Essentia
Lartillot (2014)	Roughness (after Sethares)	Audio	MIRtoolbox
Lartillot (2014)	Roughness (after Vassilakis)	Audio	MIRtoolbox

Table 3.2 continued

Reference	Original name	Input	Implementation
Weisser and Lartillot (2013)	Roughess (after Sethares)	Audio	MIRtoolbox
Culture			
Johnson-Laird et al. (2012)	Tonal dissonance	Symbolic	incon (jl12)
This chapter	Corpus dissonance	Symbolic	incon (corpdiss)

Note. ‘Reference’ identifies the literature where the model or relevant software package was originally presented. ‘Original name’ corresponds to the name of the model (or corresponding psychological feature) in the reference literature. ‘Input’ describes the input format for the model implementations used in this chapter. ‘Implementation’ describes the software used for each model implementation, with ‘incon’ referring to the *incon* package that accompanies this chapter, and ‘Essentia’ and ‘MIRtoolbox’ corresponding to the software presented in Bogdanov et al. (2013) and Lartillot et al. (2008) respectively. Terms in parentheses identify the low-level R packages that underpin the *incon* package, and that provide extended access to individual models.

3.6.1 Evaluating models individually

We begin by evaluating each consonance model individually on the Bowling et al. (2018) dataset (Figure 3.4A).[16] Our performance metric is the partial correlation[17] between model predictions and average consonance ratings, controlling for the number of notes in each chord, with the latter treated as a categorical variable. We control for number of notes to account for a design-related confound in Bowling et al. (2018) where stimulus presentation was blocked by the number of notes in each chord, potentially allowing participants to recalibrate their response scales for each new number of notes. We use predictive performance as an initial indicator of a model's cognitive validity and practical utility.

Competing theories of consonance

The three best-performing models represent three different theories of consonance perception: interference (r = .77, 95% CI: [.72, .81]), periodicity/harmonicity (r = .72, 95% CI: [.66, .77]), and cultural familiarity (r = .72, 95% CI: [.66, .77]). This similarity in performance is consistent with the idea that these three phenomena all contribute to consonance perception. Later we describe a regression analysis that provides a more principled test of this hypothesis.

Periodicity/harmonicity models

The most detailed periodicity/harmonicity model tested is that of Parncutt & Strasburger (1994), which incorporates various psychoacoustic phenomena including hearing thresholds, masking, and audibility saturation. However, this model's performance (r = .56, 95% CI: [.47, .63]) is matched or beaten by four periodicity/harmonicity models with essentially no psychoacoustic modelling (r = .62, .65, .72, .72). This suggests that these psychoacoustic details may be largely irrelevant to the relationship between periodicity/harmonicity and consonance.

Interference models

The interference models display an interesting trend in performance: Since Hutchinson & Knopoff (1978), performance has generally decreased, not increased. This is surprising, since each successive model typically incorporates a more detailed psychoacoustic understanding of the physics of ampli-

[16] See Section 3.9.3 for more information about this dataset.

[17] All correlations in this chapter are computed as Pearson correlation coefficients, except where stated otherwise.

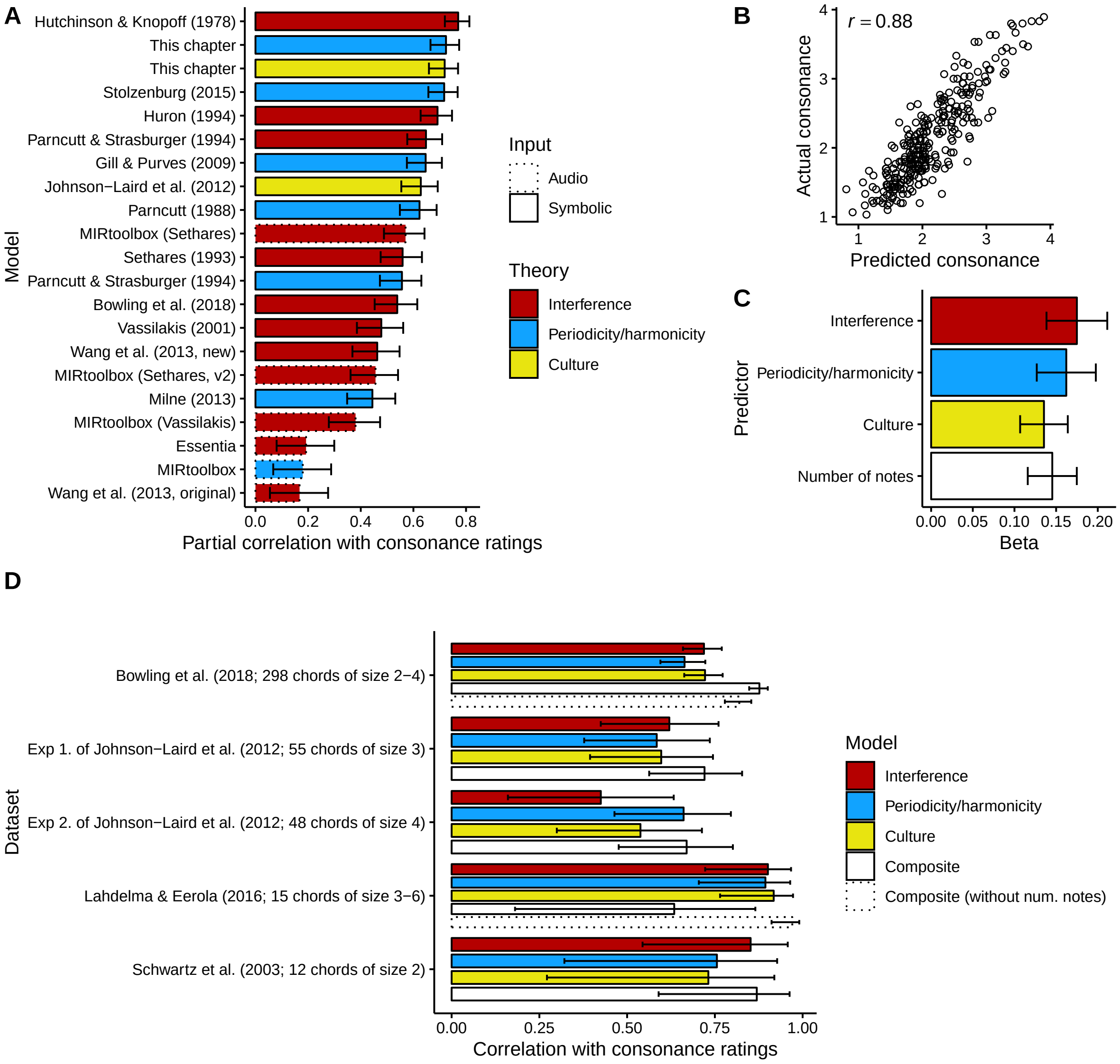

Figure 3.4: Results of the perceptual analyses. All error bars denote 95% confidence intervals. **A**: Partial correlations between model outputs and average consonance ratings in the Bowling et al. (2018) dataset, after controlling for number of notes. **B**: Predictions of the composite model for the Bowling et al. (2018) dataset. **C**: Standardised regression coefficients for the composite model. **D**: Evaluating the composite model across five datasets from four studies (Bowling et al., 2018; Johnson-Laird et al., 2012; Lahdelma & Eerola, 2016; Schwartz et al., 2003).

tude fluctuation (exceptions are the complex-dyad models of Bowling et al., 2018, and Huron, 1994, and the masking model of Parncutt & Strasburger, 1994). This trend deserves to be explored further; an interesting possibility is that amplitude-fluctuation models fail to capture the potential contribution of masking to consonance (see Section 3.3).

Cultural models

The new corpus-based consonance model (r = .72, 95% CI: [.66, .77]) outperformed the rule-based consonance model (Johnson-Laird et al., 2012, r = .63, 95% CI: [.55, .69]) (95% CI for the difference in correlations: [.012, .017], after Zou, 2007).[18]

Symbolic versus audio models

Many of the algorithms evaluated here take symbolic inputs, reducing each stimulus to a few numbers representing its constituent pitches. The other algorithms take audio inputs, and therefore have access to the full spectral content of the stimulus. Given that consonance is sensitive to spectral content, one might expect the audio algorithms to outperform the symbolic algorithms. However, Figure 3.4A shows that this is not the case: Generally speaking, the symbolic algorithms outperformed the audio algorithms. Particularly bad results were seen for MIRtoolbox's periodicity/harmonicity measure (r = .18, 95% CI: [.07, .29]) and Essentia's interference measure (r = .19, 95% CI: [.08, .30]). Fairly good results were seen for MIRtoolbox's interference measure, which performed best using its default settings (original Sethares model; r = .57, 95% CI: [.49, .64]). Nonetheless, this model was still outperformed by several simple symbolic models (e.g. Huron, 1994; Parncutt, 1988).

Wang et al.'s (2013) model

The original model of Wang et al. (2013) performed rather poorly (r = .17, 95% CI: [.05, .28]). This poor performance was surprising, given the sophisticated nature of the model and its position in a well-established modelling tradition (Aures, 1985c; Daniel & Weber, 1997). Experimenting with the model, we found its performance to improve significantly upon disabling the 'phase impact factors' component, whereby signal correlations between adjacent cochlear filters increase roughness (resulting partial correlation: r = .46, 95% CI: [.37, .55]).

[18] All statistical comparisons of correlation coefficients reported in this chapter were conducted using the *cocor* package (Diedenhofen & Musch, 2015).

3.6.2 A composite consonance model

We constructed a linear regression model to test the hypothesis that multiple psychological mechanisms contribute to consonance perception. We fit this model to the Bowling et al. (2018) dataset, using four features representing interference, periodicity/harmonicity, cultural familiarity, and number of notes. The first three features corresponded to the three best-performing models in Figure 3.4A: Hutchinson & Knopoff's (1978) roughness model, the new harmonicity model, and the new cultural familiarity model. The fourth feature corresponded to the number of notes in the chord. All features were treated as continuous predictors.

The predictions of the resulting model are plotted in Figure 3.4B. The predictions correlate rather well with the ground truth ($r = .88$, 95% CI: [.85, .90]), significantly outperforming the individual models in Figure 3.4A.

The resulting standardised regression coefficients are plotted in Figure 3.4C, with signs equated for ease of comparison. All four features contributed significantly and substantially to the model, each with broadly similar regression coefficients. As expected, interference was negatively related to consonance, whereas periodicity/harmonicity and cultural familiarity were positively related to consonance. Number of notes also contributed significantly, presumably reflecting participants recalibrating their response scales for blocks with different numbers of notes.

This pattern of regression coefficients supports our proposition that consonance is jointly determined by interference, periodicity/harmonicity, and cultural familiarity. Moreover, it implies that the effect of cultural familiarity on consonance perception is not solely mediated by learned preferences for periodicity/harmonicity (McDermott et al., 2010, 2016). However, the contribution of cultural familiarity should be taken with caution: It might alternatively reflect a non-cultural contributor to consonance that is not captured by our periodicity/harmonicity or interference models, but that influences chord prevalences in music composition, and therefore correlates with our corpus-based cultural model. Future work could test this possibility by modelling individual differences in consonance perception as a function of the listener's musical background.

3.6.3 Generalising to different datasets

A good predictive model of consonance should generalise outside the specific paradigm of Bowling et al. (2018). We therefore tested the new composite model on four additional datasets from the literature (Johnson-Laird et al., 2012; Lahdelma & Eerola, 2016; Schwartz et al., 2003), keeping the regression weights fixed to their previous estimates.[19] These datasets are relatively small, preventing model performance from being assessed with much reliability; nonetheless, they provide a useful initial test of the model's generalisability. In each case, we assessed predictive performance by correlating model predictions with averaged consonance judgments for each stimulus, and benchmarked the composite model's performance against that of its constituent sub-models. For datasets varying the number of notes in each chord, we evaluated the composite model twice: once in its original form, and once removing the number of notes predictor, which we thought might be a design-related artefact from Bowling et al. (2018).

Johnson-Laird et al. (2012) provide two relevant datasets of consonance judgments, one for three-note chords (Experiment 1, 27 participants, 55 chords), and one for four-note chords (Experiment 2, 39 participants, 48 chords). Modelling these datasets, we found a trend for the composite model to outperform the individual sub-models (Figure 3.4D). This trend is less clear in the second dataset, however, where interference performs particularly badly and periodicity/harmonicity performs particularly well, almost on a par with the composite model.[20] A possible explanation is the fact that Johnson-Laird et al. (2012) purposefully undersampled chords containing adjacent semitones, thereby restricting the variation in interference.

Lahdelma & Eerola (2016) provide a dataset of consonance judgments from 410 participants for 15 chords in various transpositions, with the chords ranging in size from three to six notes. As transposition information was missing from the published dataset, we averaged consonance judgments over transpositions before computing the performance metrics. The composite model performed considerably worse ($r = .63$, 95% CI [.18, .87]) than the sub-models ($r > .89$). This implied that the number-of-notes predictor was sabotaging predictions, and indeed, removing this predictor improved performance substantially ($r = .97$, 95% CI [.91, .99]). This pattern of results is consistent with the hypothesis that the number of notes effect observed in the Bowling et al. (2018) dataset was a design-related confound.

Schwartz et al. (2003) present data on the perceptual consonance of two-note chords as compiled from seven historic studies of consonance perception. The composite model performs well here ($r = .87$, 95% CI [.59, .96]), seemingly outperforming the sub-models ($.73 < r < .85$), but the small dataset size limits the statistical power of these comparisons.

In a subsequent exploratory analysis, we benchmarked the composite model's

[19] See Section 3.9.3 for more information about these datasets.

[20] In conducting these analyses, we detected several apparent errors in the roughness values reported by Johnson-Laird et al. (2012). Here we use roughness values as computed by our new *incon* package.

performance against the 10 best-performing models from Figure 3.4A. Model performance varied across datasets, and in some cases individual models achieved higher correlation coefficients than the composite model. However, no model significantly outperformed the composite model at a $p < .05$ level in any given dataset, even without correcting for multiple comparisons.

These evaluations provide qualified support for the composite model's generalisability across datasets. Predictive performance is generally good, with the composite model typically matching or improving upon the performance of pre-existing models. However, these inferences are constrained by the small dataset sizes of previous studies, which limit the precision of performance evaluations. A further limitation is that most previous studies do not manipulate the number of notes in the chord, which makes it difficult to test the generalisability of the number-of-notes effect observed in the Bowling et al. (2018) dataset. These limitations should be addressed in subsequent empirical work.

3.6.4 Recommendations for model selection

Figure 3.4A shows that consonance models representing similar psychological theories can vary widely in performance. This highlights the danger of testing psychological theories with single computational models, especially when those models are relatively unvalidated. For example, Lahdelma & Eerola (2016) found that MIRtoolbox's inharmonicity measure failed to predict consonance judgments, and concluded that periodicity/harmonicity does not contribute much to consonance. Our analyses replicate the low predictive power of MIRtoolbox's inharmonicity measure (partial $r < .2$), but they show that other periodicity/harmonicity measures can predict consonance much better (partial $r > .7$). If Lahdelma & Eerola (2016) had selected a different periodicity/harmonicity model, their conclusions might therefore have been very different.

Figure 3.4A provides useful information for model selection. All else aside, models with higher predictive performance are likely to be better instantiations of their respective psychological theories. Here we selected the three best-performing models in Figure 3.4A, which usefully represent three different consonance theories: interference, periodicity/harmonicity, and cultural familiarity. However, several models reached similar levels of performance, and should be retained as good candidates for consonance modelling. Stolzenburg's (2015) model performed especially well on the validation datasets, and should be considered a recommended alternative to the harmonicity model introduced in this chapter. Likewise, if it is desirable for the model to be hand-computable, Huron's (1994) model and Parncutt's (1988) model both perform remarkably well given their simplicity. When only audio information is available, our results suggest that MIRtoolbox's roughness measure is the best candidate for estimating consonance. In contrast, none of the audio-based periodicity/harmonicity measures were able to predict consonance.

There are some applications, such as emotion research, music information retrieval, or algorithmic music composition, where a composite model of consonance may be more useful than models representing individual consonance mechanisms. The composite model presented here would be well-suited for this role. However, the model would benefit from further tuning and validation, ideally on datasets varying chord spacing, tone spectra, and the number of notes in the chord.

3.7 Corpus analyses

We have argued that chord prevalences can provide a proxy for a listener's musical exposure, and therefore can be used to model the contribution of cultural familiarity to consonance perception. However, these chord prevalences may themselves be partly determined by non-cultural aspects of consonance perception, such as periodicity/harmonicity and interference.

A recent study by Parncutt et al. (2018) addressed these potential predictors of chord prevalences. The authors compiled a corpus of vocal polyphonic music spanning seven centuries of Western music, and correlated chord prevalences in this corpus with four features: interference, periodicity/harmonicity, diatonicity, and evenness. They predicted that interference and periodicity/harmonicity should respectively be negatively and positively related to chord prevalence, on account of these features' respective contributions to perceptual consonance. They predicted that diatonic chords – chords played within the Western diatonic scale – should be more common, because the familiarity of the diatonic scale induces consonance in Western listeners. They also predicted that chord prevalences should be higher for chords whose notes are approximately evenly spaced, because even spacing is associated with efficient voice leading (Tymoczko, 2011).

Parncutt and colleagues tested these hypotheses by counting occurrences of 19 different three-note chord types in their dataset. They compiled a selection of formal models for each feature, and correlated model outputs with chord counts in their musical corpus, splitting the analysis by different musical periods. The observed correlations were generally consistent with the authors' predictions, supporting the notion that perceptual consonance contributes to Western chord prevalences.

While a useful contribution, this study has several important limitations. First, restricting consideration to just 19 chord types results in very imprecise parameter estimates. For example, a correlation coefficient of $r = .5$ has a 95% confidence interval ranging from .06 to .78; it is difficult to draw reliable inferences from such information. Second, pairwise correlations are unsuitable for quantifying causal effects when the outcome variable potentially depends on multiple predictor variables. Third, pairwise correlations can only capture linear relationships, and therefore cannot test more complex relationships between chord usage and consonance, such as the proposition that chord usage is biased towards intermediate levels of consonance (Lahdelma & Eerola, 2016). Fourth, the consonance models are simple note-counting models, which often lack specificity to the feature being analysed. For example, interference is modelled using the dyadic consonance model of Huron (1994), but this model is built on dyadic consonance judgments which have recently been attributed to periodicity/harmonicity, not interference (McDermott et al., 2010; Stolzenburg, 2015).

Here we address these limitations, analysing chord occurrences in three large corpora spanning the last thousand years of Western music: a corpus of classical scores with composition dates ranging from 1198 to 2011 (Viro, 2011), a corpus of jazz lead sheets sourced from an online forum for jazz musicians (Broze & Shanahan, 2013), and a corpus of popular songs sampled from years 1958–1991 of the Billboard charts and transcribed by expert musicians (Burgoyne, 2011).[21] Instead of restricting consideration to 19 chord types, we tabulated prevalences for all 2,048 possible pitch-class chord types. Instead of pairwise correlations, we constructed polynomial regression models capable of capturing nonlinear effects of multiple simultaneous predictors. Instead of simple note-counting models, we used the best-performing consonance models from Figure 3.4A: Hutchinson & Knopoff's (1978) interference model, and this chapter's new harmonicity model.

We were particularly interested in how interference and periodicity/harmonicity contributed to chord prevalence. However, we also controlled for the number of notes in the chord, reasoning that this feature is likely to have constrained chord usage on account of practical constraints (e.g. the number of instruments in an ensemble).

Analysing interference and periodicity/harmonicity allows us to revisit recent claims that consonance is primarily determined by periodicity/harmonicity and not interference (Cousineau et al., 2012; McDermott et al., 2010, 2016). If consonance is indeed predicted primarily by periodicity/harmonicity, we would expect periodicity/harmonicity to be an important predictor of Western chord prevalences, and that interference should have little predictive power after controlling for periodicity/harmonicity. Conversely, if consonance derives from

[21] See Section 3.9.4 for more details about these datasets.

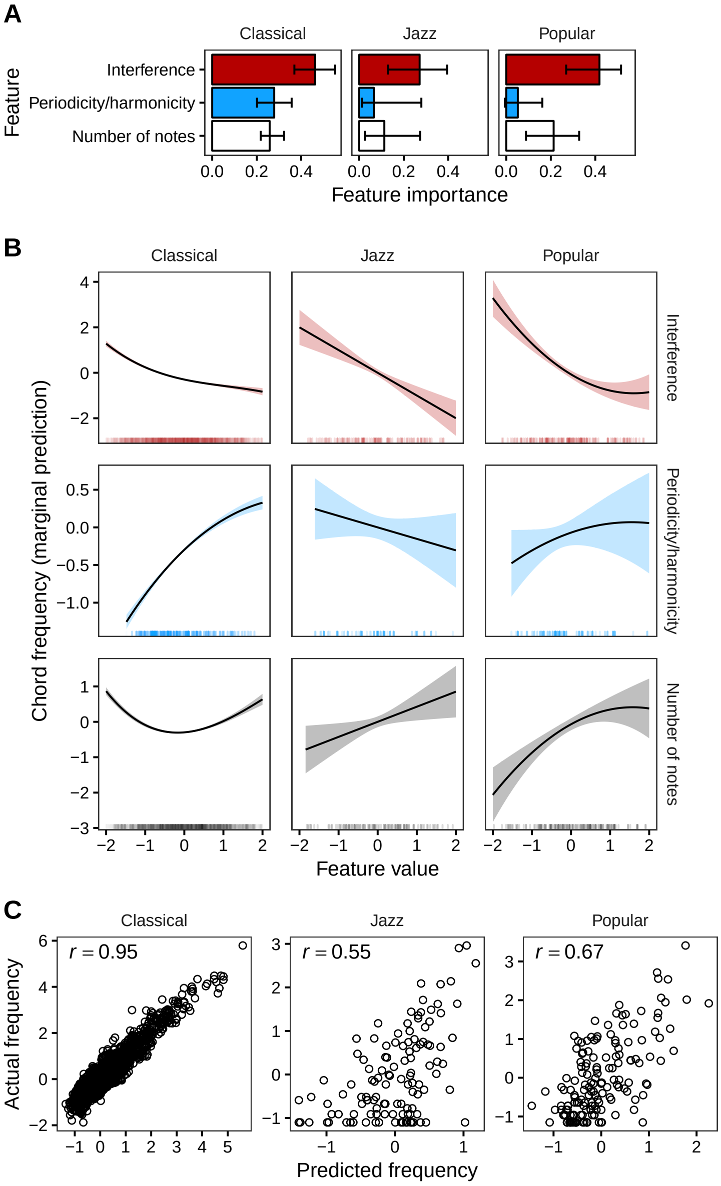

Figure 3.5: Results of the corpus analyses. **A**: Permutation-based feature importance (Breiman, 2001; Fisher, Rudin, & Dominici, 2018), with error bars indicating 95% confidence intervals (bias-corrected and accelerated bootstrap, 100,000 replicates, DiCiccio & Efron, 1996). **B**: Marginal effects of each feature, calculated using z-scores for feature values and for chord frequencies. The shaded areas describe 95% confidence intervals, and distributions of feature observations are plotted at the bottom of each panel. Distributions for the 'number of notes' feature are smoothed to avoid overplotting. **C**: Predicted and actual chord-type frequencies, alongside corresponding Pearson correlation coefficients.

both interference and periodicity/harmonicity, then we might expect both features to contribute to chord prevalences.

Compiling chord prevalences requires a decision about how to categorise chords into chord types. Here we represented each chord as a *pitch-class chord type*, defined as a pitch-class set expressed relative to the bass pitch class. This representation captures the perceptual principles of octave invariance (the chord type is unchanged when chord pitches are transposed by octaves, as long as they do not move below the bass note) and transposition invariance (the chord type is unchanged when all the chord's pitches are transposed by the same interval).

Hutchinson & Knopoff's model requires knowledge of precise pitch heights, which are not available in pitch-class chord type representations. We therefore assigned pitch heights to each chord type by applying the automatic chord voicing algorithm developed in Chapter 5 (see Section 3.9 for details).

Chord type prevalences could be operationalised in various ways. Ideally, one might sum the temporal duration of each chord type over all of its occurrences, perhaps weighting compositions by their popularity to achieve the best representation of a given musical style. However, chord durations and composition popularity were not available for our classical and jazz datasets. We therefore operationalised chord type prevalences as the total number of occurrences of each chord type, excluding immediate repetitions of the same chord (see Section 3.9).

We constructed three orthogonal polynomial regression models predicting log-transformed chord counts from interference, periodicity/harmonicity, and number of notes. The classical, jazz, and popular corpora contributed 2,048, 118, and 157 data points respectively, corresponding to the unique chord types observed in each corpus and their respective counts. Each corpus was assigned its own polynomial order by minimising the Bayesian Information Criterion for the fitted model; the classical, jazz, and popular datasets were thereby assigned third-order, first-order, and second-order polynomials respectively.

Figure 3.5A quantifies each predictor's importance using a permutation-based feature-importance metric (Breiman, 2001, see Section 3.9 for details). Across the three genres, interference was consistently the most important predictor, explaining c. 20–50% of the variance in chord prevalences. Periodicity/harmonicity was also an important predictor for classical music, but not for popular or jazz music. Number of notes predicted chord prevalences in all three genres, explaining about half as much variance as interference.

Figure 3.5B plots the marginal effects of each predictor, showing how feature values map to predictions. Interference had a clear negative effect on chord prevalence in all three genres, consistent with the notion that interference evokes dissonance, causing it to be disliked by listeners and avoided by composers.

Periodicity/harmonicity had a clear positive effect on chord prevalence in the classical dataset, consistent with the idea that periodicity/harmonicity evokes consonance and is therefore promoted by composers (Figure 3.5B). The effect of periodicity/harmonicity was less strong in the popular and jazz datasets, taking the form of a weak positive effect in the popular dataset and a weak negative effect in the jazz dataset.

Figure 3.5C summarises the predictive performances of the three regression models. Generally speaking, predictive performances were high, indicating that consonance and number of notes together explain a large part of Western chord prevalences. However, the strength of this relationship varied by musical style, with the classical dataset exhibiting the strongest relationship and the jazz dataset the weakest relationship.

In sum, these results weigh against the claim that consonance is primarily determined by periodicity/harmonicity and not interference (Bowling & Purves, 2015; Bowling et al., 2018; McDermott et al., 2010). Across musical genres, interference seems to have a strong and reliable negative effect on chord prevalences. Periodicity/harmonicity also seems to influence chord prevalences, but its effect is generally less strong, and the nature of its contribution seems to vary across musical genres.

3.8 Discussion

Recent research argues that consonance perception is driven not by interference but by periodicity/harmonicity, with cultural differences in consonance perception being driven by learned preferences for the latter (Cousineau et al., 2012; McDermott et al., 2010, 2016). We reassessed this claim by reviewing a wide range of historic literature, modelling perceptual data from four previous empirical studies, and conducting corpus analyses spanning a thousand years of Western music composition. We concluded that interference contributes significantly to consonance perception in Western listeners, and that cultural aspects of consonance perception extend past learned preferences for periodicity/harmonicity. Instead, consonance perception in Western listeners seems to be jointly determined by interference, periodicity/harmonicity perception, and learned familiarity with particular musical sonorities.

This multicomponent account of consonance is broadly consistent with several previous claims in the literature. Terhardt (1974, 1984) has emphasised the role of roughness and harmonicity in determining consonance, and Parncutt and colleagues have argued that consonance depends on roughness, harmonicity, and familiarity (Parncutt & Hair, 2011; Parncutt et al., 2018). Scientific preferences for parsimony may have caused these multicomponent accounts to be neglected in favour of single-component accounts, but our analyses demonstrate the necessity of the multicomponent approach.

This consolidation of multiple psychological mechanisms makes an interesting parallel with historic pitch perception research, where researchers strove to demonstrate whether pitch perception was driven by place coding or temporal coding (see de Cheveigné, 2005 for a review). It proved difficult to falsify either place coding or temporal coding theories, and many researchers now believe that both mechanisms play a role in pitch perception (e.g. Bendor, Osmanski, & Wang, 2012; Moore & Ernst, 2012).

Like most existing consonance research, our analyses were limited to Western listeners and composers, and therefore we can only claim to have characterised consonance in Westerners. Previous research has identified significant cross-cultural variation in consonance perception (Florian, 1981; Maher, 1976; McDermott et al., 2016); we suggest that this cross-cultural variation might be approximated by varying the regression coefficients in our composite consonance model. For example, listeners familiar with beat diaphony seem to perceive interference as consonant, not dissonant (Florian, 1981); this would be reflected in a reversed regression coefficient for interference. While the regression coefficients might vary cross-culturally, it seems plausible that the model's underlying predictors – interference, periodicity/harmonicity, familiarity – might recur cross-culturally, given the cross-cultural perceptual salience of these features (McDermott et al., 2016).

Our conclusions are not inconsistent with vocal-similarity theories of consonance perception (Bowling & Purves, 2015; Bowling et al., 2018; Schwartz et al., 2003). According to these theories, certain chords sound consonant because they particularly resemble human vocalisations. These theories usually emphasise periodicity/harmonicity as a salient feature of human vocalisations, but they could also implicate interference as a feature avoided in typical vocalisations (Bowling et al., 2018) but used to convey distress in screams (Arnal, Flinker, Kleinschmidt, Giraud, & Poeppel, 2015). It seems plausible that these mechanisms contribute a universal bias to perceive periodicity/harmonicity as pleasant and interference as unpleasant. Nonetheless, these biases must be subtle enough to allow cultural variation, if we are to account for musical cultures that lack preferences for periodicity/harmonicity (McDermott et al., 2016) or that consider interference to be pleasant (Florian, 1981).

Our perceptual analyses were limited by the available empirical data. We paid particular attention to the perceptual dataset of Bowling et al. (2018) on account of its large sample size, yet this dataset limits consideration to chords comprising no more than four notes, all drawn from a small pitch range (one octave), and tuned using an idiosyncratic approximation to just intonation (see Section 3.9). Future work should expand these datasets, with particular emphasis on varying voicing, tone spectra, tuning systems, and number of notes in the chord. Such datasets would be essential for testing the generalisability of our models.

Our perceptual analyses marginalised over participants, producing an average consonance rating for each chord. This approach neglects individual differences, which can provide an important complementary perspective on consonance perception (McDermott et al., 2010). When suitable empirical datasets become available, it would be interesting to investigate how the regression weights in Figure 3.4C vary between participants.

Our corpus analyses presented very broad approximations to musical genres, aggregating over a variety of musical styles and time periods. It would be interesting to apply these methods to more specific musical styles, or indeed to individual composers. It would also be interesting to investigate the evolution of consonance treatment over time. As we analyse music compositions dating further back in history, we should expect the chord distributions to reflect consonance perception in historic listeners rather than modern listeners. Such analyses could potentially shed light on how consonance perception has changed over time (Parncutt et al., 2018).

Our three corpora were constructed in somewhat different ways. The classical corpus was derived from published musical scores; the jazz corpus constitutes a collection of lead sheets; the popular corpus comprises expert transcriptions of audio recordings. This heterogeneity is both an advantage, in that it tests the generalisability of our findings to different transcription techniques, and a disadvantage, in that it reduces the validity of cross-genre comparisons. Future work could benefit from corpora with both stylistic diversity and consistent construction.

Our analyses were limited by the computational models tested. It would be interesting to develop existing models further, perhaps producing a version of Bowling et al.'s (2018) periodicity/harmonicity model that accepts arbitrary tunings, or a version of Parncutt & Strasburger's (1994) model without discrete-pitch approximations. It would also be interesting to test certain models not evaluated here, such as Boersma's (1993) model and Vencovsky's (2016) model. Lastly, we would like to dissuade future researchers from permanently discarding specific consonance models based on our results. In particular, the audio-based models depend on many customisable parameters such as sample rates, windowing lengths, and peak-detection thresholds, and it is quite possible that further exploration of these parameters could yield better performance on these datasets.

The psychological community often uses 'consonance' as a synonym for tonal ‘pleasantness’ (e.g. Bowling et al., 2018; Cousineau et al., 2012; McDermott et al., 2010, 2016), and this convention is also followed by the present chapter. From a linguistic perspective, however, consonance and pleasantness have subtly different connotations. In particular, a musician is likely to evaluate consonance with respect to the prototypical consonant sonorities of music theory, even if they personally consider certain non-traditional sonorities to sound pleasant. Moreover, their pleasantness judgements may be guided by phenomena other than those traditionally associated with consonance: for example, the sense in which the chord possesses some kind of interesting tonal implications, elicits an unusual acoustic effect, or evokes a powerful emotion. These and other related-yet-distinct dimensions of perceptual evaluation (e.g. tension, harmoniousness, preference) provide useful and complementary perspectives on the acoustic and psychological mechanisms underlying harmony perception. Despite early interest in the topic (e.g. Guernsey, 1928; Geer et al., 1962), these semantic distinctions have been often neglected in subsequent work, and it is good to see this area being revisited in recent empirical studies (e.g. Lahdelma & Eerola, 2016, 2019; Popescu et al., 2019).

We hope that our work will facilitate future psychological research into consonance. Our *incon* package makes it easy to test diverse consonance models on new datasets, and it can be easily extended to add new models. Our *inconData* package compiles the perceptual datasets analysed here, making it easy to test new consonance models on a variety of perceptual data.

This work should also have useful applications in computational musicology and music information retrieval. Our composite consonance model provides a principled way to operationalise the net consonance of a musical chord, while our model evaluations provide a principled way to operationalise individual consonance theories. Our software provides a consistent and easy-to-use interface to these models, facilitating their application to new datasets.

3.9 Methods

3.9.1 Models

The models evaluated in this chapter are available from three software sources: the *incon* package, *MIRtoolbox*[22], and *Essentia*[23]. Unless otherwise mentioned, all *incon* models represent unaltered versions of their original algorithms as described in the cited literature, with the exception that all idealised harmonic spectra comprised exactly 11 harmonics (including the fundamental frequency),

[22] `https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/materials/mirtoolbox`

[23] `https://essentia.upf.edu`

with the ith harmonic having an amplitude of i^{-1}, and assuming incoherence between tones for the purpose of amplitude summation. We clarify some further details below.

Milne (2013); New harmonicity model

These algorithms have three free parameters: the number of harmonics modelled in each complex tone, the harmonic roll-off rate (ρ), and the standard deviation of the Gaussian smoothing distribution (σ). We set the number of harmonics to 11 (including the fundamental frequency), and set the other two parameters to the optimised values in Milne & Holland (2016): a roll-off of $\rho = 0.75$, and a standard deviation of $\sigma = 6.83$ cents.

Hutchinson & Knopoff (1978)

Our implementation is based on Mashinter (2006), whose description includes a parametric approximation for the relationship between interval size and pure-dyad dissonance (see also Bigand et al., 1996).

Sethares (1993)

Our implementation is primarily based on Sethares (1993), but we include a modification suggested in later work (Sethares, 2005; Weisser & Lartillot, 2013) where pure-dyad consonance is weighted by the minimum amplitude of each pair of partials, not the product of their amplitudes.

Wang et al. (2013)

Our implementation of Wang et al.'s (2013) algorithm takes symbolic input and expresses each input tone as an idealised harmonic series. Time-domain analyses are conducted with a signal length of 1 s and a sample rate of 44,000. Frequency-domain analyses are conducted in the range 1–44,000 Hz with a resolution of 1 Hz. An interactive demonstration of the algorithm is available at `http://shiny.pmcharrison.com/wang13`.

Essentia: Interference

We used version 2.1 of Essentia. We analysed each audio file using the 'essentia_streaming_extractor_music' feature extractor, and retained the mean estimated dissonance for each file.

MIRtoolbox: Interference

We used version 1.6.1 of MIRtoolbox, and computed roughness using the 'mirroughness' function. The function was applied to a single window spanning the entire length of the stimulus.

We evaluated this model in several configurations (see Figure 3.4A):

a) 'Sethares' denotes the default model configuration, which implements the dissonance model of Sethares (2005), but with pure-tone dyad contributions being weighted by the *product* of their amplitudes (see Sethares, 1993);

b) 'Sethares, v2' denotes the 'Min' option in MIRtoolbox, where pure-tone dyad contributions are weighted by the *minimum* of their amplitudes, after Weisser & Lartillot (2013; see also Sethares, 2005);

c) 'Vassilakis' denotes MIRtoolbox's implementation of Vassilakis's (2001) model.

Johnson-Laird et al. (2012)

Johnson-Laird et al.'s (2012) algorithm may be separated into a cultural and an interference component, with the latter corresponding to Hutchinson & Knopoff's (1978) model. The cultural model assigns each chord to a consonance category, where categories are ordered from consonant to dissonant, and chords within a category are considered to be equally consonant. In our implementation, these consonance categories are mapped to positive integers, such that higher integers correspond to greater dissonance. These integers constitute the algorithm's outputs.

Corpus-based model of cultural familiarity

This model estimates a listener's unfamiliarity with a given chord type from its rarity in a musical corpus. Here we use the Billboard dataset (Burgoyne, 2011), a corpus of popular songs sampled from the Billboard magazine's 'Hot 100' chart in the period 1958–1991. This corpus is used as a first approximation to an average Western listener's prior musical exposure. We represent each chord in this corpus as a *pitch-class chord type*, defined as the chord's pitch-class set expressed relative to the chord's bass note. For example, a chord with MIDI note numbers {66, 69, 74} has a pitch-class chord type of {0, 3, 8}. We count how many times each of the 2,048 possible pitch-class chord types occurs in the corpus, and add 1 to the final count. Unfamiliarity is then estimated as the negative natural logarithm of the chord type's count.

