

Prediction in polyphony: modelling musical auditory scene analysis

by
Sarah A. Sauvé

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

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

September 2017

Statement of Originality

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

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

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

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

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

Signature: Sarah A Sauv 

Date: 1 September 2017

Details of collaboration and publication

One journal article currently in review and one paper uploaded to the ArXiv database contain work presented in this thesis. Two conference proceedings papers contain work highly related to, and fundamental to the development of the work presented in Chapters 5 and 7. Details of these as well as their location in the thesis and details of the collaborations, where appropriate, are presented here:

Chapter 5

- Sauv , S. A., Pearce, M. T. & Stewart, L. (2014). The effect of musical training on auditory grouping. In *Proceedings of the ICMPC-APSCOM 2014 Joint Conference*. Seoul, Korea.
- Sauv , S. A., Pearce, M. T. & Stewart, L. (2017). Attention but not musical training affect auditory grouping. *ArXiv*.

Chapter 6

- Sauv , S. A., Sayed, A., Dean, R. T. & Pearce, M. T. (in review). Effects of pitch and timing expectancy on musical emotion. *Psychomusicology*.

In this project, Sayed designed the experiment with Pearce and collected the data as part of a master's thesis. I, with support from Dean, carried out the analysis presented in this thesis. I also, with the support of Dean and Pearce and final approval of Sayed, wrote the manuscript submitted to Psychomusicology and currently under review.

Chapter 7

- Bussi, M., Sauvé, S. & Pearce, M. (2016). Relative Saliency in Polyphonic Music. In *Proceedings of the 14th International Conference for Music Perception and Cognition*. San Francisco, USA.

In this project, Bussi and myself designed the experiment, with support from Pearce. Bussi collected data and analysed it with my support. He also wrote the manuscript with support from myself and Pearce. This project inspired the second experiment of Chapter 7, where the same concept is applied to newly designed stimuli.

Abstract

How do we know that a melody is a melody? In other words, how does the human brain extract melody from a polyphonic musical context? This thesis begins with a theoretical presentation of musical auditory scene analysis (ASA) in the context of predictive coding and rule-based approaches and takes methodological and analytical steps to evaluate selected components of a proposed integrated framework for musical ASA, unified by prediction. Predictive coding has been proposed as a grand unifying model of perception, action and cognition and is based on the idea that brains process error to refine models of the world. Existing models of ASA tackle distinct subsets of ASA and are currently unable to integrate all the acoustic and extensive contextual information needed to parse auditory scenes. This thesis proposes a framework capable of integrating all relevant information contributing to the understanding of musical auditory scenes, including auditory features, musical features, attention, expectation and listening experience, and examines a subset of ASA issues – timbre perception in relation to musical training, modelling temporal expectancies, the relative salience of musical parameters and melody extraction – using probabilistic approaches. Using behavioural methods, attention is shown to influence streaming perception based on timbre more than instrumental experience. Using probabilistic methods, information content (IC) for temporal aspects of music as generated by IDyOM (information dynamics of music; Pearce, 2005), are validated and, along with IC for pitch and harmonic aspects of the music, are subsequently linked to perceived complexity but not to salience. Furthermore, based on the hypotheses that a melody is internally coherent and the most complex voice in a piece of polyphonic music, IDyOM has been extended to extract melody from symbolic representations of chorales by J.S. Bach and a selection of string quartets by W.A. Mozart.

Acknowledgements

I am endlessly grateful to my supervisor, Dr. Marcus T. Pearce, for his unwavering support and patience throughout this journey, particularly as I learned to code in LISP. Thank you also to my remaining supervisory panel, Matthias Mauch, Elaine Chew, and Simon Dixon, for helpful feedback as I progressed through critical stages of this PhD.

Thank you to all the members of the Music Cognition Lab who have offered friendship, feedback and support over the past three years. Thank you also to the Music Performance and Expression Lab, and the Centre for Digital Music and the Cognitive Science research groups for friendships, advice, and making this musician feel welcome in a computer science department.

A specific thank you must go to Peter Harrison, lab mate, flatmate and friend, without whom I would be lost in code and who has been such an encouragement in the last few weeks of this process.

Thanks to CISV, and all my friends in the organization, who provided invaluable emotional support to an expat far away from her home, and encouraged me to always continue to challenge myself.

Finally, thanks to my family and my closest friends, who have generously tolerated the incessant always-working student mode both when I was at home and from abroad, and supported me through some significant life and lifestyle decisions.

I could not have done this without all of you – thank you.

Table of Contents

Statement of Originality	2
Details of collaboration and publication	3
Abstract	5
Acknowledgements	6
Table of Contents	7
List of Figures	14
List of Tables	17
1 Introduction.....	20
2 Auditory streaming: understanding the special case of music perception	25
2.1 Auditory features.....	28
2.2 Musical features	32
2.3 Attention	36
2.4 Expectation	38
2.5 Musical training	42
2.6 Predictive coding.....	45
3 Materials	52
3.1 IDyOM	52
3.2 Musical corpora.....	59
3.2.1 Montreal Billboard Corpus	60

3.2.2 Bach Chorale Dataset	60
3.2.3 String Quartet Dataset	61
3.2.4 Nova Scotia Folk Songs	61
3.2.5 Essen Folksong Collection Subset	61
3.2.6 Bach Soprano	62
3.3 Goldsmiths Musical Sophistication Index	62
3.3.1 Perceptual tests	63
3.3.2 Self-report questionnaire	63
4 An integrated framework for musical auditory streaming	65
4.1 General models of auditory streaming	66
4.1.1 Neural-based models	66
4.1.2 Temporal coherence models	69
4.1.3 Prediction-based models	70
4.1.4 Hybrid models	71
4.2 Musical auditory streaming models	73
4.2.1 Perceptual principles	73
4.2.2 Score-based models	76
4.2.3 Data-driven models	78
4.2.4 Summary	80
4.3 A new framework for musical ASA	81
4.3.1 Basic framework function	82

4.4 Streaming information sources in the framework.....	85
4.4.1. Feature-based modules	85
4.4.2 Including attention.....	87
4.4.3 Timbre	88
4.4.4 Including musical training	89
4.5 Training data	90
4.6 Model comparisons	91
4.7 Sample use cases	94
4.8 Limitations	95
4.9 Conclusion	96
5 Attention but not musical training affects auditory grouping.....	98
5.1 The study of timbre perception	99
5.2 Study 1: Timbre as a streaming cue	101
5.2.1 Participants.....	102
5.2.2 Stimuli	103
5.2.3 Procedure	103
5.2.4 Analysis	104
5.2.5 Results	105
5.2.6 Discussion.....	109
5.3 Study 2: Expectancy control	111
5.3.1 Participants.....	111

5.3.2 Stimuli	112
5.3.3 Procedure	112
5.3.4 Results	113
5.3.5 Discussion.....	115
5.4 Study 3: Ecological extension.....	116
5.4.1 Participants.....	117
5.4.2 Stimuli	117
5.4.3 Procedure	119
5.4.4 Results	120
5.4.5 Discussion.....	122
5.5 General Discussion.....	122
6 Modelling temporal expectations alongside pitch expectations	125
6.1 Rhythm and Predictability	126
6.2 Emotion and Predictability	128
6.2.1 Current study.....	131
6.3 Method.....	133
6.3.1 Participants.....	133
6.3.2 Stimuli	134
6.3.3 Procedure	139
6.3.4 Data collection	140
6.3.5 Statistical analysis	141

6.4 Results	143
6.4.1 Melody level analysis	143
6.4.2 Cross-sectional time series analysis	149
6.4.3 Relative salience.....	158
6.5 Discussion.....	159
6.5.1 Relative salience.....	166
6.6 Conclusion	168
7 Relative salience in polyphonic music	170
7.1 Defining Saliency.....	171
7.2 Materials	178
7.2.1 Simple target stimuli	179
7.2.2 Complex target stimuli	180
7.2.3 Melodic complexity levels.....	180
7.2.4 Harmonic complexity levels	181
7.2.5 Rhythmic complexity levels	181
7.3 Linking information content to complexity.....	186
7.3.1 Participants.....	186
7.3.2 Procedure	186
7.3.3 Analysis	187
7.3.4 Results	188
7.3.5 Discussion.....	194

7.4 Linking complexity to salience.....	197
7.4.1 Participants.....	201
7.4.2 Procedure	201
7.4.3 Analysis	202
7.4.4 Results	204
7.4.5 Discussion.....	216
7.5 General Discussion.....	219
7.6 Conclusion	223
8 Prediction-based melody extraction.....	224
8.1 Melody: definition and literature	225
8.1.1 Melody extraction from audio	230
8.1.2 Melody extraction from symbolic data.....	233
8.2 Melody extraction: a prediction-based approach	238
8.3.1 Model implementation.....	239
8.3 Model evaluation.....	244
8.3.1 Bach chorales	244
8.3.2 String quartets	254
8.4 Discussion.....	2577
8.4.1 Bach chorales	2577
8.4.2 String Quartets	26363
8.4.3 General Discussion.....	2755

8.5 Conclusion	2799
9 Conclusion	281
9.1 Summary	28282
9.2 Outcomes	2844
9.3 Limitations	2855
9.4 Framework evaluation	2866
9.4.1 Timbre, musical training, attention and expectation	2866
9.4.2 Temporal viewpoints	2877
9.4.3 Relative salience	2877
9.4.4 Melody extraction	29090
9.4.5 Considering Context	29191
9.4.6 Considering Heuristics	29191
9.5 Future directions	2944
Appendix A	297
Bibliography	2999