Table 3.3: Unstandardised regression coefficients for the composite consonance model.

Term	Coefficient
Intercept	0.628434666589357
Number of notes	0.422267698605598
Interference	−1.62001025973261
Periodicity/harmonicity	1.77992362857478
Culture	−0.0892234643584134

Note. These regression coefficients are presented to full precision for the sake of exact reproducibility, but it would also be reasonable to round the coefficients to c. 3 significant figures. When generalising outside the dataset of Bowling et al. (2018), we recommend setting the number of notes coefficient to zero.

Composite model

The composite model's unstandardised regression coefficients are provided to full precision in Table 3.3. Consonance is estimated by computing the four features listed in Table 3.3, multiplying them by their respective coefficients, and adding them to the intercept coefficient. Number of notes corresponds to the number of distinct pitch classes in the chord; interference is computed using Hutchinson & Knopoff's (1978) model; periodicity/harmonicity is computed using this chapter's new harmonicity model; culture corresponds to the new corpus-based cultural model.

It is unclear whether the effect of number of notes generalises outside the dataset of Bowling et al. (2018) (see Section 3.6). We therefore recommend setting the number of notes coefficient to zero when applying the model to new datasets.

3.9.2 Software

We release two top-level R packages along with this chapter. The first, *incon*, implements the symbolic consonance models evaluated in this chapter (Table 3.2).[24] The second, *inconData*, compiles the perceptual datasets that we analysed.[25] Tutorials are available alongside these packages.

The *incon* package depends on several low-level R packages that we also release along with this chapter, namely *bowl18*, *corpdiss*, *dycon*, *har18*, *hcorp*, *hrep*, *jl12*, *parn88*, *parn94*, *stolz15*, and *wang13*. These packages provide detailed

[24] `https://github.com/pmcharrison/incon`

[25] `https://github.com/pmcharrison/inconData`

interfaces to individual consonance models and tools for manipulating harmony representations.

Our software, analyses, and manuscript were all created using the programming language R (R Core Team, 2017), and benefited in particular from the following open-source packages: *bookdown*, *boot*, *checkmate*, *cocor*, *cowplot*, *dplyr*, *ggplot2*, *glue*, *gtools*, *hht*, *knitr*, *jsonlite*, *magrittr*, *margins*, *memoise*, *numbers*, *papaja*, *phonTools*, *plyr*, *purrr*, *Rdpack*, *readr*, *rmarkdown*, *testthat*, *tibble*, *tidyr*, *usethis*, *withr*, and *zeallot*. Our analysis code is freely available online.[26]

3.9.3 Perceptual datasets

The following datasets are all included in our *inconData* package.

Bowling et al. (2018)

This study collected consonance judgments for all possible 12 two-note chord types, 66 three-note chord types, and 220 four-note chord types that can be formed from the Western chromatic scale within a one-octave span of the bass note.[27] An advantage of this dataset is its systematic exploration of the chromatic scale; a disadvantage is its restricted range of voicings.

Each chord tone was pitched as a just-tuned interval from the bass note.[28] This approach was presumably chosen because Bowling et al.'s (2018) periodicity/harmonicity model requires just tuning, but it should be noted that just tuning itself is not commonly adopted in Western music performance (e.g. Karrick, 1998; Kopiez, 2003; Loosen, 1993). It should also be noted that tuning a chord in this way does not ensure that the intervals between non-bass notes are just-tuned, and certain chords can sound unusually dissonant as a result compared to their equal-tempered equivalents.

Each chord type was assigned a bass note such that the chord's mean fundamental frequency would be equal to middle C, approximately 262 Hz. The resulting chords were played using the 'Bosendorfer Studio Model' synthesised piano in the software package 'Logic Pro 9'.

The participant group numbered 30 individuals. Of these, 15 were students at a Singapore music conservatory, each having taken weekly formal lessons in Western tonal music for an average of 13 years ($SD = 3.8$). The remaining 15 participants were recruited from the University of Vienna, and averaged less than a year of weekly music lessons prior to the study ($SD = 1.1$).

[26] `https://github.com/pmcharrison/inconPaper`

[27] As before, a *chord type* represents a chord as a set of intervals above an unspecified bass note (Chapter 2).

[28] Just tuning means expressing pitch intervals as small-integer frequency ratios. In Bowling et al. (2018), the eleven intervals in the octave were expressed as the following frequency ratios: 16:15, 9:8, 6:5, 5:4, 4:3, 7:5, 3:2, 8:5, 5:3, 9:5, 15:8, and 2:1.

Participants were played single chords, and asked to rate consonance on a four-point scale, where consonance was defined as 'the musical pleasantness or attractiveness of a sound'. Participants were free to listen to the same chord multiple times before giving a rating. Stimulus presentation was blocked by the number of notes in each chord, with stimulus presentation randomised within blocks. This presents an unfortunate potential confound; if consonance differed systematically across chords containing different numbers of notes, this may have caused participants to recalibrate their scale usage across blocks.

Johnson-Laird et al. (2012), Experiment 1

This experiment collected consonance ratings for all 55 possible three-note *pitch-class chord types*, where a pitch-class chord type is defined as a chord's pitch-class set expressed relative to the bass pitch class (Chapter 2). These chords were voiced so that each chord spanned approximately 1.5 octaves. All chords were played with synthesised piano using the 'Sibelius' software package.

The participant group numbered 27 individuals from the Princeton University community. Some were nonmusicians, some were musicians, but all were familiar with Western music.

Participants were played single chords, and asked to rate dissonance on a seven-point scale, where dissonance was defined as 'unpleasantness'. Each chord was only played once, with presentation order randomised across participants.

Johnson-Laird et al. (2012), Experiment 2

This experiment collected consonance ratings for 43 four-note pitch-class chord types. The rationale for chord selection is detailed in Johnson-Laird et al. (2012); particularly relevant is the decision to undersample chords containing three adjacent semitones, which may have mitigated contributions of interference to their results.

The participant group numbered 39 individuals from the Princeton University community. All other aspects of the design were equivalent to Experiment 1.

Lahdelma & Eerola (2016)

This experiment collected consonance ratings for 15 different *pitch chord types*, where a pitch chord type is defined as a chord's pitch set expressed relative to its bass pitch (Chapter 2). These chords ranged in size from three to six notes. The full rationale for chord selection is detailed in Lahdelma & Eerola (2016), but the main principle was to select chords with high consonance according to Huron's (1994) dyadic consonance model, and with varying levels of cultural familiarity according to Tymoczko (2011). Since Huron's model primarily captures interference (see Section 3.5), this approach is likely to minimise between-stimulus variation in interference, potentially reducing the predictive power of interference models within this dataset. All chords were played using the synthesised 'Steinway D Concert Grand' piano in the software package 'Ableton Live 9' with the 'Synthogy Ivory Grand Pianos II' plug-in.

The participant group was tested online, and numbered 418 individuals after quality-checking. These participants represented 42 different nationalities, with 91.7% coming from Europe and the Americas.

Each participant was played 30 stimuli comprising the 15 chord types each at a 'low' and a 'high' transposition, with the precise transpositions of these chord types randomly varying within an octave for each transposition category. Unfortunately, precise transposition information seems not to be preserved in the published response data. For the purpose of estimating interference, we therefore represented each chord type with a bass note of G4 (c. 392 Hz), corresponding to the middle of the range of bass notes used in the original study.

Participants were instructed to rate each chord on five five-point scales; here we restrict consideration to the 'consonance' scale. Curiously, 'consonance' was defined as 'How smooth do you think the chord is', with the scale's extremes being termed 'rough' and 'smooth'. This definition resembles more a definition of roughness than consonance, a potential problem for interpreting the study's results.

Schwartz et al. (2003)

This dataset provides consonance ratings for the 12 two-note chord types in the octave, aggregated over seven historic studies. Each study produced a rank ordering of these two-note chords; these rank orderings were then summarised by taking the median rank for each chord.

3.9.4 Musical corpora

Classical scores

The classical dataset was derived from the Peachnote music corpus (Viro, 2011).[29] This corpus compiles more than 100,000 scores from the Petrucci Music Library (IMSLP, `http://imslp.org`), spanning several hundred years of Western art music (1198–2011). Each score was digitised using optical music recognition software; such digitisation processes will surely bring some inaccuracies, but we anticipated that such inaccuracies would be unlikely to produce

[29] In particular, we downloaded the 'Exact 1-gram chord progressions' file from `http://www.peachnote.com/datasets.html` on July 2nd, 2018.

systematic confounds in the perceptual analyses. In the resulting dataset, each datum represents a distinct 'vertical slice' of the score, with new slices occurring at new note onsets, and including sustained notes sounded at previous onsets. We preprocessed this dataset to a pitch-class chord-type representation, where each chord is represented as a pitch-class set expressed relative to its bass pitch class. The resulting dataset numbered 128,357,118 chords.

Jazz lead sheets

The jazz dataset was derived from the iRb corpus (Broze & Shanahan, 2013). The iRb corpus numbers 1,186 lead sheets for jazz compositions, where each lead sheet specifies the underlying chord sequence for a given composition. These lead sheets were compiled from an online forum for jazz musicians. In the original dataset, chords are represented as textual tokens, such as 'C7b9'; we translated all such tokens into a prototypical pitch-class chord-type representation, such as $\{0, 1, 4, 7, 10\}$. This process misses the improvisatory chord alterations that typically happen during jazz performances, but nonetheless should provide a reasonable first approximation to the performed music. Chord counts were only incremented on chord changes, not chord repetitions; section repeats were omitted. The resulting dataset numbered 42,822 chords.

Popular transcriptions

The popular dataset was derived from the McGill Billboard corpus (Burgoyne, 2011), which comprised chord sequences for 739 unique songs as transcribed by expert musicians. As with the iRb dataset, we translated all chord tokens into prototypical pitch-class chord-type representations, omitting section repeats, and only incrementing chord counts on each chord change. The resulting dataset numbered 74,093 chords.

3.9.5 Corpus analyses

We transformed each of our corpora to *pitch-class chord type* representations, where each chord is represented as a pitch-class set relative to the chord's bass note (Chapter 2). We then counted occurrences of pitch-class chord types in our three corpora.

For the purpose of applying Hutchinson & Knopoff's (1978) interference model, we assigned pitch heights to each chord type using the automatic chord voicing algorithm described in Chapter 5. This model is primarily designed for voicing chord sequences, but it can also be applied to individual chords. Its purpose is to find an idiomatic assignment of pitch heights to pitch classes that reflects the kind of psychoacoustic considerations implicitly followed by traditional Western composers (e.g. Huron, 2001). As applied here, the model minimised the following linear combination of features:

$$\begin{aligned} 8.653 &\times \text{interference} \\ &+ 1.321\times \mid 5 - \text{number of notes } \mid \\ &+ 0.128\times \mid 60 - \text{mean pitch height } \mid \end{aligned} \tag{3.12}$$

where ‘interference’ refers to the raw output of Hutchinson & Knopoff’s model, ‘number of notes’ refers to the number of unique pitches in the chord voicing, and ‘mean pitch height’ corresponds to the mean of the chord’s pitches as expressed in MIDI note numbers.[30] In other words, the model minimised the chord’s interference while preferring chords containing (close to) five discrete pitches with a mean pitch height close to middle C (c. 262 Hz). These model parameters correspond to the optimal parameters derived in Chapter 5 from a dataset of 370 chorale harmonisations by J. S. Bach, but with the target number of notes changed from four to five to reflect the richer harmonic vocabularies of the three datasets. Chord voicings were restricted to the two octaves surrounding middle C, and were permitted to contain no more than five notes or the number of pitch classes in the chord type, whichever was greater.

We used polynomial regression to capture nonlinear relationships between chord features and chord prevalences. We used orthogonal polynomials, as computed by the R function ‘poly’, to avoid numerical instability, and we used the R package ‘margins’ to compute marginal predictions for the resulting models.

Standardised regression coefficients become harder to interpret as the polynomial degree increases. We therefore instead calculate a permutation-based feature-importance metric commonly used for assessing feature importance in random forest models (Breiman, 2001; see also Fisher et al., 2018). This metric may be calculated by computing two values: the model’s original predictive accuracy, and the model’s predictive accuracy after randomly permuting the feature of interest (without refitting the model). The metric is then defined as the difference in these accuracies: the greater the difference, the more the model relies on the feature of interest. Here we used R^2 as the accuracy metric, and computed confidence intervals for our feature-importance estimates using bias-corrected accelerated bootstrapping with 100,000 replicates (DiCiccio & Efron, 1996).

[30] A frequency of f Hz corresponds to a MIDI note number of $69 + 12\log_2(f/440)$.

Chapter 4

Harmonic expectation

4.1 Introduction

Probabilistic theories of music cognition hold that listeners experience music by continually generating predictions for upcoming musical events, and relating these predictions to the musical events as they occur (Huron, 2006; Koelsch, Vuust, & Friston, 2019; Meyer, 1957; Pearce, 2018; Salimpoor, Zald, Zatorre, Dagher, & McIntosh, 2015). These predictions reflect the listener's internal model of the musical style, developed through automatic processes of statistical learning. The resulting dynamics of fulfilled and denied expectations are thought to be integral to the emotional experience of music (Egermann, Pearce, Wiggins, & McAdams, 2013; Huron, 2006; Meyer, 1957; Pearce, 2018; Sauvé, Sayed, Dean, & Pearce, 2018; Sloboda, 1991).

One musical dimension that elicits strong predictions, or 'expectations', is harmony. Harmony is a fundamental structuring principle in Western music, describing how musical notes combine to form chords, and how these chords combine to form chord sequences. The expectedness of a given chord is context-dependent: for example, an E major chord may be incongruent when preceded by a C major chord, but expected when preceded by a B major chord. Harmonic expectation therefore bears similarity with linguistic syntax, where the expectedness of a word depends on the preceding words in the sentence, as well as the individual's internal model of the grammatical structure of the language. Correspondingly, the structuring principles underlying the construction of chord progressions are often termed harmonic 'syntax' (Patel, 2003; Pearce & Rohrmeier, 2018; Rohrmeier, 2011; Rohrmeier & Pearce, 2018).

Some research has shown how harmonic expectation may be driven by low-level psychoacoustic cues. For example, Bigand et al. (2014) have shown that the expectedness of musical chords may be predicted by the acoustic similarity between a chord and its musical context, where the context is accumulated in an echoic memory buffer with temporal decay (see also Craton, Lee, & Krahe, 2019; Bigand et al., 1996; Leman, 2000a; Milne & Holland, 2016; Parncutt, 1989). Other candidate cues for harmonic expectation include pitch distance (Bigand et al., 1996; Parncutt, 1989; Tymoczko, 2006), harmonicity (Bowling et al., 2018; McDermott et al., 2010; Stolzenburg, 2015; Terhardt, 1974), and interference between partials (Hutchinson & Knopoff, 1978; Plomp & Levelt, 1965). These different cues may combine to determine the overall expectedness of a given chord (e.g. Bigand et al., 1996).

Other research has shown how harmonic expectations may be driven by high-level cognitive processes, similar to those involved in parsing the syntax of natural language. Some evidence for this phenomenon comes from harmonic priming studies, which analyse how the speed and accuracy of perceptual judgments are affected by harmonic expectedness. Several such studies have shown a perceptual facilitation effect for syntactically expected chords, even when accounting for various psychoacoustic cues that might explain such a facilitation effect (Bigand et al., 2003; Sears, Pearce, Spitzer, Caplin, & McAdams, 2019; Tekman & Bharucha, 1998). More evidence for high-level harmony cognition comes from studies showing that listeners can be sensitive to nonadjacent syntactic dependencies, which are difficult to capture with simple psychoacoustic models (Koelsch, Rohrmeier, Torrecuso, & Jentschke, 2013; Rohrmeier & Cross, 2009; Spyra, Stodolak, & Woolhouse, 2019; Woolhouse, Cross, & Horton, 2016). A particular relationship with linguistic syntax processing is suggested by studies showing mutual interference between linguistic and harmonic syntax processing (Hoch et al., 2011; Koelsch et al., 2005; Kunert et al., 2016; Slevc et al., 2009; but see Poulin-Charronnat et al., 2005; Escoffier & Tillmann, 2008; Perruchet & Poulin-Charronnat, 2013; Slevc & Okada, 2015).

Surveying this evidence, it seems likely that both low-level and high-level perceptual processes contribute to harmonic expectation. Nonetheless, it remains unclear how these different mechanisms combine in practice. Previous studies of harmonic expectation typically use artificial stimuli specially constructed to probe a particular component of harmonic expectation; these studies may show that a given mechanism can contribute to harmonic expectation, but they cannot quantify how much the mechanism typically contributes to the perception of typical music compositions.

Here we address this challenge, seeking to investigate how different prediction mechanisms contribute to harmonic expectation in typical Western music compositions. We therefore avoid the specially composed chord sequences used by previous studies, and instead derive our stimuli from the Billboard corpus, a large dataset of chord sequences representing compositions randomly sampled from the Billboard 'Hot 100' charts between 1958 and 1991 (Burgoyne, 2011). These compositions are by definition 'popular' music, and hence should be broadly representative of the kind of music commonly heard by Western listeners on a day-to-day basis.

We use a computational modelling approach to simulate the predictive processing of these harmonic sequences. We begin by developing a flexible model of harmony prediction that is capable of representing both low-level and high-level perspectives on harmony. This model is feature-based, able to leverage statistical regularities in both low-level psychoacoustic features (e.g. harmonicity, spectral similarity) and high-level cognitive features (e.g. chord root intervals, pitch-class sets). Furthermore, the model is capable of learning statistical regularities at various Markov orders, where the Markov order determines how many preceding feature observations are used to condition the prediction of the next chord. By optimising and interrogating this model, we can examine how these different statistical mechanisms might contribute to harmony prediction in human listeners.

In our first set of analyses, we apply the model to a dataset of chord progressions from the Billboard popular music corpus, and investigate what strategy an ideal listener might use for making predictions. This gives us some insight into the kind of statistical structure underlying Western popular music. It also allows us to formulate hypotheses about human cognition: in particular, if we suppose that most humans are proficient music listeners, then we might hypothesise that harmony prediction in human listeners should resemble harmony prediction in the ideal-listener model.

In our second set of analyses, we test the ideal-listener model against data from human listeners. We play the same chord progressions to 50 Western non-musicians, and ask these participants to report the surprisingness of particular chords in these sequences. We then compare these surprisal ratings to the surprisal values delivered by our ideal-listener model, investigating whether this model provides a good account of participant behaviour, and using the model to examine what kinds of statistical structure drive the variation in surprisal ratings.

These analyses are accompanied by an open-source R package, *hvr*, which implements the computational models evaluated here.[1] This should constitute a useful resource to future researchers interested in harmony prediction.

[1] This package is available at `https://github.com/pmcharrison/hvr`.

4.2 Perceptual features

Listeners have access to a rich variety of perceptual features that might contribute useful information for harmony prediction. In this section we review these features, discuss how they may be simulated computationally, and survey their statistical properties (Figure 4.1, Table 4.1).

These features span a broad variety of psychological processes. For convenience, we group these features into two broad categories: 'low-level' features and 'high-level' features. The low-level features reflect relatively early stages of auditory processing, whereas the high-level features reflect relatively late stages of auditory processing. These two feature categories are also distinguished by their feature values: the low-level features take continuous values on numeric scales, whereas the high-level features take discrete values on categorical scales. This transition between continuous and discrete domains reflects the cognitive phenomenon of categorical perception, whereby perceptual objects in continuous psychophysical space are assigned to discrete categories (Harnard, 2003).

These features are grounded in the representation network defined in Chapter 2. All are ultimately derived from the 'pitch-class chord' representation, which forms the basic representation scheme for the musical corpus, as well as the event space over which the simulated listeners generate their predictions. Several of these features are in fact identical to representations from Chapter 2, such as the 'pitch-class set' feature and the 'pitch-class set relative to bass' feature. Various new features are also included that extend past the representation network from Chapter 2, with some features including additional psychoacoustic modelling (e.g. the 'interference between partials' feature) and others simulating higher-level cognitive processes such as root finding, relative pitch perception, and similarity judgment.

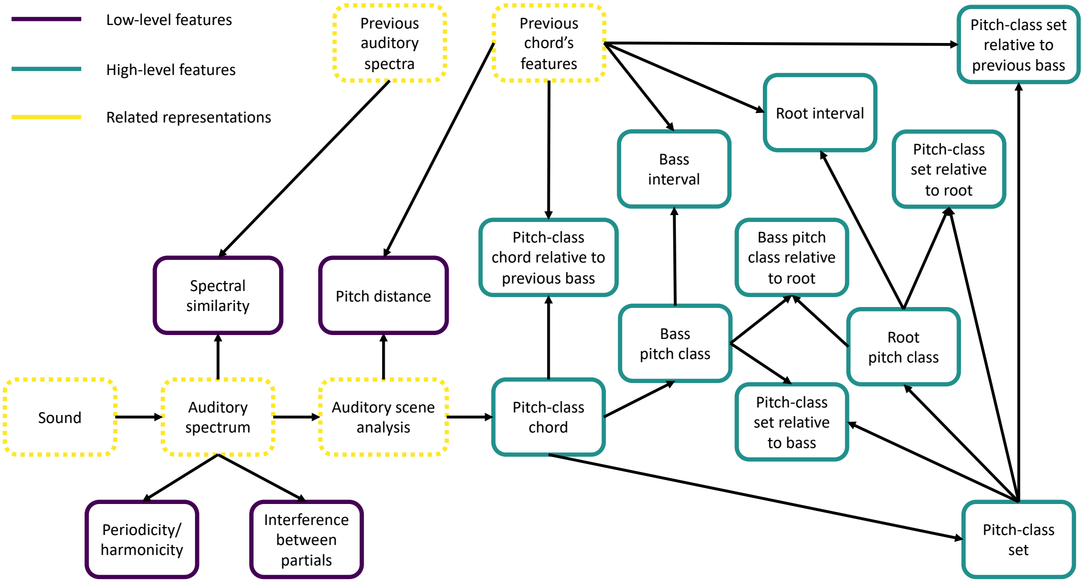
Low-level features
High-level features
Related representations
Previous auditory spectra
Previous chord's features
Pitch-class set relative to previous bass
Root interval
Bass interval
Pitch-class set relative to root
Pitch-class chord relative to previous bass
Bass pitch class relative to root
Spectral similarity
Pitch distance
Bass pitch class
Root pitch class
Sound
Auditory spectrum
Auditory scene analysis
Pitch-class chord
Pitch-class set relative to bass
Periodicity/ harmonicity
Interference between partials
Pitch-class set

Figure 4.1: Schematic network illustrating the psychological relationships between the 15 features used in the present study. An arrow indicates that a particular feature is psychologically derived from another feature.

Table 4.1: The 15 harmony features used in the present study.

Name				
Textual	Computational	Domain	Pitch representation	Alphabet size
Low-level				
Interference between partials	`hutch_78`	Continuous	-	-
Periodicity/harmonicity	`har_18_harmonicity`	Continuous	-	-
Pitch distance	`pi_dist`	Continuous	-	-
Spectral similarity	`spec_sim_3`	Continuous	-	-
High-level				
Bass pitch class	`bass_pc`	Categorical	Absolute	12
Root pitch class	`root_pc`	Categorical	Absolute	12
Bass interval	`bass_int`	Categorical	Relative	12
Root interval	`root_int`	Categorical	Relative	12
Pitch-class set	`pc_set`	Categorical	Absolute	4,095
Pitch-class chord	`pc_chord`	Categorical	Absolute	24,576
Pitch-class set relative to bass	`pc_set_rel_bass`	Categorical	Relative	2,048
Bass pitch class relative to root	`bass_pc_rel_root`	Categorical	Relative	12
Pitch-class set relative to root	`pc_set_rel_root`	Categorical	Relative	457
Pitch-class chord relative to previous bass	`pc_chord_rel_prev_bass`	Categorical	Relative	24,576
Pitch-class set relative to previous bass	`pc_set_rel_prev_bass`	Categorical	Relative	4,095

4.2.1 Low-level features

Low-level features reflect relatively early parts of the auditory processing pathway, and are hence particularly dependent on the acoustic properties of the chord. They precede perceptual categorisation, and hence take continuous values; they may therefore be termed *continuous features*. We review four such features: interference between partials, periodicity/harmonicity, spectral similarity, and pitch distance.

Interference between partials is a key predictor of consonance, the sense in which a chord's notes 'sound well together' (Chapter 3). A chord's acoustic spectrum may be represented as a set of discrete partials, each associated with a frequency and an amplitude; pairs of partials create interference as a function of their proximity, with interference being maximised by close but non-overlapping partials. This interference is thought to comprise beating and mutual masking effects, both of which are perceived as unpleasant, detracting from the chord's consonance. We operationalise interference between partials with Hutchinson & Knopoff's (1978) model, which provided the best account of consonance perception out of the interference models evaluated in Chapter 3.

Periodicity and harmonicity are also key predictors of consonance (Chapter 3; Cousineau et al., 2012; McDermott et al., 2010; Stolzenburg, 2015; Terhardt, 1974). Periodicity means that a sound has a repetitive waveform, whereas harmonicity means that a sound's partials can be approximated as integer multiples of a fundamental frequency. Both periodicity and harmonicity are positively associated with consonance. It turns out that periodicity and harmonicity are essentially equivalent mathematically speaking, and so we refer to both with the joint term 'periodicity/harmonicity'. We operationalise periodicity/harmonicity with the model presented in Chapter 3, which provided the best account of consonance perception out of the periodicity/harmonicity models evaluated in Chapter 3.

The similarity of the acoustic spectra of neighbouring sonorities seems to be a useful feature for explaining various aspects of music perception (Bigand et al., 2014, 1996; Craton et al., 2019; Dean et al., 2019; Leman, 2000a; Milne & Holland, 2016). Previous work has found it useful to preprocess these acoustic spectra in various ways to simulate human auditory processing. Parncutt's (1989) model simulates auditory masking and harmonicity-based pitch perception; Leman's (2000a) model simulates the transmission properties of the auditory periphery, and periodicity-based pitch perception; Milne et al.'s (2011) model incorporates octave invariance, and a smoothing mechanism to account for inaccuracies in pitch perception. Previous work has successfully modelled musical expectations by accumulating the resulting spectra in an exponentially-decaying echoic memory buffer, against which incoming spectra are compared for spectral similarity (Bigand et al., 2014; Leman, 2000a). Here we operationalise spectral similarity with a new model that combines the echoic memory buffer of Leman's (2000a) model with the octave invariance and smoothing mechanism of Milne et al's (2011) model. We have made this model available in the *specdec* R package.[2]

Pitch distance, often termed 'voice-leading distance' by music theorists, describes how far the tones in one chord must move to reach the tones in the next chord. Music theorists have emphasised pitch-distance minimisation as a core principle in Western harmony (Cohn, 2012; Tymoczko, 2011), and psychological studies have shown that pitch distance is a useful predictor of psychological percepts such as similarity and tension (Bigand et al., 1996; Milne & Holland, 2016). The psychological percept of pitch distance presumably derives from auditory scene analysis (Bregman, 1990), whereby the auditory spectra of successive chords are organised into perceptual streams on the basis of pitch similarity. We do not simulate auditory scene analysis in detail, but instead approximate the process using Tymoczko's (2006) minimal voice-leading algorithm, which takes a pair of chords (each defined in terms of pitch classes, not absolute pitches) and finds the best way of organising the chords into melodic lines so as to minimise the sum pitch distance moved by the melodic lines. Importantly, the number of melodic lines is not constrained in advance, and note doubling is permitted. The algorithm then returns the pitch distance achieved by the minimal voice leading. In its original form, the algorithm is defined only for pitch-class sets; we have implemented a variant that instead operates over pitch-class chords. For example, the optimal voice leading between the pitch-class chords $(4, \{0, 4, 7\})$ and $(3, \{0, 3, 6, 8\})$ according to this algorithm is found to be $4 \rightarrow 3$, $7 \rightarrow 6$, $7 \rightarrow 8$, and $0 \rightarrow 0$, which corresponds to a voice-leading distance of $1 + 1 + 1 + 0 = 3$ semitones. The algorithm's implementation is available in the *minVL* R package. Alternative pitch-distance algorithms have been presented by Parncutt (1989) and Parncutt & Strasburger (1994); we have implemented the latter algorithm in our *parn94* R package.[3]

4.2.2 High-level features

High-level features reflect higher stages of auditory cognition. Unlike the continuous low-level features, the high-level features take discrete values, reflecting the cognitive phenomenon of categorical perception; they may therefore be termed *categorical features*. This categorical nature is useful for labelling purposes, and

[2] `https://github.com/pmcharrison/specdec`

[3] `https://github.com/pmcharrison/parn94`

correspondingly many of these features can be transcribed from musical scores or musical performances by experienced musicians. We review 11 such features, derived from a combination of three primary concepts from music theory: pitch classes, bass notes, and root notes. In the process we will briefly recapitulate some definitions from Chapter 2.

Musical chords correspond to collections of tones produced by pitched musical instruments. The pitch of a given tone depends on its fundamental frequency: higher frequencies correspond to higher perceived pitches. Frequencies may be mapped to pitches using the following equation:

$$p = 69 + 12 \log_2 (f/f_{ref}) \tag{4.1}$$

where p is the pitch, expressed as a MIDI note number, f is the frequency (Hz), and f_{ref} is the reference frequency (typically 440 Hz).

The notes on the Western piano keyboard are enumerated by the integers of the MIDI scale; for example, middle C is given the number 60. Here we will assume that all chords are drawn from this integer-based MIDI scale. The pitch distance between two notes on the MIDI scale may be calculated by subtracting their corresponding integers, producing an interval in semitones. By virtue of the logarithmic relationship between frequency and pitch, each pitch interval corresponds to a particular frequency ratio: for example, an interval of 12 semitones corresponds to a 2:1 frequency ratio.

This 2:1 frequency ratio has a privileged function in many musical styles. Psychologically speaking, tones separated by (approximately) 2:1 frequency ratios tend to share some kind of perceptual quality termed *chroma* (Bachem, 1950); this notion of chroma seems particularly relevant for determining the harmonic function of a musical chord. Western music theorists capture this perceptual phenomenon by defining *pitch classes* as sets of pitches that are related to each other by octaves. The numeric representation of a pitch class is typically defined as the remainder after dividing the MIDI note number by 12; for example, middle C has a MIDI note number of 60 and a pitch class of 0. A pitch-class interval may then be computed by subtracting two pitch classes modulo 12.

Some of the chord's pitch classes are thought to have particular psychological prominence for harmony perception. One such case is the *bass pitch class*, defined as the pitch class of the lowest tone in the chord. A second such case is the *root pitch class*, defined more subjectively as the pitch class that best summarises the tonal content of the chord. Music psychologists have explained chord roots as virtual pitches arising from pattern-matching processes of periodicity/harmonicity detection (Parncutt, 1988; Parncutt et al., 2019; Terhardt, 1974, 1982). Here we simulate chord root perception using the psychoacoustic algorithm of Parncutt (1988), which identifies candidate chord roots by applying a template-matching algorithm to the chord's pitch-class set. The bass and root pitch classes constitute our first two high-level features.

The bass pitch class and root pitch class features both imply *absolute pitch* perception: they assume that the listener represents pitch classes in a context-independent manner. However, it is well-established that most Western listeners cannot easily access absolute-pitch representations (though see Eitan, Ben-Haim, & Margulis, 2017; Frieler et al., 2013; Levitin, 1994; Miyazaki, 1988 for exceptions). Instead, listeners typically perceive pitch classes relative to the local musical context (e.g. Schneider, 2018). Correspondingly, we define the *bass interval* feature as the pitch-class interval from the previous chord's bass pitch class to the current chord's bass pitch class, and likewise define the *root interval* feature as the interval from the previous root to the current root.

The remaining pitch classes in the chord also contribute to its perceptual identity. The *pitch-class set* feature specifies the set of pitch classes in the chord; for example, a C major triad would be represented by {0, 4, 7}. The *pitch-class chord* feature specifies both the bass pitch class and the pitch-class set; for example, a C major triad in first inversion would be specified as (4, {0, 4, 7}), where 4 is the bass pitch class and {0, 4, 7} is the pitch-class set (see Chapter 2).

Applying the principle of relative pitch perception, the bass and root pitch classes can be used as references for expressing other pitch classes. As before, this involves subtracting the reference pitch class from the original pitch classes using modulo 12 arithmetic. First, we define a feature called *pitch-class set relative to bass*, which expresses the pitch-class set relative to the bass note.[4] Second, we define two features that relate pitch classes to the root pitch class: *bass pitch class relative to root*, and *pitch-class set relative to root*. Lastly, we define two features that relate the current chord to the bass pitch class of the previous chord: *pitch-class chord relative to previous bass*, and *pitch-class set relative to previous bass*. These relative-pitch features are each transposition-invariant, meaning that they are unaffected when the chord sequence is transposed (shifted) by some pitch-class interval.

Each categorical feature has a finite alphabet that enumerates the different categories that the feature can take. The sizes of these alphabets are listed in Table 4.1. Features corresponding to single pitch classes (bass pitch class, root pitch class, bass interval, root interval, bass pitch class relative to root) inherit

[4] This is equivalent to the pitch-class chord type representation in Chapter 2. We write 'pitch-class set relative to bass' here to emphasise the representation's relationship with the pitch-class set relative to root and pitch-class set relative to previous bass representations.

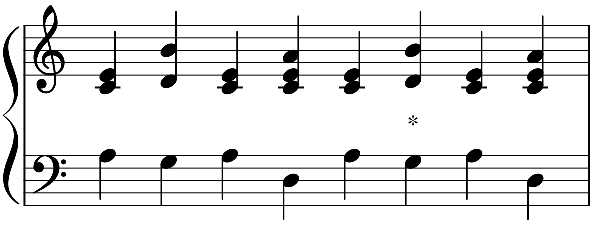

Figure 4.2: Example eight-chord sequence corresponding to an extract from the song *Super Freak* by Rick James (1981). Such sequences were used in both the ideal-listener analysis and the behavioural study, in which computational models and human listeners generated predictions for the target chord marked by an asterisk. The chords are voiced using the heuristic method described in Section 4.7.3. Feature values are provided in Table 4.2.

the standard pitch-class alphabet, namely the integers from 0 to 11. For pitch-class sets and pitch-class chords, we use the alphabets defined in Chapter 2 and implemented in the *hrep* package for the programming language R.[5] These alphabets are also inherited by the features that express pitch-class sets and pitch-class chords relative to the previous bass pitch class. For the pitch-class set relative to bass representation, we use the pitch-class chord type alphabet defined in Chapter 2 and implemented in the *hrep* package. For pitch-class set relative to root, we initialise the alphabet as an empty list, then iterate through the alphabet of pitch-class chords, expressing each chord as a pitch-class set relative to the root pitch class, and appending the result to the alphabet if it hasn't already occurred.

4.2.3 Feature distributions in popular music

Table 4.2 displays feature values for each chord of the chord sequence notated in Figure 4.2. Such tables have been termed *solution arrays* (Conklin & Witten, 1995; Ebcioğlu, 1988). Inspecting this solution array we can immediately see some characteristic structure: the 'bass pitch class relative to root' feature is always zero, all chords have a periodicity/harmonicity greater than 0.8, there is a repeating pattern of root pitch classes (9, 7, 9, 2; 9, 7, 9, 2), and so on. Such structures could prove useful for generating predictions about upcoming chords. We now examine these feature distributions more systematically through an analysis of the Billboard corpus, a large dataset of chord sequences from the Billboard 'Hot 100' music charts (Burgoyne, 2011), preprocessed so that all symbols correspond to chord changes.

[5] `https://github.com/pmcharrison/hrep`

Table 4.2: Feature values for the chord sequence notated in Figure 4.2.

	Chord							
Feature	1	2	3	4	5	6	7	8
Low-level								
Interference	0.18	0.14	0.18	0.22	0.18	0.14	0.18	0.22
Periodicity/harmonicity	0.93	0.93	0.93	0.85	0.93	0.93	0.93	0.85
Pitch distance	NA	5	5	10	10	5	5	10
Spectral similarity	NA	0.34	0.77	0.95	0.94	0.55	0.88	0.97
High-level								
Bass PC	9	7	9	2	9	7	9	2
Root PC	9	7	9	2	9	7	9	2
Bass interval	NA	10	2	5	7	10	2	5
Root interval	NA	10	2	5	7	10	2	5
PC set	{0, 4, 9}	{2, 7, 11}	{0, 4, 9}	{0, 2, 4, 9}	{0, 4, 9}	{2, 7, 11}	{0, 4, 9}	{0, 2, 4, 9}
PC chord	(9, {0, 4})	(7, {2, 11})	(9, {0, 4})	(2, {0, 4, 9})	(9, {0, 4})	(7, {2, 11})	(9, {0, 4})	(2, {0, 4, 9})
PC set rel. bass	{0, 3, 7}	{0, 4, 7}	{0, 3, 7}	{0, 2, 7, 10}	{0, 3, 7}	{0, 4, 7}	{0, 3, 7}	{0, 2, 7, 10}
Bass PC rel. root	0	0	0	0	0	0	0	0
PC set rel. root	{0, 3, 7}	{0, 4, 7}	{0, 3, 7}	{0, 2, 7, 10}	{0, 3, 7}	{0, 4, 7}	{0, 3, 7}	{0, 2, 7, 10}
PC chord rel. prev. bass	NA	(10, {2, 5})	(2, {5, 9})	(5, {0, 3, 7})	(7, {2, 10})	(10, {2, 5})	(2, {5, 9})	(5, {0, 3, 7})
PC set rel. prev. bass	NA	{2, 5, 10}	{2, 5, 9}	{0, 3, 5, 7}	{2, 7, 10}	{2, 5, 10}	{2, 5, 9}	{0, 3, 5, 7}

Note. ‘PC’ is an abbreviation of ‘pitch class’ and ‘rel.’ is an abbreviation of ‘relative’. All intervals are expressed modulo 12.