List of Figures

2.1 An example of pseudo-polyphony in Bach’s Violin Partita No. 3, BWV 1006.....	27
2.2 Illustration of the integrated and segregated percepts and fission and temporal coherence boundaries.....	30
2.3 Illustration of the frequency dimension in stimuli used by Bendixen et al. (2010).	39
3.1 Context tree for the word <i>abracadabra</i>	54
3.2 An example IDyOM analysis.....	55
3.3 Illustration of the expansion possibilities in IDyOM.	57
3.4 The musical sophistication sub-scale of the Gold-MSI.....	64
4.1.Illustration of a framework module’s working process.....	83
4.2.Illustration of the overall framework’s working process	84
5.1 Illustration of ABA_ paradigm, ascending and descending, modifying timbre only.	102
5.2 Percent target timbre contained in the morphing stream at the point of a change in percept.	107
5.3 Percent target timbre of the morphed stream at the point of a change in percept for brass and string instrumental family groups.....	108
5.4 Timbre dissimilarity ratings.....	114
5.5 Mean of initial and matching variables, both significantly different from chance.....	114
5.6 Single and double click tracks presented to participants alongside the relevant audio files.	120

6.1 Excerpts from one melody from each of the four different predictability-based types of experimental stimuli.....	135
6.2 Mean pitch IC and onset IC of each melody.	139
6.3 Mean expectancy, arousal and valence ratings for each melody.	147
6.4 Box plots illustrating important mean comparisons.....	148
6.5 Intercept and slope values of random effects on Participant and MelodyID for expectancy, arousal and valence models.	156
6.6 Expectancy, arousal and valence ratings for single randomly selected participants.....	157
7.1 Simple target stimuli.	178
7.2 Complex target stimuli.	179
7.3 The mean IC of each voice plotted for each of the 48 stimuli by type of manipulation and type of target.	183
7.4 Scores of excerpts for complex target stimuli.....	184
7.5 Mean complexity ratings for each level by parameter.	187
7.6 Theoretical task performance patterns according to similarity and complexity hypotheses.	198
7.7 Mean hit rate by objective complexity level.....	203
8.1 Illustration of window formation in Madsen & Widmer’s (2007) streaming algorithm.	236
8.2 Illustration of the iteration process of the first and second model implementations.	240
8.3 Illustration of the iteration process of the third model implementation.	242
8.4 First measures of Bach Chorale No. 6.....	259
8.5 Opening three measures of Mozart K590, movement 1.....	266

8.6 Examples of pattern detection in Mozart string quartet movement K428-1.	267
8.7 Examples of pattern detection in Mozart string quartet movement K428-2.	268
8.8 Examples of pattern detection in Mozart string quartet movement K458-3.	268
8.9 The principle melodic pattern of Mozart's movement K464-2.	269
8.10 Examples of pattern detection in Mozart string quartet movement K499-3.	270
8.11 Examples of pattern detection in Mozart string quartet movement K575-2.	270
8.12 Examples of pattern detection in Mozart's string quartet movement K590-1.....	272
8.13 Mozart's string quartet K428, movement 1, mm34-8.	273

List of Tables

3.1 All basic and derived viewpoints of the musical surface implemented in IDyOM and used in this thesis.	56
3.2 Summary of viewpoints included in optimal models in the re-analysis of three data sets.	58
3.3 Catalogue of movement details of the seven Mozart string quartets from the test set of the String Quartet Dataset.	61
5.1 Mean percent of trial duration by standard and target timbre, and by instrumentalist. ..	106
5.2 Experimental design.	118
5.3 Details of mixed effects binomial logistic regression.	121
6.1 Details of the datasets used in stimulus selection.	136
6.2 Details of the training sets used to train IDyOM.	136
6.3 Mean pitch IC, mean onset IC, mean MIDI pitch and mean IOI for all melody types...	137
6.4 Summary of 16 original melodies.	138
6.5 CSTSA modelling of expectancy ratings for all melodies.	151
6.6 CSTSA modelling of arousal ratings for all melodies.	152
6.7 CSTSA modelling of valence ratings for all melodies.	154
6.8 Coefficients of sub-models for expectancy ratings.	158
6.9 Coefficients of sub-models for arousal ratings.	161
6.10 Coefficients of sub-models for valence ratings.	162
7.1 Summary of the fixed effects manipulation model.	189
7.2 Summary of the maximally fitted manipulation model.	190

7.3 Summary of the fixed effects information content model.....	191
7.4 Summary of the maximally fitted information content model.	191
7.5 Summary of the fixed effects musical properties model.....	193
7.6 Summary of the maximally fitted musical properties model.....	193
7.7 Illustration of experimental design.....	200
7.8 Summary of the fixed effects manipulation model for complex targets.	205
7.9 Summary of the maximally fitted manipulation model for complex targets.....	206
7.10 Summary of the fixed effects information content model for complex targets.	207
7.11 Summary of the maximally fitted manipulation model for complex targets.....	207
7.12 Summary of the fixed effects musical properties model for complex targets.	209
7.13 Summary of the maximally fitted manipulation model for complex targets.....	209
7.14 Summary of the fixed effects complexity ratings model for complex targets.....	210
7.15 Summary of the maximally fitted complexity ratings model for complex targets.	210
7.16 Summary of the fixed effects manipulation model for simple targets.	211
7.17 Summary of the maximally fitted manipulation model for simple targets.....	212
7.18 Summary of the fixed effects information content model for simple targets.	213
7.19 Summary of the maximally fitted information content model for simple targets.....	213
7.20 Summary of the fixed effects musical properties model for simple targets.	214
7.21 Summary of the maximally fitted musical properties model for simple targets.....	214
7.22 Summary of the fixed effects complexity ratings model for simple targets.....	215
7.23 Summary of the maximally fitted complexity ratings model for simple targets.	215
8.1 Details of audio datasets used for melody extraction evaluation.....	229
8.2 Details of symbolic datasets used in melody extraction evaluation.....	230

8.3 Tabulation of all transitions between bars 1 and 2 of Mozart’s String Quartet No. 16, second movement.	240
8.4 Tabulation of all transitions between bars 5 and 6 of Mozart’s String Quartet No. 16, first movement.	242
8.5 Results of the evaluation of two model versions on the Bach chorale test dataset.....	248
8.6 Results of the evaluation of two model versions on the Bach chorale validation dataset.	249
8.7 Results of the evaluation of two model versions on the Bach chorale datasets using entropy	251
8.8 Results of the evaluation of two model versions on the Bach chorale validation dataset using entropy	252
8.9 Mean percent match for the Mozart string quartets between extracted melody and annotated melody.	256
8.10 Mean information content for each score-based voice and annotated melody.....	274
8.11 Mean information content of the extracted melody voice.....	274

1 Introduction

Take a moment to listen to your favourite song in your mind's ear. Hum the melody. Pay attention to the accompaniment. What subtleties are you remembering this time? Perhaps the bass line surfaces with an interesting lick, or the horns call your attention. Our memory for music is amazing, especially for pieces we know and love. But in order to remember them, we first have to first make sense of them; organize the mass of soundwaves hitting our eardrums into something comprehensible. Where is the melody? Is there even a melody? Is there a theme? How is it developed? What instruments are involved and how do they blend together or split apart into merging or dividing lines? Just like we understand the world around us by matching sounds to their sources through a process called auditory scene analysis (ASA; (Bregman, 1990), music can be described as a special case of ASA, where we understand music by organizing it into melody and accompaniment, such as is typical in a classical period piano

sonata, or one unified sound mass, such as in homophonic chorales by J.S. Bach, or many interacting independent voices, such as in a baroque fugue. Unlike typical day-to-day ASA, in music the auditory system is manipulated by composers into perceiving more or fewer sound sources than there are instruments. In a sense, composers are masters of auditory illusions. So how does this auditory illusion called music become organized into melody, accompaniment, counterpoint, or sound mass and phrases, movements and pieces?

This thesis has two main goals: 1) it proposes an integrated framework for musical ASA that combines multiple sources of information to create an analytical tool for future research in auditory streaming; 2) it examines a range of relevant aspects of musical ASA presented in this theoretical model using a probabilistic approach inspired by predictive coding (Clark, 2013b).

There are many sources of information contributing to the formation of an organized auditory scene. In the case of music, we can summarize these into five high-level categories, incorporating top-down and bottom-up, as well as vertical and horizontal aspects of music perception.

Auditory features. The first is auditory features, which address all the bottom-up acoustic information processed from the cochlea and the primary auditory cortex before being passed to higher-level processes. These include such cues as frequency, location, intensity, timbre and rate of occurrence (Bregman, 1990; Hartmann & Johnson, 1991; Iverson, 1995; Marozeau, Innes-Brown, & Blamey, 2013; Micheyl & Oxenham, 2010; van Noorden, 1975; Vliegen, Moore, & Oxenham, 1999). All of these cues apply to music as well as to auditory scenes in everyday life.

Musical features. The second category is musical features, which address relevant information specific to music including harmonic relationships, phrase boundary perception, repetition and similarity. Though harmonic quality is a feature of all sounds, music is

overwhelmingly harmonic and the relationship between simultaneous pitches in multi-instrument music is crucial to understanding musical scene analysis. Boundary detection is important to comprehension of both speech and music, and horizontal musical structure is important to organizing a musical auditory scene. Similarly, repetition at the scale at which it happens in music is a unique feature of this sound world, where exact repetitions are not only normal but highly enjoyed. If it weren't, no recorded music could ever be sold or consumed. Not all repetition is exact however, and similarity plays a large part in musical structure, making it also relevant to breaking down a musical auditory scene.

Attention. The third category is attention, where a listeners' attentional set lends top-down biases to their perception of the auditory scene, musical or otherwise (Barnes & Jones, 2000; E. Bigand, McAdams, & Forêt, 2000; Carlyon, Cusack, Foxton, & Robertson, 2001; Macken, Tremblay, Houghton, Nicholls, & Jones, 2003).

Expectation. The fourth category is expectation, where again a listeners' expectations are a top-down influence on their perception, and expectancy has recently been investigated as an auditory streaming cue (Bendixen, Denham, Gyimesi, & Winkler, 2010; Huron, 2006; Schröger et al., 2014; Southwell et al., 2017).

Musical training. Finally, the last category is listener experience, which here will be approximated by musical training as measured by the training sub-scale of the Goldsmiths Musical Sophistication Index (Gold-MSI). Musical expertise has been shown to affect low-level perceptual skills (Fujioka, Ross, Kakigi, Pantev, & Trainor, 2006; Fujioka, Trainor, Ross, Kakigi, & Pantev, 2004; Micheyl, Delhommeau, Perrot, & Oxenham, 2006; Skoe & Kraus, 2013; Strait, Parbery-Clark, Hittner, & Kraus, 2012), cognitive skills (Carey et al., 2015; Corrigan & Trainor, 2011; Franklin et al., 2008; Tsang & Conrad, 2011) and expectancies (Hansen & Pearce, 2014) in relation to both music and everyday listening situations.

This thesis will address all these categories at least once throughout in varying combinations, with each chapter exploring one subset of auditory streaming. As mentioned, these subsets will be investigated using primarily probabilistic methods. To do so, an existing predictive model implementing statistical learning of various aspects of the musical surface will be used. This model is called IDyOM, or information dynamics of music (Pearce, 2005), and has been validated behaviourally, computationally and neurophysiologically (Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010) and is increasingly used and discussed in the literature (Dean, 2016; Demorest & Morrison, 2016; Gingras et al., 2016; Hansen & Pearce, 2014; Hansen, Vuust, & Pearce, 2016; Pearce & Müllensiefen, 2017; Pearce, Müllensiefen, & Wiggins, 2010; van der Weij, Pearce, & Honing, 2017).

The thesis will be structured as follows. Chapter 2 will introduce the literature surrounding auditory stream analysis, both musical and non-musical, as well as provide an introduction to predictive coding. Chapter 3 will present IDyOM in more detail, as well as our chosen measurement of musical training, the Goldsmiths Musical Sophistication Index (Gold-MSI) and all musical corpora used. Chapter 4 will present a theoretical integrated framework for musical ASA that will integrate all five sources of information listed above using statistical learning models (IDyOM and extensions) to process music in both horizontal (over time) and vertical (simultaneous) aspects. These will be applied in *modules*, units of analysis each addressing a particular subset of the musical ASA problem. Together, these modules will estimate the perceived simultaneous organization of a piece of music at any given time and identify the melody. This design by module, unified by the concept of prediction is intended to allow the model to expand and to be developed collaboratively, a rare occurrence in current music cognition research. It also allows the evaluation of isolated concepts, a number of which will be addressed in this thesis. The first of these, Chapter 5, will use a standard behavioural

streaming paradigm to examine the effects of attention, expectation and musical training on streaming perception as manipulated by timbre, incorporating four of the five above-mentioned categories, namely auditory features, attention, expectation and musical training. Chapter 6 will achieve two things: it will validate some of IDyOM's temporal viewpoints, where previous validation of the model focuses on the pitch domain. It will also investigate the role of pitch and temporal predictability on the perception of musical emotion while taking into account musical training, incorporating two streaming information categories: expectation and musical training. Chapter 7 will explore the link between melodic, rhythmic and harmonic expectancy and salience, where the relative salience of musical parameters is crucial to auditory streaming. The focus of attention on pitch content as opposed to rhythmic or harmonic content for example would lead to organization of an auditory scene relying more on pitch features while attention towards rhythm would lead to a scene organized more heavily by temporal features. The chapter also links expectancy with perceived complexity while considering musical training, thus once again including three categories of streaming information. Chapter 8 presents an extension of IDyOM in the form of a melody extracting feature based on the hypotheses that a melody is both internally coherent and the most interesting stream in a piece of music. Finally, Chapter 9 will both summarize the work presented in this thesis and re-evaluate the theorized integrated framework for ASA in the context of the findings presented in Chapters 5 through 8.

2 Auditory streaming

Understanding the special case of music perception

Auditory scene analysis (ASA) is the process of parsing the jumble of messy sound waves that reach our eardrums into coherent, sensible and interpretable sound sources. Coined by Albert Bregman, ASA has many parallels with visual scene analysis in the use of Gestalt psychology to explain grouping and in the consideration of the top-down influences of attention and expectation in scene perception. Parsing perceptual scenes can be looked at as a segregation problem, where the challenge is to separate sound sources from each other in a ‘messy’ scene, or an integration problem, where the challenge is to bind features together to form a coherent sound source from diverging input. The problem can also be divided into vertical and horizontal aspects of sound organization, attempting to understand grouping of

simultaneous, or sequential sounds respectively. Either way, a successful computer model of ASA – one that can parse an auditory scene like a human does – is the ultimate goal of investigating ASA, where the assumption is that a successful model provides support for a given hypothesis regarding how the brain parses auditory scenes. This problem is also called stream segregation, where it is preferred to separate a full scene by its constituent sources rather than build up a source from parts of the scene. For the remainder of this thesis, the terms auditory scene analysis and stream segregation will be used interchangeably. By simulating stream segregation, problems like declining speech-in-noise perception in aging adults and improving music perception for cochlear implant users can be tackled, increasing the quality of life of thousands of individuals.

Bregman presented an excellent analogy of the kind of challenge the brain is facing when performing ASA:

“Imagine that you are on the edge of a lake and a friend challenges you to play a game. The game is this: Your friend digs two narrow channels up from the side of the lake. Each is a few feet long and a few inches wide and they are spaced a few feet apart. Halfway up each one, your friend stretches a handkerchief and fastens it to the sides of the channel. As waves reach the side of the lake they travel up the channels and cause the two handkerchiefs to go into motion. You are allowed to look only at the handkerchiefs and from their motions to answer a series of questions: How many boats are there on the lake and where are they? Which is the most powerful one? Which one is closer? Is the wind blowing? Has any large object been dropped suddenly into the lake?” (Bregman, 1990, p. 5-6)

Answering these questions seems impossible, but his analogy is strictly representative of our auditory system, where the lake is the air, the channels are ear canals and the handkerchiefs are ear drums, vibrating with the multitude of incoming sound waves from



Figure 2.1. An example of pseudo-polyphony in Bach’s Violin Partita No. 3, BWV 1006. The lower notes are perceptually segregated from the higher notes, creating an illusion of two voices from the single instrument.

objects in the environment. This analogy clearly illustrates how difficult understanding the auditory environment is, though it is taken for granted at every moment of every day.

This thesis is concerned with the special case of music. Though not a universally accepted definition, music is often described as organized sound (Varèse & Wen-Chung, 1966, p.). It can also be described as a case of auditory scene illusions where instruments are blended to simulate a common source or played to simulate separate sources (called pseudo-polyphony as found in Bach solo string music, Figure 2.1). Despite this trickery, humans love music; there is no known human society that did not have music as part of its culture. Is music perception a special case of perception or can more general perceptual principles successfully be applied to music? This thesis proposes an integrated framework for auditory streaming based on the general perceptual principle of prediction, where the brain learns about patterns in music just like it can learn about anything else, such as language. In this framework, the specificity of music perception, and similarly any other domain, again such as language, comes only from the types of patterns it contains. This pattern learning can be simulated by a computational model and applied in turn to simulate how a typical listener would organize a piece of music while listening to it.

As a complex process, there is much to consider when attempting to model stream segregation, especially in the context of music. First, the basic auditory features of the scene being heard must be included (Section 2.1); second, considering the musical relationships between parts on the score is important in this particular context (Section 2.2); third, attention

may be applied – an endogenous process such as choosing an instrument to listen to – or drawn – an exogenous process, for example by a loud outburst from the horns (Section 2.3); fourth, every listener develops expectations about what will come next (Section 2.4); and finally, the listener’s experience with that particular type of scene will affect how it is perceived (Section 2.5). For example, a mechanic might be able to diagnose an engine problem just by listening to it, and a conductor can identify exactly who in the orchestra played a wrong note! Bringing all these considerations together is a daunting task, but one overarching theory of brain function is capable of doing so: predictive coding (Section 2.6).

Before beginning any discussion of the streaming literature, it is worth defining the terms *voice* and *stream* as they will be used in this thesis. *Voice* defines a musical line either from a score (i.e., soprano, alto, tenor or bass voice in choral music) or constructed to match a line from a score. It is always monophonic. *Stream* defines a perceptual construct, such as melody or accompaniment. A stream may be monophonic or polyphonic. Furthermore, while auditory stream segregation can be applied to either audio or symbolic input, this thesis focuses exclusively on symbolic auditory streaming.

2.1 Auditory features

In this section, an overview of the most commonly researched auditory features in stream segregation is given, including frequency (or pitch), loudness, synchronicity, timbre, periodicity, location, and harmonicity.