Low-level feature distributions

The low-level features are all continuous, and hence their distributions may be visualised using kernel density estimators, which provide nonparametric estimates of the feature's sampling distributions. To understand how these features may be used for predicting chord progressions, we compare feature distributions in the corpus to baseline feature distributions computed by randomly sampling from the full alphabet of possible chords. Divergence between corpus and baseline distributions indicates that a feature should be useful for generating predictions.

Figure 4.3 plots the resulting feature distributions. The largest divergences between corpus and baseline distributions occur for the two consonance features: interference between partials, and periodicity/harmonicity. This suggests that these two features should be particularly useful for predicting chord progressions. Examining these divergences, it is clear that the corpus exhibits relatively high periodicity/harmonicity and relatively low interference between partials; this is consistent with the idea that composers tend to promote consonance in their compositions.

The two remaining features, spectral similarity and pitch distance, also exhibit divergences between corpus and baseline distributions, indicating some potential predictive power. The corpus exhibits relatively high spectral similarity, indicating that composers tend to prefer chords with high spectral similarity to their recent context. Conversely, the corpus exhibits relatively low pitch distance, indicating that composers tend to prefer chords that possess proximal pitch content to their predecessors.

Correlations between these features are listed in Table 4.3. The two consonance features – interference between partials and periodicity/harmonicity – correlate strongly and negatively, indicating that both features provide similar yet not identical information. Spectral similarity and voice-leading distance are essentially uncorrelated, indicating that they provide mostly non-overlapping information.

Interestingly, spectral similarity proves to be moderately negatively correlated with consonance. It turns out that most chords with high spectral similarity to their predecessors tend to be created by adding additional notes to the preceding chords: adding notes in this way tends to increase dissonance.

Similarly, pitch distance also proves to be moderately negatively correlated with consonance. This effect seems also driven by the number of notes in the chord: chords with high pitch distance from their predecessors tend to contain many notes, and chords containing many notes tend to be dissonant.

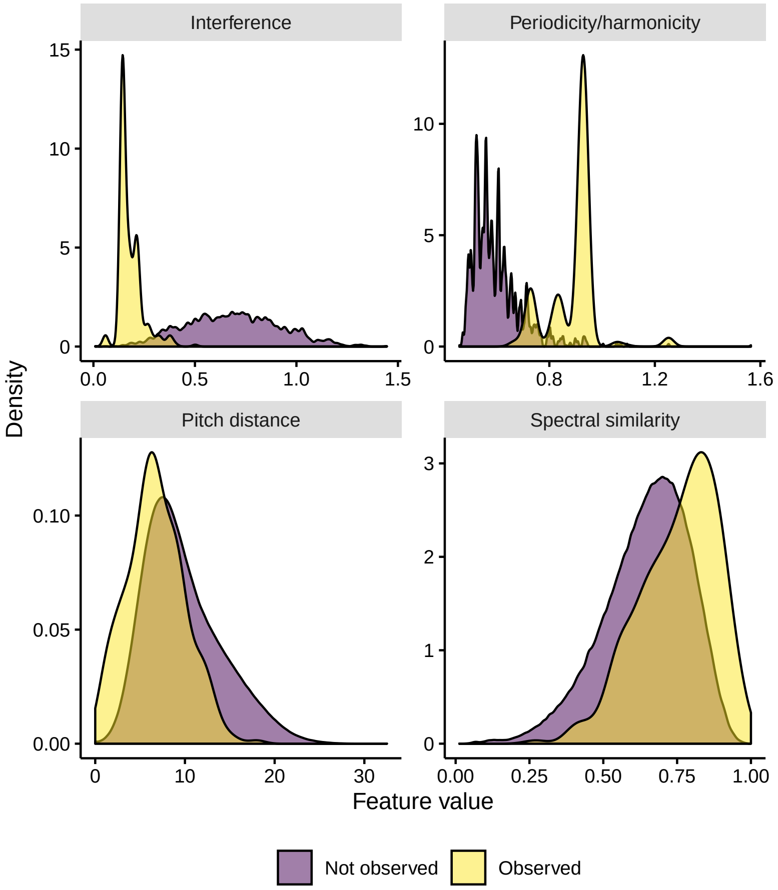

Figure 4.3: Distributions for continuous features, comparing observed chords with unobserved chords at each position in the popular music corpus. These distributions are plotted as kernel density estimates with Gaussian kernels and bandwidths estimated using Silverman's (1986) 'rule of thumb'.

Table 4.3: Pearson correlation matrix for the low-level features.

	Interference	Spectral similarity	Pitch distance
Periodicity/harmonicity	-0.84	-0.39	-0.31
Interference		0.36	0.36
Spectral similarity			0.04

Note. N = 7,372,800; see Section 4.7 for details.

High-level feature distributions

The categorical structure of the high-level features makes them well-suited to expressing sequential dependencies. Sequential dependencies can be visualised by plotting transition matrices, which illustrate how successive observations depend on preceding observations.

Here we focus on structural regularities in chord roots. Traditional harmonic analysis emphasises chord roots and the transitions between these chord roots (Hedges & Rohrmeier, 2011; Meeus, 2000; Rameau, 1722; Tymoczko, 2003); we would therefore expect to find useful sequential structure in these viewpoints that should help predict chord progressions.

Figure 4.4A plots transition matrices for the root pitch class feature. The 0th-order probability distribution captures the relative frequency of different roots in the corpus. Some roots are relatively common, others less common, but generally speaking most roots receive moderate exposure. The 0th-order distribution is therefore not particularly informative. The 1st-order distribution, meanwhile, captures the relative frequencies of different roots conditioned upon the immediately preceding root, normalised by the corresponding 0th-order distributions. There are two clear diagonal lines in the 1st-order distribution, indicating that root progressions of ascending perfect fourths (5 semitones) and fifths (7 semitones) are particularly common.

The diagonal symmetry of this transition matrix implies that root progressions depend less on the absolute pitch classes and more on the intervals between the pitch classes. This motivates inspection of the root interval feature (Figure 4.4B). The 0th-order distribution is now much more informative: it indicates high probabilities for descending and ascending perfect fifths, moderate probabilities for descending and ascending major seconds, and low probabilities for the remaining progressions. These patterns are broadly similar to those reported in de Clercq & Temperley's (2011) corpus analysis of rock harmony. Further structure can be identified in the 1st-order transition matrix, such as the high probability of an ascending major second after a descending major third, and the high probability for successive pairs of root intervals to sum to an octave. The analytic power of the root interval feature presumably reflects how Western listeners mostly possess relative rather than absolute pitch representations, incentivising composers to structure music around relative-pitch features, such as root intervals, rather than absolute-pitch features, such as root pitch classes (e.g. Schneider, 2018). Correspondingly, we might hypothesise that the root interval feature will prove to be more useful than the root pitch class feature for generating harmony predictions. Similarly, we might also expect the other relative-pitch features to prove particularly useful compared to their absolute-pitch equivalents.

Figure 4.5 plots analogous transition matrices for the bass pitch class and bass interval features. These matrices are mostly similar to their root-based partners (Figure 4.4), reflecting the fact that in popular music the root pitch class is commonly situated in the bass line. One salient difference however is that unison intervals are more common in the bass than in the root: put another way, if a chord change occurs and the bass note stays the same, then the chord root typically changes.

It is impractical to visualise such transition matrices for high-level features with alphabet sizes much larger than 12. However, the sense in which different features convey similar information can be quantified by correlating the predictions generated by these features. Here we train a variable-order Markov model on each feature, use these models to generate predictions for chord sequences in the popular music corpus, and then correlate the log probabilities of the observed chords according to the different features (see Section 4.7.2 for details). We then apply hierarchical clustering to the resulting correlation matrix, defining the distance measure as one minus the correlation coefficient.

Figure 4.6 shows that the hierarchical clustering clearly recovers the theoretical relationships between the different features. Root interval and root pitch class are clustered together, as are bass interval and bass pitch class, reflecting how root intervals and bass intervals are respectively derived from root pitch classes and bass pitch classes. The next cluster combines pitch-class set relative to bass and pitch-class set relative to root, reflecting how both features express the chord's pitch-class content in a transposition-invariant manner. The next cluster combines two features that express the chord's pitch-class content relative to the bass note of the previous chord; the final cluster combines two features that express the chord's pitch-class content using absolute pitch-class representations. This clustering analysis could be useful for developing a simplified feature set for a harmony prediction model: features within the same cluster deliver similar information, and hence each cluster could potentially be replaced with a single representative feature without losing much information.

4.3 Model

This section introduces a new model for predicting chord progressions on the basis of these different perceptual features, termed *viewpoint regression*. This technique advances the multiple viewpoint approach (Conklin & Witten, 1995; Hedges & Wiggins, 2016a, 2016b; Pearce, 2005; Whorley & Conklin, 2016; Whorley, Wiggins, Rhodes, & Pearce, 2013) by adding support for continuous

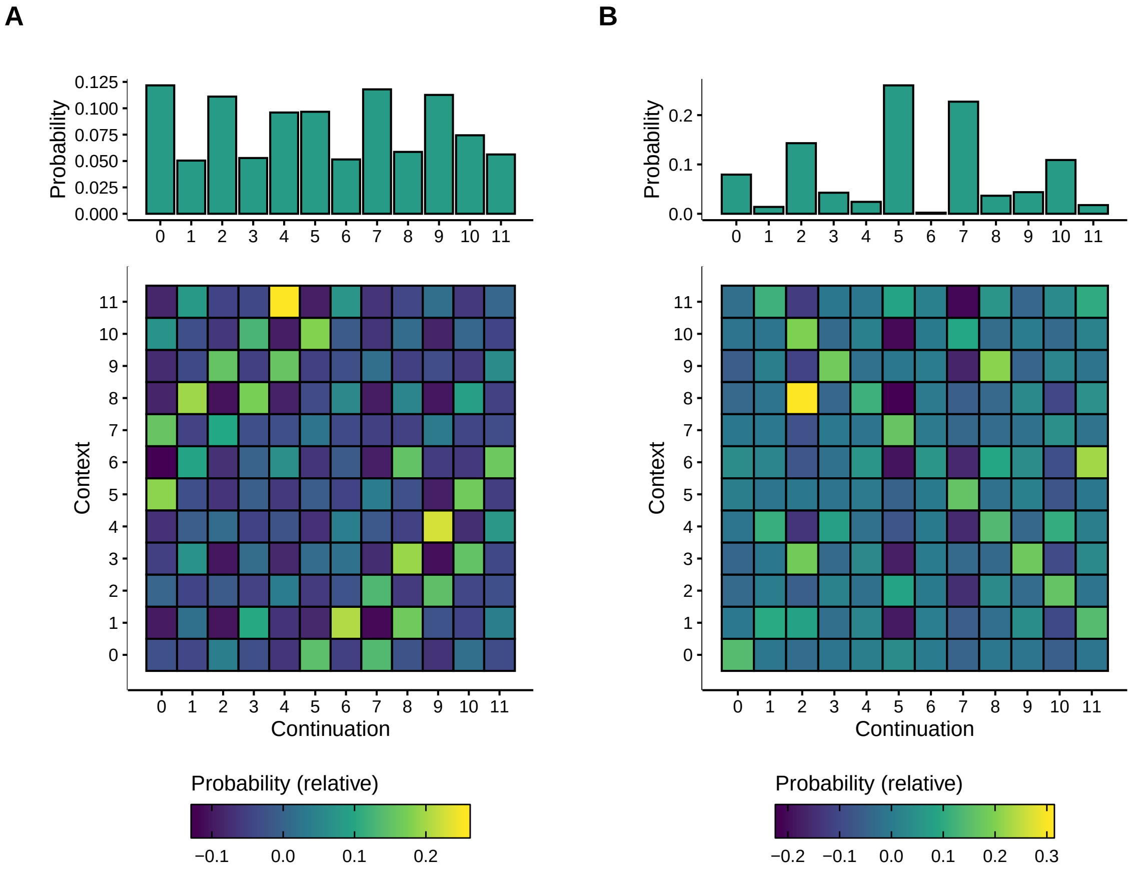

Figure 4.4: 0th- and 1st-order probability distributions for (**A**) root pitch class and (**B**) root interval as derived from the Billboard popular music corpus (Burgoyne, 2011). 1st-order probabilities are expressed relative to the corresponding 0th-order probabilities.

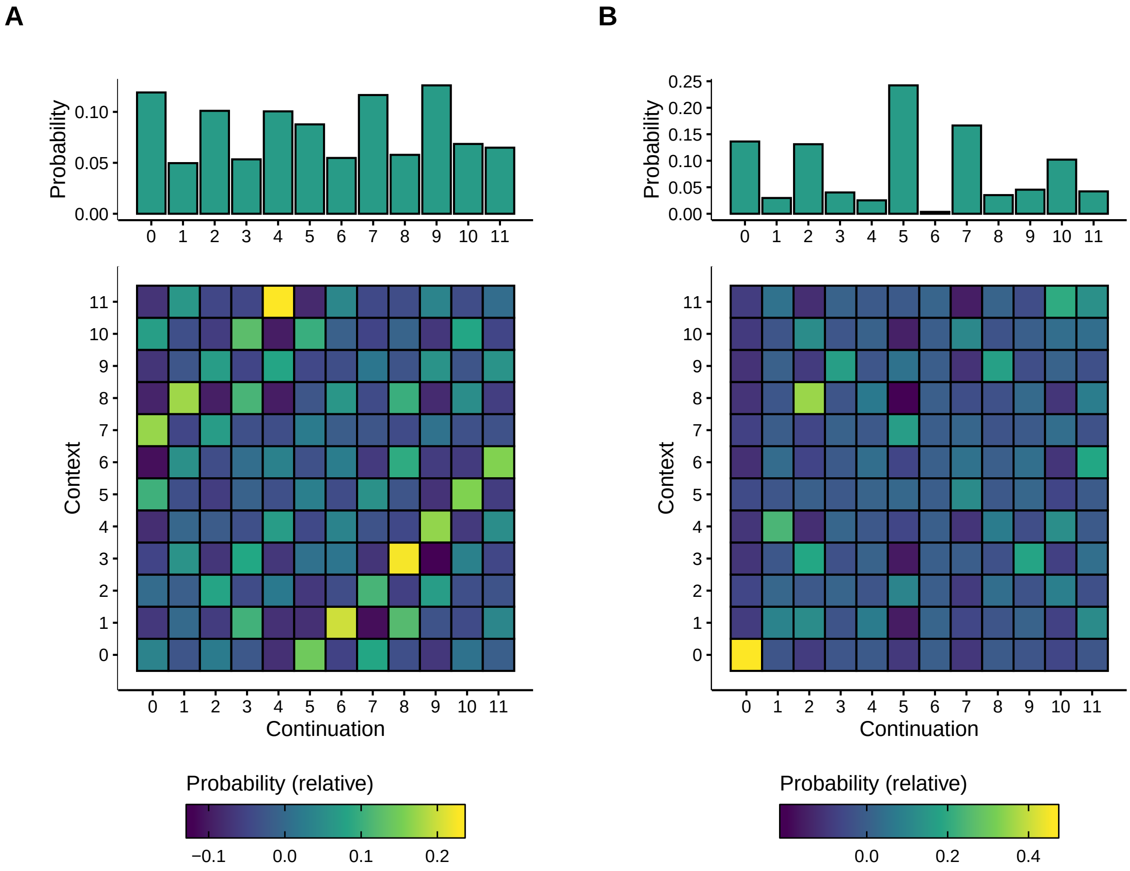

Figure 4.5: 0th- and 1st-order probability distributions for (**A**) bass pitch class and (**B**) bass interval as derived from the Billboard popular music corpus (Burgoyne, 2011). 1st-order probabilities are expressed relative to the corresponding 0th-order probabilities.

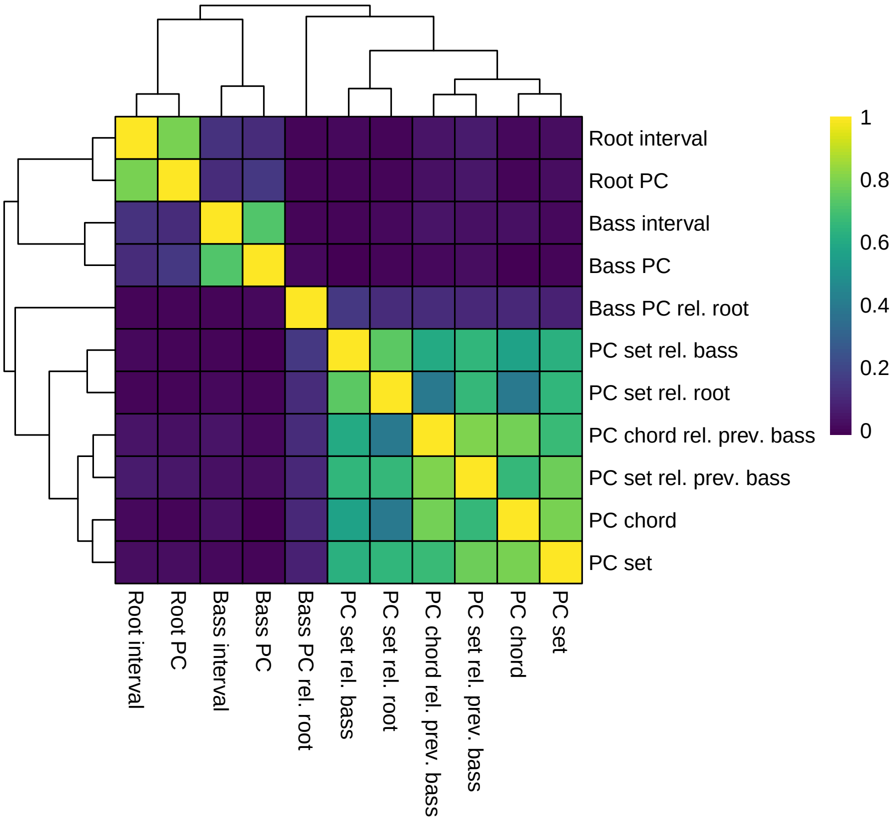

Figure 4.6: Correlation matrix for log probabilities of observed chords as computed by variable-order Markov models trained on different high-level features (see Section 4.7 for details). Features are hierarchically clustered using a distance measure of one minus the correlation coefficient.

features and providing interpretable feature weights. We will briefly summarise the approach here, but the reader is directed to Section 4.7.2 for a more detailed exposition.

The proposed technique embodies the following principles:

1. Chords should be more likely if they represent likely continuations of categorical feature sequences;

2. Chords should be more likely if they represent likely combinations of continuous feature values.

These concerns are quantified in a *metafeature vector*. Each categorical feature contributes two terms to the metafeature vector, each corresponding to a type of *expectedness*: *short-term expectedness* and *long-term expectedness*. For a given feature, the two expectedness terms correspond to log probabilities as estimated by variable-order Markov models trained on that feature. Short-term expectedness is estimated by training the model on feature observations from the portion of the sequence seen so far; long-term expectedness is estimated by training the model on feature observations from a large musical corpus representing the listener's prior musical exposure. In both cases, the Markov model has no access to the other features in the model, and therefore its predictions solely reflect the currently selected feature. The Markov model outputs probability distributions over feature values, which are converted to probability distributions over chords by uniformly distributing the probability assigned to each feature value over all chords that map to that feature value. Each continuous feature meanwhile contributes n_p terms to the metafeature vector, corresponding to an orthogonal polynomial representation with degree n_p. Here $n_p = 4$, meaning that continuous feature effects are estimated as quartic polynomials.[6]

The probability of a given chord is modelled as being proportional to an exponentiated weighted sum of the elements of the metafeature vector (f_i), with regression weights (w_i) that are optimised during model training:

$$P(x) \propto \exp\left(\sum_i w_i f_i\right). \tag{4.2}$$

Since chord repetitions are excluded from the present corpus, we fix $P(x)$ to zero for chord repetitions, and normalise the resulting probability distribution over the remaining 24,575 pitch-class chords.

The functional form of Equation 4.2 appears in the literature under various names, including the log-linear model, the multinomial logit model, and the

[6] Higher-order polynomials achieve more flexibility (i.e. less bias) at the cost of reduced stability (i.e. more variance). Quartic polynomials seemed to be a reasonable compromise between these two goals.

conditional logit model (McFadden, 1974). The model is closely related to the geometric multiple-viewpoint system of Pearce et al. (2005); Pearce's model can be reproduced by removing all continuous features and fixing the metafeature weights to be a function of the Markov model's predictive entropy.

4.4 Modelling an ideal listener

We now apply this model to an ideal-listener analysis. In an ideal-listener analysis, the goal is to understand what the optimal strategy might be for performing a particular cognitive task. This can be useful for two reasons: a) providing insight into the structure of the task domain, in this case harmony, and b) generating hypotheses about how humans might perform the task.

For complex cognitive tasks such as harmony prediction, an ideal-listener analysis will generally take the form of a computational simulation. Here the goal is to simulate how a listener might generate predictions for upcoming chords in a chord sequence, potentially taking advantage of various perceptual features such as spectral similarity and chord roots, and potentially taking advantage of musical knowledge learned from prior exposure to a musical style.

If the musical style were defined by a known probabilistic model, then it might be possible to derive a provably optimal ideal-listener model. However, musical styles are not typically defined by explicit probabilistic models, and hence it is difficult to imagine recovering the 'true' ideal-listener model. However, we can approximate this model by searching a sufficiently broad family of models – in this case, viewpoint regression models – for the model that achieves the best performance, with the understanding that the best-performing model should be the closest approximation to this 'true' ideal-listener model.

Here we formalise harmony prediction through the following paradigm. First, we simulate enculturation with Western popular music by exposing the ideal-listener model to harmonic transcriptions of 439 songs from the Billboard popular music corpus (Burgoyne, 2011) with all rhythmic and metrical content stripped. Second, we play the model 300 unfamiliar chord sequences, extracted from the 300 remaining songs in the Billboard corpus, and evaluate the model's ability to predict chords in these chord sequences (Figure 4.2). The current analyses are computer simulations, and hence we could theoretically evaluate the model on the full 300 songs; however, since we later administer the same paradigm to human listeners, we need to keep the scale of the task practical. Correspondingly, we limit the 300 novel chord sequences to comprise eight chords each, corresponding to a randomly selected extract of the composition, and instruct the model to predict the sixth chord in each sequence, termed the target chord. See Section 4.7 for a more detailed description of the methods.

Table 4.4: Cross entropy as a function of Markov order bound.

	Markov order bound					
	0	1	2	3	4	5
Cross entropy (bits)	3.63	**3.48**	3.50	3.51	3.49	3.49

Note. Lower cross entropy indicates better predictive performance. The best-performing model configuration is given in bold.

It is important to understand that the restricted nature of these stimuli will limit the kinds of harmonic syntax that our analysis can capture. Since the predictions are generated with just five chords of context, the listener will not have access to any high-level structure within the composition. Instead, we expect our stimuli to tell us about *local* harmonic syntax, the sense in which the few chords immediately preceding the target chord determine the predictions that are made for that chord.

The viewpoint regression model can learn statistical regularities for a variety of Markov orders. A Markov order bound of n means that the model takes into account up to n previous feature observations when calculating the expectedness of a given categorical feature. For example, an order bound of two means that the expectedness of a given root interval is conditioned on the previous two root intervals. Because the root interval feature is itself defined with reference to the previous chord, a root interval feature with an order bound of two will take into account the three previous chords. Higher order bounds are theoretically able to capture more complex syntactic structure.

We begin the ideal-listener analysis by examining the necessity of higher Markov orders for generating successful predictions. If high Markov orders were to prove necessary, this would suggest that popular music possesses relatively complex harmonic syntax, and that human listeners would require relatively complex statistical abilities to understand this syntax.

We evaluate the model's predictive performance using the cross entropy error metric. This metric is standard in the machine-learning literature, and corresponds to the mean negative log probability of each observed chord according to the model. Better performing models assign higher probability to observed chords, and correspondingly receive lower cross entropy. Here all logarithms are computed to base two, meaning that cross entropy is expressed in units of bits.

Table 4.4 lists cross entropies for the six Markov order bounds. Increasing the order bound from zero to one provides a 4.23% improvement in predictive performance; interestingly, however, further increasing the order bound does not improve performance, and if anything diminishes it. This implies that the

Table 4.5: Cross entropy and Akaike Information Criteria (AIC) for different feature combinations.

	Features			
	None	Low-level only	High-level only	All
Cross entropy (bits)	14.58	6.37	3.69	**3.49**
AIC	6065.72	2682.40	1578.38	**1529.14**

Note. Lower cross entropy and lower AIC indicates better predictive performance. The best-performing model configuration is given in bold.

local harmonic syntax of this musical style can be mostly characterised by simple feature statistics, specifically the relative frequencies of different categorical features (e.g. 'major triads are more common than seventh chords') and the relative frequencies of different feature transitions (e.g. 'seventh chords are likely to be followed by major triads').

Next we examine the types of features available to the model. While the full model uses both low-level and high-level features, one can contemplate alternative models that only have access to low-level features (e.g. consonance, spectral similarity) or alternatively are restricted to high-level features (e.g. root interval, pitch-class set). By comparing the predictive performances of these models, we can examine the sufficiency of different feature sets for generating effective predictions.

Table 4.5 compares these three models – the full model, the low-level model, and the high-level model – to a baseline model containing no features. The low-level model performs substantially better than the baseline model, halving the cross entropy from 14.58 to 6.37, and the high-level model performs even better, reducing the cross entropy to 3.69. The best performance, however, is achieved by the combined model, with a cross entropy of 3.49. Examining the Akaike Information Criterion, we see that added complexity of the combined model is justified by the performance improvement compared to the high-level model. To summarise: neither low-level or high-level features are sufficient by themselves to achieve optimal performance, but the high-level features can achieve almost optimal performance.

Next we investigate how individual features contribute to the full model. In particular, we are interested in seeing whether the model is dominated by a small number of features, or whether it simultaneously depends on many features. We approach this question by computing a permutation-based feature-importance metric, which quantifies feature importance by taking the evaluation dataset, randomly permuting feature values, and assessing how much the model's predictive performance drops (Breiman, 2001; Fisher et al., 2018).[7]

Feature importances for the full and low-level models are plotted in Figure 4.7. Examining feature importances for the full model (Figure 4.7A), we see that the two most important features are two low-level features: interference between partials and periodicity/harmonicity, both aspects of consonance. This is initially surprising: we previously saw that the low-level model was substantially outperformed by the high-level model, implying that low-level features are considerably less useful than high-level features. However, this incongruence can be explained by feature redundancy: while the two consonance features can effectively predict which chord types will be more common than others (see Chapter 3), this information can also be approximated with 'brute force' by a high-level viewpoint (such as 'pitch-class set relative to bass') that simply memorises the prevalence of different chord types. This redundancy between the feature sequences allows the high-level model to perform well in the absence of the low-level features, even if the low-level features are preferred by the combined model. Consistent with this view, removing the high-level features causes the model to rely even more on the low-level features (Figure 4.7B).

Beside the two consonance features, we see that many other low-level and high-level features also contribute to the full model (Figure 4.7A). This is consistent with the idea that Western popular harmony carries structure along a variety of different features, many of which can be leveraged by the listener to help predict chord progressions.

The contributions of the low-level features can be visualised as marginal effects, showing how different levels of a given feature are associated with different levels of expectedness while holding the other features constant (Figure 4.8). First, we see that the marginal effects differ between the full model and the low-level model. This presumably reflects how certain information is shared between the high-level features and the low-level features; partialling out the contributions of high-level features therefore significantly changes the shape of the low-level features. Second, it is apparent that the marginal effects are not always straightforward reflections of the univariate distributions plotted in Figure 4.3. For example, Figure 4.3 shows that observed chords tend to exhibit relatively low pitch distances, yet Figure 4.8 indicates that the models associate high pitch distance with greater expectedness. This phenomenon is further ev-

[7] This permutation process produces a dataset where feature values may be inconsistent with each other, for example imagining a chord where the pitch-class set is {0, 4, 7} but the bass pitch class is 2. It may seem counterintuitive to imagine such impossible cases, and if the model is very sensitive to interactions between features then the technique might give misleading results. However, by construction the viewpoint regression technique does not model interactions between features, and so its performance on these permuted datasets should relate meaningfully to its performance on real datasets.

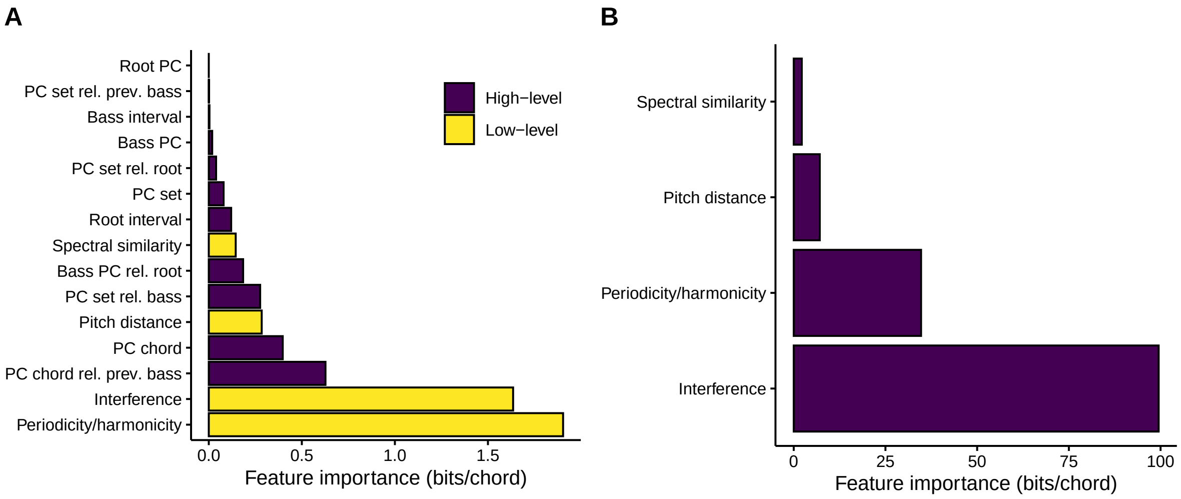

Figure 4.7: Permutation-based feature importances for (**A**) the full ideal-listener model and (**B**) the low-level model, after Breiman (2001; see also Fisher et al., 2018).

idence for shared information between features. It makes it difficult to draw strong conclusions about the causal contribution of a given feature to compositional practice: a given feature may seem to be avoided by composers when considered in isolation, but preferred once other features are accounted for.

The relative contributions of the high-level features can be summarised by their regression weights (Figure 4.9). A regression weight of w_i means that a one-unit increase in feature-based expectedness corresponds to a w_i increase in overall expectedness, where expectedness is measured as log probability. Greater regression weights therefore correspond to greater feature importance.

Examining the regression weights, we see that the model takes advantage of a variety of high-level features. There is a general trend for relative-pitch features (e.g. root interval, bass interval) to be preferred over absolute-pitch features (e.g. root pitch class, bass pitch class): this is consistent with the notion that Western listeners predominantly possess relative pitch perception, which is then reflected by transpositional invariances in music composition (see also Figure 4.4).

Comparing short-term and long-term regression weights, we see that the model pays particular attention to short-term regularities in the feature sequences. This is interesting because, when predicting the sixth chord in a sequence, the model only has five prior chords to learn from, so one might expect this context not to be particularly informative. Were the model to progress fur-

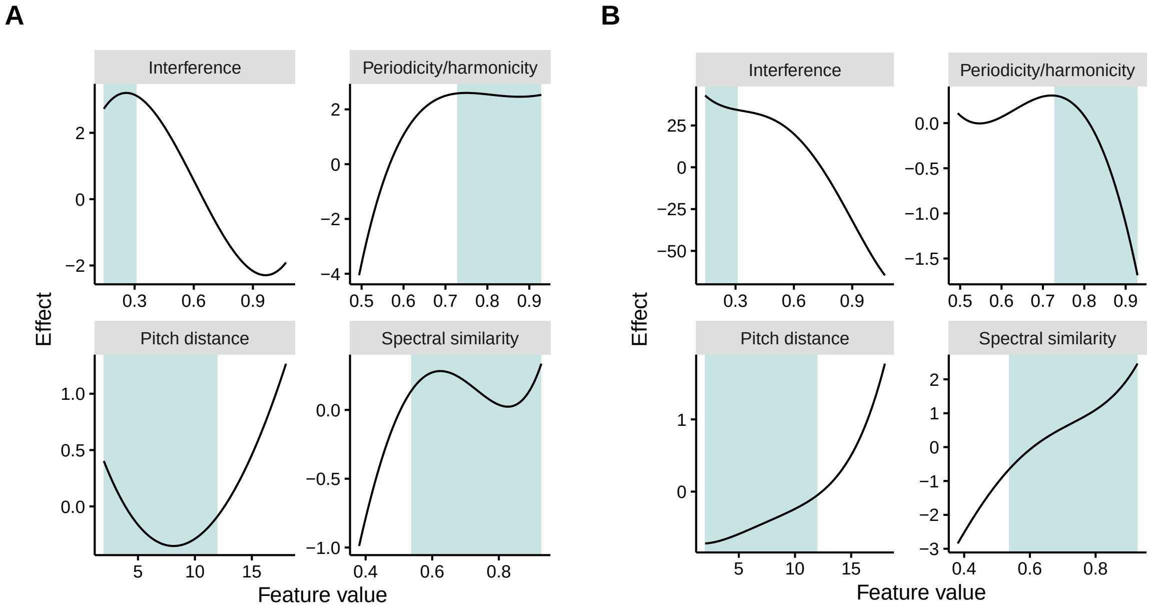

Figure 4.8: Marginal effects of continuous features in the (**A**) full and (**B**) low-level ideal-listener models. The horizontal axes span from the 2.5th to the 97.5th percentiles of all theoretically possible feature values, computed over the full chord alphabet ($N = 24{,}576$) at each of the 300 chord positions being modelled. The shaded regions identify the 2.5th and 97.5th percentiles of observed feature values (i.e. those that were actually observed in the chord sequences).

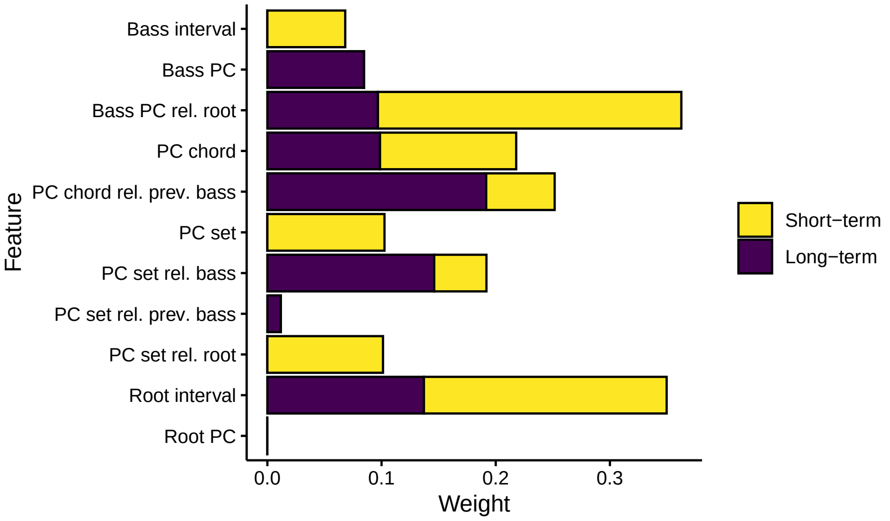

Figure 4.9: Stacked bar plot of categorical feature weights in the full ideal-listener model, split by short-term and long-term features.

ther into the music compositions, we might expect these short-term regularities to become still more informative.

These analyses give some useful insights into the nature of harmony prediction in Western popular music. The headline results may be summarised as follows:

1. Markov orders greater than one do not contribute significantly to the ideal-listener model; in other words, the local harmonic syntax of popular music can be largely explained by first-order Markovian relationships.
2. The ideal model takes advantage of many perceptual features when generating predictions. This is consistent with the idea that harmonic syntax is shaped by many perceptual principles.
3. The ideal model pays particular attention to two low-level perceptual features: interference between partials, and periodicity/harmonicity. However, if these features are not available, their loss can be largely compensated for by certain high-level features, such as 'Pitch-class set relative to bass'.

4.5 Modelling human listeners

Human listeners' responses to music seem to be determined, in large part, by the phenomenon of surprisal, whereby certain musical events sounds more or less expected than others. The musical events within a given musical piece typically vary in surprisal, and this variation is associated with various psychological phenomena including emotional experience (Egermann et al., 2013; Huron, 2006; Meyer, 1957; Sauvé et al., 2018; Sloboda, 1991), perceptual facilitation (Omigie, Pearce, & Stewart, 2012), and phrase-boundary perception (Pearce et al., 2010a), as well as prominent electrophysiological markers (Omigie, Pearce, Williamson, & Stewart, 2013; Pearce et al., 2010b).

Here we investigate human listeners' perception of surprisal in chord sequences from popular music compositions. To generate our behavioural data, we presented 50 participants with the 300 chord sequences from the ideal-listener analysis (see Figure 4.2 for an example), and asked the participants to rate the surprisingness of the sixth chord in each sequence on a scale from one to nine. We normalised these responses within participants to z-scores, and then averaged over participants to produce one mean surprisal rating for the sixth chord in each of the 300 sequences. See Section 4.7.3 for more detail on the experimental methods.

We then use our probabilistic ideal-listener models to analyse the way in which our participants made their surprisal judgments. Given a probabilistic model, the surprisal of a given chord may be operationalised as an information-theoretic quantity termed *information content*, defined as the negative log probability of the observed chord conditioned on its previous context. Our basic strategy for evaluating a given ideal-listener model is to correlate these information content values with the participants' surprisal judgments: high correlations indicate that the model provides a good account of human behaviour. By evaluating different configurations of this model, and comparing it to alternative perceptual models, we can gain insights into the way in which human listeners make their surprisal judgments.

Table 4.6: Benchmarking the ideal-listener model against low-level models of harmonic expectation.

			Pearson r		
Model	Reference	Adj. R^2	Raw (95% CI)	Disattenuated	Spearman ρ
Ideal listener					
Full model	This chapter	.43	.66 [.59, .72]	.73	.70
Low-level models					
Consonance	Chapter 3	.19	−.44 [−.52, −.34]	−.49	−.38
Spectral similarity	Leman (2000)	.11	−.33 [−.43, −.22]	−.37	−.33
Spectral similarity	Parncutt & Strasburger (1994)	.02	.15 [.04, .26]	.17	.11
Pitch distance	Parncutt & Strasburger (1994)	.01	.09 [−.02, .20]	.10	.07
Pitch distance	Tymoczko (2006)	.01	−.12 [−.23, −.01]	−.14	−.14
Pitch-class distance	Tymoczko (2006)	.00	−.07 [−.18, .05]	−.07	−.04
Spectral similarity	Collins et al. (2014)	−.00	−.00 [−.12, .11]	−.00	−.04
Spectral similarity	This chapter	−.00	−.05 [−.17, .06]	−.06	−.03
Combined low-level models					
Linear regression	This chapter	.34	.60 [.52, .66]	.67	.61

Note. Prior to performing this comparison, the half-life parameters for Leman's (2000a) spectral similarity model and the new spectral similarity model were optimised on the behavioural data (see Section 4.7.3 for details). We computed disattenuated correlation coefficients to correct for the measurement error of the listeners' mean surprisal ratings, estimated by taking the mean standard error under the Central Limit Theorem.

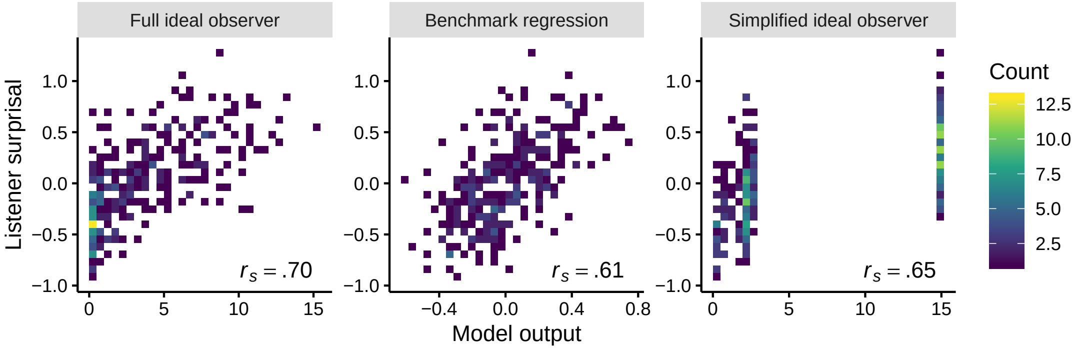

Figure 4.10: Density plots of associations between model outputs and listener surprisals for three models: the full ideal-listener model, the benchmark regression model, and the simplified ideal-listener model. 'Count' refers to the number of stimuli present in the respective region of the plot.

We begin by applying the full ideal-listener model (all features, no order bound) to our perceptual data. The resulting information content values predict participant surprisal ratings with a Spearman correlation of .70 (Table 4.6, Figure 4.10). This indicates that the model provides a fairly successful account of participant responses, and suggests that probabilistic prediction is a good metaphor for how participants make surprisal judgments.

We then compare this ideal-listener model to several other perceptual models, including a consonance model, four spectral similarity models, and three pitch/pitch-class distance models (Table 4.6). We see from Table 4.6 that none of these models compete individually with the ideal-listener model; the best-performing competitor is the consonance model from Chapter 3, with a Spearman correlation of $-.38$. A better approximation to surprisal ratings can be achieved by combining these alternative models using linear regression; the resulting model achieves a Spearman correlation of .61, not so distant from the ideal listener's score of .70. The ideal-listener model and the linear-regression model are similar in that they both generate their predictions using a weighted combination of musical features, but the conceptual advantage of the ideal-listener model is that it explains how the regression weights originate, namely in optimising for the prediction of musical events.

The full ideal-listener model is rather complex: it incorporates many different perceptual features, learns statistical regularities across many Markov orders, and is pretrained on a large dataset of representative music. By testing various simplified versions of the model, we can examine the relative importance of

Table 4.7: Comparing the performance of different simplified versions of the ideal-listener model.

	Model configuration			Listener correlations		
	Features	Order	Pretrain	Pearson	Spearman	Cross entropy
1	**All**	**5**	**Yes**	**.66 [.59, .72]**	**.70**	**3.48 [3.09, 3.86]**
2	All	0	Yes	.65 [.58, .71]	.69	3.62 [3.24, 4.00]
3a	High-level	0	Yes	.66 [.59, .72]	.69	3.89 [3.47, 4.32]
3b	Low-level	0	No	.51 [.43, .59]	.50	6.32 [5.93, 6.71]
4	PC chord	0	Yes	.66 [.60, .72]	.67	4.64 [4.20, 5.08]
5	**PC chord**	**0**	**No**	**.60 [.52, .67]**	**.65**	**6.44 [5.73, 7.15]**

Note. Cross entropy is given in bits; lower cross entropy indicates that the model performs better at predicting chord progressions. 95% confidence intervals are reported in square brackets.

different model attributes for explaining participant behaviour (Table 4.7). First we reduce the Markov order to 0, and find that the model's performance is essentially unchanged (model version 2, Spearman correlation = .69). This indicates that the model can successfully predict listener behaviour without learning transition probabilities, instead just learning that certain feature values (e.g. root progressions of a fifth, or high spectral similarity) are more common than others.

Next we try restricting the feature set to either all low-level features or all high-level features (model versions 3a, 3b). We see that performance is essentially unchanged when the low-level features are dropped (model version 3a, Spearman correlation = .69), indicating that the low-level features are not necessary to explain performance. Nonetheless, the model with only low-level features still performs moderately well (model version 3b, Spearman correlation = .50).

Next we remove all derived features from the model, so that it only has access to the original pitch-class chord representations of the chord sequences. In other words, the model no longer has access to perceptual principles such as relative pitch and chord roots. Again, there is no real diminution in performance (model version 4, Spearman correlation = .67).

Lastly we disable the pretraining component of the model, so that it only learns statistical regularities from the preceding five chords in the chord sequence, with these statistical regularities then being forgotten upon presentation of the next chord sequence. Predictive performance still remains fairly high (model version 5, Spearman correlation = .65).

This pattern of results is quite remarkable. It indicates that many of the complexities of the ideal-listener model can be removed without much affecting the model's ability to predict participants' surprisal judgments. In particular, the simplest model version can be interpreted as a basic repetition-priming model, where the probability of observing a given chord is approximately proportional to the number of times that it has been observed in the portion of the sequence heard so far. The fact that this radically simplified model performs so well indicates that our listeners' predictions were dominated by short-term repetition priming.

It is tempting to take the strong performance of the simplified model as evidence that our participants were suboptimal listeners. However, we should remember that the full model predicted participant performance just as well, if not better (95% confidence interval for the difference in Pearson correlation coefficients: $[-.00, .11]$, Zou, 2007; Diedenhofen & Musch, 2015). The fact that simplifying the model makes little difference suggests that the full and simplified models deliver similar predictions for our stimuli, which we can verify by correlating these sets of predictions (Pearson $r(298) = .78$, 95% CI = [.73, .82]). In other words, the relative surprisal of different chords in the popular music extracts is mostly predicted by simple repetition effects, not by higher-order statistics. However, if we enter both the full and the simplified model into a linear regression model predicting surprisal judgments, the resulting model (adjusted $R^2 = .45$) identifies substantive contributions from both the full model ($\beta = 0.48$, 95% CI = [0.34, 0.62], $p < .001$) and the simple model ($\beta = 0.23$, 95% CI = [0.09, 0.36], $p = .001$). This suggests that, while our listeners' responses can largely be explained by short-term repetition priming, higher-level statistics also contribute.

To illustrate by example, Figure 4.11 displays the full model's output for the first composition in the popular music corpus, *I don't mind* by James Brown (1961). It is clear that repetition plays an important role in this piece. The first two chords (A minor, C major) repeat five times, and then move to a new chord, F major. This progression from C major to F major is completely unsurprising from a syntactic perspective – in fact, it corresponds to the most common root progression in Figure 4.4 – but it is surprising from a local perspective because it disrupts the regular pattern that precedes it. The full model successfully captures this surprisingness, returning an information content of 12.2 bits versus the 0.0 bits of the preceding chord. Importantly, however, a simple repetition-priming model can also capture this effect, simply working on the principle that novel chords are more surprising than recently heard chords. Looking further through Figure 4.11, it is evident that the importance of local repetition continues throughout the composition. The dominance of these local repetition effects is consistent with the conclusion of the wider quantitative analyses: the dynamics of harmonic expectation in popular music are driven less by high-level syntactic structures and more by low-level statistical or psychoacoustic features.

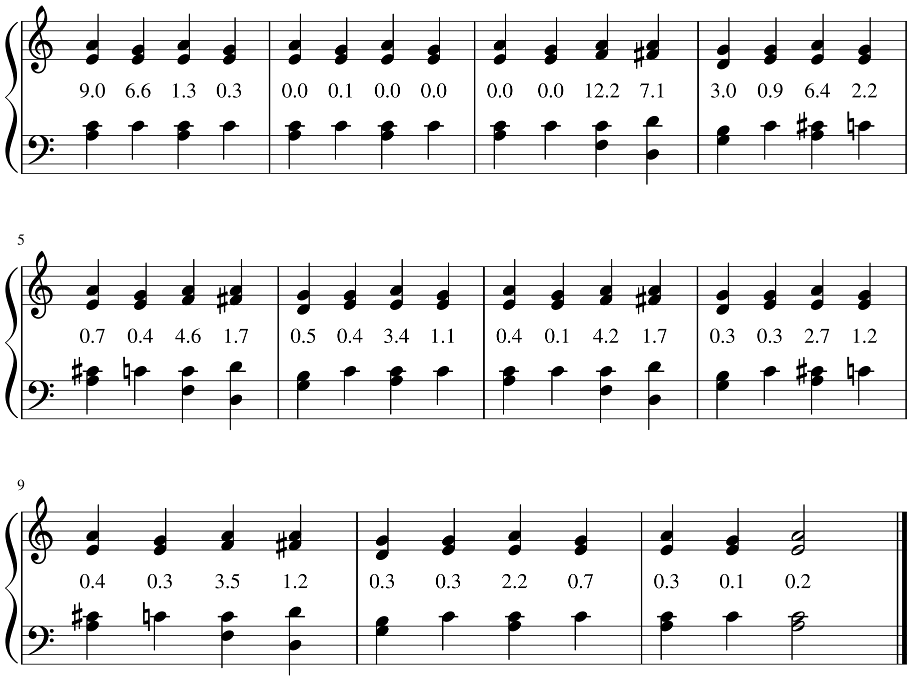

Figure 4.11: Analysis of the first composition in the popular music corpus, *I don't mind* by James Brown (1961), using the full ideal-listener model from Section 4.4. Each number corresponds to a chord's information content, or surprisal, expressed to one decimal place. The regression weights are preserved from the ideal-listener analysis, but the categorical feature models are retrained on the last 700 compositions from the popular music corpus. Durations, pitch heights, and bar lines are arbitrary.

4.6 Discussion

Music listening is thought to be an inherently probabilistic process, with the listener constantly generating predictions for upcoming musical events on the basis of learned statistical knowledge. Here we investigated the probabilistic prediction of musical harmony, developing a new computational model that predicts chord sequences from statistical regularities manifested in a broad collection of perceptual features. We applied our model to a corpus of chord sequences from Western popular music (Burgoyne, 2011), seeking to understand how an ideal listener would predict these chord sequences. We then applied this ideal-listener model to a new behavioural experiment investigating how human listeners perceive surprisal within these chord sequences.

These chord sequences were eight chords in length, sampled from random locations in the original corpus. The listeners were instructed to evaluate the sixth chord in these sequences, with only five preceding chords to contextualise these evaluations. Such short contexts give high-level syntactic structure little opportunity to contribute to predictions, especially since any such structure will be obfuscated if the stimulus starts partway through a musical phrase. Our stimuli instead provide a perspective on *local* harmonic syntax, the sense in which chord progressions are influenced by the few immediately preceding chords.

Listeners derive rich feature representations for musical chords. One example feature is harmonicity: high harmonicity occurs when a chord's acoustic spectrum resembles a harmonic series, and typically has positive aesthetic connotations. Another such feature is the chord root, corresponding to the brain's hypothesis about the fundamental frequency of the musical chord. A third such feature is the root interval, corresponding to the pitch interval between the previous chord root and the current chord root. In our corpus analyses, we showed that the harmonic syntax of popular music expresses salient statistical regularities across many of these features, and that an ideal listener would take advantage of many such features when predicting chord sequences. The optimised model paid particular attention to two low-level consonance features, namely interference between partials (after Hutchinson & Knopoff, 1978) and periodicity/harmonicity (after Chapter 3), as well as a high-level feature that captured relative pitch perception by expressing each chord relative to the bass note of the previous chord. However, the ideal model also took advantage of various other features (e.g. spectral similarity, pitch distance, root interval, chord inversion).

Traditional feature-based models of music prediction cannot incorporate continuous features such as interference between partials and spectral similarity (e.g. Hedges & Wiggins, 2016a, 2016b; Pearce, 2005, 2018; Rohrmeier & Graepel, 2012; Whorley & Conklin, 2016; Whorley et al., 2013). Our viewpoint regression technique is a novel contribution in this regard, generating predictions from both continuous and categorical features. This technique seems particularly appropriate for harmony, but could well generalise to other sequential domains. One example is melody, where categorical statistical-learning models have dominated in the last two decades (Pearce, 2005; Pearce & Müllensiefen, 2017; Pearce et al., 2010b; Pearce & Wiggins, 2006), yet recent modelling research indicates that continuous features (pitch proximity, central pitch tendency) are also needed to explain listener behaviour (Morgan, Fogel, Nair, & Patel, 2019).

Our corpus analysis indicated that (near)-optimal predictive performance for popular harmony within this modelling framework can be achieved using a model with a Markov order bound of one. An order bound of one means that the model generates predictions using just two types of information: the frequency of particular feature values (zeroth-order structure), and the frequency of particular feature values conditioned on the immediately preceding feature value (first-order structure). Since several high-level features are based on interval representations (e.g. root interval, bass interval), this means in practice that the optimal model generates predictions conditioned on the two previous chords. The fact that optimal performance can be achieved with such a low order bound implies that the local harmonic syntax of popular music is dominated by relatively low-level statistical regularities. This phenomenon is likely to differ between musical styles; more complex styles may require higher Markov order bounds or more complex modelling approaches. Previous studies have found optimal Markov orders of two or three for modelling jazz harmony (Hedges, Roy, & Pachet, 2014; Rohrmeier & Graepel, 2012) and an optimal order of one for modelling Beethoven string quartets (Landsnes et al., 2019).

Our behavioural study investigated how chords in popular harmony vary in perceived surprisal. This moment-to-moment variation in musical surprisal is considered to be a key part of musical aesthetics (Egermann et al., 2013; Huron, 2006; Meyer, 1957; Sauvé et al., 2018; Sloboda, 1991), but little is known about how these dynamics develop in Western harmony. Applying our full ideal-listener model to the behavioural data, we found a Spearman correlation of .70 between model surprisal and listener surprisal, suggesting that probabilistic prediction provides a good account of our participants' behaviour. We then analysed different versions of the model to investigate the statistical processes underlying participants' surprisal judgments, and found that these judgments could be largely explained by a radically simplified model that predicts chords solely on the basis of their number of occurrences in the preceding five-chord context (Spearman correlation of .65). This implies that, in popular music, the moment-to-moment variation in harmonic surprisal does not reflect complex syntactic phenomena to any great extent, but instead primarily reflects simple repetition effects.

It is important to note that probing chord-to-chord surprisal variation is not the same as probing listeners' full predictive distributions. It is quite possible that simple repetition statistics are sufficient to explain chord-to-chord surprisal variation, but insufficient to explain the listener's full probability distributions. For example, suppose that a musical style universally uses highly consonant chords: in this case, consonance would not predict variations in surprisal within the corpus, but it would still be useful for predicting what chords are likely or unlikely to come next. Nonetheless, moment-to-moment surprisal dynamics are thought to be particularly important for aesthetic responses to music (e.g. Egermann et al., 2013; Huron, 2006; Meyer, 1957; Sauvé et al., 2018; Sloboda, 1991), motivating their study here.

After adjusting for measurement error, the full ideal-listener model predicted mean surprisal judgments with a Pearson correlation coefficient of .73. This correlation is quite strong, but there remains clear room for improvement. A more reliable account of listener judgments might be achieved by adding computational models of other musical dimensions, in particular melody and voice leading. We might expect higher surprisal ratings when the melody line (comprising the highest note in each chord) makes unexpected progressions, for example jumping by large intervals. We might also expect higher surprisal ratings when the chord progression involves an unidiomatic voice leading, for example one with parallel octaves. Future work could address these possibilities using computational models of melodic expectation (e.g. Pearce, 2018; Temperley, 2008) and voice leading (e.g. Chapter 5).

This study's findings are restricted to popular music and (relatively) untrained Western music listeners. Popular music is appealing to study because, by definition, it corresponds to a large proportion of real-world music listening; likewise, untrained listeners are appealing to study because they represent the majority of the world's population. Nonetheless, it would certainly be interesting to study other musical styles and other participant groups. It is quite possible that certain musical styles, such as jazz music, carry more complex syntactic structure than the popular music studied here. It is also quite possible that more musically experienced listeners will be sensitive to more complex statistical structure in this music (e.g. Hansen & Pearce, 2014).

It is worth considering what might happen if we were to repeat this study with longer chord sequences. On the one hand, longer chord sequences provide more scope for high-level syntactic structure to contribute to predictions. On the other hand, longer sequences provide more exposure for repeated patterns specific to a given composition, and these repeated patterns may well come to dominate prediction performance at the expense of harmonic syntax. These repeating passages may often be relatively long, meaning that the model's ideal Markov order may increase past one. Such effects should be taken into account when comparing results between different music prediction studies (Hedges et al., 2014; Landsnes et al., 2019; Rohrmeier & Graepel, 2012; Sears et al., 2018a). In particular, evaluating models on entire musical compositions is likely to favour models that effectively detect local repetitions, whereas evaluating models on short chord sequences is likely to favour models that effectively capture harmonic syntax.

These sequences were synthesised as minimal sequences of isochronous piano chords. This approach is useful for isolating harmony from other parts of the musical experience, but it creates a somewhat impoverished listening experience compared to naturalistic listening. Harmonic expectation does seem to be influenced by external cues such as metre (Hedges & Wiggins, 2016b) and timbre (Vuvan & Hughes, 2019), and such effects should ultimately be incorporated into computational models of harmonic expectation. The present study could therefore be complemented by studies that synthesise the stimuli using more realistic musical textures, or that use audio from the original musical compositions.

Our viewpoint regression model of harmonic expectation was trained on a corpus of 439 popular music compositions, intended as a first approximation to an average Western listener's musical exposure. Better listener simulations might be achieved by tailoring this corpus more closely to the listening experiences of individual participants, perhaps to capture the mix of musical styles that they have been exposed to, or to capture the amount of musical exposure that they have received. The outcomes of the ideal-listener analyses may also be affected by corpus selection; in particular, larger training sets support the acquisition of more complex statistical regularities. As a result, an ideal-listener model trained on larger music corpora might take more advantage of high-order Markov models and more expressive features, in particular features with large alphabets (e.g. pitch-class sets).

We have developed an open-source software package, *hvr*, for supporting these future research possibilities.[8] This package defines formal models for many perceptual features thought to be important for harmony perception, such as interference between partials, spectral similarity, and chord roots. It provides methods for simulating how listeners mine statistical regularities from these perceptual features, and for simulating how statistical regularities from

[8] This package is available at `https://github.com/pmcharrison/hvr`.

different features might be combined to generate predictions for chord sequences. Importantly, it is straightforward to constrain various aspects of the model, such as the feature set and the Markov order bound, to investigate how different statistical regularities contribute to prediction generation.

One possibility is to use this model as trained in the present study, and use it to generate predictions for new chord sequences. These predictions may be summarised using information-theoretic quantities such as information content, which quantifies the surprisingness of a given chord, and predictive entropy, which quantifies the uncertainty of the model's predictions for a given chord (e.g. Figure 4.11).

A second possibility is to retrain the model on a new music corpus, allowing the researcher to simulate musical enculturation. In this latter case, the trained model itself has music-theoretic applications: in particular, the feature importance metrics and continuous feature effects can be examined to identify what kinds of psychological features are structurally important for a given musical style, and the ways in which these different features contribute to determine the choice of the next chord.

In music-theoretic applications such as these, it is important to note that the optimised viewpoint weights will depend both on the available features and the amount of training data given to the categorical viewpoint models. For example, a given feature may be useful in isolation, but be made redundant by the addition of a more informative feature. Furthermore, a given feature might be predictively useful given lots of training data, but uninformative given insufficient training data; this commonly applies to categorical features with large alphabets (e.g. pitch-class sets), whose models contain lots of parameters and hence need lots of data to estimate accurately. Correspondingly, such features can receive low weights with small training datasets, but high weights with large training datasets.

One modelling assumption that deserves interrogation is the idea that listeners compute full probability distributions over all 24,576 possible chord choices, as required to compute the normalisation constant in Equation 4.2. This seems intuitively unrealistic. One possibility is that the listener does not compute the full probability distribution, but instead computes probabilities for a small number of chords representing the most likely continuations. A second possibility is that the listener approximates the normalisation constant in some way. Implementing such mechanisms could significantly reduce the computational demands of the model's implementation.

A second assumption that deserves interrogation is the way in which viewpoint weights are optimised on many observations simultaneously. For studies of online learning, it would be more cognitively plausible to adopt an online optimisation algorithm, where viewpoint weights are incrementally updated after the observation of each musical event. This could also reduce the computational demands of the model's implementation.

There are several other ways in which the model might be developed further. A limitation of the PPM algorithm, used to model the categorical features, is that it cannot learn explicit representations of latent states. Latent states could be useful for capturing the music-theoretic concept of local key. It might therefore be interesting to trial alternative sequence prediction models, such as hidden Markov models (Rabiner, 1989) and long short-term memory recurrent neural networks (Hochreiter & Schmidhuber, 1997). A second limitation is that polynomial feature expansion, used to model the continuous features, can perform poorly when extrapolating outside the range of feature values observed in the training set. It might therefore be worthwhile to trial alternative approaches such as penalised spline regression (Kneib, Baumgartner, & Steiner, 2007). A third limitation is that the model cannot respond to the local distribution of continuous features within a given composition, unlike the categorical component of the model, which incrementally learns the distribution despite being able learn local distributions of categorical features.

Here we prespecified the model's perceptual features on the basis of previous research in music psychology and music theory, but it would also be possible to have the model learn its features in a data-driven manner (e.g. Langhabel, Lieck, Toussaint, & Rohrmeier, 2017; Whorley et al., 2013). From a cognitive perspective, it does seem likely that learning contributes to the formation of harmonic representations; for example, the 12-tone chromatic scale is unlikely to be innate, and is more likely acquired through long-term exposure to Western music. Other features, such as interference between partials and harmonicity, are less likely to be learned through solely musical exposure. The ideal model would simulate how these different representations arise through a combination of biological predispositions and auditory experience.

Here we have focused on cognitive modelling, but the proposed model also has interesting applications in other fields. It is well-suited to music generation, defining a probability distribution over successive chords that can be sampled from to generate new chord sequences. An appealing feature of the model is the interpretability of its parameters, which composers could take advantage of to explore new stylistic spaces. For example, it would be straightforward to manipulate the low-level feature effects, creating for example a musical style that minimises spectral similarity while maximising voice-leading efficiency. It would also be straightforward to train different feature models on different musical corpora, allowing the composer to create compositions that combine one aspect of one musical style with another aspect of a second musical style (e.g. combining the harmonic vocabulary of the composer Claude Debussy with the root progressions of popular music). The model also has potential applications in automatic music transcription, providing a prior distribution that biases the transcription process towards chord sequences that have high plausibility within a given musical style. These possibilities deserve to be explored in future work.

4.7 Methods

4.7.1 Corpus

The musical corpus corresponded to version 2.0 of the McGill Billboard corpus (Burgoyne, 2011), which comprises expert musicians' transcriptions of 739 unique songs sampled from the Billboard 'Hot 100' charts between 1958 and 1991. We used the pitch-class chord version of the corpus developed in Chapter 2, which omits repeated compositions, omits section repeats, and omits successive repetitions of the same chord. In the pitch-class chord representation, the chord is defined as a combination of a bass pitch class and an unordered set of non-bass pitch classes. This dataset is available in the *hcorp* package under the label `popular_1`.

4.7.2 Modelling

New spectral similarity model

We used a new spectral similarity model that compares each incoming chord against an auditory buffer representing the last few chords heard by the listener. Each chord is first represented as a smooth pitch-class spectrum, after Milne et al. (2011): this involves representing the chord as a pitch-class set, expanding each pitch class into its implied harmonics, and circularly convolving the resulting spectrum with a Gaussian function to simulate inaccuracies in pitch perception. Following the perceptual optimisations of Milne & Holland (2016), we represent each pitch class as 12 harmonics with weights declining as $1/n^{0.75}$, and use a Gaussian function with standard deviation 6.83. We then simulate an auditory buffer by cumulatively summing smooth pitch-class spectra from successive chords, with the contribution of a given pitch-class spectrum exponentially decaying as a function of the number of chords played since its original presentation. The model defaults to a half life of three chords, meaning that a given pitch-class spectrum contributes half as much after three chords have elapsed, and a quarter as much after six chords. A computationally efficient way of implementing this exponential decay is to multiply the cumulative sum by $2^{-1/h}$ after adding each new chord, where h is the half life. Spectral similarity is then operationalised as the cosine similarity between the new chord's pitch-class spectrum and the auditory buffer's pitch-class spectrum at the point when the new chord was presented. This model is implemented in the *specdec* package for the programming language R.[9]

Analysing feature distributions

For analysing low-level feature distributions, we constructed a dataset corresponding to the target chords at the sixth position of each of the 300 behavioural stimuli, enumerating each of the possible 24,576 pitch-class chords that could have been played at that position, and calculating feature values for each of these hypothetical chords. This produced 7,372,800 data points in total, 300 of which corresponded to observed chords, and 7,372,500 of which corresponded to unobserved chords.

For analysing high-level feature distributions, we tabulated feature transitions from the 439 compositions in the Billboard corpus that did not overlap with our behavioural stimuli, and plotted the results as transition matrices (Figures 4.4, 4.5). We then trained *viewpoint models* on these 439 compositions, and used them to generate predictions for the 300 target chords in the behavioural stimuli. A viewpoint model generates predictions for successive chords derived from a sequence prediction model trained on a given categorical feature (see *Viewpoint models* for details). Here the underlying sequence prediction model is Prediction by Partial Matching (PPM; Cleary & Witten, 1984) as updated by Moffat (1990) and Bunton (1997) and configured by Pearce & Wiggins (2004; Pearce, 2005) (see *Sequence prediction models* for details). Following Pearce & Wiggins (2004; Pearce, 2005), the PPM model is configured to use escape method 'C' (Moffat, 1990) and has update exclusion disabled. The viewpoint model's predictions are converted to log probabilities and then correlated in Figure 4.6, with hierarchical clustering performed on the rows and columns of the correlation matrix using the R function `hclust` with the default 'complete linkage' method. The log probabilities correspond to the 'long-term expectedness' features of the viewpoint regression model described later in this section.

Sequence prediction models

A *sequence prediction model* is a statistical model that defines a probability distribution over the continuations of a sequence conditioned on the preceding portion of the sequence. Here we are particularly interested in sequence prediction models that operate over symbolic sequences, where a symbolic sequence is defined as an ordered series of symbols drawn from a finite alphabet $\mathcal{A}$. Let e_0^n

[9] `https://github.com/pmcharrison/specdec`

denote a sequence of length n, let e_i denote the ith element in this sequence, with e_0 corresponding to an unwritten placemarker for the start of the sequence, and let e_i^j denote the subsequence $(e_i, e_{i+1}, \ldots, e_j)$. Let $e_0^i :: x$ denote the sequence produced by appending the symbol x to the sequence e_0^i. A sequence prediction model is then tasked with estimating the probability distribution $P(e_i \mid e_0^{i-1})$.

One such sequence prediction model is the Prediction by Partial Matching (PPM) algorithm, a variable-order Markov model that was originally developed for data compression (Cleary & Witten, 1984) but subsequently proved well-suited to cognitive modelling (see Pearce, 2018 for a review). PPM generates predictions by blending together predictions for multiple n-gram models of different orders, where an n-gram model generates predictions from the empirical distribution of subsequences of length n (n-grams) in the training corpus. This blending process allows PPM to take advantage of the strengths of both low-order and high-order n-gram models. Low-order models have a smaller alphabet of n-gram frequencies to estimate, and hence require less training data to produce reliable predictions. High-order models have larger n-gram alphabets, allowing the models to capture more complex sequential structure as long as sufficient training data are available. PPM initially favours low-order models and gradually moves to high-order models with increasing training data, a strategy that yields competitive performance on both small and large datasets. The precise mechanics of this strategy have developed over the years. Our implementation reproduces that of Pearce (2005, 2018), which incorporates various improvements including update exclusion (Moffat, 1990) and interpolated smoothing (Bunton, 1997).

Viewpoint models

The PPM algorithm can be applied directly to pitch-class chords to estimate conditional probabilities for sequence continuations. However, such an approach would not incorporate the kinds of perceptual features that are thought to be important for harmony cognition. This challenge is addressed by *viewpoint models*, which can generate predictions for pitch-class chords using statistical regularities learned from perceptual feature spaces. Viewpoint models were first introduced by Conklin & Witten (1995) in the context of multiple viewpoint systems, and were subsequently developed by other researchers (Hedges & Wiggins, 2016a; Pearce, 2005). The technique was originally introduced as a heuristic procedure for developing feature-based predictive models of melody, but here we give viewpoint modelling an explicit generative formalisation in the context of harmony modelling.

Viewpoint models generate predictions using *categorical features*, features that take values from a discrete alphabet. A categorical feature may be formalised as a pair of functions: an *availability* function $f_{\text{available}}$, and an *extractor* function f_{extract}. The availability function takes a context sequence of chords as input, returning `TRUE` if the feature will be defined for the next chord in the sequence, and `FALSE` otherwise. The extractor function takes as input both a context sequence and a continuation, and returns a symbol from the feature's alphabet that characterises the continuation in the context of the preceding chord sequence. For example, the root interval feature would be defined by an availability function that only returns `TRUE` if the context sequence is non-empty, and an extractor function that computes the pitch-class interval from the root pitch class of the previous chord to the root pitch class of the continuation.

The viewpoint model then supposes that the structure of the chord sequence is wholly driven by the syntactic structure of the categorical feature, which is itself determined by some latent generative model. The composer is modelled as choosing each chord in a chord progression through the following process: first sample a feature value from the feature's generative model, then randomly choose a chord that reproduces this feature value. This process is specified more formally as follows:

1. Initialise the chord sequence and the feature sequence as empty lists.
2. For $i = 1, 2, \ldots, n$:
    (a) Apply the availability function ($f_{\text{available}}$) to the portion of the chord sequence generated so far.
    (b) If the availability function returns `FALSE`, randomly sample the next chord, x, from the set of all pitch-class chords.
    (c) Otherwise:
        i. Sample a new feature value, y, from a generative model conditioned on the current portion of the feature sequence.
        ii. Find the set $S = x : f_{\text{extract}}(\text{context}, x) = y$.
        iii. If S is empty, backtrack to step i.
        iv. Otherwise, randomly sample the next chord, x, from S.
        v. Append y to the feature sequence.
    (d) Append x to the chord sequence.

To train the viewpoint model, it is sufficient to approximate the feature's underlying generative model. This model can be approximated by fitting a generative model to the observed feature sequences. This approximation should be unbiased if backtracking never occurs, which is guaranteed if, whenever a feature is 'available', all feature values always have at least one chord to represent them. This condition is satisfied by all the categorical features used in the present work.

The resulting probabilistic model may be expressed algebraically as follows:

$$P(e_i \mid e_0^{i-1}) \begin{cases} \propto P\left(e_i \mid F\left(e_0^i\right)\right)\ P\left(F\left(e_0^i\right) \mid F\left(e_0^{i-1}\right)\right) & \text{if } f_{\text{available}}\left(e_0^{i-1}\right), \\ = 1/\mathcal{A} & \text{otherwise,} \end{cases} \tag{4.3}$$

where $F(e_0^i)$ is the sequence of feature values observed for the sequence e_0^i, defined recursively as

$$F(e_0) = (), \tag{4.4}$$

$$F(e_0^j) = \begin{cases} F(e_0^{j-1}) :: f_{\text{extract}}(e_0^{j-1}, e_j) & \text{if } f_{\text{available}}(e_0^{j-1}), \\ F(e_0^{j-1}) & \text{otherwise,} \end{cases} \tag{4.5}$$

and $P\left(e_i \mid F\left(e_0^i\right)\right)$, the probability of observing a given chord conditioned upon the feature sequence $F\left(e_0^i\right)$, is computed as a uniform distribution over all possible chords consistent with that feature sequence:

$$P\left(e_i \mid F\left(e_0^i\right)\right) = \frac{1}{\left|\left\{x \in \mathcal{A} : F\left(e_0^{i-1} :: x\right) = F\left(e_0^i\right)\right\}\right|}. \tag{4.6}$$

As discussed above, the conditional distribution over feature symbols $(P\left(F\left(e_0^i\right) \mid F\left(e_0^{i-1}\right)\right))$ may be approximated by fitting a sequence prediction model to observed feature sequences. Note that the proportionality relationship in Equation 4.3 becomes an equivalence relationship if and only if all feature observations with non-zero probability can be realised with at least one pitch-class chord, which is always the case with the features used in this paper.

The resulting viewpoint model generates predictions for pitch-class chords on the basis of single categorical features derived from these chords. The next section shows how viewpoint regression extends this technique to model multiple continuous and categorical features simultaneously.

Viewpoint regression

As discussed in Section 4.3, the viewpoint regression technique defines a sequence prediction model that combines information from both categorical and continuous features. The categorical and continuous characteristics of a given chord are summarised in a metafeature vector. Categorical features contribute short-term and long-term expectedness terms to this vector, defined as the chord's log probability according to viewpoint models trained on either the portion of the sequence heard so far (short-term expectedness) or on a large musical corpus representing musical enculturation (long-term expectedness). Continuous features, meanwhile, contribute orthogonal polynomial terms to the metafeature vector.

Equation 4.2, the estimated probability of a chord, can be written more formally as

$$P(e_i \mid e_0^{i-1}, \mathbf{w}) \propto \begin{cases} 0 & \text{if } e_i = e_{i-1}, \\ \exp\left(\sum_{j=1}^{m} w_j f_j \left(e_0^{i-1}, e_i\right)\right) & \text{otherwise.} \end{cases} \tag{4.7}$$

where $f_i \left(e_0^{i-1}, e_i\right)$ is the ith term of the metafeature vector as computed for the observation e_i with the context e_0^{i-1}, and m is the dimensionality of this vector.