The concept of streaming focuses foremost on studying sound segregation rather than sound integration. Integration is typically considered the default percept where the gathering of evidence from various auditory cues informs source separation (Bregman, 1990); therefore research focus tends to be on the point at which cues cause scenes to break apart into their constituent sources. Some of the earliest work in sound segregation began several decades ago,

when Miller and Heise (1950) presented participants with an alternating sequence of A and B tones to form an ABAB pattern. When A and B were close in pitch, participants heard something akin to a musical trill, while as the pitch difference between the A and B tones grew, the percept changed to two isochronous streams of tones, one with just As and one with just Bs. This was the first controlled demonstration of the role of pitch as a cue for sound segregation, where the magnitude of the pitch difference dictated whether a sequence of tones would be heard as coming from one integrated source or two distinct sources. Leon van Noorden (1975) extended this by showing that both an increase in pitch distance and an increase in tempo led to a streaming percept. van Noorden also introduced a clever alteration to the paradigm, where A tones sounded at twice the rate of the B tones, producing the pattern ABA-, where the '-' represents a silence. This small change causes two distinct percepts: when integrated, the sequence will sound like 'galloping' triplets; when segregated, a listener will hear two series of isochronous tones where one series is at twice the tempo of the other. These alternate percepts are illustrated in Figure 2.2a. Two thresholds were proposed in van Noorden's work, replacing the single trill threshold previously identified by Miller and Heise (1950) and thus offering more defined perceptual options: the fission boundary and the temporal coherence boundary. The fission boundary marks the point below which segregation is impossible while the temporal coherence boundary marks the point above which integration is impossible. This is illustrated in Figure 2.2b. These thresholds differ based on the perceptual cue being manipulated. For example, manipulating pitch alone, tempo alone and pitch and tempo together each produce a unique threshold profile describing when a listener would perceive the A and B tones as integrated or segregated. This paradigm has recently been extended to more complex scenes containing three different frequencies, or ABC tones rather than only A and B tones (Thomassen & Bendixen, 2017). The role of frequency distance, and

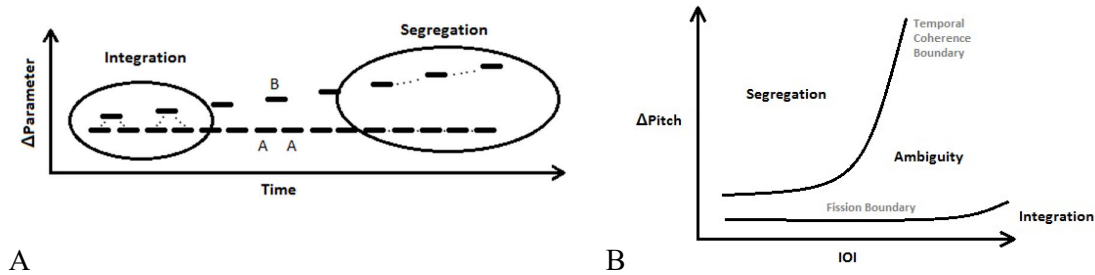


Figure 2.2. Illustration of the integrated and segregated percepts (A) and fission and temporal coherence boundaries (B) and how these separate integrated and segregated percepts based on pitch distance and inter-onset-interval (IOI).

additionally location was demonstrated in this more complex context. Furthermore, outer voices were more often perceived in the foreground as compared to the middle voice.

Since the pioneering work that studied the effects of frequency and tempo on stream segregation, many other parameters have been studied as potential streaming cues. These include: synchrony, harmonicity, timbre, loudness, location, attention and predictability. In the first case, sequences of tones presented in synchrony promote integration (Duane, 2013), sometimes even at large frequency separation (Micheyl, Kreft, Shamma, & Oxenham, 2013). For example, Micheyl and colleagues (2013) presented participants with a 9-tone target sequence surrounded by multiple channels of masking tones that were either synchronous or asynchronous with the target sequence and asked participants to identify whether the 9th target tone was higher or lower than the other eight. While also investigating harmonicity (here harmonic relationship between streams) as a segregating cue, the results indicated that both cues influenced results, where asynchrony and inharmonicity (unharmonic relationship) between target and distractors promoted segregation. Using a very different approach, Duane (2013) calculated the percent of synchronous onsets and offsets between pairs of instruments in 18th and 19th century string quartet expositions and found that these were strong predictors of integration, outperforming pitch comodulation and harmonic overlap.

Studies of timbre's effect on streaming break down into manipulations of spectral, temporal and periodicity sub-cues, where bigger differences between A and B tones, or target and distractor melodies on any of these cues, leads to increased segregation (Marozeau et al., 2013; Singh & Bregman, 1997; Vliegen et al., 1999). Spectral cues tend to be more important than temporal and periodicity cues, requiring less contrast to cause segregation, for example in a target/distractor paradigm where more similar distractors masked the target more effectively (Marozeau et al., 2013). Larger differences in loudness also promote segregation (Marozeau et al., 2013; van Noorden, 1975). Finally, humans are very good at locating sounds relative to each other, unless they are directly in front of or behind the listener, with conductors developing particularly good location-based segregation through training (Münste, Kohlmetz, Nager, & Altenmüller, 2001; Nager, Kohlmetz, Altenmüller, Rodriguez-Fornells, & Münste, 2003).

To complicate matters further, segregation and integration are not the only perceptual options. Recall the fission and temporal coherence boundaries from van Noorden's work. When first introducing these, only one side of each threshold was described, leaving out the space between them; this is addressed now. Just as visual stimuli can have alternating percepts (i.e. the vase/face illusion), so can auditory stimuli and this, it has been argued, is particularly helpful for studying the neural correlates of ASA (Pressnitzer, Suied, & Shamma, 2011) as it allows the matching of reported percepts to signals emitted in the brain. Bi-stability has been shown to occur in conditions spanning combinations of frequency difference from 1 to 24 semi-tones and tempo of 75-250ms stimulus onset asynchrony (SOA), with features investigated limited to frequency difference and tempo (Denham, Gyimesi, Stefanics, & Winkler, 2013). While the organization of these features affects the initial percept, bi-stability is present across all contexts and it is the rate of switching between percepts that is affected. Researchers who have considered this difficult problem conclude that bi-stability is necessary to any complete

model of stream segregation (Pressnitzer et al., 2011; Winkler, Denham, Mill, Bohm, & Bendixen, 2012). As previously mentioned, music is a special case of ASA and is in itself an elaborate auditory illusion, making it easy to agree with this conclusion: a model of stream segregation should consider multiple plausible percepts and select the most likely one, while allowing the possibility of changing to another. It is worth noting that calling music an auditory illusion implies that music, with perhaps the exception of some solo instrumental music, would tend not to include cues that would only fall either below the fission boundary or above the temporal coherence boundary but a mix of these in addition to cues that fall in between the boundaries, in the realm of ambiguity.

2.2 Musical features

While studies involving stream segregation in music do consider some of the basic auditory features introduced in Section 2.1 above (Cambouropoulos, 2008; Chew & Wu, 2004; Duane, 2013), some higher-level musical features must also be taken into consideration. These include harmony, phrase boundaries, repetition and similarity.

Harmony can be considered on a number of different levels. A combination of sine tones at different frequencies with specific ratio relationships between each consecutive ascending pair is considered a harmonic sound, or a complex tone, and is perceived as one single sound, a pitch, defined by the lowest frequency, the fundamental. If one of these sine tones is mistuned, it segregates from its harmonic complex and two tones are heard: one complex tone with pitch corresponding to the fundamental frequency and one sine tone at the mistuned frequency. The combination of complex tones with high overlap between component frequencies sound both pleasant and even richer. The most pleasant ratios, or the most consonant, are 2:1, 3:2 and 4:3, equivalent to the octave, fifth and fourth intervals in music theory. The most unpleasant ratio, or the most dissonant, is 45:32, the tritone. Musical

harmony is based on these relationships between complex tones, or pitches. Western classical music is highly consonant overall, where simultaneously sounding notes tend to have close harmonic relationships and these relationships change over time to form different types of simultaneous sonorities called chords: combinations of pitches with different ratios between them form different chord identities such as major, minor, diminished and augmented. This high degree of consonance favours the integration of simultaneous pitches into a single piece of music (Duane, 2013). It is possible to use harmony to group simultaneous pieces of music together, but this is a relatively rare compositional device used in the 20th century by composers such as Charles Ives, Igor Stravinsky, Béla Bartók and Raymond Murray-Schafer. In terms of musical auditory scene analysis, the harmonic nature of music implies strong vertical integration of a musical piece into one source. The balance between this overall vertical integration and the lower-level division of the piece into melody, counter-melody, accompaniment or texture is specific to music ASA.

Phrase boundary identification is an important part of music perception, lending structure to musical works and typically equating to sentences in speech. Research in this domain has explored the effects of musical training (Neuhaus, Knösche, & Friederici, 2006), cultural knowledge (Nan, Knösche, & Friederici, 2006; Nan, Knösche, Zysset, & Friederici, 2008), timing (Istók, Friberg, Huotilainen, & Tervaniemi, 2013; Palmer & Krumhansl, 1987; Silva, Barbosa, Marques-Teixeira, Petersson, & Castro, 2014) and Gestalt principles (Bod, 2002) on boundary identification in an effort to understand how listeners perceive these structural points in the music. Some prediction-based approaches have also been explored, where points of high predictability followed by low predictability signal the end, and beginning of a phrase, respectively (Aslin, Saffran, & Newport, 1999; Lattner, Grachten, Agres, & Chacón, 2015; Marcus T. Pearce, Müllensiefen, et al., 2010). Boundary perception works hand

in hand with auditory scene analysis, addressing the horizontal processing aspect of the problem (Bregman, 1990), where music is grouped into phrases over time. Phrase boundaries inform vertical auditory scene analysis by pinpointing potential locations of change in streaming structure, while a vertical streaming organization will tend to stay the same throughout a phrase.

Repetition is another defining characteristic of music, even labelled a music universal as it is present in all known music cultures (Nettl, 2010). It is particularly noticeable in popular music but also incredibly prevalent in Western art music. Not only is music internally repetitive, but listeners enjoy listening to repetitive music repeatedly. A speaker repeating an exact sentence two or three times in a row would be nonsensical, whereas phrase repetition in music is both normal and pleasant, particularly for unfamiliar music (Margulis, 2013). It has been proposed that this is because of the rewarding aspect of recognition, and correctly predicting the completion of a pattern once it has begun (Huron, 2006). Repetitions in music can be either exact, or varied in some way, while still being similar enough to the original to be recognized as related: these are variations. Algorithmically detecting repetition in music is a challenge that has been approached by a number of researchers over the past few decades, where aims include compression (how to more efficiently encode music; (Meredith, Lemström, & Wiggins, 2002; Rolland, 1999), generation (how to best incorporate repetition in music generation; Rolland, 1999) or cognitive modelling (how can human repetition detection be simulated; Cambouropoulos, 2006). Margulis has probed the processing of repetition (Margulis, 2012, 2014), finding that listeners are good at detecting repetition at an optimal segment length (with pattern length resulting in an inverted U-curve pattern of performance) and with increased exposure. Repetition also helps auditory scene analysis by emulating the principle of common fate, where similar sounds are presumed to come from one same source

(Bregman, 1990). Whether this source is an orchestra or a soloist depends on contextual information, which will be discussed further in Chapter 9.