Here we compute expectedness values using PPM-based viewpoint models. Following Pearce (2005), the PPM models are configured slightly differently for short-term versus long-term models. In particular, the short-term models use update exclusion (Moffat, 1990) and escape method 'AX' (Moffat, Neal, & Witten, 1998), whereas the long-term models disable update exclusion and use escape method 'C' (Moffat, 1990). Our PPM implementation is available in the R package *ppm* (Harrison et al., 2020).[10]

The probability of an individual sequence may then be written as

$$P(e_0^n) = \prod_{i=1}^{n} P(e_i \mid e_0^{i-1}) \tag{4.8}$$

where n is the length of the sequence; the log likelihood of the sequence is then

$$\log P(e_0^n) = \sum_{i=1}^{n} \log P(e_i \mid e_0^{i-1}). \tag{4.9}$$

The log likelihood of a corpus of sequences may then be written as a function of the weight vector, $\mathbf{w}$:

$$\ell(\mathbf{w}) = \sum_{k=1}^{N} \sum_{i=1}^{n_k} \log P(e_{i,k} \mid e_{0,k}^{i-1}, \mathbf{w}) \tag{4.10}$$

where N is the number of sequences in the corpus, n_k is the number of symbols in the kth sequence, $e_{i,k}$ denotes the ith symbol in the kth sequence, and $e_{0,k}^{i}$

[10] `https://github.com/pmcharrison/ppm`

denotes a subsequence corresponding to the first i symbols of the kth sequence in the corpus.

Following standard results for the log-linear model, the gradient of this log likelihood function is then

$$\frac{d\ell}{d\mathbf{w}}(\mathbf{w}) = \sum_{k=1}^{N} \sum_{i=1}^{n_k} \left(\mathbf{f}(e_{0,k}^{i}) - \sum_{x \in \mathcal{A} : x \neq e_{i-1,k}} P(e_{i,k} \mid e_{0,k}^{i-1}, \mathbf{w}) \; \mathbf{f}\left(e_{0,k}^{i-1} :: x\right) \right) \tag{4.11}$$

where $\mathbf{f}$ computes a vector of feature values. This expression may be interpreted as the difference between each feature's sum observed values and each feature's sum expected values according to the model.

These expressions for the log likelihood and its gradient may be plugged into a generic optimiser to find the maximum-likelihood weight vector for a given corpus. Here we used the limited-memory BFGS optimiser of Byrd, Lu, Nocedal, & Zhu (1995), as implemented in the R function `optim`. Categorical feature weights were constrained to be non-negative, reflecting the intuition that any associations between feature expectedness and chord expectedness should be positive.

We used polynomial functions to model contributions from continuous features. Using polynomial functions instead of linear functions enables the model to capture nonlinear relationships between features and chord probabilities. The higher the polynomial degree, the more complex relationships can be discovered, but the greater risk of the model overfitting to the training dataset. We used quartic polynomials as a compromise between these two issues; further increasing the polynomial order yielded little quantitative change in the shape of the feature effects, implying that these polynomials were sufficient to capture the primary trends in the data. For the sake of numerical stability, all polynomial features were computed as orthogonal polynomials using the R function `poly`.

We have implemented the resulting model as a freely available package, *hvr*, written for the programming language R. The core of the statistical model is written in C++ for speed. The package depends on five other packages of ours: *hrep* for representing and manipulating chords, *hvrmap* for precomputing feature derivations, *specdec* for implementing the new spectral similarity model, *ppm* for implementing the PPM algorithm, and *minVL* for computing pitch distance. All of these packages are available on GitHub and permanently archived on Zenodo.[11]

[11] `https://github.com/pmcharrison`; `https://zenodo.org`

Table 4.8: Participants' responses to the question "How often do you listen to popular music?"

Rarely	Sometimes	Often	Very often
6	10	17	17

Note. Each cell corresponds to the number of participants who selected a given response option.

4.7.3 Behavioural experiment

Participants

The participant group numbered fifty psychology undergraduates (6 male, 44 female) who participated in exchange for course credit or monetary compensation. The participants had a mean age of 18.7 years (*SD* = 1.7), and mostly self-reported as frequent listeners to popular music (Table 4.8).[12] Musical training was assessed using the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report questionnaire (Müllensiefen, Gingras, Musil, & Stewart, 2014), with participants receiving a mean score of 15.1 (*SD* = 8.0), which corresponds to the 22nd percentile of the original Gold-MSI calibration sample.

Stimuli

The stimuli numbered 300 eight-chord sequences from the popular music corpus (see Section 4.7.1), randomly sampled under the constraint that no song occurred more than once. Each chord sequence was synthesised with piano timbre using the audio software Timidity and SoX, with chords played at a tempo of 60 beats per minute without metrical cues. Chords were voiced using the following heuristic procedure: bass notes were assigned to the octave below middle C, whereas non-bass notes were assigned to the octave above middle C.

Procedure

Participants were individually tested in a quiet room with stimuli played over headphones. Stimulus administration and response recording was managed by software created using the Javascript package *jsPsych* (Leeuw, 2015) and the

[12] Participants were asked the following question: "How often do you listen to popular music? By popular music we mean music that many people listen or listened to, as reflected in the popular music charts. This includes both current popular music and popular music from previous years/decades." They were given the following response options: "Never", "Rarely", "Sometimes", "Often", "Very often", and "Don't know".

Table 4.9: Optimising local and global half lives for Leman's (2000a) spectral similarity model.

Half life (s)		Correlation with behavioural data	
Local	Global	Pearson	Spearman
0.1	1.5	−.04 [−.15, .07]	−.05
0.1	2.5	−.09 [−.20, .03]	−.09
0.1	4.0	−.11 [−.22, −.00]	−.12
0.5	1.5	−.25 [−.36, −.14]	−.26
0.5	2.5	−.30 [−.40, −.20]	−.31
0.5	**4.0**	**−.33 [−.43, −.22]**	**−.33**

Note. Correlations are conducted between raw model outputs and listeners' mean surprisal ratings. The selected configuration is marked in bold.

R package *psychTestR* (Harrison, 2020). This software was run on a desktop computer, using keyboard and mouse as input devices.

Each participant was presented 150 unique stimuli, each randomly selected from the pool of 300 stimuli, with the total number of presentations of each stimulus balanced across participants. Upon presentation of each stimulus, the participant was instructed to rate the surprisingness of the sixth chord on a one to nine scale using the keyboard. This sixth chord, termed the target chord, was cued using a visual clock-like animation that incremented continuously throughout the stimulus. Stimuli were administered in 25-trial blocks, between which participants were given 10-second breaks.

Participants were introduced to the task using a standardised training routine that included three practice trials. After stimulus presentation, participants completed a short questionnaire concerning basic demographics and their familiarity with popular music, followed by the musical training subscale of the Gold-MSI (Müllensiefen et al., 2014). The entire procedure took approximately 40 minutes.

Optimising spectral similarity models

Prior to making the behavioural evaluations reported in Section 4.5, we performed a parameter search for Leman's (2000a) model and the new spectral similarity model, selecting the configurations with the highest Pearson correlations between model outputs and listeners' mean surprisal ratings (Tables 4.9 and 4.10). Leman's model proved particularly sensitive to the local half life parameter: a local half life of 0.1 s produced no statistically significant correlations, but a local half life of 0.5 s produced statistically significant correlations

Table 4.10: Optimising the half life of the new spectral similarity model.

	Correlation with behavioural data	
Half life (chords)	Pearson	Spearman
1	.04 [−.07, .16]	.06
2	−.02 [−.13, .09]	−.00
3	−.04 [−.15, .07]	−.02
4	**−.05 [−.17, .06]**	**−.03**

Note. Correlations are conducted between raw model outputs and listeners' mean surprisal ratings. The configuration selected for the behavioural analyses is marked in bold. Note that the selected 4-chord half life is slightly longer than the 3-chord half life used in the corpus analyses (Sections 4.2.3 and 4.4). Note also that the optimised model still does not substantively predict surprisal ratings.

of about $r = -.3$. This sensitivity should be noted in future studies using this model; recently Sears et al. (2019) applied Leman's model to simulate listener responses to tonal cadences, and found that the model performed poorly, yet better results might have been achieved had they tried local half lives other than 0.1 s.[13] In contrast, the new spectral similarity model proved unable to predict listener judgments, despite its conceptual similarity to Leman's (2000a) model. This difference might be attributed to the fact that the new model incorporates octave invariance, unlike Leman's model. Further work could test this interpretation by implementing and evaluating a version of the model without octave invariance.

Data preprocessing

One participant was excluded because they gave the same surprisal rating for all 150 stimuli. Each remaining participants' ratings were standardised to z-scores to account for individual differences in scale usage, and then averaged across participants to produce one mean surprisal rating for each target chord.

The viewpoint regression models predict surprisal in terms of the log probability of the target chord. We capped surprisal estimates at 15.0 to prevent outliers from driving the parametric analyses. In practice this threshold was very rarely exceeded.

[13] This criticism also applied to an early version of the present work (Harrison & Pearce, 2018).

Correlation attenuation

Table 4.6 includes Pearson correlation coefficients that are disattenuated to correct for the measurement error in the mean surprisal ratings. Writing $\hat{x}_i$ for the ith mean surprisal rating, we calculate the mean standard error of mean surprisal ratings as $\overline{SE} = \frac{1}{n}\sum_{i=1}^{n} SE(\hat{x}_i)$, with n denoting the number of chords, and the standard error of each mean surprisal rating being estimated using the Central Limit Theorem.[14] We then calculate the reliability coefficient as $1 - \overline{SE}^2 / Var(\widehat{\mathbf{x}})$, where $Var(\widehat{\mathbf{x}})$ is the sample variance of the mean surprisal ratings. The disattenuated correlation coefficient is then computed by dividing the original correlation coefficient by the square root of this reliability coefficient.

4.7.4 Software

Our software and our analysis scripts depend heavily on open-source software. Particularly useful were *purrr*, *dplyr*, *ggplot2*, *plyr*, *magrittr*, *checkmate*, *egg*, *cowplot*, *pheatmap*, *shiny*, and *jsPsych.* Full dependency lists can be found in the source code.[15]

[14] By the Central Limit Theorem, the standard error for a given chord's mean surprisal rating is computed as $SD(\mathbf{x})/\sqrt{|\mathbf{x}|}$, where $\mathbf{x}$ is the vector of surprisal ratings for a given chord and SD is the sample standard deviation.

[15] `https://github.com/pmcharrison/hvr-analyses`

Chapter 5

Voice leading

5.1 Introduction

Western music pedagogy traditionally emphasises two aspects of compositional practice: harmony and voice leading. Harmony specifies a vocabulary of harmonic units, termed ‘chords’, alongside conventions for combining these chords into chord sequences; voice leading describes the art of realising these chords as collections of individual voices, with a particular emphasis on the progression of individual voices from chord to chord.

Huron (2001, 2016) has argued that Western voice-leading practice is largely driven by the goal of manipulating the listener’s psychological processes of auditory scene analysis. Auditory scene analysis describes how the listener organises information from the acoustic environment into perceptually meaningful elements, typically corresponding to distinct auditory sources that can be related to real-world objects (Bregman, 1990). In Baroque music, voice-leading practice is often consistent with the principle of promoting the perceptual independence of the different musical voices. For example, Baroque composers tended to avoid parallel octaves between independent voice parts, presumably because parallel octaves cause the two voice parts to temporarily ‘fuse’ into one perceptual voice, an incongruous effect when the voices are elsewhere perceived as separate voices (Huron, 2001, 2016). However, perceptual independence is not a universal musical goal: for example, octave doubling has long been accepted in Western music as a technique for creating the percept of a single voice with a reinforced timbre. This approach was taken further by composers such as Debussy, who often constructed entire musical textures from parallel motion while freely disregarding traditional prohibitions against parallel fifths and octaves (e.g. *La Cathédrale Engloutie*, 1910, L. 117/10). In such cases, we might hypothesise that Debussy purposefully adopted parallelism to minimise the perceptual independence of the underlying voices, hence creating a unitary textural stream (Huron, 2016).

Here we seek to develop a computational cognitive model of voice leading. This model is intended to simulate how a composer might choose between various candidate voice leadings on the basis of their consequences for music perception. One goal of constructing such a model is to create a formal basis for testing voice-leading theories on large datasets of music compositions. A second goal is to create a tool for generating voiced versions of unseen chord sequences, with potential applications in music composition and music cognition research.

A computational cognitive model of voice leading could adopt various levels of explanatory depth. For example, a researcher might introduce a model that takes the musical surface as input, simulates the process of auditory scene analysis, and quantifies the extent to which individual voices are recognised as independent auditory streams. If this model successfully predicted composers' decisions, this would support the hypothesis that voice leading is ultimately driven by the goal of maximising the perceptual independence of musical voices. A second researcher might agree that voice-leading practices were originally shaped by perceptual principles, but hypothesise that experienced composers pay little attention to auditory scene analysis in practice, and instead construct their voice leadings from knowledge of voice-leading practice accrued through musical experience. Correspondingly, this second researcher might build a data-driven model that learns to construct voice leadings by emulating voice-leading practice in representative musical corpora, without any reference to auditory scene analysis.

Neither of these approaches is necessarily more 'correct' than the other, but both do serve different goals. From a cognitive modelling perspective, the auditory scene analysis model better addresses the ultimate causes of voice-leading practices, explaining how compositional practice may have been shaped by general perceptual principles. In contrast, the data-driven model might better simulate the psychological processes of an individual composer. From a music generation perspective, the auditory scene analysis model is unlikely ever to approximate a particular musical style perfectly, since it neglects cultural contributions to voice-leading practice. In contrast, the data-driven model might effectively approximate a given musical style, but fail to distinguish perceptually grounded principles from culturally grounded principles, and hence fail to generalise usefully to other musical styles.

Here we adopt an approach intermediate to these two extremes. We do not try to build a comprehensive model of auditory scene analysis, and we do not construct a solely data-driven model. Instead, we construct a model that characterises voice-leading acceptability as an interpretable function of various features that might reasonably be considered by an experienced composer, such as voice-leading distance, parallel octaves, and interference between partials. This level of abstraction is useful for interpretation: it means that we can inspect the model and understand what it has learned about voice-leading practice. This interpretability is also useful for music generation, as it allows the user to manipulate particular aspects of the model to achieve particular musical effects.

Following Huron (2001, 2016), we ground our model's features in both music theory and auditory perception. Music theory tells us about voice-leading rules that composers may have been explicitly taught during their musical training, as well as voice-leading rules that analysts have inferred from their study of musical practice. Auditory perception tells us what implications these features may have for the listener, and helps to explain why particular musical styles adopt particular voice-leading practices.

The resulting model is well-suited to both corpus analysis and music generation. Applied to a music corpus, the model provides quantitative estimates of the importance of different voice-leading principles, as well as p-values for estimating the statistical reliability of these principles. Applied to novel chord progressions, the model can generate voice leadings based on these different voice-leading principles, with the user having the freedom to use parameters derived from a reference corpus or alternatively to use hand-specified parameters in order to achieve a desired musical effect.

Importantly, the model does not assume a universal ideal for voice-leading practice. According to the model, voice-leading practice in a particular musical style is characterised by a set of regression weights that determine the extent to which composers promote or avoid certain musical features, such as parallel octaves and interference between partials. Depending on the musical style, the contribution of a given feature might reverse entirely; for example, parallel octaves are avoided in Bach chorales, but are commonplace in certain compositions by Debussy. The model's main assumption is that a common set of perceptual features underpins voice leading in diverse musical styles, an assumption that seems plausible in the context of the proposed relationship between voice-leading practice and auditory scene analysis (Huron, 2001, 2016).

In its broader definitions, the art of voice leading includes processes of embellishment and elaboration, whereby an underlying harmonic skeleton is extended through the addition of musical elements such as passing notes, neighbor notes, suspensions, and appoggiaturas (Huron, 2016). These additions can contribute much to the interest of a musical passage. However, they add a whole layer of complexity to the voice-leading task, potentially contributing a new 'surface' harmonic progression that should itself obey certain syntactic conventions. It is difficult to model such processes while maintaining a strict division between harmony and voice leading. In this chapter, therefore, we omit processes of embellishment and instead formalise voice leading as the task of assigning pitch heights to pitch classes prescribed by a fixed harmonic progression. This process might also be termed ‘voicing’; we retain the term ‘voice leading’ to emphasise how we are interested not only in the construction of individual chord voicings but also in the way that these voicings lead consecutively from one voicing to the next.

Voice leading is typically taught in the context of musical styles where each note is explicitly assigned to a particular voice part, such as Baroque chorale harmonisations. However, voice leading can also be important in other styles: for example, effective voice leading is considered essential to jazz music, despite the fact that jazz harmony is often played on the piano or guitar, where explicit voice assignment is lacking (Tymoczko, 2011). We wish for our model to generalise to such styles, and therefore we do not include explicit voice assignment in the algorithm. Instead, the algorithm infers voice assignments solely from the pitch content of the musical passage, and uses these inferred assignments to evaluate voice-leading rules.

There are several published precedents for voice-leading modelling. Models specifically of voice leading are quite rare (see Hörnel, 2004 for one such model), but many models do exist for melody harmonisation, a compositional task that often involves a voice-leading component (see Fernández & Vico, 2013 for a review). Generally speaking, these models are grounded more in artificial intelligence research than cognitive science research; as a result, there is little emphasis on auditory perception, model interpretability, or corpus analysis. Many of the models are neural networks, which can potentially capture very complex musical principles but typically possess low interpretability (Hild, Feulner, & Menzel, 1984; Hörnel, 2004). Others are rule-based, providing a formal instantiation of the researcher’s music-theoretic knowledge without necessarily testing this knowledge against musical practice (Ebcioğlu, 1988; Emura, Miura, & Yanagida, 2008). Both the neural-network approaches and the rule-based approaches seemed ill-suited to our cognitive modelling goals. Moreover, the models generally lack publicly available implementations, which restricts their utility to potential users. We address these concerns in the present work, developing a cognitively motivated voice-leading model and releasing a publicly available implementation in the form of *voicer*, an open-source software package for the R programming language (R Core Team, 2017).

5.2 Model

We suppose that a chord sequence can be represented as a series of N tokens, $(x_1, x_2, \ldots, x_N)$, where each token constitutes a *pitch-class chord*, defined as a pitch-class set with known bass pitch class (Chapter 2). For example, a IV-V-I cadence in C major might be written as $((5, \{0, 5, 9\}), (7, \{2, 7, 11\}), (0, \{0, 4, 7\}))$. Further, we suppose that we have a candidate generation function, C, which generates a set of candidate voicings for a given pitch-class chord. For example, we might have $C((0, \{0, 4, 7\})) = \{\{48, 52, 55\}, \{48, 52, 67\}, \{48, 64, 67\}, \ldots\}$, where each voicing is expressed as a set of MIDI note numbers. Our aim is to model the process by which the musician assigns each pitch-class chord x_i a voicing $X_i \in C(x_i)$.

We suppose that the probability of choosing a voicing X_i varies as a function of certain features of X_i as evaluated with respect to the previous voicing, X_{i-1}. We write f_j for the jth of these features, and define a *linear predictor* $L(X_i, X_{i-1})$ as a weighted sum of these features, where the regression weight of feature f_j is denoted w_j.

$$L(X_i, X_{i-1}) = \sum_j w_j f_j(X_i, X_{i-1}) \tag{5.1}$$

The linear predictor summarises the desirability of a particular voicing, aggregating information from the different features. As with traditional regression models, the regression weights determine the contribution of the respective features; for example, a large positive value of w_j means that voicings are preferred when they produce large positive values of f_j, whereas a large negative value of w_j means that large negative values of f_j are preferred. In this work the regression weights will stay fixed for different compositions within a music corpus; however, the method could easily be adjusted to allow regression weights to vary between compositions, or even within compositions, for example to capture different voice-leading behaviour at cadences.

We suppose that the probability of sampling a given chord voicing is proportional to the exponentiated linear predictor, with the normalisation constant being computed by summing over the set of candidate voicings, $C(x_i)$:

$$P(X_i \mid X_{i-1}, x_i) = \begin{cases} \frac{e^{L(X_i, X_{i-1})}}{\sum_{X \in C(x_i)} e^{L(X, X_{i-1})}} & \text{if } X_i \in C(x_i), \\ 0 & \text{otherwise.} \end{cases} \tag{5.2}$$

This is a log-linear model with a similar functional form to the harmonic expectation model from Chapter 4. The probability of the full sequence of voicings can then be expressed as a product of these expressions:

$$P(X_1, X_2, \ldots, X_N \mid x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} P(X_i \mid X_{i-1}, x_i). \tag{5.3}$$

where X_0 is a fixed start symbol for all sequences.

Once the candidate voicing generation function C and the features f_i are defined, the regression weights w_i can be optimised on a corpus of chord sequences using maximum-likelihood estimation. Here we perform this optimisation using iteratively reweighted least squares as implemented in the *mclogit* package (Elff, 2018). The resulting regression weights quantify the contribution of each feature to voice-leading practice.

5.3 Features

Our feature set comprises 12 features that we hypothesised should be useful for the voice-leading model. We designed these features to cover the 13 traditional rules reviewed in Huron's (2001) perceptual account of voice leading (see also Huron, 2016).

5.3.1 Voice-leading distance

The voice-leading distance between two chords may be defined as the sum distance moved by the implied voice parts connecting the two chords. A chord progression that minimises voice-leading distance is said to have 'efficient' voice leading. Efficient voice leading promotes auditory stream segregation through the pitch proximity principle, which states that the coherence of an auditory stream is improved when its tones are separated by small pitch distances (Huron, 2001, 2016). Correspondingly, we expect our voice-leading model to penalise voice-leading distance when applied to common-practice Western music. We compute voice-leading distance using the minimal voice-leading algorithm of Tymoczko (2006) with a taxicab norm, modified to return pitch distances instead of pitch-class distances. This algorithm generalises effectively to chords with different numbers of pitches by supposing that several voices can start or end on the same pitch. For example, the optimal voice-leading between {C4, E4, G4} and {B3, D4, F4, G4} is found to be C4 → B3, C4 → D4, E4 → F4, G4 → G4, which corresponds to a voice-leading distance of $1+2+1 = 4$ semitones.

5.3.2 Melodic voice-leading distance

The uppermost voice may be a special case when it comes to voice-leading efficiency. On the one hand, the uppermost voice is particularly salient to listeners

(Trainor, Marie, Bruce, & Bidelman, 2014), implying that any voice-leading inefficiencies in this line may also be particularly salient to listeners. On the other hand, the uppermost voice is often tasked with creating an interesting melodic line, potentially incentivising the occasional use of large intervals. We avoid committing to a hypothesis about the direction of this particular effect, but create a feature capable of capturing both possibilities, termed *melodic voice-leading distance*, defined as the distance in semitones between the uppermost voices of successive chords.

Efficient voice leading is likely to be particularly salient for the uppermost voice, on account of the high voice superiority effect (Trainor et al., 2014). We capture this hypothesis with a feature termed *melodic voice-leading distance*, defined as the distance between the uppermost voices of successive chords, measured in semitones. We expect our model to penalise melodic voice-leading distance when applied to common-practice Western music.

5.3.3 Pitch height

Harmonic writing in common-practice Western music commonly uses pitches drawn from a three-octave span centered on middle C (C4, 261.63 Hz) (Huron, 2001). This three-octave span corresponds approximately to the combined vocal range of male and female voices, and to the frequency range for which complex tones elicit the clearest pitch percepts (Huron, 2001). We address this phenomenon with three features. *Mean pitch height* computes the absolute difference between the chord's mean pitch height, defined as the mean of its MIDI note numbers, and middle C, corresponding to a MIDI note number of 60. *Treble pitch height* is defined as the distance that the chord's highest note spans above C5 (523.25 Hz), expressed in semitones, and returning zero if the chord's highest note is C5 or lower. Similarly, *bass pitch height* is defined as the distance that the chord's lowest note spans below C3 (130.81 Hz), expressed in semitones, and returning zero if the chord's highest note is C3 or higher. We expect our model to penalise each of these features.

5.3.4 Interference between partials

Any given chord may be realised as an acoustic spectrum, where the spectrum defines the amount of energy present at different oscillation frequencies. The peaks of this spectrum are termed *partials*, and typically correspond to integer multiples of the fundamental frequencies of the chord's constituent tones. Partials separated by small frequency differences are thought to elicit interference effects, in particular *masking* and *roughness* (Section 3.3.2). Masking, the auditory counterpart to visual occlusion, describes the way in which the auditory system struggles to resolve adjacent pitches that are too similar in frequency. Roughness describes the amplitude modulation that occurs from the superposition of two tones of similar frequencies. Both masking and roughness are thought to have negative aesthetic valence for Western listeners, potentially contributing to the perceptual phenomenon of 'dissonance'. Correspondingly, musicians may be incentivised to find voice leadings that minimise these interference effects.

Corpus analyses have shown that interference between partials provides a good account of chord spacing practices in Western music, in particular the principle that lower voices should be separated by larger pitch intervals than upper voices (Huron & Sellmer, 1992). Correspondingly, we introduce interference between partials as a voice-leading feature, operationalised using the computational model of Hutchinson & Knopoff (1978) as implemented in the *incon* package (Chapter 3). This model expands each chord tone into its implied harmonics, and sums over all pairs of harmonics in the resulting spectrum, modelling the interference of a given pair of partials as a function of their critical bandwidth distance and the product of their amplitudes. We expect our voice-leading model to penalise high values of this interference feature.

5.3.5 Number of pitches

The number of distinct pitches in a chord voicing must be greater than or equal to the size of the chord's pitch-class set. Larger chords can be produced by mapping individual pitch classes to multiple pitches. Instrumental forces place absolute constraints on this process; for example, a four-part choir cannot produce voicings containing more than four pitches, but can produce voicings with fewer than four pitches by assigning multiple voices to the same pitches. Other stylistic principles place weaker constraints on this process, which we aim to capture with our voice-leading model. First, we suppose that the musical style defines an ideal number of pitches, and that this ideal can be deviated from with some penalty; for example, a four-part chorale preferentially contains four pitches in each chord voicing, but it is permissible occasionally to use voicings with only three pitches. We operationalise this principle with a feature called *Number of pitches (difference from ideal)*. Second, we suppose that there may be some additional preference for keeping the number of pitches consistent in successive voicings, and operationalise this principle with a feature called *Number of pitches (difference from previous chord)*. We expect the voice-leading model to penalise both of these features.

5.3.6 Parallel octaves/fifths

Octaves and fifths are pitch intervals spanning 12 semitones and 7 semitones respectively. Parallel octaves and parallel fifths occur when two voice parts separated by octaves or fifths both move by the same pitch interval in the same direction. Parallel motion tends to promote perceptual fusion, and this effect is particularly strong for harmonically related tones, such as octaves and fifths (Huron, 2016). The avoidance of parallel octaves and fifths in common-practice voice leading may therefore be rationalised as a mechanism for promoting the perceptual independence of the voices. Conversely, extended sequences of parallel octaves and fifths in the music of Debussy (e.g. *La Cathédrale Engloutie*, 1910, L. 117/10) may encourage listeners to perceive these sequences as single textural streams (Huron, 2016).

We capture this phenomenon using a Boolean feature termed *Parallel octaves/fifths (any parts)* that returns 1 if parallel octaves or fifths (or compound versions of these intervals[1]) are detected between any two parts and 0 otherwise. Voice assignments are computed using Tymoczko's (2006) algorithm, meaning that the feature remains well-defined in the absence of notated voice assignments.

As noted by Huron (2001), parallel octaves and fifths are particularly salient and hence particularly prohibited when they occur between the outer parts. We capture this principle with a Boolean feature termed *Parallel octaves/fifths (outer parts)*, which returns 1 if parallel octaves or fifths are detected between the two outer parts and 0 otherwise.

5.3.7 Exposed octaves (outer parts)

Exposed octaves, also known as 'hidden octaves' or 'direct octaves', occur when two voices reach an interval of an octave (or compound octave) by moving in the same direction. Injunctions against exposed octaves appear in many voice-leading textbooks, but the nature of these injunctions differs from source to source. For example, some say that the rule against exposed octaves applies to any pair of voice parts, whereas others say that the rule only applies to the outer parts; likewise, some say that exposed octaves are acceptable when either of the voices move by step, whereas others say that exposed octaves are only excused when the top line moves by step (see Arthur & Huron, 2016 for a review).

Auditory scene analysis provides a useful perspective on this debate. Like parallel octaves, exposed octaves combine similar motion with harmonic pitch intervals, and are hence likely to promote fusion between the constituent voices.

[1] A compound interval is produced by adding one or more octaves to a standard interval.

Approaching the interval with stepwise motion may counteract this fusion effect by introducing a competing cue (pitch proximity) that helps the listener differentiate the two voice parts (Huron, 2001, 2016). This provides a potential psychological explanation for why exposed octaves might be excused if they are approached by stepwise motion.

Arthur & Huron (2016) investigated the perceptual basis of the exposed octaves rule, and found that stepwise motion had little effect on perceptual fusion. However, they did find tentative evidence that stepwise motion reduces fusion in the specific case of the uppermost voice moving by step. They explained this effect by noting that fusion comes from the listener interpreting the upper tone as part of the lower tone, resulting in a single-tone percept at the lower pitch. Approaching the lower pitch with stepwise motion presumably reinforces this lower pitch, and therefore has limited consequences for the fusion effect. In contrast, approaching the higher pitch with stepwise motion may encourage the listener to 'hear out' this upper pitch, therefore reducing the fusion effect (Arthur & Huron, 2016).

Further work is required before the perceptual basis of exposed octaves is understood fully. For now, we implement a Boolean feature that captures the most consistently condemned form of exposed octaves: those that occur between the outer parts with no stepwise motion in either part. We term this feature *Exposed octaves (outer parts)*. Future work could implement different variants of this feature to capture the different nuances discussed above.

5.3.8 Part overlap

Ascending part overlap occurs when a voice moves to a pitch above that of a higher voice from the preceding chord. Similarly, descending part overlap occurs when a voice moves to a pitch below that of a lower voice from the preceding chord. According to Huron (2001), composers avoid part overlap because it interferes with pitch-based auditory stream segregation, making it harder for listeners to identify the constituent voices in a chord progression. Correspondingly, we define a Boolean feature termed *Part overlap* that returns 1 when part overlap is detected and 0 otherwise. This feature uses Tymoczko's (2006) algorithm to determine voice assignments for each pitch.

5.4 Analysis

We now use our model to analyse a dataset of 370 chorale harmonisations by J. S. Bach, sourced from the virtual music library *KernScores* (Sapp, 2005).[2] These chorales provide a useful baseline application for the model: they are relatively stylistically homogeneous, they have a consistent texture of block chords, and they are considered to be a touchstone of traditional harmonic practice.

These chorales were originally notated as four independent voices. For our analyses, it is necessary to translate these independent voices into sequences of vertical sonorities. We achieve this using *full expansion* (Conklin, 2002): we create a new sonority at each timepoint when a new note onset occurs, with this sonority comprising all pitches already sounding or starting to sound at that timepoint. Because of embellishments such as passing notes and appoggiaturas, these sonorities do not correspond to chords in the conventional sense; deriving a conventional chord sequence would require the services of either a music theorist or a harmonic reduction algorithm (e.g. Pardo & Birmingham, 2002; Rohrmeier & Cross, 2008). We therefore use the term 'sonority' to identify the collections of pitch classes identified by the full-expansion algorithm.

Our sequential features (e.g. voice-leading distance) are undefined for the starting sonority in each chorale. We therefore omit all starting sonorities from the model-fitting process. An alternative approach would be to set all sequential features to zero for these starting sonorities.

One of our features – 'Number of pitches (difference from ideal)' – is intended to capture the default number of pitches in each voicing for a particular musical style. Since all the chorales in our dataset have four voices, all of which tend to sing throughout the chorale, we set the ideal number of pitches to four.

The model supposes that each sonority has a finite set of candidate voicings. For a given sonority, we enumerate all candidate voicings that satisfy the following conditions:

a) All pitches must range between C2 (65.41 Hz) and B5 (987.77 Hz) inclusive;[3]

b) The voicing must represent the same pitch-class set as the original sonority;

c) The voicing and the original sonority must share the same bass pitch class;

[2] The collection was originally compiled by C. P. E. Bach and Kirnberger, and later encoded by Craig Sapp. The encoded dataset omits chorale no. 50, the only chorale not in four parts. This dataset is available as the `bach_chorales_1` dataset in the *hcorp* package (`https://github.com/pmcharrison/hcorp`). Source code for our analyses is available at `https://doi.org/10.5281/zenodo.2613563`.

[3] This range needs to be generous enough to include all reasonable voicing candidates, but conservative enough to keep the analysis computationally tractable.

d) The voicing must contain between one and four distinct pitches, reflecting the fact that the chorales were originally written for four voice parts.

Before beginning the analysis, it is worth acknowledging two simplifications we have made when modelling Bach's composition process. First, Bach took his soprano lines from pre-existing chorale melodies, and only composed the lower parts; in contrast, our model recomposes the melody line as well as the lower parts. Correspondingly, our model is not really a simulation of chorale harmonisation, but rather a simulation of the Bach chorale style itself. Second, our model assumes that the sonorities are fixed in advance of constructing the voice leadings, which is arguably unrealistic given that the sonorities derived from full expansion include embellishments that are themselves motivated by voice leading, such as passing notes. This simplification is useful for making the analysis tractable, but future work could investigate ways of modelling interactions between harmony and voice leading.

5.4.1 Performance

Having fitted the voice-leading model to the corpus, we assess its performance by iterating over each sonority in the corpus and assessing the model's ability to reproduce Bach's original voicings. Different performance metrics can be defined that correspond to different methods for sampling voicings. One approach is to select the voicing with the maximum probability according to the model: in this case, the model retrieves the correct voicing 63.05% of the time. A second approach is to sample randomly from the model's probability distribution: in this case, the model has an average success rate of 44.63%. A third approach is to sample voicings from the model in descending order of probability, until the correct voicing is recovered: on average, this takes 2.55 samples, corresponding to 2.14% of the available voicings. Given that there are on average 102.96 available voicings for each sonority, these figures suggest fairly good generative choices.

5.4.2 Moments

The 'Moments' portion of Table 5.1 describes feature distributions in the original corpus. For example, the first entry indicates that the mean voice-leading distance between successive voicings is 5.96, with a standard deviation of 4.29. Given that these chorales are each voiced in four parts, this implies that each voice part moves on average by 1.49 semitones between each voicing.

It is interesting to examine features corresponding to strict rules that we might expect never to be violated in Bach's work. For example, parallel oc-

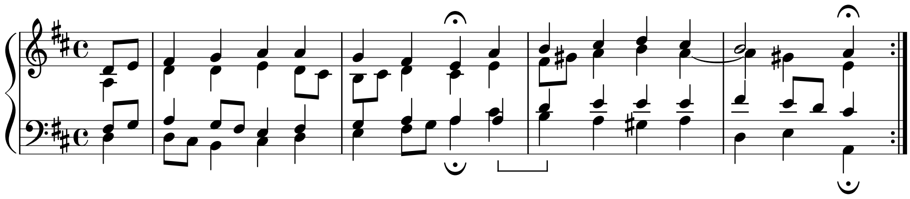

Figure 5.1: J. S. Bach, *Mach's mit mir, Gott, nach deiner Güt'*, BWV 377, bb. 1–4. The two chords immediately after the first fermata imply parallel fifths and octaves that have been only partly mitigated by swapping the bass and tenor parts.

taves and fifths are often taught to music students as unacceptable violations of common-practice style, yet our analysis identifies such voice leadings in 1.09% of Bach's progressions. These cases often correspond to passages where Bach introduced voice crossings to avoid parallel progressions (e.g. Figure 5.1); such voice crossings have no impact on our algorithm, which recomputes all voice leadings using Tymoczko's (2006) algorithm. We decided not to remove such cases, because voice reassignment arguably only partially eliminates the aesthetic effect of these parallel progressions, and because we wish the algorithm to generalise to textures without explicit voice assignment.

5.4.3 Weights

The 'Weights' portion of Table 5.1 lists optimised weights for each feature, alongside the corresponding standard errors and p-values. Consider the voice-leading distance weight, which takes a value of -0.37: this means that increasing voice-leading distance by one semitone modifies a voicing's predicted probability by a factor of $\exp(-0.37) = 0.69$. Similarly, the melodic voice-leading distance weight takes a value of -0.24: this means that increasing melodic voice-leading distance by one semitone modifies predicted probability by a factor of $\exp(-0.24) = 0.79$, in addition to the penalisation induced by the overall voice-leading distance measure.

Similar reasoning applies to Boolean features, which can only take two values: 'true' (coded as 1) or 'false' (coded as 0). For example, part overlap has a weight of -0.67, meaning that a voicing with overlapping parts is $\exp(-0.67) = 0.51$ times less likely to occur than an equivalent voicing without overlapping parts. Part overlap is therefore a moderate contributor to voice-leading decisions: something to be avoided but not prohibited. Parallel octaves and fifths, meanwhile, are almost prohibited. Parallel progressions between outer parts are penalised particularly heavily; such progressions reduce a voicing's probability by a factor of $\exp(-2.49 - 2.32) = 0.01$.

5.4.4 Feature importance

It is difficult to make meaningful comparisons between the weights of continuous features, because each must be expressed in the units of the original feature. This problem is addressed by the permutation-based feature importance metrics in Table 5.1. These metrics operationalise feature importance as the drop in model performance observed when the trained model is evaluated on a dataset (in this case the Bach chorale corpus) where the feature is randomly permuted (Breiman, 2001; Fisher et al., 2018).[4] Table 5.1 presents two feature importance metrics corresponding to two previously presented performance metrics: the accuracy of maximum-probability samples and the accuracy of random samples. Both metrics indicate that voice-leading efficiency, particularly in the melody line, is the primary contributor to model performance.