Closely related to repetition is similarity: how similar must two passages be to be termed variations, and how is similarity encoded in the brain? Work in perceptual similarity attempts to understand what aspects of a sequence, such as contour, rhythm and mode, influence similarity judgments most. Halpern, Bartlett & Dowling (1998) found that young and old, musically trained and untrained listeners rating similarity of pairs of sequences (on a Likert scale) almost always based their ratings on mode first, followed by contour and finally by rhythm, so that a pair of sequences that differed only in mode were rated as more similar than a pair that differed only in rhythm. The only exception was that older musically trained listeners judged rhythms as more similar than contour, though rhythm was still a stronger predictor of dissimilarity than contour. Bartlett & Dowling (1980; 1981) found that differences between sequences were well recognized when pitch interval was modified, but contour and rhythm were constant, suggesting that interval patterns supersede these other measures of similarity. Modelling musical similarity is challenging, and has been the topic of special journal issues (I. Deliège, 2003; Toiviainen, 2007), specialized workshops (Benetos, 2015; Volk, Chew, Hellmuth Margulis, & Anagnostopoulou, 2016), the MIR community (Aucouturier, Pachet, & others, 2002; Berenzweig, Logan, Ellis, & Whitman, 2004; Bogdanov, Serra, Wack, Herrera, & Serra, 2011; Flexer, Schnitzer, & Schlüter, 2012; Li & Ogihara, 2004; Mardirossian & Chew, 2005, 2006; Pampalk, Flexer, Widmer, & others, 2005; Pohle, Schnitzer, Schedl, Knees, & Widmer, 2009; Slaney, Weinberger, & White, 2008; West & Lamere, 2007) and psychology research (Cambouropoulos, 2009; Irène Deliège, 2001, 2007; Harrison, Müllensiefen, & Collins, 2017; M. Pearce & Müllensiefen, 2017). It is relevant to

musical auditory streaming in a similar way to repetition, where similar sounds tend to group together as a common source.

The general concept to be drawn from the body of literature presented in the last two sections is that dissimilarity lends itself to sound segregation. This makes sense intuitively, as an object tends to produce sounds that are more similar than they are different, particularly when compared to other objects. Complications arise because sounds can be similar and dissimilar in various ways. Given a piece of music where the violin and the oboe play in unison (or at any other interval), their timbres alone are considered dissimilar (McAdams, Winsberg, Donnadieu, De Soete, & Krimphoff, 1995), encouraging segregation, while their absolute rhythmic unity – perfect similarity – encourages integration. In this case, it is most likely that a single musical line with an interesting timbre would be perceived rather than two separate lines that happen to be played at the same time. This is just one example of music as a special case of ASA where composers constantly and deliberately manipulate ASA processes to create captivating auditory illusions. This challenging question of relative salience and complex interplay of acoustic and musical parameters will be studied experimentally in Chapter 7 and addressed again in Chapter 9.

2.3 Attention

A particularly important, but immensely tricky aspect of auditory scene analysis to consider is attention. Attention has been a particularly interesting area of debate in the literature, where the principal question of interest has been whether attention is needed for the creation of auditory streams, or whether auditory streams are formed automatically, and attention is simply allocated to these streams. Researchers investigating this question often rely on the idea of auditory stream build-up, which claims that the default percept is to perceive everything as coming from one sound source, and divide it into multiple sound sources as

evidence for these builds up over time (Bregman, 1978). This build-up effect is observed as a gradual increase in the probability of hearing a sequence as segregated over time. Evidence for the necessity of attention for stream build-up is given by Carlyon et al. (Carlyon et al., 2001), where participants who were instructed to switch their attention from a different auditory task to a streaming task – for which an integrated percept made the task more difficult – performed worse than participants who had focused on the streaming task for the duration of the trial. However, more recent evidence suggests it is possible to initially hear a segregated percept (Deike, Heil, Böckmann-Barthel, & Brechmann, 2012; Susan L. Denham et al., 2013) and it has even been argued that this concept of streaming buildup is an artefact of averaging data across subjects (Pressnitzer & Hupé, 2006). This evidence calls into question previous assumptions made in attention research and highlights attention as a highly relevant and active research area.

One major criticism of work involving the investigation of attention is that auditory streaming paradigms require a listener's attention to complete the task, and attention is simply directed. For example, participants may be asked to identify an error or irregularity in one stream while ignoring the other (Bigand et al., 2000). Macken et al. (2003) use the irrelevant sound effect paradigm to address this problem, where task-irrelevant sounds distract from a serial recall task when they are integrated into a single stream but not when segregated, as integrated sequences contain a greater range of sounds while multiple segregated sequences each contain more similar sounds that can be grouped and therefore more easily ignored.

A particularly interesting attentional strategy proposed especially for music listening has been called prioritized integrated attention, and it includes a mix of integration and segregation (e.g. Uhlig, Fairhurst, & Keller, 2013). It is prioritized because a listener can choose to mainly integrate the music as a whole piece or segregate the music, focusing on one

or a few parts. Neither a perfect balance of the two strategies nor the use of only one strategy is considered possible. It is integrated because it includes both integrated and segregated percepts. This combination lends itself particularly well to music listening or music performance: hearing the melody clearly does not exclude hearing the integrated musical context and a performer must listen to what is happening around them while also focusing on their own part. Bigand and colleagues (Bigand et al., 2000) propose a similar mechanism when alternate possibilities including integration, divided attention, figure-ground and attentional switching are not supported by an experiment asking participants to detect wrong notes in simultaneously presented popular French folk songs (an octave apart). The discussion surrounding the effect and importance of attention for auditory streaming is far from over, but it is important to acknowledge that attention has an effect on perception and something as simple as the choice of instructions (i.e. try to segregate vs try to integrate) has an influence on the results obtained in a given experiment (van Noorden, 1975).

2.4 Expectation

Where this thesis will deal extensively with expectation, it is worth defining some recurring terms and concepts before diving in. *Expectation* will refer to the general concept of having an idea of what will happen in the future, near or far. *Expectancy* will be used when referring to the subjective feeling of expectation, i.e., specific expectancies in relation to a particular event or note. Prediction and uncertainty are also closely tied to the concept of expectation, where a *prediction* refers to a specific expected event (i.e., a tonic chord will follow the dominant chord in the last bar of this piece of music) and *uncertainty* reflects how sure an agent is about their prediction. It is important to note that expectancy and uncertainty can be completely contradictory. For example, it is possible to be absolutely certain and be

wrong. Therefore it is important to consider expectancy and uncertainty separately; this thesis focuses on expectancy.

Furthermore, there are multiple sources of expectancy (Huron, 2006). *Schematic* influences reflect general stylistic patterns acquired through extensive musical listening to many pieces of music while *veridical* influences reflect specific knowledge of a familiar piece of music. *Dynamic* influences reflect dynamic learning of structure within an unfamiliar piece of music (e.g. recognizing a repeated motif). Listening to new, unfamiliar music in a familiar style engages schematic and dynamic mechanisms, the former reflecting long-term learning over years of musical exposure and the latter short-term learning within an individual piece of music. Listening to familiar music engages both of these in addition to veridical expectations, though veridical expectation may be weighed more heavily.

While it is clear that expectations are developed during music processing, the role of predictability as a streaming cue itself has received little attention in the literature until recently (Andreou, Kashino, & Chait, 2011; Bendixen et al., 2010; S. L. Denham & Winkler, 2006; István Winkler, Denham, & Nelken, 2009). Presumably, regularity binds a sequence together into one stream while simultaneously ‘removing’ it from the rest of the auditory scene as it becomes a unified source, segregated from its context. An early study by French-St. George and Bregman (1989) in which two streams were manipulated to have either repeated 4-note patterns or random pitch patterns and isochronous or non-isochronous timing and where participants were asked to integrate the sequences into one stream found no effect of predictability on streaming percept. Similarly, Rogers and Bregman (1993) found that the regularity of an *induction* sequence did

Regular

Random

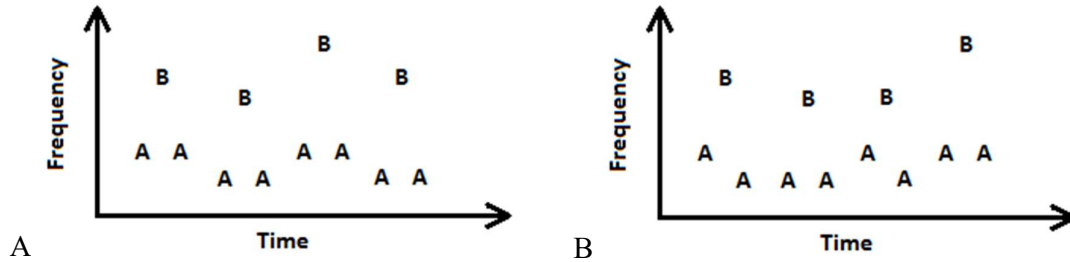


Figure 2.3. Illustration of the frequency dimension in stimuli used by Bendixen et al. (2010). There are two different A tones and 3 different B tones organized so that both sequences were predictable (A) or random (B) or a combination of the two, where one sequence was predictable and one random (not pictured).

not affect streaming. However, authors in both cases acknowledged that predictability in the streams translated to predictability of the overall sequence and that these two levels of predictability should be controlled for in the future. This was indeed picked up by Bendixen and her colleagues (Bendixen et al., 2010), who modified the typical ABA paradigm so that patterns could be regular only within certain streams, or across both. To do so, either frequency or intensity were random or regular in the A stream (two different tones), the B stream (three different tones), or both (Figure 2.3). This resulted in ten conditions with varying degrees of regularity over two parameters for the two streams. Participants reported a larger proportion and longer periods of segregated percept for increasingly regular conditions and a smaller proportion but equal periods of integrated percept for increasingly random conditions. The lack of an effect of predictability on the phase length of the integrated percept was interpreted as evidence that regularity helps stabilize a segregated percept once it has occurred and regularity can be detected, but does not have an effect on its formation. If regularity had an effect on stream formation, there would be a decrease in phase length of the integrated percept as it would switch to a segregated percept more quickly; this was not the case. Additionally, regularity in both intensity and frequency increased integration for random patterns and segregation for regular patterns more than regularity in one or no features, supporting a

proportional link between regularity and perception. Finally, attention seems to have an influence as well, where regularity in A the stream had a greater impact on perception than regularity in the B stream. It was suggested that the A stream was most likely chosen as the foreground stream because of its faster rate of presentation (twice that of the B stream), drawing attention more strongly than the B stream. Thus, regularity has the greatest impact in stream sustainability when multiple cues are regular and these regularities are present in the foreground stream.