It is worth noting that a large feature weight can accompany a small feature importance. For example, parallel fifths/octaves between the outer parts yields a relatively large weight of -2.32, but a relatively small feature importance of 0.01 (maximum-probability sampling). This can be rationalised by the observation that parallel fifths/octaves between the outer parts is essentially prohibited in common-practice voice leading (hence the large weight), but this rule only excludes a tiny proportion of possible voice leadings (hence the small feature importance).

It is also worth noting how each feature's importance will necessarily depend on which other features are present. For example, the weight attributed to 'mean pitch height (distance from C4)' is likely to be attenuated by voice-leading distance, because if the previous voicing already had an good mean pitch height, and the next voicing only differs by a small voice-leading distance, then the next voicing is guaranteed to have a fairly good mean pitch height. As a result, the 'mean pitch height' feature only needs to give a slight nudge in the appropriate direction to prevent mean pitch height from wandering over time.

5.4.5 Statistical significance

Two features received regression weights that did not differ statistically significantly from zero: *Exposed octaves (outer parts)* and *Number of pitches (difference from previous)*. The lack of statistical significance for the exposed-octaves feature is particularly interesting, given how commonly Western music pedagogy prohibits these progressions. Examining the *Moments* column of Table 5.1, it is clear that such progressions are extremely rare in the chorale dataset, which is surprising given the minimal contribution of the corresponding feature. This suggests that these progressions are being penalised by other features. Three

[4] Note that the feature is only permuted in the test dataset, not the training dataset.

such features seem particularly relevant: *Voice-leading distance*, *Melodic voice-leading distance*, and *Mean pitch height.* According to our definitions, exposed octaves only occur when both outer parts move by three or more semitones; such large movements are likely to be heavily penalised by the voice-leading distance features. Furthermore, the two voices must progress in similar motion, thereby inducing a significant change in mean pitch height. Assuming that the previous voicing was already at a suitable mean pitch height, this is likely to take the voicing to an unsuitable mean pitch height, resulting in penalisation by the *Mean pitch height* feature. In sum, therefore, it seems plausible that the exposed-octaves feature is made redundant by the other features.

The non-significant contribution of the feature *Number of pitches (difference from previous)* is arguably unsurprising given the corpus being modelled. Each of these chorales is written for four voices, and so the primary pressure on the number of pitches in the sonority is likely to be the goal of providing these four voices with distinct lines; deviations from this four-pitch norm are generally rare and quickly resolved. This phenomenon can be captured by the feature *Number of pitches (difference from ideal)*, making the feature *Number of pitches (difference from previous)* unnecessary. However, this latter feature may become more important in corpora where the number of voices is less constrained, such as in keyboard music.

Table 5.1: Descriptive and inferential statistics for the 12 voice-leading features as applied to the Bach chorale dataset.

	Moments		Weights			Feature importance	
Feature	*M*	*SD*	Value	*SE*	*p*	Max. probability	Random sample
Voice-leading distance	5.961	4.295	−0.375	0.003	< .001	.553	.379
Melodic voice-leading distance	1.133	1.486	−0.240	0.006	< .001	.164	.100
Treble pitch height (distance above C5)	0.742	1.430	−0.237	0.008	< .001	.103	.052
Bass pitch height (distance below C3)	0.801	1.814	−0.173	0.006	< .001	.072	.038
Interference (Hutchinson & Knopoff, 1978)	0.189	0.076	−8.653	0.231	< .001	.068	.037
Parallel octaves/fifths (Boolean)	0.011	0.104	−2.489	0.062	< .001	.056	.033
Number of pitches (difference from ideal)	0.080	0.278	−1.321	0.035	< .001	.056	.032
Mean pitch height (distance from C4)	2.506	1.793	−0.128	0.005	< .001	.031	.018
Part overlap (Boolean)	0.028	0.164	−0.669	0.041	< .001	.013	.013
Parallel octaves/fifths (outer parts; Boolean)	0.001	0.023	−2.323	0.270	< .001	.008	.004
Exposed octaves (outer parts; Boolean)	0.001	0.037	(0.204)	(0.164)	.214	.000	.000
Number of pitches (difference from previous)	0.125	0.336	(0.008)	(0.034)	.822	.000	.000

Note. Moments provides the mean and standard deviation of feature values in the Bach chorale dataset. *Weights* provides the regression weights for each feature, alongside corresponding standard errors and *p*-values. *Feature importance* provides permutation-based importance metrics for each feature.

5.5 Generation

The probabilistic model developed in the previous section can be directly applied to the automatic generation of voice leadings for chord sequences. Given a prespecified chord sequence, the model defines a probability distribution over all possible voice leadings for that chord sequence, which factorises into probability distributions for each chord voicing conditioned on the previous chord voicing. It is straightforward to sample from this factorised probability distribution: simply iterate from the first to the last chord in the sequence, and sample each voicing according to the probability distribution defined by the log-linear model, using the sampled voicing at position i to define the feature set for chord voicings at position $i + 1$.

If our goal is to approximate a target corpus as well as possible, then this random sampling is a sensible approach. However, if our goal is to generate the best possible voice leading for a chord sequence, then we must identify some objective function that characterises the quality of a chord sequence's voice leading and optimise this objective function.

Here we propose optimising the sum of the model's linear predictors. As defined previously, the linear predictor characterises a given chord voicing as a weighted sum of feature values, with this linear predictor being exponentiated and normalised to estimate the probability of selecting that voicing. The linear predictor might be interpreted as the attractiveness of a given voicing, as inversely related to features such as voice-leading distance and interference between partials.

Optimising the sum of the linear predictors is subtly different to optimising for probability. Optimising for probability means maximising the ratio of the exponentiated linear predictors for the chosen voicing to the exponentiated linear predictors for the alternative voicings. This maximisation does not necessarily entail high values of the linear predictor; in perverse cases, high probabilities may be achieved when the chosen voicing is simply the best of a very bad set of candidates. We wish to avoid such cases, and to identify chord voicings that possess good voice-leading attributes in an absolute sense, not simply relative to their local competition.

The space of all possible voice leadings is large: given 100 candidate voicings per chord, a sequence of 80 chords has 10^{160} potential voice-leading solutions. It is clearly impractical to enumerate these voice leadings exhaustively. A simple 'greedy' strategy would be to choose the chord voicing with the highest linear predictor at each chord position; however, this is not guaranteed to maximise the sum of linear predictors across all chord positions. Instead, we take a dynamic-programming approach that deterministically retrieves the optimal voice-leading solution while restricting the number of linear predictor evaluations to approximately a^2n, where a is the number of candidate voicings for each chord and n is the number of chords in the sequence. This approach simplifies the computation by taking advantage of the fact that none of our features look back beyond the previous chord's voicing. See Algorithm 1 for details.

Several of the features, such as voice-leading distance and part overlap, are undefined for the first chord in the sequence. Correspondingly, the first chord of each sequence was excluded from the model-fitting process described in Section 5.4. When generating from the model, however, it is inappropriate to exclude these chords from the optimisation. Instead, we set all context-dependent features to zero for the first chord of each sequence (in fact, any numeric constant would have the same effect). The initial chord voicings are then optimised according to the context-independent features, such as interference between partials and mean pitch height.

Figure 5.2 demonstrates the algorithm on the first ten sonorities of the chorale dataset: *Aus meines Herzens Grunde*, BWV 269.[5] For comparison purposes, Figure 5.2A displays J. S. Bach's original voice leading, and Figure 5.2B displays a heuristic voice leading where the bass pitch class is played in the octave below middle C and the non-bass pitch classes are played in the octave above middle C, as in Chapter 4. Figure 5.2C displays the voice leading produced by the new algorithm, using regression weights as optimised on the original corpus, and generating candidate chords according to the same procedure as described in Section 5.4. Unlike the heuristic algorithm, the new algorithm consistently employs four notes in each chord, similar to the original chorale harmonisation. The new algorithm successfully avoids the two parallel fifths produced in the last two bars by the heuristic algorithm, and achieves considerably more efficient voice leading throughout.

In chorale harmonisations the soprano line is typically constrained to follow the pre-existing chorale melody. We can reproduce this behaviour by modifying the candidate voicing generation function so that it only generates voicings with the appropriate soprano pitches. Figure 5.2D displays the voice leading produced when applying this constraint. Our implementation also supports further constraints such as forcing particular chord voicings to contain particular pitches, or alternatively fixing entire chord voicings at particular locations in the input sequence.

We were interested in understanding how the trained model would generalise to different musical styles. In harmony perception studies, it is often desirable to present participants with chord sequences derived from pre-existing music

[5] Source code is available at `https://doi.org/10.5281/zenodo.2613563`. Generated voice leadings for all 370 chorales are available at `https://doi.org/10.5281/zenodo.2613646`.

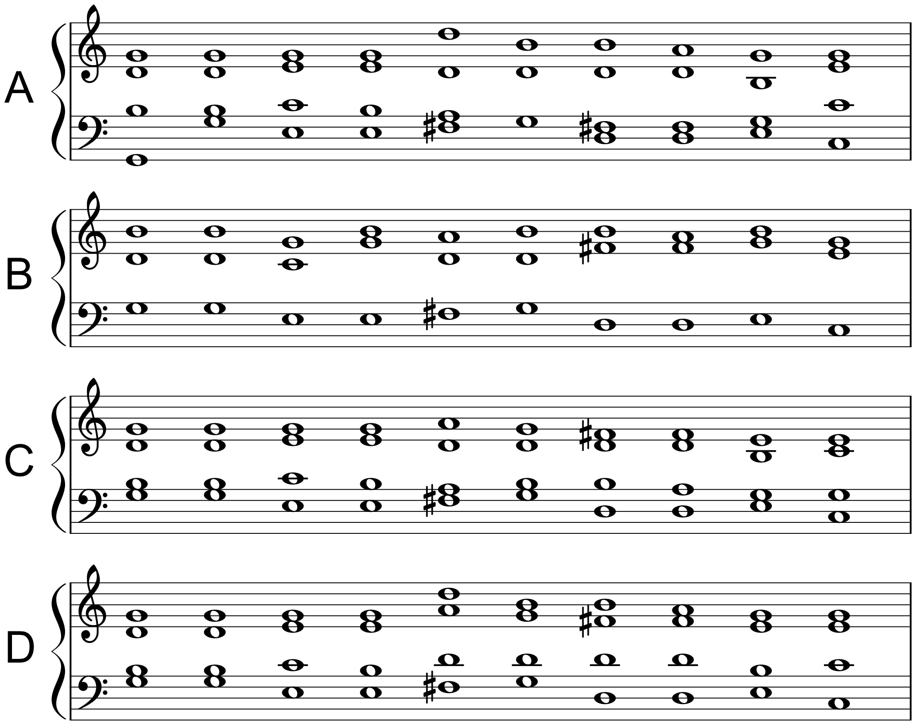

Figure 5.2: Example voice leadings for J. S. Bach's chorale *Aus meines Herzens Grunde* (BWV 269), chords 1–10. **A**: Bach's original voice leading. **B**: Heuristic voice leading. **C**: New algorithm. **D**: New algorithm with prespecified melody.

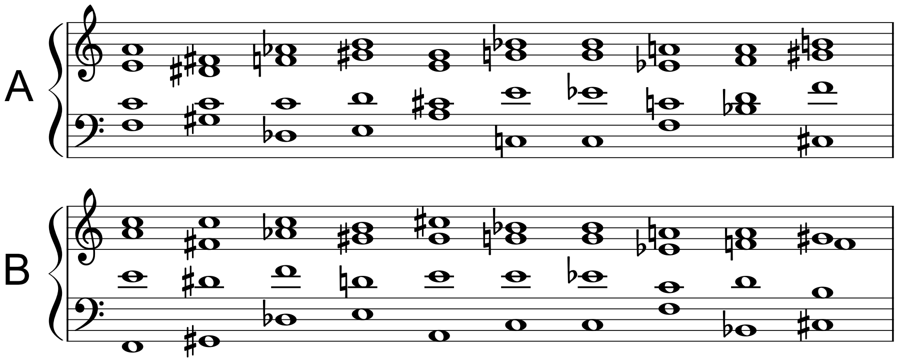

Figure 5.3: Example voice leadings for the first 10 chords of John Coltrane's *26-2*. **A**: Heuristic voice leading. **B**: New algorithm.

corpora, such as the McGill Billboard corpus (Burgoyne, 2011) and the iRb corpus (Broze & Shanahan, 2013). Unfortunately, these corpora just provide chord symbols, not fully voiced chords, and so the researcher is tasked with creating voice leadings for these chord sequences. We had yet to identify suitable

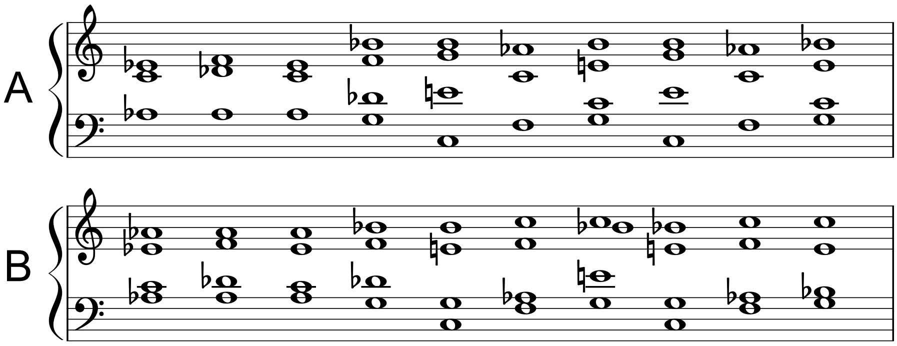

Figure 5.4: Example voice leadings for the first 10 chords of *You've got a friend* by Roberta Flack and Donny Hathaway. **A**: Heuristic voice leading. **B**: New algorithm.

datasets of voiced chord sequences for popular or jazz music, and therefore wished to understand whether Bach chorales would be sufficient for training the algorithm to generate plausible voice leadings for these musical styles.

From an auditory scene analysis perspective, there are clear differences between Bach chorales and popular/jazz harmony. The chorales consistently use four melodically independent voices, and Bach's voice-leading practices are consistent with the compositional goal of maximising the perceptual independence of these voices while synchronising text delivery across the vocal parts (Huron, 2001, 2016). In contrast, harmony in popular and jazz music is often delivered by keyboards or guitars, both of which produce chords without explicit voice assignment, with the number of distinct pitches in each chord often varying from chord to chord. Correspondingly, voice independence seems likely to be less important in popular/jazz harmony than in Bach chorales. Nonetheless, we might still expect popular/jazz musicians to pay attention to the perceptual independence of the outer parts, since these voices are particularly salient to the listener even when the voice parts are not differentiated by timbre. We might also expect popular/jazz listeners to prefer efficient voice leadings, even if they are not differentiating the chord progression into separate voices, because efficient voice leading helps create the percept of a stable textural stream (Huron, 2016). In summary, therefore, there are reasons to expect some crossover between voice-leading practices in Bach chorales and voice-leading practices in popular/jazz music.

Figures 5.3 and 5.4 demonstrate the application of the chorale-trained model to examples from two such corpora: the iRb jazz corpus (Broze & Shanahan, 2013) and the Billboard popular music corpus (Burgoyne, 2011). We use both datasets as translated to pitch-class notation in Chapter 2, and use the same model configuration as for the Bach chorale voicing.

Figures 5.3A and 5.3B correspond to the first ten bars of the jazz corpus, from the composition *26-2* by John Coltrane. Figure 5.3A displays the heuristic algorithm described earlier, and Figure 5.3B displays the new algorithm's voicing. Informally, the new algorithm seems to generalise sensibly to this extract, despite the extract's radical harmonic differences from the Bach chorale harmonisations. The clearest difference from the heuristic algorithm seems to be in the spacing of the voices; the new algorithm enforces a wide gap between the lower voices, presumably on account of the *Interference between partials* feature.

Figures 5.4A and 5.4B correspond to the first ten bars of *You've got a friend* by Roberta Flack and Donny Hathaway, the second composition in the popular corpus.[6] As before, Figures 5.4A and 5.4B correspond to the heuristic and new algorithms respectively. Unlike the heuristic algorithm, the new algorithm maintains four-note voicings at all times, arguably producing a richer and more consistent sound as a result. The voice-leading efficiency is also considerably improved, particularly in the melody line. At first sight, the new algorithm does also produce some unusual voice leadings: for example, the tenor part jumps by a tritone between the fourth chord and the fifth chord. One might expect this inefficient voice leading to be heavily penalised by the model. However, the model considers this voice leading to be relatively efficient, as Tymoczko's (2006) algorithm connects the two voicings by approaching the lower two notes of the fifth chord (C, G) from the bass note of the previous chord (G), and approaching the second-from-top note in the fifth chord (E) from the second-from-bottom note in the fourth chord (D flat). This suggested voice assignment is indeed plausible when the extract is performed on a keyboard instrument, but it could not be realised by a four-part ensemble of monophonic instruments. For such applications, it would be worth modifying Tymoczko's (2006) algorithm to set an upper bound on the number of inferred voices.

5.6 Implementation

We have implemented these algorithms in an open-source software package called *voicer*, written for the R programming language, and coded in a mixture of R and C++. The source code is available from the open-source repository `https://github.com/pmcharrison/voicer` and permanently archived at `https://doi.org/10.5281/zenodo.2613565`.

[6] We originally tried the first composition in this corpus, but it was too repetitive to give much insight into the new algorithm.

Having installed the *voicer* package, the following code instructs the package to voice a perfect (or authentic) cadence:

```
library(voicer)
library(hrep)
library(magrittr)
# Each chord is represented as a sequence of MIDI note numbers.
# The first number is the bass pitch class.
# The remaining numbers are the non-bass pitch classes.
list(pc_chord("0 4 7"), pc_chord("5 0 2 9"),
     pc_chord("7 2 5 11"), pc_chord("0 4 7")) %>%
  vec("pc_chord") %>%
  voice(opt = voice_opt(verbose = FALSE)) %>%
  print(detail = TRUE)
```

```
## [[1]] Pitch chord: 48 64 67 72
## [[2]] Pitch chord: 53 62 69 72
## [[3]] Pitch chord: 55 62 65 71
## [[4]] Pitch chord: 48 55 64 72
```

By default, *voicer* uses the same regression weights and voicing protocol as presented in the current chapter. However, it is easy to modify this configuration, as demonstrated in the following example:

```
library(voicer)
library(hrep)
library(magrittr)
chords <- list(pc_chord("0 4 7"), pc_chord("5 0 9"),
               pc_chord("7 2 11"), pc_chord("0 4 7")) %>%
  vec("pc_chord")
# Modify the default weights to promote parallel fifths/octaves
weights <- voice_default_weights
weights["any_parallels"] <- 100
voice(chords, opt = voice_opt(verbose = FALSE,
                              weights = weights,
                              min_notes = 3,
                              max_notes = 3)) %>%
  print(detail = TRUE)
```

```
## [[1]] Pitch chord: 48 55 64
## [[2]] Pitch chord: 53 60 69
## [[3]] Pitch chord: 55 62 71
## [[4]] Pitch chord: 60 67 76
```

The *voicer* package also exports functions for deriving regression weights from musical corpora, and using these new weights to parametrise the voicing algorithm. The following example derives regression weights from the first two pieces in the Bach chorale dataset, and uses these weights to voice a chord sequence.

```
if (!requireNamespace("hcorp"))
  devtools::install_github("pmcharrison/hcorp")
library(voicer)
library(hrep)
# Choose the features to model
features <- voice_features()[c("vl_dist", "dist_from_middle")]
# Compute the features
corpus <- hcorp::bach_chorales_1[1:2]
corpus_features <- voicer::get_corpus_features(
  corpus, min_octave = -2, max_octave = 1, features = features,
  revoice_from = "pc_chord", min_notes = 1, max_notes = 4,
  verbose = FALSE)
# Model the features
mod <- model_features(corpus_features,
                      perm_int = FALSE,
                      verbose = FALSE)
as.data.frame(mod$weights)
```

```
##            feature   estimate    std_err          z            p
## 1          vl_dist -0.5320266 0.03441282 -15.460129 6.446884e-54
## 2 dist_from_middle -0.1910925 0.06838340  -2.794428 5.199166e-03
```

```
# Voice a chord sequence
chords <- list(pc_chord("0 4 7"), pc_chord("5 0 9"),
               pc_chord("7 2 11"), pc_chord("0 4 7")) %>%
  vec("pc_chord")
voice(chords, opt = voice_opt(weights = mod,
                              features = features,
                              verbose = FALSE)) %>%
  print(detail = TRUE)
```

```
## [[1]] Pitch chord: 36 52 72 79
## [[2]] Pitch chord: 41 53 72 81
## [[3]] Pitch chord: 43 50 71 79
## [[4]] Pitch chord: 48 52 72 79
```

5.7 Discussion

We have introduced a new model for the analysis and generation of voice leadings. This model uses perceptually motivated features to predict whether a given voice leading will be considered appropriate in a particular musical context. Applied to a dataset of 370 chorale harmonisations by J. S. Bach, this model delivered quantitative evidence for the relative importance of different musical features in determining voice leadings. Applied to generation, the model demonstrated an ability to create plausible voice leadings for pre-existing chord sequences, and to generalise to musical styles dissimilar to the Bach chorales upon which it was trained.

Combining analysis with generation provides a powerful way to examine which principles are sufficient to explain voice-leading practice. While the analysis stage provides quantitative support for the importance of different musical features in voice leading, the generation stage can provide a litmus test for the sufficiency of the resulting model. Examining the outputs of the model, we can search for ways in which the model deviates from idiomatic voice leading, and test whether these deviations can be rectified by incorporating additional perceptual features into the model. If so, we have identified an additional way in which voice leading may be explained through auditory perception, after Huron (2001, 2016); if not, we may have identified an important cultural component to voice-leading practice. To this end, we have released automatically generated voicings for the full set of 370 Bach chorale harmonisations;[7] we hope that they will provide useful material for identifying limitations and potential extensions of the current approach.

The existing literature already suggests several additional features that might profitably be incorporated into the voice-leading model. One example is the ‘leap away rule’, which states that large melodic intervals (leaps) are better situated in the outer voices than the inner voices, and that these intervals should leap away from the other voices rather than towards the other voices (Huron, 2016). This should be straightforward to implement computationally. A second example is the ‘follow tendencies rule’, which states that the progressions of individual voices should follow the listener’s expectations, which may themselves derive from the statistics of the musical style (*schematic*

[7] https://doi.org/10.5281/zenodo.2613646

expectations), the statistics of the current musical piece (*dynamic expectations*), or prior exposure to the same musical material (*veridical expectations*) (Huron, 2016). Schematic expectations could be operationalised by using a dynamic key-finding algorithm to represent the sonority as scale degrees (e.g. Huron & Parncutt, 1993), and then evaluating the probability of each scale-degree transition with respect to a reference musical corpus; dynamic expectations could be operationalised in a similar manner, but replacing the reference corpus with the portion of the composition heard so far. Veridical expectations would require more bespoke modelling to capture the particular musical experience of the listener. An interesting possibility would be to unite these three types of expectation using Pearce's (2005) probabilistic model of melodic expectation (see also Sauvé, 2017). Further rules that could be implemented include the 'semblant motion rule' (avoid similar motion) and the 'nonsemblant preparation rule' (avoid similar motion where the voices employ unisons, octaves, or perfect fifths/twelfths) (Huron, 2016).

Our model also has practical applications in automatic music generation. For example, a recurring problem in music psychology is to construct experimental stimuli representing arbitrary chord sequences, which often involves the time-consuming task of manually constructing voice leadings. Our model could supplant this manual process, bringing several benefits including a) *scalability*, allowing the experimental design to expand to large stimulus sets; b) *objectivity*, in that the voice leadings are created according to formally specified criteria, rather than the researcher's aesthetic intuitions; c) *reproducibility*, in that the methods can be reliably reproduced by other researchers.

Interpreted as a model of the compositional process, the model assumes that chords are determined first and that voice leading only comes later. This may be accurate in certain musical scenarios, such as when performers improvise from figured bass or from lead sheets, but it is clearly not a universal model of music composition. A more universal model might include some kind of alternation between composing the harmonic progression and composing the voice leading, so that the composer can revise the harmonic progression if it proves impossible to find a satisfactory voice leading.

The model also assumes a one-to-one mapping between the chords of the underlying harmonic progression and the chord voicings chosen to represent it. While this assumption may hold true for certain musical exercises, it is not universally valid for music composition. For example, an improviser playing from figured bass may choose to extend a single notated chord into multiple vertical sonorities, for example through arpeggiation or through the introduction of passing notes. It would be interesting to model this process explicitly. One approach would be to use the original model to generate block chords at the level of the harmonic rhythm, and then to post-process these block chords with an additional algorithm to add features such as passing notes and ornamentation.

While the model deserves further extension and validation, it seems ready to support ongoing research in music psychology, music theory, and computational creativity. Our R package, *voicer*, should be useful in this regard: It provides a convenient interface for analysing voice leadings in musical corpora and for generating voice leadings for chord sequences. The ongoing development of this package may be tracked at its open-source repository (`https://github.com/pmcharrison/voicer`).

Algorithm 1: A dynamic programming algorithm for maximising the sum of the linear predictors f over all chord transitions.

```
input  : $candidates$, a list of length $N$; $candidates[i]$ lists the candidate
         voicings for chord $i$
output: $chosen$, a list of length $N$; $chosen[i]$ identifies the chosen voicing for
         chord $i$
$best\_scores \leftarrow list(N)$
$best\_prev\_states \leftarrow list(N)$
$best\_scores[1] \leftarrow vector(length(candidates[1]))$
for $j \leftarrow 1$ to $length(candidates[1])$ do
    $best\_scores[1][j] \leftarrow f(\mathsf{NULL}, candidates[1][j])$
end
for $i \leftarrow 2$ to $N$ do
    $best\_scores[i] \leftarrow vector(length(candidates[i]))$
    for $j \leftarrow 1$ to $length(candidates[i])$ do
        $best\_prev\_states[i][j] \leftarrow 1$
        $best\_scores[i][j] \leftarrow f(candidates[i-1][1], candidates[i][j])$
        for $k \leftarrow 2$ to $length(candidates[i-1])\}$ do
            $new\_score \leftarrow f(candidates[i-1][k], candidates[i][j])$
            if $new\_score > best\_scores[i][j]$ then
                $best\_prev\_states[i][j] \leftarrow k$
                $best\_scores[i][j] \leftarrow new\_score$
            end
        end
    end
end
$chosen \leftarrow vector(N)$
$chosen[N] \leftarrow which\_max_j(best\_scores[N][j])$
for $n \leftarrow N-1$ to $1$ do
    $chosen[n] \leftarrow best\_prev\_states[n+1][chosen[n+1]]$
end
return $chosen$
```

Chapter 6

Conclusion

6.1 Overview

The primary goal of this thesis was to develop an improved understanding of harmony cognition through the computational modelling of large datasets of perceptual data and music compositions. A secondary goal was to design, implement, and evaluate computational models of harmony that subsequent scientists and music theorists could take advantage of in future research. These goals motivated the research described in Chapters 2–5.

Each of the computational studies depends on a shared collection of low-level cognitive representations for harmony, introduced in Chapter 2. Low-level representations correspond to the early stages of cognitive processing, and are linked relatively unambiguously to the musical score and the audio signal. We defined three classes of low-level representations: *symbolic* representations, defined as succinct and categorical descriptions of chords, *acoustic* representations, which characterise the musical sound, and *sensory* representations, which reflect the listener's perceptual images of the resulting sound. We defined a network of representations organised into these three categories and explained the computational operations involved in translating between these different representations. We also defined methods for mapping four of the symbolic representations to integer encodings, an important prerequisite for many statistical modelling techniques. Finally, we discussed ways of deriving these representations from common corpus encodings, and reviewed methods for deriving high-level cognitive representations such as chord roots and tonality from these low-level representations. The chapter is accompanied by the R package *hrep*, which implements the various representations and conversion methods described in the chapter, as well as the *hcorp* package, which provides several musical corpora encoded using these representation schemes.

The following chapters – Chapters 3, 4, and 5 – addressed three core phenomena in harmony cognition: consonance, harmonic expectation, and voice leading. A recurring research question across these chapters was as follows:

> "To what extent is the perception and composition of Western harmony determined by low-level psychoacoustic processes versus high-level cognitive processes?"

Chapter 3 addressed simultaneous consonance, the sense in which certain combinations of tones sound 'well' together. We began by defining three classes of consonance theories with particular support from the existing literature: interference between partials, periodicity/harmonicity, and cultural familiarity. Interference between partials is the most low-level of these phenomena, relating primarily to the resolution of partials on the basilar membrane of the inner ear (Daniel & Weber, 1997; Vencovský, 2016). Periodicity/harmonicity is a mid-level phenomenon, relating to pitch detection processes within the auditory cortex (Wang et al., 2013). Cultural familiarity is a relatively high-level cognitive phenomenon, corresponding to the listener's internalised knowledge of the prevalences of different chord types within a given musical style. We compiled computational operationalisations of each of these classes of consonance theory, numbering 20 models in total, and evaluated these models on a) four behavioural studies of consonance judgments representing more than 500 participants and b) three large music corpora representing more than 100,000 compositions. While recent work has argued that consonance perception is independent of interference between partials, we identified a substantial contribution of this interference phenomenon in both our perceptual and our corpus analyses. These analyses also identified sizeable contributions of periodicity/harmonicity and cultural familiarity, suggesting that consonance is not a unitary phenomenon but rather a composite phenomenon deriving from a combination of low-level, mid-level, and high-level psychological processes. This composite account is represented by a three-factor computational model that combines Hutchinson & Knopoff's interference model, a new harmonicity detection model, and a new data-driven cultural familiarity model. The chapter is accompanied by the *incon* package, which implements this composite consonance model alongside 15 other consonance models evaluated in the chapter.

Chapter 4 addressed harmonic expectation, the way in which certain chord progressions set up expectations for the chords that follow them. We formalised harmonic expectation probabilistically, supposing that the brain continually constructs predictive probability distributions over the alphabet of all possible chords, with these probability distributions being conditioned on the preceding chords in the progression. Noting the rich variety of perceptual features that listeners can derive from chord progressions – for example, harmonicity, chord roots, root intervals, and pitch-class sets – we modelled harmonic prediction using a new technique termed 'viewpoint regression', which generates predictions for chord symbols using statistical regularities acquired from an assortment of categorical and continuous perceptual features. We then used this model to investigate what strategies an ideal listener would adopt when predicting chord progressions, and how these strategies might manifest in everyday music listening. We found that the ideal listener paid particular attention to two low-level consonance features, namely interference between partials (after Hutchinson & Knopoff, 1978) and periodicity/harmonicity (the new model presented in Chapter 3), as well as a high-level feature that captured relative pitch perception by expressing each chord relative to the bass note of the previous chord. We followed this ideal-listener analysis with a human-listener analysis, where we conducted a behavioural experiment to elicit surprisal ratings for 300 chord sequences excerpted from real popular music compositions, and compared the resulting data to our ideal-listener model. The ideal-listener model reproduced human performance fairly well, predicting mean surprisal ratings with a correlation coefficient of .70. Interrogating this model further, we found that variations in perceived surprisal were primarily driven by a simple repetition-priming process, whereby listeners expect chords to recur if they have already occurred in the recent musical context. The simplicity of this repetition-priming process makes a remarkable contrast with the complexity of recent language-inspired theories of harmonic syntax, but is consistent with Bigand et al.'s (2014) assertion that much of music syntax processing can be explained by the accumulation and comparison of information in auditory short-term memory.

Chapter 5 addressed voice leading, which describes how a chord's pitch classes are assigned pitch heights, and how the pitches in successive chords connect to form simultaneous melodies. We developed a cognitive model to characterise this process, which uses various perceptually motivated features to predict whether a composer will choose a given voice leading in a given musical context. These features were motivated by Huron's (2001, 2016) perceptual account of voice leading, which relates Western voice-leading practice to the goal of manipulating psychological processes of auditory scene analysis, by which the listener organises the acoustic environment into perceptually meaningful elements (Bregman, 1990). We applied this model to a corpus of 370 chorale harmonisations by J. S. Bach, and derived quantitative estimates for the relative importance of various perceptual features such as interference between partials, voice-leading distance, and part overlap. We found that almost all of the proposed features contributed meaningfully to voice-leading practice, with the exception of exposed octaves, which seemed to be made redundant by a combination of the voice-leading distance and pitch height features. We then examined the model's ability to generate new voice leadings for pre-existing chord sequences, and found that the model seemed to perform rather effectively, even when applied to quite dissimilar musical styles to the original Bach chorales, such as popular music and jazz music. These results imply that, as argued by Huron (2001, 2016), much of voice-leading practice can be effectively summarised by a few perceptual principles related to auditory scene analysis, rather than being dominated by arbitrary cultural conventions.

A common conclusion from these studies is clear: many aspects of harmony cognition depend substantially on relatively low-level psychoacoustic processes. Consonance is primarily driven by interference between partials and periodicity/harmonicity (Chapter 3); harmonic expectation in Western popular music is strongly influenced by basic repetition-priming effects (Chapter 4); voice leading reflects psychological processes of auditory scene analysis (Chapter 5).

Nonetheless, higher-level cognitive processes clearly also contribute to harmony cognition. Existing psychoacoustic models could only explain part of the variance in our consonance perception data, and a cultural familiarity model proved useful in explaining an additional portion of this variance. Likewise, psychoacoustic models were an important part of the ideal-listener model of harmony prediction, but the ideal model also took advantage of learned transition probabilities between high-level chord representations. Our voice-leading model only addressed relatively low-level contributions to voice-leading practice, but Huron (2016) has argued that voice-leading practice also depends on culturally learned expectations for melodic progressions.

It seems, therefore, that computational models of harmony cognition must ultimately combine low-level psychoacoustic processing with high-level cognitive processing. Our consonance model achieves this through a linear regression approach, which combines two low-level consonance models (interference between partials, periodicity/harmonicity) with a high-level model of cultural familiarity. Our harmonic expectation model achieves this through a viewpoint regression approach, which generates predictive probability distributions over chord progressions on the basis of statistical regularities learned from both low-level psychoacoustic features (e.g. interference between partials, spectral similarity) and high-level cognitive features (e.g. chord root progressions). Our voice-leading model is currently limited to low-level psychoacoustic features, but future work could add higher-level features to this model, such as a feature capturing culturally determined transition probabilities between scale degrees.

We have implemented these different computational models as interoperable open-source software packages, written for the programming language R. We hope that these packages will prove useful for subsequent research into harmony cognition. Each of these packages depends on the *hrep* package, presented in Chapter 2, which defines a collection of harmonic representation schemes within an object-oriented framework. This *hrep* package should be useful to future researchers wishing to develop new harmony models within the R programming language. Chapter 3 presented the *incon* package, which implements a variety of consonance models from the literature, alongside the composite consonance model developed in the present work. These implementations should be useful to those wishing to conduct further research into the nature of consonance, and to those wishing to find a ready-made operationalisation of consonance for practical applications such as music classification or music generation. Chapter 4 presented the *hvr* package, which implements the harmonic viewpoint regression model; this should be useful to researchers wishing to model the probabilistic processing of harmonic progressions. Chapter 5 presented the *voicer* package, which implements the feature-based statistical model of voice leading practice. One application of this model is corpus analysis: given a symbolically encoded musical corpus, the model will quantify the contributions of different perceptually motivated features (e.g. interference between partials, pitch height) to voice-leading practice. A second application is generative: given a chord progression, the model can generate a voiced version of that chord progression using feature weights that have either been derived from corpus analysis or hand-specified by the user. We expect this generative system to be particularly useful for future studies of harmony perception, allowing researchers to take chord sequences from preexisting musical corpora (e.g. the Billboard corpus, Burgoyne, 2011) and convert them to block chords that can be synthesised as stimuli for perceptual experiments.

6.2 Limitations and future directions

An important priority in this work was to maximise the interpretability of our computational models. Interpretability is crucial if we wish to gain cognitive insights from our models. However, interpretability typically requires simplistic assumptions that can reduce the model's verisimilitude and predictive power. We made several such assumptions in this work. First, we analysed harmony in isolation from metre, yet metre is known to contribute to harmonic structure. Second, our models take symbolic musical representations as input, and hence neglect the ways in which varying acoustic properties of different musical instruments contribute to harmony perception. Third, we adopted relatively restricted statistical models such as linear models, Markov models, and log-linear models, which cannot capture as complex statistical patterns as their connectionist equivalents. Fourth, we glossed over the way in which chord sequences are derived from full musical textures. Music theorists are used to the idea of reducing passages of polyphonic music to their underlying chord sequences, but this is a nontrivial task that has yet to receive a fully satisfying cognitive analysis. These different simplifications would be worthwhile to address in future research.

This research was primarily restricted to modelling Western listeners and Western composers. The motivation for this restriction was twofold: first, harmony is considered to be particularly characteristic of Western music, and second, the majority of publicly available perceptual data and music corpora come from Western listeners and Western composers. However, many cultures across the world also produce music that involves chordal sonorities, and hence presumably evokes many of the psychological mechanisms studied in the present work. Studying such cultures can be particularly useful for differentiating biological, environmental, and cultural contributions to music cognition. A few cross-cultural scientific studies have been conducted into consonance in music perception and composition (Butler & Daston, 1968; Gill & Purves, 2009; Maher, 1976; McDermott et al., 2016; Sethares, 2005), but very little cross-cultural work has been conducted into harmonic expectation and voice leading, particularly from a computational perspective. This is an important avenue for future research.