Andreou et al. (2011) took the paradigm modification one step further and modified the temporal regularity of the A and B tones, while keeping frequency and intensity constant. Based on previous literature suggesting that listeners keep track of acoustic regularities, humans are quick to form temporal expectancies in order to optimize behaviour (Honing, Ladinig, Háden, & Winkler, 2009; Mari R. Jones, 1976; Lange, 2010; McAuley, Jones, Holub, Johnston, & Miller, 2006). Using an objective streaming measure where participants detected frequency modulation patterns (in either the A or the B stream, counterbalanced), the influence of regularity of the unattended sequence was investigated. The unattended sequence was either random, or regular for three tempos: faster, equal to, or slower than the mean inter-tone-interval (ITI) of the attended sequence. While there was no effect of unattended sequence regularity when the frequency difference between A and B tones was four semitones, an effect was clear when the frequency difference between tones was only two semitones, suggesting that while temporal regularity may aid segregation, this mechanism may only be engaged when alternative cues provide minimal evidence for segregation.

While research on the role of predictability and expectation in auditory processing is progressing with an increasing number of publications (i.e., (Bendixen, Denham, & Winkler, 2014; Bendixen, SanMiguel, & Schröger, 2012; S. L. Denham & Winkler, 2006; Vuust,

Ostergaard, Pallesen, Bailey, & Roepstorff, 2009; Vuust & Witek, 2014; I. Winkler et al., 2012; István Winkler et al., 2009), addressing predictability of frequency (Bendixen et al., 2010; Pearce, Müllensiefen, & Wiggins, 2010; Pearce & Wiggins, 2006), intensity (Bendixen et al., 2010), time (Andreou et al., 2011; Pearce, 2005; Chapter 5), and simple (Bendixen, Roeber, & Schröger, 2007; Bendixen & Schröger, 2008; Schröger et al., 2014) and complex patterns (Bendixen et al., 2014; Bendixen, Prinz, Horváth, Trujillo-Barreto, & Schröger, 2008), there is still much room for development. In particular, neural markers (Bendixen et al., 2012) and computational implementations (Pearce, 2005; Schröger et al., 2014) of prediction-based cognitive mechanisms are in their infancy and there is a need for new methodologies to test prediction-based processing hypotheses (Clark, 2013). While only an introduction was provided here, a comprehensive review of prediction as it relates to auditory streaming will be given in Chapter 4. This line of research is very closely aligned to the concept of predictive coding, introduced in more detail in Section 2.6 below.

2.5 Musical training

This section discusses the literature investigating the various effects of musical training, in this thesis an approximation of listening background, on perceptual and cognitive abilities as well as existing tools for measuring musical sophistication. Together, this body of literature demonstrates that musical training has a substantial effect on music perception, including stream segregation.

The most reliable and well understood effects of musical training are its effects on perceptual abilities, using techniques including both brain imagery and behavioural paradigms. A number of studies suggest that musicians encode sound more accurately and at a finer level of detail than those without musical training. For example, musicians are typically faster and more accurate in tasks of pitch discrimination (Brattico, Näätänen, & Tervaniemi, 2001),

contour and interval processing (Fujioka et al., 2004), temporal discrimination (T. Rammsayer & Altenmüller, 2006; T. H. Rammsayer, Buttkus, & Altenmüller, 2012), temporal performance (Repp & Doggett, 2007), timbre discrimination (Marozeau et al., 2013), processing harmonic information (Aksentijevic, Smith, & Elliott, 2014; Brattico et al., 2009; Spada, Verga, Iadanza, Tettamanti, & Perani, 2014) and stream segregation (François, Jaillet, Takerkart, & Schön, 2014; Fujioka, Trainor, Ross, Kakigi, & Pantev, 2005; Zendel & Alain, 2009). It is worth noting that the majority of these studies do not demonstrate causal effects of musical training but rather correlations between musical training and the given ability being investigated. It is entirely possible that musicians have better perceptual abilities prior to training and this aided their training. However, a few studies compare perceptual abilities before and after training, demonstrating positive effects of training (Fujioka et al., 2006; Menning, Roberts, & Pantev, 2000). Together with the large volumes of correlational evidence, these studies support a link between musical training and perceptual abilities.

One problem with treating musicians as a single category is that differences between musicians may be missed, either based on their instrumental (Tervaniemi, 2009) or stylistic (Nan et al., 2006) speciality. For example, one might suppose that orchestral instrumentalists have a more sensitive ear to tuning and percussionists to rhythm, and that jazz or folk musicians might have more developed improvisation skills. Pioneers of this type of investigation, Pantev and colleagues (Pantev, Roberts, Schulz, Engelen, & Ross, 2001) found that certain instrumentalists were more sensitive to the timbre of their own instrument than to others, as measured by auditory evoked fields (AEF). Violinists and trumpet players were presented with trumpet, violin and sine tones while MEG was recorded. Both instrumentalists presented stronger AEFs for complex over sine tones, and stronger AEFs still for their own instrument. In a similar study (Shahin, Roberts, Chau, Trainor, & Miller, 2008), professional violinists and

amateur pianists as well as young piano students and young non-musicians were presented with piano, violin and sine tones while reading or watching a movie while EEG was recorded. Gamma band activity (GBA) was more robust in professional musicians for their own instruments and young musicians showed more robust GBA to piano tones after their one year of musical training. Furthermore, Drost, Rieger, & Prinz (2007) found that pianists and guitarists' musical performance was negatively affected by auditory interference, but only if this interference came from the same instrument. Taking a step further and using more ecological stimuli, Margulis, Mlsna, Uppunda, Parrish, & Wong (2009) explored neural expertise networks in violinists and flautists as they listened to excerpts from partitas for violin and flute by J. S. Bach. Increased sensitivity to syntax, timbre and sound-motor interactions (activity in the motor cortex in response to the instrumentalist' timbre) were seen for musicians when listening to their own instrument. This effect of musical training on timbre will be explored in a behavioural streaming paradigm in Chapter 5.

Beyond its effects on perception and the brain at the neural and structural levels, musical training has also been studied in relation to cognitive skills such as memory (Chan, Ho, & Cheung, 1998; Franklin et al., 2008; Strait et al., 2012), spatial-temporal skill (Gromko & Poorman, 1998; Hurwitz, Wolff, Bortnick, & Kokas, 1975; Rauscher & Hinton, 2011) and general IQ (Bilhartz, Bruhn, & Olson, 1999; Phillips, 1976), with links between musical training and reading comprehension (Corrigall & Trainor, 2011) or reading skill (Anvari, Trainor, Woodside, & Levy, 2002; Moreno, Friesen, & Bialystok, 2011; Tsang & Conrad, 2011) and music and speech processing (Strait & Kraus, 2011) also explored. This body of literature has a much more complex output of results than the sections discussed above: some find benefits of musical training (Chan et al., 1998), others do not (Haimson, Swain, & Winner, 2011; Steele, Ball, & Runk, 1997) and some find improvements to some abilities (i.e., auditory

psychophysical measures) but not to others (i.e., multi-modal sequence processing; Carey et al., 2015). Careful consideration and appropriate controls should therefore be taken when evaluating effects of musical training on cognitive tasks.

Musical expectations are also influenced by training, where musicians form more specific expectations in relation to many aspects of music, including pitch (Granot & Donchin, 2002; Habibi, Wirantana, & Starr, 2013; Hansen & Pearce, 2014), chords (Bigand, Parncutt, & Lerdahl, 1996) and stream segregation (François et al., 2014) due to their extensive exposure to music and its patterns throughout their training. However, extensive listening can be enough to develop these expectations (Bigand & Poulin-Charronnat, 2006), which contributes to the ongoing discussions about the extent of effects of musical training on the brain and behaviour.

It is easy to see how modelling polyphonic music as opposed to tone sequences can become very complicated very fast, what with all the unique streaming threshold profile combinations possible for pitch, tempo, rhythm, timbre and harmony to name just a few and how these interact between multiple voices – often more than two. Add attentional bias and individual differences in musical training and listening history and a model becomes practically impossible. Though not currently computationally viable, a model that processes all of these features and considers all these issues within one, integrated framework should be the focus of future research. This proposed framework is predictive coding.