We addressed three core aspects of harmony cognition: consonance, expectation, and voice leading. There are many other aspects of harmony cognition that would be interesting to study using similar computational methods, including the development of aesthetic preferences, the expression and induction of emotion, and the articulation of large-scale structure. It seems likely that consonance and expectation should be particularly useful for understanding such phenomena; our computational models of consonance and expectation should prove useful for research into this area.

The software documented in this thesis was mostly written for the programming language R. The decision to work in R was guided primarily by its wide collection of statistical packages and its popularity among music psychologists. However, Python is a strong competitor language for this kind of work, in that it already possesses a strong collection of music-related open-source packages such as *music21* and *Essentia*. It would be worthwhile to port some of the software developed here to Python in order to take advantage of this ecosystem.

Our perceptual studies were observational in the sense that they used unmanipulated musical stimuli (chords or chord sequences synthesised using a piano timbre) played to participants with no prior interventions, and tested different psychological theories using correlation- or regression-based analyses. This approach is appealing because it facilitates the use of large and ecologically valid stimulus sets. However, such observational approaches struggle to provide definitive proofs of causation, and so it would be worthwhile to supplement these observational approaches with interventional paradigms. These interventions could happen at the level of the stimulus, for example by introducing timbral manipulations (e.g. Geary, 1980; Nordmark & Fahlén, 1988; Pierce, 1966; Sethares, 2005; Vos, 1986), or at the level of the participant, for example by introducing passive exposure to an artificial harmonic language (e.g. Jonaitis & Saffran, 2009; Loui et al., 2009; Rohrmeier & Cross, 2009). Such interventions could help to distinguish theories that deliver correlated predictions, as is the case in consonance perception (Chapter 3).

Our perceptual studies relied on behavioural methods and did not incorporate any neuroscientific measures. Behavioural methods typically benefit from more straightforward interpretation and greater statistical power than neuroscientific methods, but they are typically mediated by conscious introspection and decision-making. In contrast, neuroscientific methods can complement behavioural methods by providing more direct access to the listener's internal perceptual processes, and helping to reveal the neural substrates underpinning these processes. The computational methods developed in this thesis should be useful for supporting such research. In particular, the computational models should be useful for conducting studies with naturalistic stimuli, providing quantitative operationalisations of different perceptual or cognitive features that can be incorporated into model-based analyses. We are currently conducting several collaborations to explore these possibilities.

The behavioural studies were limited to two response modes: consonance/pleasantness judgements (Chapter 3) and surprisal judgements (Chapter 4). We argued that these two approaches capture core perceptual phenomena in harmony cognition: consonance has long been understood to play a foundational role in Western listeners' aesthetic and emotional responses to music (e.g. Helmholtz, 1863; Stumpf, 1898), and more recently the psychological processes of expectation and surprise have also been recognised as important drivers of musical aesthetics and emotion (Egermann et al., 2013; Huron, 2006; Meyer, 1956; Pearce, 2005). However, our behavioural paradigms were nonetheless fairly distant from how people usually engage with music in the real world. In particular, evaluating music on rating scales is an unusual musical behaviour; music theorists often speak of the consonance of individual intervals or chords, but typically use only two or three categories (e.g. 'perfect consonance', 'imperfect consonance', 'dissonance') rather than the granular numeric scales used here. Moreover, although musical surprisal seemed to be an intuitive concept to our participants, explicitly rating surprisal on a chord-by-chord basis is not a common musical behaviour outside of the laboratory. Limiting consideration to these two behavioural indices of harmony cognition inevitably limits the generalisations that can be made from the present research. An important future goal is to generalise these computational investigations to a broader range of behavioural paradigms that better represent everyday musical behaviours.

Our corpus analyses were limited to four specific musical corpora: the Peachnote corpus (Viro, 2011), the Billboard corpus (Burgoyne, 2011), the iRb corpus (Broze & Shanahan, 2013), and a dataset of Bach chorale harmonisations (Sapp, 2005). The conclusions of these analyses may yet prove to be sensitive to the sampling methods used for constructing these corpora and the techniques used for digitising them. The best way to cement these conclusions will be to follow up these initial investigations with further analyses that explore different musical corpora constructed using alternative sampling and encoding methods.

Chapters 4 and 5 presented cognitive models of harmonic expectation and voice leading. Both are formulated as probabilistic models of musical structure: harmonic expectation is modelled as conditional probabilities of future chords conditioned on previous chords, whereas voice leading is modelled as conditional probabilities of chord voicings conditioned on a prespecified chord sequence. These two models could be combined to form a complete probabilistic model of voiced chord progressions. Such a model could function as a cognitive model of predictive processing, explaining how listeners predict upcoming musical sonorities on the basis of both harmonic and voice-leading principles. The model could also be used for generative applications, automating the generation of fully voiced chord sequences on the basis of statistics learned from a given musical corpus. This remains an interesting prospect for future work.

References

Aarden, B. J. (2003). *Dynamic melodic expectancy* (PhD thesis). Ohio State University, Columbus, OH.

Albrecht, J., & Shanahan, D. (2013). The use of large corpora to train a new type of key-finding algorithm: An improved treatment of the minor mode. *Music Perception*, *31*, 59–67. https://doi.org/10.1525/MP.2013.31.1.59

Ambrazevičius, R. (2017). Dissonance/roughness and tonality perception in Lithuanian traditional Schwebungsdiaphonie. *Journal of Interdisciplinary Music Studies*, *8*(1&2), 39–53. https://doi.org/10.4407/jims.2016.12.002

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051–2056. https://doi.org/10.1016/j.cub.2015.06.043

Arthur, C. (2018). A perceptual study of scale-degree qualia in context. *Music Perception*, *35*, 295–314. https://doi.org/10.1525/MP.2018.35.3.295

Arthur, C., & Huron, D. (2016). The direct octaves rule: Testing a scene-analysis interpretation. *Musicae Scientiae*, *20*, 495–511. https://doi.org/10.1177/1029864915623093

Arthurs, Y., Beeston, A. V., & Timmers, R. (2018). Perception of isolated chords: Examining frequency of occurrence, instrumental timbre, acoustic descriptors and musical training. *Psychology of Music*, *46*(5), 662–681. https://doi.org/10.1177/0305735617720834

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*(4), 321–324. https://doi.org/10.1111/1467-9280.00063

Aures, W. (1984). *Berechnungsverfahren für den Wohlklang beliebiger Schallsignale, ein Beitrag zur gehörbezogenen Schallanalyse* (PhD thesis). Technical University of Munich, Germany.

Aures, W. (1985a). Berechnungsverfahren für den sensorichen Wohlklang beliebiger Schallsignale (A procedure for calculating the sensory consonance of any sound). *Acustica*, *59*(2), 130–141.

Aures, W. (1985b). Der sensorische Wohlklang als Funktion psychoakustischer Empfindungsgrößen (Sensory consonance as a function of psychoacoustic parameters). *Acta Acustica United with Acustica*, *58*(5), 282–290.

Aures, W. (1985c). Ein Berechnungsverfahren der Rauhigkeit. *Acustica*, *58*(5), 268–281.

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain*, *125*(2), 238–251. https://doi.org/10.1093/brain/awf028

Babbitt, M. (1965). The use of computers in musicological research. *Perspectives of New Music*, *3*(2), 74–83.

Bachem, A. (1950). Tone height and tone chroma as two different pitch qualities. *Acta Psychologica*, *7*, 80–88. https://doi.org/10.1016/0001-6918(50)90004-7

Balaguer-Ballester, E., Denham, S. L., & Meddis, R. (2008). A cascade autocorrelation model of pitch perception. *The Journal of the Acoustical Society of America*, *124*, 2186–2195. https://doi.org/10.1121/1.2967829

Barthélemy, J., & Bonardi, A. (2001). Figured bass and tonality recognition. In *Proceedings of the Second International Symposium on Music Information Retrieval* (pp. 129–136). Bloomington, IN.

Bellmann, H. (2005). About the determination of key of a musical excerpt. In R. Kronland-Martinet, T. Voinier, & S. Ystad (Eds.), *Proceedings of the Third International Symposium of Computer Music Modeling and Retrieval* (pp. 76–91). Berlin, Germany: Springer.

Bendor, D., Osmanski, M. S., & Wang, X. (2012). Dual-pitch processing mechanisms in primate auditory cortex. *Journal of Neuroscience*, *32*, 16149–16161. https://doi.org/10.1523/jneurosci.2563-12.2012

Benetos, E., & Dixon, S. (2013). Multiple-instrument polyphonic music transcription using a temporally constrained shift-invariant model. *The Journal of the Acoustical Society of America*, *133*, 1727–1741. https://doi.org/10.1121/1.4790351

Bernstein, J. G. W., & Oxenham, A. J. (2005). An autocorrelation model with place dependence to account for the effect of harmonic number on fundamental frequency discrimination. *The Journal of the Acoustical Society of America*, *117*, 3816–3831. https://doi.org/10.1121/1.1904268

Bharucha, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. *Music Perception*, *5*, 1–30. https://doi.org/10.2307/40285384

Bharucha, J. J., & Pryor, J. H. (1986). Disrupting the isochrony underlying rhythm: An asymmetry in discrimination. *Perception & Psychophysics*, *40*(3), 137–141. https://doi.org/10.3758/bf03203008

Bidelman, G. M., & Heinz, M. G. (2011). Auditory-nerve responses predict pitch attributes related to musical consonance-dissonance for normal and impaired hearing. *The Journal of the Acoustical Society of America*, *130*, 1488–1502. https://doi.org/10.1121/1.3605559

Bidelman, G. M., & Krishnan, A. (2009). Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem.

Journal of Neuroscience, *29*, 13165–13171. https://doi.org/10.1523/JNEUROSCI.3900-09.2009

Bigand, E., Delbé, C., Poulin-Charronnat, B., Leman, M., & Tillmann, B. (2014). Empirical evidence for musical syntax processing? Computer simulations reveal the contribution of auditory short-term memory. *Frontiers in Systems Neuroscience*, *8*. https://doi.org/10.3389/fnsys.2014.00094

Bigand, E., Madurell, F., Tillmann, B., & Pineau, M. (1999). Effect of global structure and temporal organization on chord processing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 184–197. https://doi.org/10.1037/0096-1523.25.1.184

Bigand, E., Parncutt, R., & Lerdahl, F. (1996). Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, *58*(1), 124–141. https://doi.org/10.3758/BF03205482

Bigand, E., & Pineau, M. (1997). Global context effects on musical expectancy. *Perception & Psychophysics*, *59*(7), 1098–1107. https://doi.org/10.3758/BF03205524

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., & Adamo, D. A. D. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *29*, 159–171. https://doi.org/10.1037/0096-1523.29.1.159

Bilsen, F. A. (1977). Pitch of noise signals: evidence for a ”central spectrum”. *The Journal of the Acoustical Society of America*, *61*, 150–161. https://doi.org/10.1121/1.381276

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, *17*(1193), 97–110.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., ... Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In *Proceedings of the 14th International Society for Music Information Retrieval Conference*. Curitiba, Brazil.

Boomsliter, P., & Creel, W. (1961). The long pattern hypothesis in harmony and hearing. *Journal of Music Theory*, *5*, 2–31.

Borchgrevink, H. M. (1975). Musical consonance preference in man elucidated by animal experiments (Norwegian). *Tidsskrift for Den Norske Laegeforening*, *95*(6), 356–358.

Boulanger-Lewandowski, N., Bengio, Y., & Vincent, P. (2013). Audio chord recognition with recurrent neural networks. In *Proceedings of the 14th International Society for Music Information Retrieval Conference* (pp. 335–340). Curitiba, Brazil.

Boulanger-Lewandowski, N., Vincent, P., & Bengio, Y. (2012). Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In J. Langford & J. Pineau (Eds.), *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. Edinburgh, Scotland. Retrieved from `http://arxiv.org/abs/1206.6392`

Bowling, D. L., & Purves, D. (2015). A biological rationale for musical consonance. *Proceedings of the National Academy of Sciences*, *112*(36), 11155–11160. `https://doi.org/10.1073/pnas.1505768112`

Bowling, D. L., Purves, D., & Gill, K. Z. (2018). Vocal similarity predicts the relative attraction of musical chords. *Proceedings of the National Academy of Sciences*, *115*(1), 216–221. `https://doi.org/10.1073/pnas.1713206115`

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brooks, D. I., & Cook, R. G. (2010). Chord discrimination by pigeons. *Music Perception*, *27*, 183–196. `https://doi.org/10.1525/mp.2010.27.3.183`

Broze, Y., & Shanahan, D. (2013). Diachronic changes in jazz harmony: A cognitive perspective. *Music Perception*, *31*, 32–45. https://doi.org/10.1525/rep.2008.104.1.92

Bunton, S. (1997). Semantically motivated improvements for PPM variants. *The Computer Journal*, *40*(2/3), 76–93. https://doi.org/10.1093/comjnl/40.2_and_3.76

Burgoyne, J. A. (2011). *Stochastic Processes & Database-Driven Musicology* (PhD thesis). McGill University, Montréal, Québec, Canada. https://doi.org/10.1145/957013.957041

Butler, J. W., & Daston, P. G. (1968). Musical consonance as musical preference: A cross-cultural study. *The Journal of General Psychology*, *79*, 129–142.

Buus, S. (1997). Auditory masking. In M. J. Crocker (Ed.), *Encyclopedia of Acoustics, Volume Three* (pp. 1427–1445). John Wiley & Sons. https://doi.org/10.1002/9780470172537.ch115

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *6*(5), 1190–1208.

Callender, C., Quinn, I., & Tymoczko, D. (2008). Generalized voice-leading. *Science*, *320*, 346–348. https://doi.org/10.1126/science.1153021

Cambouropoulos, E. (2010). The musical surface: Challenging basic assumptions. *Musicae Scientiae*, *14*, 131–147. https://doi.org/10.1177/10298649100140S209

Cambouropoulos, E. (2016). The harmonic musical surface and two novel chord representation schemes. In D. Meredith (Ed.), *Computational music analysis* (pp. 31–56). Cham, Switzerland: Springer International Publishing.

Cambouropoulos, E., Kaliakatsos-Papakostas, M., & Tsougras, C. (2014). An idiom-independent representation of chords for computational music analysis and generation. In *Proceedings of the joint 11th Sound and*

Music Computing Conference (SMC) and 40th International Computer Music Conference (ICMC). Athens, Greece.

Cariani, P. A. (1999). Temporal coding of periodicity pitch in the auditory system: An overview. *Neural Plasticity*, *6*(4), 147–172.

Cariani, P. A., & Delgutte, B. (1996). Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of Neurophysiology*, *76*(3), 1698–1716.

Carter, E. (2002). *Harmony Book.* (N. Hopkins & J. F. Link, Eds.). New York, NY: Carl Fischer.

Chen, T.-P., & Su, L. (2018). Functional harmony recognition of symbolic music data with multi-task recurrent neural networks. In E. Gómez, Hu Xiao, E. Humphrey, & E. Benetos (Eds.), *Proceedings of the 19th International Society for Music Information Retrieval Conference* (pp. 90–97). Paris, France.

Chew, E. (2000). *Towards a mathematical model of tonality* (PhD thesis). MIT, Cambridge, MA.

Chew, E. (2002). The Spiral Array: An algorithm for determining key boundaries. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and Artificial Intelligence. ICMAI 2002. Lecture Notes in Computer Science, vol. 2445.* Berlin, Germany: Springer. https://doi.org/10.1007/3-540-45722-4

Chiandetti, C., & Vallortigara, G. (2011). Chicks like consonant music. *Psychological Science*, *22*(10), 1270–1273. https://doi.org/10.1177/0956797611418244

Chuan, C.-H., & Chew, E. (2011). Generating and evaluating musical harmonizations that emulate style. *Computer Music Journal*, *35*(4), 64–82. https://doi.org/10.1162/COMJ_a_00091

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

Clercq, T. de, & Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music*, *30*(1), 47–70. https://doi.org/10.1017/S026114301000067X

Cohen, M. A., Grossberg, S., & Wyse, L. L. (1995). A spectral network model of pitch perception. *The Journal of the Acoustical Society of America*, *98*, 862–879. https://doi.org/10.1121/1.413512

Cohn, R. (2012). *Audacious euphony: Chromatic harmony and the triad's second nature.* New York, NY: Oxford University Press.

Collins, T., Tillmann, B., Barrett, F. S., Delbé, C., & Janata, P. (2014). A combined model of sensory and cognitive representations underlying tonal expectations in music: From audio signals to behavior. *Psychological Review*, *121*, 33–65. https://doi.org/10.1037/a0034695

Conklin, D. (2002). Representation and discovery of vertical patterns in music. In C. Anagnostopoulou, M. Ferrand, & A. Smaill (Eds.), *Music and Artificial Intelligence: Proc. ICMAI 2002* (pp. 32–42). Berlin, Germany: Springer-Verlag.

Conklin, D., & Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, *24*, 51–73. https://doi.org/10.1080/09298219508570672

Cook, N. D. (2009). Harmony perception: Harmoniousness is more than the sum of interval consonance. *Music Perception*, *27*, 25–41. https://doi.org/10.1525/MP.2009.27.1.25

Cook, N. D. (2017). Calculation of the acoustical properties of triadic harmonies. *Journal of the Acoustical Society of America*, *142*, 3748–3755. https://doi.org/10.1121/1.5018342

Cook, N. D., & Fujisawa, T. (2006). The psychophysics of harmony perception: Harmony is a three-tone phenomenon. *Empirical Musicology Review*, *1*(2), 106–126.

Cousineau, M., McDermott, J. H., & Peretz, I. (2012). The basis of musical consonance as revealed by congenital amusia. *Proceedings of* *the National Academy of Sciences*, *109*(48), 19858–19863. https://doi.org/10.1073/pnas.1207989109

Cramer, E. M., & Huggins, W. H. (1958). Creation of pitch through binaural interaction. *The Journal of the Acoustical Society of America*, *30*, 413–417. https://doi.org/10.1121/1.1909628

Craton, L. G., Lee, J. H. J., & Krahe, P. M. (2019). It's only rock 'n roll (but I like it): Chord perception and rock's liberal harmonic palette. *Musicae Scientiae*. https://doi.org/10.1177/1029864919845023

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1119–1130. https://doi.org/10.1037/0278-7393.30.5.1119

Crespo-Bojorque, P., & Toro, J. M. (2015). The use of interval ratios in consonance perception by rats (Rattus norvegicus) and humans (Homo sapiens). *Journal of Comparative Psychology*, *129*(1), 42–51. https://doi.org/10.1037/a0037991

Crowder, R. G., Reznick, J. S., & Rosenkrantz, S. L. (1991). Perception of the major/minor distinction: V. Preferences among infants. *Bulletin of the Psychonomic Society*, *29*(3), 187–188. https://doi.org/10.3758/BF03335230

Dahlhaus, C. (1990). *Studies on the origin of harmonic tonality.* Princeton, NJ: Princeton University Press.

Daniel, P., & Weber, R. (1997). Psychoacoustical roughness: Implementation of an optimized model. *Acta Acustica United with Acustica*, *83*(1), 113–123.

Dean, R. T., Milne, A. J., & Bailes, F. (2019). Spectral pitch similarity is a predictor of perceived change in sound- as well as note-based music. *Music & Science*, *2*. https://doi.org/10.1177/2059204319847351

de Cheveigné, A. (1998). Cancellation model of pitch perception. *The Journal of the Acoustical Society of America*, *103*, 1261–1271. https://doi.org/10.1121/1.423232 de Cheveigné, A. (2005). Pitch perception models. In C. J. Plack & A. J. Oxenham (Eds.), *Pitch: Neural coding and perception* (pp. 169–233). New York, NY: Springer. https://doi.org/10.1007/0-387-28958-5_6

DeWitt, L. A., & Crowder, R. G. (1987). Tonal fusion of consonant musical intervals: The oomph in Stumpf. *Perception and Psychophysics*, *41*(1), 73–84.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–212.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0121945

Dienes, Z., & Longuet-Higgins, C. (2004). Can musical transformations be implicitly learned? *Cognitive Science*, *28*, 531–558. https://doi.org/10.1016/j.cogsci.2004.03.003

Di Giorgi, B., Dixon, S., Zanoni, M., & Sarti, A. (2017). A data-driven model of tonal chord sequence complexity. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 2237–2250. https://doi.org/10.1109/TASLP.2017.2756443

Dillon, G. (2013). Calculating the dissonance of a chord according to Helmholtz. *European Physical Journal Plus*, *128*(90). https://doi.org/10.1140/epjp/i2013-13090-4

Di Stefano, N., Focaroli, V., Giuliani, A., Formica, D., Taffoni, F., & Keller, F. (2017). A new research method to test auditory preferences in young listeners: Results from a consonance versus dissonance perception study. *Psychology of Music*, *45*(5), 699–712. https://doi.org/10.1177/0305735616681205

Duifhuis, H., Willems, L. F., & Sluyter, R. J. (1982). Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *The Journal of the Acoustical Society of America*, *71*, 1568–1580. https://doi.org/10.1121/1.387811

Ebcioğlu, K. (1988). An expert system for harmonizing four-part chorales. *Computer Music Journal*, *12*(3), 43–51.

Ebeling, M. (2008). Neuronal periodicity detection as a basis for the perception of consonance: A mathematical model of tonal fusion. *The Journal of the Acoustical Society of America*, *124*, 2320–2329. https://doi.org/10.1121/1.2968688

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*(3), 533–553. https://doi.org/10.3758/s13415-013-0161-y

Eitan, Z., Ben-Haim, M. S., & Margulis, E. H. (2017). Implicit absolute pitch representation affects basic tonal perception. *Music Perception*, *34*, 569–584. https://doi.org/10.1525/MP.2017.34.5.569

Elff, M. (2018). *mclogit: Mixed conditional logit models.* Retrieved from https://CRAN.R-project.org/package=mclogit

Emura, N., Miura, M., & Yanagida, M. (2008). A modular system generating Jazz-style arrangement for a given set of a melody and its chord name sequence. *Acoustical Science and Technology*, *29*(1), 51–57. https://doi.org/10.1250/ast.29.51

Endress, A. D. (2010). Learning melodies from non-adjacent tones. *Acta Psychologica*, *135*(2), 182–90. https://doi.org/10.1016/j.actpsy.2010.06.005

Escoffier, N., & Tillmann, B. (2008). The tonal function of a task-irrelevant chord modulates speed of visual processing. *Cognition*, *107*, 1070–1083. https://doi.org/10.1016/j.cognition.2007.10.007

Euler, L. (1739). *Tentamen novae theoria musicae.* Saint Petersburg, Russia: Academiae Scientiarum.

Fedorenko, E., Patel, A., Casasanto, D., Winawer, J., & Gibson, E. (2009). Structural integration in language and music: Evidence for a shared system. *Memory and Cognition*, *37*(1), 1–9. https://doi.org/10.3758/MC.37.1.1

Fernández, J. D., & Vico, F. (2013). AI methods in algorithmic composition: A comprehensive survey. *Journal of Artificial Intelligence Research*, *48*, 513–582. https://doi.org/10.1613/jair.3908

Fisher, A., Rudin, C., & Dominici, F. (2018). *Model Class Reliance: Variable importance measures for any machine learning model class, from the "Rashomon" perspective.* Retrieved from https://arxiv.org/pdf/1801.01489.pdf

Fletcher, H. (1924). The physical criterion for determining the pitch of a musical tone. *Physical Review*, *23*(3), 427–437. https://doi.org/10.1103/PhysRev.23.427

Florian, G. (1981). The two-part vocal style on Baluan Island Manus Province, Papua New Guinea. *Ethnomusicology*, *25*(3), 433–446.

Forte, A. (1973). *The structure of atonal music.* New Haven, CT: Yale University Press.

Frieler, K., Fischinger, T., Schlemmer, K., Lothwesen, K., Jakubowski, K., & Müllensiefen, D. (2013). Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs. *Musicae Scientiae*, *17*, 334–349. https://doi.org/10.1177/1029864913493802

Geary, J. M. (1980). Consonance and dissonance of pairs of inharmonic sounds. *The Journal of the Acoustical Society of America*, *67*, 1785–1789. https://doi.org/10.1121/1.384307

Gebhardt, R. B., Davies, M. E. P., & Seeber, B. U. (2016). Psychoacoustic approaches for harmonic music mixing. *Applied Sciences*, *6*. https://doi.org/10.3390/app6050123

Geer, W. J. van de, Levelt, W. J. M., & Plomp, R. (1962). The connotation of musical consonance. *Acta Psychologica*, *20*(4), 308–319. https://doi.org/10.1016/0001-6918(62)90028-8

Gill, K. Z., & Purves, D. (2009). A biological rationale for musical scales. *PLoS ONE*, *4*(12). https://doi.org/10.1371/journal.pone.0008144

Goldstein, J. L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. *The Journal of the Acoustical Society of America*, *54*, 1496–1516. https://doi.org/10.1121/1.1914448

Grose, J. H., Buss, E., & Hall III, J. W. (2012). Binaural beat salience. *Hearing Research*, *285*(1-2), 40–45. https://doi.org/10.1016/j.heares.2012.01.012

Guernsey, M. (1928). The rôle of consonance and dissonance in music. *The American Journal of Psychology*, *40*(2), 173–204.

Haas, B. de, Magalhães, J. P., & Wiering, F. (2012). Improving audio chord transcription by exploiting harmonic and metric knowledge. In A. de Souza Britto Jr., F. Gouyon, & S. Dixon (Eds.), *Proceedings of the 13th International Society for Music Information Retrieval Conference* (pp. 295–300). Porto, Portugal.

Hansen, N. C., & Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01052

Harnard, S. (2003). Categorical perception. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*. New York, NY: Nature Publishing Group. Retrieved from https://eprints.soton.ac.uk/257719

Harrison, P. M. C. (2020). psychTestR: An R package for designing and conducting behavioural psychological experiments. *PsyArXiv*. https://doi.org/10.31234/osf.io/dyar7

Harrison, P. M. C., Bianco, R., Chait, M., & Pearce, M. T. (2020). PPM-Decay: A computational model of auditory prediction with memory decay. *bioRxiv*. https://doi.org/10.1101/2020.01.09.900266

Harrison, P. M. C., & Pearce, M. T. (2018). Dissociating sensory and cognitive theories of harmony perception through computational modeling. In R. Parncutt & S. Sattmann (Eds.), *Proceedings of*

ICMPC15/ESCOM10. Graz, Austria. https://doi.org/10.31234/osf.io/wgjyv

Harte, C., Sandler, M., Abdallah, S., & Gómez, E. (2005). Symbolic representation of musical chords: A proposed syntax for text annotations. In *Proceedings of the 6th International Conference on Music Information Retrieval* (pp. 66–71). London, UK.

Härmä, A., & Palomäki, K. (1999). HUTear – a free Matlab toolbox for modeling of human auditory system. In *Proceedings of the 1999 MATLAB DSP Conference*. Espoo, Finland. Retrieved from http://legacy.spa.aalto.fi/software/HUTear/

Hedges, T., & Rohrmeier, M. A. (2011). Exploring Rameau and beyond: A corpus study of root progression theories. In C. Agón, M. Andreatta, G. Assayag, E. Amiot, J. Bresson, & J. Mandereau (Eds.), *Mathematics and Computation in Music – Third International Conference, MCM 2011* (pp. 334–337). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-642-21590-2_27

Hedges, T., Roy, P., & Pachet, F. (2014). Predicting the composer and style of jazz chord progressions. *Journal of New Music Research*, *433*, 276–290. https://doi.org/10.1080/09298215.2014.925477

Hedges, T., & Wiggins, G. A. (2016a). Improving predictions of derived viewpoints in multiple viewpoint systems. In *Proceedings of the 17th International Society for Music Information Retrieval Conference* (pp. 420–426). New York, NY.

Hedges, T., & Wiggins, G. A. (2016b). The prediction of merged attributes with multiple viewpoint systems. *Journal of New Music Research*, *45*, 314–332. https://doi.org/10.1080/09298215.2016.1205632

Heffernan, B., & Longtin, A. (2009). Pulse-coupled neuron models as investigative tools for musical consonance. *Journal of Neuroscience Methods*, *183*(1), 95–106. https://doi.org/10.1016/j.jneumeth.2009.06.041

Helmholtz, H. (1863). *On the sensations of tone*. New York, NY: Dover.

Hild, H., Feulner, J., & Menzel, W. (1984). HARMONET: A neural net for harmonizing chorales in the style of J. S. Bach. In *Advances in Neural Information Processing Systems* (pp. 267–274). Los Altos, CA: Morgan Kaufmann.

Hindemith, P. (1945). *The craft of musical composition.* New York, NY: Associated Music Publishers.

Hoch, L., Poulin-Charronnat, B., & Tillmann, B. (2011). The influence of task-irrelevant music on language processing: Syntactic and semantic structures. *Frontiers in Psychology*, *2*. `https://doi.org/10.3389/fpsyg.2011.00112`

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hoeschele, M., Cook, R. G., Guillette, L. M., Brooks, D. I., & Sturdy, C. B. (2012). Black-capped chickadee (Poecile atricapillus) and human (Homo sapiens) chord discrimination. *Journal of Comparative Psychology*, *126*(1), 57–67. `https://doi.org/10.1037/a0024627`

Houtsma, A. J. M., & Goldstein, J. L. (1972). The central origin of the pitch of complex tones: Evidence from musical interval recognition. *The Journal of the Acoustical Society of America*, *51*, 520–529. `https://doi.org/10.1121/1.1912873`

Hörnel, D. (2004). CHORDNET: Learning and producing voice leading with neural networks and dynamic programming. *Journal of New Music Research*, *33*, 387–397. `https://doi.org/10.1080/0929821052000343859`

Hu, D. J., & Saul, L. K. (2009). A probabilistic topic model for unsupervised learning of musical key-profiles. In K. Hirata, G. Tzanetakis, & K. Yoshii (Eds.), *Proceedings of the 10th International Society for Music Information Retrieval Conference.* Kobe, Japan.

Hulse, S. H., Bernard, D. J., & Braaten, R. F. (1995). Auditory discrimination of chord-based spectral structures by European starlings (Sturnus vulgaris). *Journal of Experimental Psychology: General*, *124*, 409–423. `https://doi.org/10.1037/0096-3445.124.4.409`

Huron, D. (1991). Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception*, *9*, 135–154.

Huron, D. (1994). Interval-class content in equally tempered pitch-class sets: Common scales exhibit optimum tonal consonance. *Music Perception*, *11*, 289–305. `https://doi.org/10.2307/40285624`

Huron, D. (2001). Tone and voice: A derivation of the rules of voice-leading from perceptual principles. *Music Perception*, *19*, 1–64. `https://doi.org/10.1525/mp.2001.19.1.1`

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation.* Cambridge, MA: MIT Press.

Huron, D. (2016). *Voice leading: The science behind a musical art.* Cambridge, MA: MIT Press.

Huron, D., & Parncutt, R. (1993). An improved model of tonality perception incorporating pitch salience and echoic memory. *Psychomusicology*, *12*, 154–171.

Huron, D., & Sellmer, P. (1992). Critical bands and the spelling of vertical sonorities. *Music Perception*, *10*, 129–149.

Hutchinson, W., & Knopoff, L. (1978). The acoustic component of Western consonance. *Journal of New Music Research*, *7*, 1–29. `https://doi.org/10.1080/09298217808570246`

Hutchinson, W., & Knopoff, L. (1979). The significance of the acoustic component of consonance in Western triads. *Journal of Musicological Research*, *3*(1-2), 5–22. `https://doi.org/10.1080/01411897908574504`

Immerseel, L. V., & Martens, J. (1992). Pitch and voiced/unvoiced determination with an auditory model. *The Journal of the Acoustical Society of America*, *91*, 3511–3526.

Izumi, A. (2000). Japanese monkeys perceive sensory consonance of chords. *The Journal of the Acoustical Society of America*, *108*, 3073–3078.

Jackendoff, R. (1987). *Consciousness and the Computational Mind*. Cambridge, MA: MIT Press.

Jacoby, N., Tishby, N., & Tymoczko, D. (2015). An information theoretic approach to chord categorization and functional harmony. *Journal of New Music Research*, *44*, 219–244. `https://doi.org/10.1080/09298215.2015.1036888`

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The cortical topography of tonal structures underlying Western music. *Science*, *298*, 2167–2170. `https://doi.org/10.1126/science.1076262`

Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, *4*, 227–241. `https://doi.org/10.1109/TSSC.1968.300117`

Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On musical dissonance. *Music Perception*, *30*, 19–35.

Jonaitis, E. M., & Saffran, J. R. (2009). Learning harmony: The role of serial statistics. *Cognitive Science*, *33*, 951–968. `https://doi.org/10.1111/j.1551-6709.2009.01036.x`

Ju, Y., Condit-Schultz, N., Arthur, C., & Fujinaga, I. (2017). Non-chord tone identification using deep neural networks. In K. Page (Ed.), *Proceedings of the 4th International Workshop on Digital Libraries for Musicology* (pp. 13–16). Shanghai, China. `https://doi.org/10.1145/3144749.3144753`

Kaestner, G. (1909). Untersuchungen über den Gefühlseindruck unanalysierter Zweiklänge. *Psychologische Studien*, *4*, 473–504.

Kameoka, A., & Kuriyagawa, M. (1969a). Consonance theory part I: Consonance of dyads. *The Journal of the Acoustical Society of America*, *45*, 1451–1459. `https://doi.org/10.1121/1.1911623`

Kameoka, A., & Kuriyagawa, M. (1969b). Consonance theory Part II: Consonance of complex tones and its calculation method. *The Journal* *of the Acoustical Society of America*, *45*, 1460–1469. https://doi.org/10.1121/1.1911624

Karrick, B. (1998). An examination of the intonation tendencies of wind instrumentalists based on their performance of selected harmonic musical intervals. *Journal of Research in Music Education*, *46*(1), 112–127.

Kneib, T., Baumgartner, B., & Steiner, W. J. (2007). Semiparametric multinomial logit models for analysing consumer choice behaviour. *AStA Advances in Statistical Analysis*, *91*, 225–244. https://doi.org/10.1007/s10182-007-0033-2

Koda, H., Nagumo, S., Basile, M., Olivier, M., Remeuf, K., Blois-Heulin, C., & Lemasson, A. (2013). Validation of an auditory sensory reinforcement paradigm: Campbell's monkeys (Cercopithecus campbelli) do not prefer consonant over dissonant sounds. *Journal of Comparative Psychology*, *127*(3), 265–271. https://doi.org/10.1037/a0031237

Koelsch, S., Gunter, T. C., Schröger, E., Tervaniemi, M., Sammler, D., & Friederici, A. D. (2001). Differentiating ERAN and MMN: An ERP study. *NeuroReport*, *12*(7), 1385–1389. https://doi.org/10.1063/1.4973814

Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: An ERP study. *Journal of Cognitive Neuroscience*, *17*(10), 1565–1577. https://doi.org/10.1162/089892905774597290

Koelsch, S., Gunter, T., Friederici, A. D., & Schröger, E. (2000). Brain indices of music processing: "Nonmusicians" are musical. *Journal of Cognitive Neuroscience*, *12*(3), 520–541. https://doi.org/10.1162/089892900562183

Koelsch, S., Rohrmeier, M. A., Torrecuso, R., & Jentschke, S. (2013). Processing of hierarchical syntactic structure in music. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(38), 15443–15448. https://doi.org/10.1073/pnas.1300272110

Koelsch, S., Schroger, E., & Gunter, T. C. (2002). Music matters: Preattentive musicality of the human brain. *Psychophysiology*, *39*, 38–48. `https://doi.org/10.1111/1469-8986.3910038`

Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences*, *23*(1), 63–77. `https://doi.org/10.1016/j.tics.2018.10.006`

Kopiez, R. (2003). Intonation of harmonic intervals: Adaptability of expert musicians to equal temperament and just intonation. *Music Perception*, *20*, 383–410. `https://doi.org/10.1525/mp.2003.20.4.383`

Kröger, P., Passos, A., Sampaio, M. da S., & Cidra, G. de. (2008). Rameau: A system for automatic harmonic analysis. In *Proceedings of the 2008 International Computer Music Conference* (pp. 273–281). Belfast, Northern Ireland.

Krueger, F. (1910). Die Theorie der Konsonanz. *Psychologische Studien*, *5*, 294–411.

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

Krumhansl, C. L., Sandell, G. J., & Sergeant, D. C. (1987). The perception of tone hierarchies and mirror forms in twelve-tone serial music. *Music Perception*, *5*, 31–77.

Kunert, R., Willems, R. M., & Hagoort, P. (2016). Language influences music harmony perception: Effects of shared syntactic integration resources beyond attention. *Royal Society Open Science*, *3*. `https://doi.org/10.1098/rsos.150685`

Lahdelma, I., & Eerola, T. (2016). Mild dissonance preferred over consonance in single chord perception. *I-Perception*. `https://doi.org/10.1177/2041669516655812`

Lahdelma, I., & Eerola, T. (2019). Exposure impacts the pleasantness of consonance/dissonance but not its perceived tension. *PsyArXiv*. `https://doi.org/10.31234/osf.io/fmxpw`

Landsnes, K., Mehrabyan, L., Wiklund, V., Lieck, R., Moss, F. C., & Rohrmeier, M. A. (2019). A model comparison for chord prediction on the Annotated Beethoven Corpus. In *Proceedings of the 16th Sound & Music Computing Conference*. Málaga, Spain.