2.6 Predictive coding

Predictive coding has become a popular theory in the literature recently (over 15 000 hits in Google Scholar for “predictive coding” since 2017), including in music cognition (Agres, Abdallah, & Pearce, 2017; Bendixen et al., 2014; Globerson et al., 2017; Schröger et al., 2014; Vuust et al., 2009; Vuust & Witek, 2014; I. Winkler et al., 2012). A candidate for a grand unifying model of brain function, its attractiveness is obvious: it is a simple idea that has

been developed to explain perception (including the special cases of binocular rivalry and delusions), action, emotion, music perception, learning, inference, brain plasticity, attention, social action and multi-agent coordination among others. Its dangers are equally obvious: it is hugely ambitious, difficult to disprove in its current formulation, and may tempt forced explanations of phenomena to fit the framework so that it remains truly unifying.

First developed as a data compression strategy (Shi & Sun, 1999), predictive coding posits that the brain encodes error rather than raw sensory information. Of course for there to be an error, there needs to be some kind of comparison being made, which is where prediction comes into play. This framework describes the brain as a prediction machine that develops models of the environment and derives predictions about the environment using these models; this is the top-down aspect of the framework. These predictions, made on several timescales and cascaded from higher- to lower-order levels, are met with the incoming sensory signals – the bottom-up portion of the framework – and are either correct, a rare case in which case the model remains, or incorrect to varying degrees, where error is sent back up the necessary hierarchical levels until the model is updated. A fairly simple concept, this framework nevertheless generates large amounts of discussion in the literature (see Clark, 2013b – special issue of Behavioural and Brain Sciences – and Clark, 2013a for a fascinating, in-depth discussion) as details of its implementation, implications and scope are debated. Collecting evidence in support of this framework is particularly challenging at this time as the methodology for investigating brain function in terms of prediction error is not yet developed; thus far, there is only evidence that demonstrates the brain behaving “as if” it were employing this framework, with direct testing still to be developed.

Perception, learning and attention can be explained by the framework in a straightforward manner. It is intuitive to think of perception as a series of predictions that are

validated or corrected. Predictive processing explains change blindness (a person is not expected to be replaced by someone else when something passes between them and an observer) and binocular rivalry (two objects are not expected to occupy the same space at the same time at the same scale so while one percept becomes stable, the error generated by the other percept causes a switch, and when that percept becomes stable the error generated by the first percept causes a switch, and the bi-stability continues) with the basic idea of error encoding. Learning is explained by the continuous development of the brain's models through error incurred from interaction with the environment and attention is explained by controlling the gain, or relative importance, of uncertain input where attentional focus is reflected in higher gains.

On the other hand, the connection between predictive coding and action, emotion and the development are slightly less intuitive. The inclusion of action in the predictive processing framework began with work by Friston (2003, 2010), where he proposes that perception and action are deeply unified by making use of the same computational strategies. He proposes that perception and action (and cognition) are both results of prediction: humans perceive what they expect to perceive and do what they expect to do. As predictions are made about where fingers will be while typing for example, the error generated by a finger not being at the place it is expected to be will induce a motor sequence to eliminate that error and arrive at the place initially predicted. In Bayesian terms, the goal (prediction) is the observed state and Bayesian inference is performed to find the appropriate actions to get there (Toussaint, 2009). Furthermore, action elicits the streams of sensory information that the brain predicts, so in a sense, perceiving and moving (and thinking) continually work together in a sort of “self-fulfilling prophecy” (Friston, 2009). This is where one of the biggest criticisms of the framework is raised, and explained. If the brain's sole purpose is to reduce error between what

is predicted and what is perceived, why is it that humans do not hole themselves up in a dark room and never move from it? Clark proposes that it is because humans do not expect to be in a dark room. Hunger, thirst, light and other humans are expected and therefore sought out in order to eliminate the error caused by their absence. However, this is only true because humans are born in or evolved in such a world. Perhaps if humans only knew darkness and isolation, these would be expected. There is also the possibility that these expectation are innate due to thousands of years of living circadian rhythms and community. At this time, only speculation is possible.

In a response to Clark's Behavioural and Brain Sciences 2013 article, Seth & Critchley present evidence for the extension of predictive coding to emotion as interoceptive inference (Seth & Critchley, 2013). Their model (Critchley & Seth, 2012; Seth, Suzuki, & Critchley, 2011) explains subjective feeling states with predictions of the causes of interoceptive input. These predictions are continually updated and compared to visceral, autonomic and motor signals. Minor, low-level violations in prediction will not cause noticeable emotional reactions but larger, high-level violations can cause deep emotional trauma, with daily emotions falling somewhere between these two extremes. This type of model provides potential explanations for chronic anxiety (Paulus & Stein, 2006) and schizophrenia (Frith, 2012; Silverstein, 2013), where interoceptive error is heightened or imprecise, respectively.

A further stretch of the predictive coding concept includes cultural practice, proposing that culture is a reflection of shared predictions and shared error reduction. If everyone behaves in a certain way (i.e. start a conversation with a greeting), then predictions about a situation are validated and collective error is reduced. This applies to conversation, ritual, convention and shared practices, making life in society that little bit easier. Paton et al.'s (2013) response to Clark's Behavioural and Brain Sciences 2013 article and Clark's own response to it discuss

this further (Paton, Skewes, Frith, & Hohwy, 2013). There are many unanswered questions about how emotion, shared culture and action fit into the predictive coding framework, all highly compelling and worth investigation.

However, as mentioned previously, collecting evidence for such a unifying model is challenging as it is difficult to know whether the brain is actually producing prediction and error in a hierarchical way, or if it is acting as if it was doing so. Neural evidence will be the only way to effectively test hypotheses generated by such a model, as its details lie in neural function and not in behaviour. Specifically, an important consequence of the predictive coding framework is that it involves two neural sub-populations: one encoding expectations, and the other encoding error. These are thought to be deep and superficial pyramidal cells, respectively, though there is no direct evidence for this as of yet.

Invasive and non-invasive neural imaging techniques such as fMRI, TMS and single-cell recording will be the most informative methods in exploring and testing the predictive coding framework moving forward. Current behavioural, neural and computational evidence is indirect at best, though encouraging. For example, one fMRI study demonstrated the suppression of primary visual cortex area V1 once higher level areas had settled on an interpretation of a visual shape (Murray, Kersten, Olshausen, Schrater, & Woods, 2002), which is exactly what is expected in a case where the top-down mechanism is “explaining away” bottom-up input. Another found decreased responses to predictable stimuli (Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010) while another found that predictive processing was by far the best explanation of their results (Egner, Monti, & Summerfield, 2010), where the fusiform face area (FFA) brain region activation was indistinguishable for both house and face stimuli when faces were highly expected but differentiated when faces were unexpected. This suggests that the FFA was encoding expectation and error rather than features. In the

auditory domain, EEG and MEG evidence for pitch and pitch sequence processing has also been found in support of hierarchical, predictive processes (Furl et al., 2011; Kumar et al., 2011). One of the biggest current strengths of the predictive coding framework is that it can be effectively implemented in a computationally tractable way with neural models of the Bayesian Brain. Simulations so far can reproduce phenomena such as the non-classical receptive field effect (Rao & Sejnowski, 2002), the repetition suppression effect (Summerfield, Monti, Trittschuh, Mesulam, & Egner, 2008) and bi-phasic response profiles (Jehee & Ballard, 2009). The non-classical receptive field effect describes a situation where an active neuron will be suppressed if surrounding neurons are stimulated by an identically oriented stimulus but enhanced when the surrounding neurons are stimulated by an orthogonally oriented stimulus, demonstrating error encoding: similar stimuli, therefore more predictable, do not trigger error and therefore the central neuron is suppressed, with the opposite for a different stimulus. The repetition suppression effect describes the reduction in neural response as a result of stimulus repetition. However, if that repetition is unexpected, neural response is increased. Bi-phasic response profiles describe neurons whose optimal driving stimulus changes rapidly (as quickly as 20ms), for example in low-level visual processing centres (Jehee & Ballard, 2009). This can be explained by neurons as error detectors rather than feature detectors. While these are only a few explained phenomena and all the evidence cited is indirect, the predictive coding framework can explain all the above data and phenomena with the same basic, unifying principle of prediction formulation and error propagation and subsequent cancellation.

To summarize, predictive coding proposes a mechanism for action, perception and cognition based on prediction, where the brain generates biased models with low variance whose ultimate goal is to minimize error. Previous experience, and perhaps even innate

knowledge – though this is unclear and not much discussed – shape models of the world which are validated or violated by incoming sensory data. If violated, an error signal is generated and propagated up the neural hierarchy until the model is updated. Its scope is enormous, as it proposes to explain everything about how the brain works and that processes previously thought of as separate – action, perception, cognition – are actually intricately related and driven by the same underlying mechanism: prediction. It has quickly become a popular framework for brain function, with details being the focus of debate (Clark, 2013b). However, it is also limited by the lack of direct evidence for neural prediction and error encoding. Only further research can clarify its implementation and true scope. This thesis contributes to this effort by investigating predictive coding in the context of musical auditory scene analysis by primarily using information content as a metric.

This chapter has introduced the literature surrounding the various sources of information that require consideration to achieve a comprehensive model of auditory streaming – basic auditory features, musical features, attention, expectation and musical training – concluding with the presentation of predictive coding, a unifying framework of perception, action and cognition that will inspire the use of expectation and heuristic-like rules as the centrepieces of an integrated framework for auditory streaming. Though introduced separately, these sources of information interact in highly complex ways to decode a given auditory environment. In this thesis, some of these interactions will be explored in Chapters 5-8, investigating aspects of the proposed integrated, prediction-based framework for auditory streaming presented in Chapter 4. First, a description of the materials common to multiple chapters and of all musical corpora used is given in Chapter 3.

3 Materials

This chapter introduces tools and data that will be used repeatedly throughout the thesis: IDyOM (information dynamics of music; Pearce, 2005), a number of training and evaluation corpora of music, and the Goldsmiths Musical Sophistication Index (Gold-MSI; (Müllensiefen, Gingras, Musil, & Stewart, 2014)).