Langhabel, J., Lieck, R., Toussaint, M., & Rohrmeier, M. A. (2017). Feature discovery for sequential prediction of monophonic music. In S. J. Cunningham, Z. Duan, X. Hu, & D. Turnbull (Eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference* (pp. 649–656). Suzhou, China.

Langner, G. (1997). Temporal processing of pitch in the auditory system. *Journal of New Music Research*, *26*, 116–132. `https://doi.org/10.1080/09298219708570721`

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*, 119–159.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for Music Information Retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 261–268). Berlin, Germany: Springer.

Lee, K. M., Skoe, E., Kraus, N., & Ashley, R. (2015). Neural transformation of dissonant intervals in the auditory brainstem. *Music Perception*, *32*, 445–459. `https://doi.org/10.1525/MP.2015.32.5.445`

Lee, K., & Slaney, M. (2008). Acoustic chord transcription and key extraction from audio using key-dependent HMMs trained on synthesized audio. *IEEE Transactions on Audio, Speech and Language Processing*, *16*, 291–301. `https://doi.org/10.1109/TASL.2007.914399`

Leeuw, J. R. D. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavioral Research Methods*, *47*(1), 1–12. `https://doi.org/10.3758/s13428-014-0458-y`

Leman, M. (2000a). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception*, *17*, 481–509.

Leman, M. (2000b). Visualization and calculation of the roughness of acoustical music signals using the Synchronization Index Model. In *Proceedings of the COSTG-6 Conference on Digital Audio Effects (DAFX-00)*. Verona, Italy.

Leman, M., Lesaffre, M., & Tanghe, K. (2001). Introduction to the IPEM Toolbox for Perception-based Music Analysis. In *Conference Program and Abstracts of SMPC 2001 Kingston*. https://doi.org/10.1080/09298219708570719

Lerdahl, F. (1988). Tonal pitch space. *Music Perception*, *5*, 315–349.

Levelt, W. J. M., Geer, J. P. van de, & Plomp, R. (1966). Triadic comparisons of musical intervals. *The British Journal of Mathematical and Statistical Psychology*, *19*(2), 163–179.

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics*, *56*(4), 414–423. https://doi.org/10.3758/BF03206733

Lewin, D. (1987). *Generalized musical intervals and transformations*. New Haven, CT: Yale University Press.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia*, *7*(4), 128–134.

Loosen, F. (1993). Intonation of solo violin performance with reference to equally tempered, Pythagorean, and just intonations. *The Journal of the Acoustical Society of America*, *93*, 525–539. https://doi.org/10.1121/1.405632

Lots, I. S., & Stone, L. (2008). Perception of musical consonance and dissonance: An outcome of neural synchronization. *Journal of the Royal Society Interface*, *5*(29), 1429–1434. https://doi.org/10.1098/rsif.2008.0143

Loui, P., Wu, E. H., Wessel, D. L., & Knight, R. T. (2009). A generalized mechanism for perception of pitch patterns. *Journal of Neuroscience*, *29*, 454–459. https://doi.org/10.1523/JNEUROSCI.4503-08.2009

Lundin, R. W. (1947). Toward a cultural theory of consonance. *The Journal of Psychology*, *23*(1), 45–49. https://doi.org/10.1080/00223980.1947.9917318

Madsen, S. T., & Widmer, G. (2007). Key-finding with interval profiles. In *Proceedings of the International Computer Music Conference (ICMC)*. Copenhagen, Denmark.

Maher, T. F. (1976). "Need for resolution" ratings for harmonic musical intervals: A comparison between Indians and Canadians. *Journal of Cross-Cultural Psychology*, *7*(3), 259–276.

Marin, M. M., Forde, W., Gingras, B., & Stewart, L. (2015). Affective evaluation of simultaneous tone combinations in congenital amusia. *Neuropsychologia*, *78*, 207–220. https://doi.org/10.1016/j.neuropsychologia.2015.10.004

Martin, N. (2008). The Tristan chord resolved. *Intersections*, *28*(2), 6–30. https://doi.org/10.7202/029953ar

Martino, D. (1961). The source set and its aggregate formations. *Journal of Music Theory*, *5*, 224–273.

Masada, K., & Bunescu, R. (2017). Chord recognition in symbolic music using semi-Markov Conditional Random Fields. In S. J. Cunningham, Z. Duan, X. Hu, & D. Turnbull (Eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference* (pp. 272–278). Suzhou, China.

Masataka, N. (2006). Preference for consonance over dissonance by hearing newborns of deaf parents and of hearing parents. *Developmental Science*, *9*(1), 46–50. https://doi.org/10.1111/j.1467-7687.2005.00462.x

Mashinter, K. (2006). Calculating sensory dissonance: Some discrepancies arising from the models of Kameoka & Kuriyagawa, and Hutchinson & Knopoff. *Empirical Musicology Review*, *1*(2), 65–84.

Mauch, M. (2010). *Automatic chord transcription from audio using computational models of musical context* (PhD thesis). Queen Mary University of London, London, UK.

Mauch, M., Noland, K., & Dixon, S. (2009). Using musical structure to enhance automatic chord transcription. In K. Hirata, G. Tzanetakis, & K. Yoshii (Eds.), *Proceedings of the 10th International Society for Music Information Retrieval Conference* (pp. 231–236). Kobe, Japan.

McDermott, J., & Hauser, M. (2004). Are consonant intervals music to their ears? Spontaneous acoustic preferences in a nonhuman primate. *Cognition*, *94*, B11–B21. `https://doi.org/10.1016/j.cognition.2004.04.004`

McDermott, J. H., Lehr, A. J., & Oxenham, A. J. (2010). Individual differences reveal the basis of consonance. *Current Biology*, *20*(11), 1035–1041. `https://doi.org/10.1016/j.cub.2010.04.019`

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, *535*(7613), 547–550. `https://doi.org/10.1038/nature18635`

McFadden, D. (1974). Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York, NY: Academic Press.

McGowan, J. (2011). Psychoacoustic foundations of contextual harmonic stability in jazz piano voicings. *Journal of Jazz Studies*, *7*(2), 156–191.

McLachlan, N., Marco, D., Light, M., & Wilson, S. (2013). Consonance and pitch. *Journal of Experimental Psychology: General*, *142*, 1142–1158. `https://doi.org/10.1037/a0030830`

Mearns, L. (2013). *The computational analysis of harmony in Western art music* (PhD thesis). Queen Mary University of London, London, UK.

Meddis, R. (2011). *MATLAB auditory periphery (MAP): Model technical description.* Retrieved from `https://code.soundsoftware.ac.uk/projects/map`

Meddis, R., & Hewitt, M. J. (1991a). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: Phase sensitivity. *The*

Journal of the Acoustical Society of America, *89*, 2883–2894. https://doi.org/10.1121/1.400726

Meddis, R., & Hewitt, M. J. (1991b). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *The Journal of the Acoustical Society of America*, *89*, 2866–2882. https://doi.org/10.1121/1.400725

Meddis, R., & O'Mard, L. (1997). A unitary model of pitch perception. *The Journal of the Acoustical Society of America*, *102*, 1811–1820. https://doi.org/10.1121/1.420088

Meeus, N. (2000). Toward a post-Schoenbergian grammar of tonal and pre-tonal harmonic progressions. *Music Theory Online*, *6*(1).

Meyer, L. B. (1956). *Emotion and meaning in music.* Chicago, IL: The University of Chicago Press.

Meyer, L. B. (1957). Meaning in music and information theory. *The Journal of Aesthetics and Art Criticism*, *15*(4), 412–424.

Miles, S. A., Rosen, D. S., & Grzywacz, N. M. (2017). A statistical analysis of the relationship between harmonic surprise and preference in popular music. *Frontiers in Human Neuroscience*, *11*. https://doi.org/10.3389/fnhum.2017.00263

Milne, A. J. (2013). *A computational model of the cognition of tonality* (PhD thesis). The Open University, Milton Keynes, England.

Milne, A. J., & Holland, S. (2016). Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, *10*, 59–85. https://doi.org/10.1080/17459737.2016.1152517

Milne, A. J., Laney, R., & Sharp, D. (2015). A spectral pitch class model of the probe tone data and scalic tonality. *Music Perception*, *32*, 364–393. https://doi.org/10.1525/MP.2015.32.4.364

Milne, A. J., Laney, R., & Sharp, D. B. (2016). Testing a spectral model of tonal affinity with microtonal melodies and inharmonic spectra. *Musicae Scientiae*, *20*, 465–494. https://doi.org/10.1177/1029864915622682

Milne, A. J., Sethares, W. A., Laney, R., & Sharp, D. B. (2011). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music*, *5*, 1–20. https://doi.org/10.1080/17459737.2011.573678

Mirand, R. A., & Ullman, M. T. (2007). Double dissociation between rules and memory in music: An event-related potential study. *Neuroimage*, *38*(2), 331–345.

Miyazaki, K. (1988). Musical pitch identification by absolute pitch possessors. *Perception & Psychophysics*, *44*(6), 501–512. https://doi.org/10.3758/BF03207484

Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Transactions on Communications*, *38*, 1917–1921. https://doi.org/10.1109/26.61469

Moffat, A., Neal, R. M., & Witten, I. H. (1998). Arithmetic coding revisited. *ACM Transactions on Information Systems*, *16*(3), 256–294. https://doi.org/10.1109/DCC.1995.515510

Moore, B. C. J., & Ernst, S. M. A. (2012). Frequency difference limens at high frequencies: Evidence for a transition from a temporal to a place code. *The Journal of the Acoustical Society of America*, *132*, 1542–1547. https://doi.org/10.1121/1.4739444

Morgan, E., Fogel, A., Nair, A., & Patel, A. D. (2019). Statistical learning and Gestalt-like principles predict melodic expectations. *Cognition*, *189*, 23–34. https://doi.org/10.1016/j.cognition.2018.12.015

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0089642

Neuwirth, M., Harasim, D., Moss, F. C., & Rohrmeier, M. A. (2018). The Annotated Beethoven Corpus (ABC): A dataset of harmonic analyses of all Beethoven string quartets. *Frontiers in Digital Humanities*, *5*. https://doi.org/10.3389/fdigh.2018.00016

Nobili, R., Vetešník, A., Turicchia, L., & Mammano, F. (2003). Otoacoustic emissions from residual oscillations of the cochlear basilar membrane in a human ear model. *Journal of the Association for Research in Otolaryngology*, *4*, 478–494.

Nordmark, J., & Fahlén, L. E. (1988). Beat theories of musical consonance. *STL-QPSR*, *29*(1), 111–122.

Olsen, K. N., Thompson, W. F., & Giblin, I. (2018). Listener expertise enhances intelligibility of vocalizations in death metal music. *Music Perception*, *35*, 527–539. https://doi.org/10.1525/MP.2018.35.5.527

Omigie, D., Dellacherie, D., & Samson, S. (2017). Effects of learning on dissonance judgments. *Journal of Interdisciplinary Music Studies*, *8*(1-2), 11–28. https://doi.org/10.4407/jims.2016.12.001

Omigie, D., Pearce, M. T., & Stewart, L. (2012). Tracking of pitch probabilities in congenital amusia. *Neuropsychologia*, *50*(7), 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*(9), 1749–1762. https://doi.org/10.1016/j.neuropsychologia.2013.05.010

Oxenham, A. J. (2018). How we hear: The perception and neural coding of sound. *Annual Review of Psychology*, *69*, 27–50. https://doi.org/10.1146/annurev-psych-122216-011635

Pardo, B., & Birmingham, W. P. (2002). Algorithms for chordal analysis. *Computer Music Journal*, *26*(2), 27–49. https://doi.org/10.1162/014892602760137167

Parncutt, R. (1988). Revision of Terhardt's psychoacoustical model of the root(s) of a musical chord. *Music Perception*, *6*, 65–94.

Parncutt, R. (1989). *Harmony: A psychoacoustical approach.* Berlin, Germany: Springer-Verlag.

Parncutt, R. (1993). Pitch properties of chords of octave-spaced tones. *Contemporary Music Review*, *9*(1-2), 35–50.

Parncutt, R. (1997). A model of the perceptual root(s) of a chord accounting for voicing and prevailing tonality. In M. Leman (Ed.), *Music, gestalt, and computing: Studies in cognitive and systematic musicology* (pp. 181–199). Berlin, Germany: Springer. https://doi.org/10.1007/BFb0034114

Parncutt, R. (2006a). Commentary on Cook & Fujisawa's "The psychophysics of harmony perception: Harmony is a three-tone phenomenon". *Empirical Musicology Review*, *1*(4), 204–209.

Parncutt, R. (2006b). Commentary on Keith Mashinter's "Calculating sensory dissonance: Some discrepancies arising from the models of Kameoka & Kuriyagawa, and Hutchinson & Knopoff. *Empirical Musicology Review*, *1*(4), 201–203.

Parncutt, R., & Hair, G. (2011). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, *5*(2), 119–166. https://doi.org/10.4407/jims.2011.11.002

Parncutt, R., & Hair, G. (2018). A psychocultural theory of musical interval: Bye bye Pythagoras. *Music Perception*, *35*, 475–501.

Parncutt, R., Reisinger, D., Fuchs, A., & Kaiser, F. (2018). Consonance and prevalence of sonorities in Western polyphony: Roughness, harmonicity, familiarity, evenness, diatonicity. *Journal of New Music Research*, *48*. https://doi.org/10.1080/09298215.2018.1477804

Parncutt, R., Sattmann, S., Gaich, A., & Seither-Preisler, A. (2019). Tone profiles of isolated musical chords: Psychoacoustic versus cognitive models. *Music Perception*, *36*, 406–430.

Parncutt, R., & Strasburger, H. (1994). Applying psychoacoustics in composition: "Harmonic" progressions of "nonharmonic" sonorities. *Perspectives of New Music*, *32*(2), 88–129.

Patel, A. D. (2003). Language, music, syntax and the brain. *Nature Neuroscience*, *6*(7), 674–681. https://doi.org/10.1038/nn1082

Patel, A. D., Gibson, E., Ratner, J., Besson, M., & Holcomb, P. J. (1998). Processing syntactic relations in language and music: An event-related potential study. *Journal of Cognitive Neuroscience*, *10*(6), 717–733.

Patterson, R. D. (1986). Spiral detection of periodicity and the spiral form of musical scales. *Psychology of Music*, *14*, 44–61. https://doi.org/10.1177/0305735686141004

Patterson, R. D., & Green, D. M. (2012). Auditory masking. In E. Carterette (Ed.), *Handbook of Perception, Volume IV: Hearing* (pp. 337–361). Amsterdam, The Netherlands: Elsevier.

Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (PhD thesis). 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., Conklin, D., & Wiggins, G. A. (2005). Methods for combining statistical models of music. In *Proceedings of the second international conference on computer music modeling and retrieval* (pp. 295–312). Berlin, Germany: Springer.

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. (2010a). The role of expectation and probabilistic learning in auditory boundary perception:

A model comparison. *Perception*, *39*(10), 1365–1389. https://doi.org/10.1068/p6507

Pearce, M. T., & Rohrmeier, M. A. (2018). Musical syntax II: Empirical perspectives. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 487–505). Berlin, Germany: 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. (2010b). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302–313. https://doi.org/10.1016/j.neuroimage.2009.12.019

Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, *33*, 367–385. https://doi.org/10.1080/0929821052000343840

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*, 377–405. https://doi.org/10.1525/mp.2006.23.5.377

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., Susini, P., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *Journal of the Acoustical Society of America*, *130*, 2902–2916. https://doi.org/10.1121/1.3642604

Perani, D., Cristina, M., Scifo, P., Spada, D., Andreolli, G., & Rovelli, R. (2010). Functional specializations for music processing in the human newborn brain. *Proceedings of the National Academy of Sciences*, *107*(10), 4758–4763. https://doi.org/10.1073/pnas.0909074107

Perruchet, P., & Poulin-Charronnat, B. (2013). Challenging prior evidence for a shared syntactic processor for language and music. *Psychonomic Bulletin and Review*, *20*(2), 310–317. https://doi.org/10.3758/s13423-012-0344-5

Pierce, J. R. (1966). Attaining consonance in arbitrary scales. *The Journal of the Acoustical Society of America*, *40*, 249.

Plantinga, J., & Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. *Cognition*, *98*, 1–11. https://doi.org/10.1016/j.cognition.2004.09.008

Plantinga, J., & Trehub, S. E. (2014). Revisiting the innate preference for consonance. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 40–49. https://doi.org/10.1037/a0033471

Plomp, R. (1965). Detectability threshold for combination tones. *The Journal of the Acoustical Society of America*, *37*, 1110–1123. https://doi.org/10.1121/1.1909532

Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, *38*, 548–560. https://doi.org/10.1121/1.1909741

Ponsford, D., Wiggins, G. A., & Mellish, C. (1999). Statistical learning of harmonic movement. *Journal of New Music Research*, *28*, 150–177. https://doi.org/10.1076/jnmr.28.2.150.3115

Popescu, T., Neuser, M. P., Neuwirth, M., Bravo, F., Mende, W., Boneh, O., … Rohrmeier, M. A. (2019). The pleasantness of sensory dissonance is mediated by musical style and expertise. *Scientific Reports*, *9*. https://doi.org/10.1038/s41598-018-35873-8

Poulin-Charronnat, B., Bigand, E., Madurell, F., & Peereman, R. (2005). Musical structure modulates semantic priming in vocal music. *Cognition*, *94*, B67–B78. https://doi.org/10.1016/j.cognition.2004.05.003

Pressnitzer, D., & McAdams, S. (1999). Two phase effects in roughness perception. *Journal of the Acoustical Society of America*, *105*, 2773–2782.

Quinn, I. (2010). Are pitch-class profiles really "key for key"? *Zeitschrift Der Gesellschaft Für Musiktheorie*, *7*(2), 151–163.

Quinn, I., & Mavromatis, P. (2011). Voice-leading prototypes and harmonic function. In C. Agón, M. Andreatta, G. Assayag, E. Amiot, J. Bresson, & J. Mandereau (Eds.), *Mathematics and Computation in Music – Third*

International Conference, MCM 2011 (pp. 230–240). Berlin, Germany: Springer.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Rahn, J. (1980). *Basic atonal theory.* New York, NY: Schirmer.

Rameau, J.-P. (1722). *Treatise on harmony.* Paris, France: Jean-Baptiste-Christophe Ballard.

Raphael, C., & Stoddard, J. (2004). Functional harmonic analysis using probabilistic models. *Computer Music Journal*, *28*(3), 45–52. https://doi.org/10.1162/0148926041790676

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rohrmeier, M. A. (2011). Towards a generative syntax of tonal harmony. *Journal of Mathematics and Music*, *5*, 35–53. https://doi.org/10.1080/17459737.2011.573676

Rohrmeier, M. A., & Cross, I. (2008). Statistical properties of tonal harmony in Bach's chorales. In K. Miyazaki, Y. Hiraga, M. Adachi, Y. Nakajima, & M. Tsuzaki (Eds.), *Proceedings of the 10th International Conference on Music Perception and Cognition* (pp. 619–627). Sapporo, Japan.

Rohrmeier, M. A., & Cross, I. (2009). Tacit tonality: Implicit learning of context-free harmonic structure. In J. Louhivuori, T. Eerola, S. Saarikallio, T. Himberg, & P.-S. Eerola (Eds.), *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music (ESCOM 2009), Jyväskylä, Finland* (pp. 443–452).

Rohrmeier, M. A., & Graepel, T. (2012). Comparing feature-based models of harmony. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)* (pp. 357–370). London, UK.

Rohrmeier, M. A., & Pearce, M. T. (2018). Musical syntax I: Theoretical perspectives. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 473–486). Berlin, Germany: Springer-Verlag. https://doi.org/10.1007/978-3-662-55004-5_2

Rohrmeier, M. A., Rebuschat, P., & Cross, I. (2011). Incidental and online learning of melodic structure. *Consciousness and Cognition*, *20*(2), 214–222. https://doi.org/10.1016/j.concog.2010.07.004

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month old infants. *Science*, *274*, 1926–1928. https://doi.org/10.1126/science.274.5294.1926

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

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*(4), 606–621. https://doi.org/10.1006/jmla.1996.0032

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., & McIntosh, A. R. (2015). Predictions and the brain: How musical sounds become rewarding. *Trends in Cognitive Sciences*, *19*(2), 86–91. https://doi.org/10.1016/j.tics.2014.12.001

Sapp, C. (2007). Computational chord-root identification in symbolic musical data: Rationale, methods, and applications. *Computing in Musicology*, *15*, 99–119.

Sapp, C. S. (2005). Online database of scores in the Humdrum file format. In *Proceedings of the 6th International Society for Music Information Retrieval Conference* (pp. 664–665).

Sapp, C. S. (2011). *Computational methods for the analysis of musical structure* (PhD thesis). Stanford University, Stanford, CA.

Sauvé, S. A. (2017). *Prediction in polyphony: modelling musical auditory scene analysis* (PhD thesis). Quen Mary University of London, London, UK.

Sauvé, S. A., Sayed, A., Dean, R. T., & Pearce, M. T. (2018). Effects of pitch and timing expectancy on musical emotion. *Psychomusicology: Music, Mind, and Brain*, *28*, 17–39. https://doi.org/10.1037/pmu0000203

Scharf, B. (1971). Fundamentals of auditory masking. *Audiology*, *10*, 30–40.

Schellenberg, E. G., & Trehub, S. E. (1994). Frequency ratios and the perception of tone patterns. *Psychonomic Bulletin & Review*, *1*(2), 191–201.

Schmuckler, M. A., & Tomovski, R. (2005). Perceptual tests of an algorithm for musical key-finding. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 1124–1149. https://doi.org/10.1037/0096-1523.31.5.1124

Schneider, A. (1997). “Verschmelzung”, tonal fusion, and consonance: Carl Stumpf revisited. In M. Leman (Ed.), *Music, Gestalt, and Computing: Studies in Cognitive and Systematic Musicology* (pp. 115–143). https://doi.org/10.1007/BFb0034111

Schneider, A. (2018). Pitch and pitch perception. In R. Bader (Ed.), *Springer handbook of systematic musicology* (pp. 605–685). Berlin, Germany: Springer. https://doi.org/10.1007/978-3-662-55004-5_31

Schoenberg, A. (1978). *Theory of harmony.* Berkeley; Los Angeles, CA: University of California Press.

Schouten, J. F. (1938). The perception of subjective tones. *Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*, *41*, 1086–1093.

Schwartz, D. A., Howe, C. Q., & Purves, D. (2003). The statistical structure of human speech sounds predicts musical universals. *The Journal of Neuroscience*, *23*, 7160–7168.

Sears, D. R. W., Arzt, A., Frostel, H., Sonnleitner, R., & Widmer, G. (2017). Modeling harmony with skip-grams. In S. J. Cunningham, Z. Duan, X. Hu, & D. Turnbull (Eds.), *Proceedings of the 18th International Society for Music Information Retrieval Conference.* Suzhou, China.

Sears, D. R. W., Korzeniowski, F., & Widmer, G. (2018a). Evaluating language models of tonal harmony. In E. Gómez, Hu Xiao, E. Humphrey, & E. Benetos (Eds.), *Proceedings of the 19th International Society for Music Information Retrieval Conference*. Paris, France. Retrieved from `http://arxiv.org/abs/1806.08724`

Sears, D. R. W., Pearce, M. T., Caplin, W. E., & McAdams, S. (2018b). Simulating melodic and harmonic expectations for tonal cadences using probabilistic models. *Journal of New Music Research*, *47*, 29–52. `https://doi.org/10.1080/09298215.2017.1367010`

Sears, D. R. W., Pearce, M. T., Spitzer, J., Caplin, W. E., & McAdams, S. (2019). Expectations for tonal cadences: Sensory and cognitive priming effects. *Quarterly Journal of Experimental Psychology*, *72*(6), 1422–1438. `https://doi.org/10.1177/1747021818814472`

Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, *94*, 1218–1228.

Sethares, W. A. (2005). *Tuning, timbre, spectrum, scale.* London, UK: Springer.

Shamma, S., & Klein, D. (2000). The case of the missing pitch templates: How harmonic templates emerge in the early auditory system. *The Journal of the Acoustical Society of America*, *107*, 2631–2644. `https://doi.org/10.1121/1.428649`

Shmulevich, I., & Yli-Harja, O. (2000). Localized key finding: Algorithms and applications. *Music Perception*, *17*, 531–544.

Sigtia, S., Benetos, E., & Dixon, S. (2016). An end-to-end neural network for polyphonic music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*, 927–939.

Silverman, B. W. (1986). *Density estimation.* London, UK: Chapman & Hall.

Skovenborg, E., & Nielsen, S. H. (2002). Measuring sensory consonance by auditory modelling. In *Proceedings of the 5th International Confer-*

ence on Digital Audio Effects (DAFX-02) (pp. 251–256). Hamburg, Germany.

Slaney, M., & Lyon, R. F. (1990). A perceptual pitch detector. In *International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 357–360). https://doi.org/10.1109/ICASSP.1990.115684

Slevc, L. R., & Okada, B. M. (2015). Processing structure in language and music: a case for shared reliance on cognitive control. *Psychonomic Bulletin and Review*, *22*(3), 637–652. https://doi.org/10.3758/s13423-014-0712-4

Slevc, L. R., Rosenberg, J. C., & Patel, A. D. (2009). Making psycholinguistics musical: Self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychonomic Bulletin & Review*, *16*(2), 374–381. https://doi.org/10.3758/16.2.374

Sloboda, J. A. (1991). Music structure and emotional response: Some empirical findings. *Psychology of Music*, *19*, 110–120. https://doi.org/10.1177/0305735691192002

Smoorenburg, G. F. (1972). Combination tones and their origin. *The Journal of the Acoustical Society of America*, *52*, 615–632. https://doi.org/10.1121/1.1913152

Sorge, G. A. (1747). *Vorgemach der musicalischen Composition.* Lobenstein, Germany: Verlag des Autoris.

Spagnolo, B., Ushakov, Y. V., & Dubkov, A. A. (2013). Harmony perception and regularity of spike trains in a simple auditory model. In *AIP Conference Proceedings* (Vol. 1510, pp. 274–289). https://doi.org/10.1063/1.4776512

Spyra, J., Stodolak, M., & Woolhouse, M. (2019). Events versus time in the perception of nonadjacent key relationships. *Musicae Scientiae*. https://doi.org/10.1177/1029864919867463

Stainsby, T., & Cross, I. (2009). The perception of pitch. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 47–58). New York, NY: Oxford University Press.

Steedman, M. J. (1984). A generative grammar for jazz chord sequences. *Music Perception*, *2*, 52–77.

Stewart, L. (2011). Characterizing congenital amusia. *Quarterly Journal of Experimental Psychology*, *64*(4), 625–638. https://doi.org/10.1080/17470218.2011.552730

Stolzenburg, F. (2015). Harmony perception by periodicity detection. *Journal of Mathematics and Music*, *9*, 215–238. https://doi.org/10.1080/17459737.2015.1033024

Stolzenburg, F. (2017). Periodicity detection by neural transformation. In E. Van Dyck (Ed.), *Proceedings of the 25th Anniversary Conference of the European Society for the Cognitive Sciences of Music* (pp. 159–162). Ghent, Belgium.

Straus, J. N. (1991). A primer for atonal set theory. *College Music Symposium*, *31*, 1–26.

Stumpf, C. (1890). *Tonpsychologie.* Leipzig, Germany: Verlag S. Hirzel.

Stumpf, C. (1898). Konsonanz und dissonanz. *Beiträge Zur Akustik Und Musikwissenschaft*, *1*, 1–108.

Sugimoto, T., Kobayashi, H., Nobuyoshi, N., Kiriyama, Y., Takeshita, H., Nakamura, T., & Hashiya, K. (2010). Preference for consonant music over dissonant music by an infant chimpanzee. *Primates*, *51*, 7–12. https://doi.org/10.1007/s10329-009-0160-3

Tabas, A., Andermann, M., Sebold, V., Riedel, H., Balaguer-Ballester, E., & Rupp, A. (2017). *Early processing of consonance and dissonance in human auditory cortex.* Retrieved from http://arxiv.org/abs/1711.10991

Tartini, G. (1754). *Trattato di musica secondo la vera scienza dell'armonia.* Padova, Italy.

Tekman, H. G., & Bharucha, J. J. (1992). Time course of chord priming. *Perception & Psychophysics*, *51*(1), 33–39. https://doi.org/10.3758/BF03205071

Tekman, H. G., & Bharucha, J. J. (1998). Implicit knowledge versus psychoacoustic similarity in priming of chords. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 252–260. https://doi.org/10.1037/0096-1523.24.1.252

Temperley, D. (1997). An algorithm for harmonic analysis. *Music Perception*, *15*, 31–68.

Temperley, D. (1999). What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception*, *17*, 65–100.

Temperley, D. (2007). *Music and Probability*. Cambridge, MA: MIT Press.

Temperley, D. (2008). A probabilistic model of melody perception. *Cognitive Science*, *32*, 418–444. https://doi.org/10.1080/03640210701864089

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., & De Clercq, T. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research*, *423*, 187–204. https://doi.org/10.1080/09298215.2013.788039

Temperley, D., & Sleator, D. (1999). Modeling meter and harmony: A preference-rule approach. *Computer Music Journal*, *23*(1), 10–27.

Terhardt, E. (1974). Pitch, consonance, and harmony. *The Journal of the Acoustical Society of America*, *55*, 1061–1069.

Terhardt, E. (1982). Die psychoakustischen Grundlagen der musikalischen Akkordgrundtöne und deren algorithmische Bestimmung. In C. Dahlhaus & M. Krause (Eds.), *Tiefenstruktur*. Berlin, Germany: Technical University of Berlin.

Terhardt, E. (1984). The concept of musical consonance: A link between music and psychoacoustics. *Music Perception*, *1*, 276–295.

Terhardt, E., Stoll, G., & Seewann, M. (1982a). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, *71*, 679–688. https://doi.org/10.1121/1.387544

Terhardt, E., Stoll, G., & Seewann, M. (1982b). Pitch of complex signals according to virtual-pitch theory: Tests, examples and predictions. *The Journal of the Acoustical Society of America*, *71*, 671–678. https://doi.org/10.1121/1.387543

Thagard, P. (2019). Cognitive science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Stanford, CA: Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2019/entries/cognitive-science

Tillmann, B., Bharucha, J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review*, *107*, 885–913. https://doi.org/10.1037/0033-295X.107.4.885

Tillmann, B., & Bigand, E. (2001). Global context effect in normal and scrambled musical sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 1185–1196.

Toiviainen, P., & Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, *32*, 741–766. https://doi.org/10.1068/p3312

Toro, J. M., & Crespo-Bojorque, P. (2017). Consonance processing in the absence of relevant experience: Evidence from nonhuman animals. *Comparative Cognition & Behavior Reviews*, *12*, 33–44. https://doi.org/10.3819/CCBR.2017.120004

Trainor, L. J., & Heinmiller, B. M. (1998). The development of evaluative responses to music: Infants prefer to listen to consonance over dissonance. *Infant Behavior and Development*, *21*(1), 77–88. https://doi.org/10.1016/S0163-6383(98)90055-8

Trainor, L. J., Marie, C., Bruce, I. C., & Bidelman, G. M. (2014). Explaining the high voice superiority effect in polyphonic music: Evidence from cortical evoked potentials and peripheral auditory models. *Hearing*

Research, *308*, 60–70. https://doi.org/10.1016/j.heares.2013.07.014

Trainor, L. J., Tsang, C. D., & Cheung, V. H. W. (2002). Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception*, *20*, 187–194.

Trulla, L. L., Stefano, N. D., & Giuliani, A. (2018). Computational approach to musical consonance and dissonance. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00381

Tymoczko, D. (2003). Progressions fondamentales, fonctions, degrés: Une grammaire de l'harmonie tonale élémentaire. *Musurgia*, *10*(3/4), 35–64.

Tymoczko, D. (2006). The geometry of musical chords. *Science*, *313*, 72–74. https://doi.org/10.1126/science.1126287

Tymoczko, D. (2008). Scale theory, serial theory and voice leading. *Music Analysis*, *27*, 1–49. https://doi.org/10.1111/j.1468-2249.2008.00257.x

Tymoczko, D. (2011). *A Geometry of Music*. New York, NY: Oxford University Press.

Tymoczko, D. (2016). In quest of musical vectors. In J. B. L. Smith, E. Chew, & G. Assayag (Eds.), *Mathemusical conversations* (pp. 256–282). London, UK: World Scientific.

Ushakov, Y. V., Dubkov, A. A., & Spagnolo, B. (2010). Spike train statistics for consonant and dissonant musical accords in a simple auditory sensory model. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *81*(4). https://doi.org/10.1103/PhysRevE.81.041911

Vassilakis, P. N. (2001). *Perceptual and physical properties of amplitude fluctuation and their musical significance* (PhD thesis). UCLA, Los Angeles, CA.

Vassilakis, P. N. (2005). Auditory roughness as a means of musical expression. In R. A. Kendall & R. H. Savage (Eds.), *Selected Reports in Ethnomusicology: Perspectives in Systematic Musicology* (Vol. 12, pp.

119–144). Los Angeles, CA: Department of Ethnomusicology, University of California.

Vencovský, V. (2016). Roughness prediction based on a model of cochlear hydrodynamics. *Archives of Acoustics*, *41*(2), 189–201. https://doi.org/10.1515/aoa-2016-0019

Viro, V. (2011). Peachnote: Music score search and analysis platform. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 359–362). Miami, FL.

Virtala, P., Huotilainen, M., Partanen, E., Fellman, V., & Tervaniemi, M. (2013). Newborn infants' auditory system is sensitive to Western music chord categories. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00492

Vos, J. (1986). Purity ratings of tempered fifths and major thirds. *Music Perception*, *3*, 221–257.

Vuvan, D. T., & Hughes, B. (2019). Musical style affects the strength of harmonic expectancy. *Music & Science*, *2*. https://doi.org/10.1177/2059204318816066

Vyčinienė, D. (2002). Lithuanian Schwebungsdiaphonie and its south and east European parallels. *The World of Music*, *44*(3), 55–57.

Wang, Y. S., Shen, G. Q., Guo, H., Tang, X. L., & Hamade, T. (2013). Roughness modelling based on human auditory perception for sound quality evaluation of vehicle interior noise. *Journal of Sound and Vibration*, *332*(16), 3893–3904. https://doi.org/10.1016/j.jsv.2013.02.030

Watanabe, S., Uozumi, M., & Tanaka, N. (2005). Discrimination of consonance and dissonance in Java sparrows. *Behavioural Processes*, *70*(2), 203–208. https://doi.org/10.1016/j.beproc.2005.06.001

Weij, B. van der, Pearce, M. T., & Honing, H. (2017). A probabilistic model of meter perception: Simulating enculturation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00824

Weisser, S., & Lartillot, O. (2013). Investigating non-Western musical timbre: A need for joint approaches. In *Proceedings of the Third International Workshop on Folk Music Analysis* (pp. 33–39). Amsterdam, The Netherlands.

Wever, E. G., Bray, C. W., & Lawrence, M. (1940). The origin of combination tones. *Journal of Experimental Psychology*, *27*, 217–226.

White, C. W. (2018). Feedback and feedforward models of musical key. *Music Theory Online*, *24*(2). `https://doi.org/10.30535/mto.24.2.4`

White, C. W., & Quinn, I. (2018). Chord context and harmonic function in tonal music. *Music Theory Spectrum*, *40*, 314–335O. `https://doi.org/10.1093/mts/mty021`

Whorley, R. P., & Conklin, D. (2016). Music generation from statistical models of harmony. *Journal of New Music Research*, *45*, 160–183.

Whorley, R. P., Wiggins, G. A., Rhodes, C., & Pearce, M. T. (2013). Multiple viewpoint systems: Time complexity and the construction of domains for complex musical viewpoints in the harmonization problem. *Journal of New Music Research*, *42*, 237–266. `https://doi.org/10.1080/09298215.2013.831457`

Wightman, F. L. (1973). The pattern-transformation model of pitch. *The Journal of the Acoustical Society of America*, *54*, 407–416. `https://doi.org/10.1121/1.1913592`

Winograd, T. (1968). Linguistics and the computer analysis of tonal harmony. *Journal of Music Theory*, *12*, 2–49.

Woolhouse, M., Cross, I., & Horton, T. (2016). Perception of nonadjacent tonic-key relationships. *Psychology of Music*, *44*(4), 802–815. `https://doi.org/10.1177/0305735615593409`

Zajonc, R. B. (2001). Mere exposure: A gateway to the subliminal. *Current Directions in Psychological Science*, *10*(6), 224–228.

Zentner, M. R., & Kagan, J. (1998). Infants' perception of consonance and dissonance in music. *Infant Behavior and Development*, *21*(3), 483–492.

Zou, G. Y. (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*(4), 399–413. https://doi.org/10.1037/1082-989X.12.4.399

Zwicker, E., Flottorp, G., & Stevens, S. S. (1957). Critical band width in loudness summation. *The Journal of the Acoustical Society of America*, *29*, 548–557. https://doi.org/10.1121/1.1908963