3.1 IDyOM

IDyOM is a computational model of auditory expectation that harnesses the power of statistical learning. It learns the frequencies of variable-order musical patterns from a large corpus of music (via the long-term model, or LTM) and from the current piece of music being processed (via the short-term model, or STM) in an unsupervised manner and generates probabilistic predictions about the properties of the next note in a melody given the preceding melodic context. IDyOM is a multiple-viewpoint model, capable of learning patterns from pitch- and time-derived note properties (source viewpoints) to predict relevant note properties

(target viewpoints). These viewpoints can be use-defined or selected through optimization. Its implementation is described here in detail.

N-gram models are common in data compression and statistical language modelling (Cavnar, Trenkle, & others, 1994; Manning & Schütze, 1999; Ziv & Lempel, 1978) and are the basis of the IDyOM system. An n-gram is a sequence of n symbols and an n-gram model is a collection of sequences, each associated with a frequency count. This frequency count is obtained by *training* the model on a corpus of sequences. To date, IDyOM is limited to monophonic music. For the present purposes, this is sufficient; however, a representation scheme for polyphonic music is needed and will be introduced below. In relation to music, an n-gram model of length 1 is equivalent to a basic frequency count of every different event occurring in the particular piece or group of pieces of music analysed. An n-gram model of length 2 tabulates pairs of events while higher-order n-grams simply count longer sequences. No one order of n-gram models can explain music in a comprehensive way as the lower orders cannot capture melodic patterns specific to a particular piece of music while higher orders do not necessarily generalise to new contexts and lack statistical reliability as these particular patterns may not occur frequently, with frequency decreasing as context length increases.

To allow for this parallel use of variable n-gram orders, prediction by partial-matching (PPM) (Cleary, Teahan, & Witten, 1995) is implemented. PPM is a statistical compression technique based on prediction and context modelling. While an n-gram model can capture low- as well as high-level contextual patterns, all levels should be considered to form an accurate representation of the musical surface being analysed, where lower-order models allow generalization to new contexts and higher statistical power, and higher-order models provide

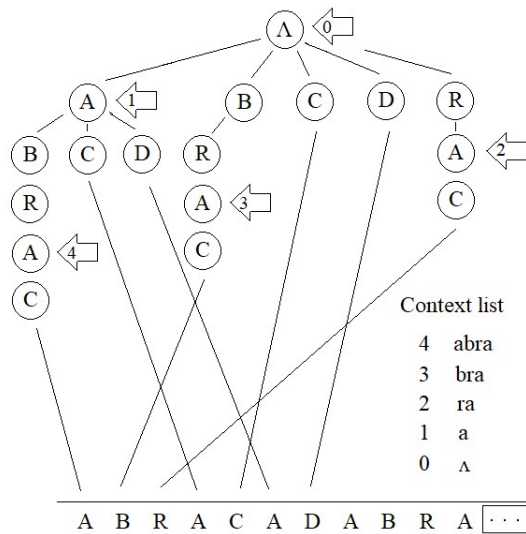


Figure 3.1. Context tree for the word *abracadabra*, where the arrows identify all possible contexts for ‘a’, the last letter of the word.

structural specificity. This particular technique allows for the efficient storage of multiple lengths of context using what are called *context trees* (see Figure 3.1) that store contexts as *leaves* on *branches* so that there is no need to store an entire context multiple times for every potential following event but instead simply update the core branch. PPM works for contexts of any lengths, and deals with new material by using *escape* probabilities: for each context length, starting from the longest, if the new event has never been seen in any recorded context, the system ‘escapes’ and looks at the next shortest context, continuing until a previously known context is found, or until the alphabet of events itself, tracking and combining all probabilities as it goes along. In this way, PPM combines probabilities at multiple levels and can handle brand new material in an elegant manner to make a final prediction of appropriate viewpoints, which include basic, derived and linked viewpoints. Basic viewpoints are data that are directly encoded at import, such as pitch, duration and mode; derived viewpoints are created from the basic data, such as interval (difference between two pitches) and scale degree (from pitch and key signature); linked viewpoints have their predictions combined at model creation rather than producing separate models for two viewpoints, allowing the representation and modelling of



Cpitch	3.36	2.01	2.30	2.20	2.49	3.85	0.78	0.78	1.14	1.12
Cpint	5.20	2.32	2.73	2.09	3.54	4.95	1.20	3.32	1.14	2.38
IOI	4.52	1.49	1.04	1.96	0.67	1.75	0.87	0.86	0.85	0.86
Cpitch x Cpint	5.20	2.30	2.26	2.02	2.14	3.90	0.78	0.76	1.07	1.10
Cpitch x IOI	7.88	3.51	3.34	4.17	3.16	5.61	1.66	1.64	2.00	2.12

Figure 3.2. An example IDyOM analysis for the first two bars (and pickup) of a British folk song (A162) from Schellenberg’s experiments (1996; see also Section 5.3.2). The LTM here was trained on the typically used set of 185 Bach chorale melodies, Nova Scotia folk songs, and the *fink* subset of the Essen Folk Song Collection’s German collection (see also Section 5.3.2). Information content for *chromatic pitch* (basic), *pitch interval* (derived), and linked viewpoint pitch and interval predicting chromatic pitch as well as *IOI* (derived) predicting onset and a linked viewpoint pitch and *IOI* predicting pitch and onset respectively. The first pitch has the highest value, as all pitches or onsets are equally likely before the piece begins. As the piece progresses, IC fluctuates, where repetition lowers IC (i.e. A to G occurs twice). Higher IC values reflect low predictability, or higher unexpectedness.

dependencies between features (see Table 3.1 for the list of IDyOM viewpoints referred to or used in this thesis). For example, a user can give a command to learn about the interval and scale degree patterns found in a training set to predict pitch in a given new piece of music. Figure 3.2 presents an example analysis of a folk song using various viewpoints and combinations of viewpoints.

Furthermore, IDyOM contains options for short-term, and long-term learning, or both. The long-term model (LTM) learns exclusively from the given training set while the short-term model (STM) learns from the piece currently being analysed. In this way, the LTM captures stylistic patterns while the STM captures patterns contained within each individual piece of music. When combined, the system contains knowledge about both the greater stylistic context, and the current piece of music, just as a human listener combines schematic and

Table 3.1. All basic and derived viewpoints of the musical surface implemented in IDyOM and used in this thesis, with a short description (linked viewpoints can be combinations of any two or three of the viewpoints listed below). The default timebase is 96, corresponding to a whole note.

Viewpoint Type	Viewpoint Name	Description
Basic	cpitch	Chromatic pitch, encoded with MIDI numbers
	onset	Event start time; depends on timebase
	dur	Event duration; depends on timebase
	deltast	Duration between last note and its predecessor
	bioi	Basic Inter Onset Interval between last note and its predecessor
	voice	Voice number as in the score
Derived	cpint	Chromatic pitch interval
	cpintfref	Chromatic interval from tonic
	ioi	Like BIOI but undefined for first event

dynamic expectations when exposed to new music (Huron, 2006). The system also offers a ‘plus’ option, which adds knowledge from each composition to the LTM as it analyses a dataset.

IDyOM accepts MIDI, kern and text files into its database for both monophonic and polyphonic music. In the case of polyphony, a system of slices is used, where a *slice* is created at every existing pitch onset, containing all the sounding notes at that time. This can be thought of as a cross-sectional view of the music. The user can choose between full expansion (Conklin, 2002), where pitches are repeated at every onset, or an extension called continuation expansion, implemented for use in this thesis, where pitches whose durations extend into a new onset are labelled as continuations rather than simply a new pitch. This has advantages for modelling melody extraction and auditory streaming, allowing the prevention of voice crossing into a pitch mid-duration. Figure 3.3 illustrates this, where red vertical lines delimit slices and blue note-heads indicate pitches sounding throughout those particular slices but whose onsets belong in a previous slice. In full expansion, these blue note-head events would not be related to the pitches in the previous slice that they are continuations of, leaving open the possibility

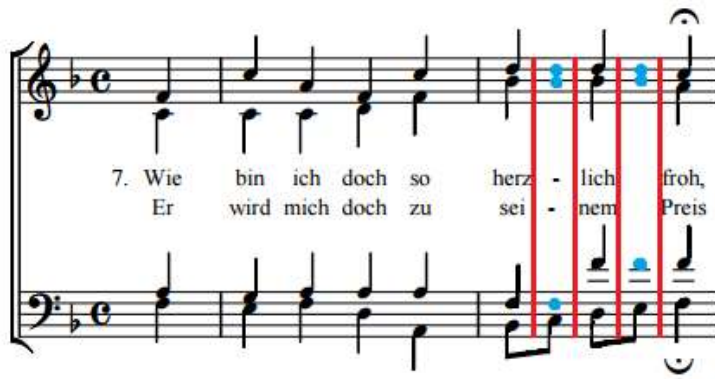


Figure 3.3. Illustration of the expansion possibilities in IDyOM using the first two measures of BWV 001 by J.S. Bach. Red vertical lines mark slices. In full expansion, blue note heads are not related to the previous event, of which they are continuations. In continuation expansion, these same note heads are marked as continuations of the previous event, forcing all events in a melody extraction or streaming model to begin at onsets corresponding to the score.

for one of these events to be considered a new pitch onset. By marking these as continuations, or part of the previous pitch onset, it is impossible for a stream to contain a pitch beginning anywhere other than its initial onset.

IDyOM has been validated by computational, behavioural and neuropsychological studies, making it an attractive cognitive model. Computationally, IDyOM was used to reanalyse data from Cuddy and Lunney (1995), where participants rated continuation of 25 chromatic pitches after a two-tone context on a scale of 1 to 7. IDyOM's long term model (contexts were too short to meaningfully employ the short term model) explained more of the participants' responses, 72%, than Schellenberg's two-factor model, which explained 68% (Schellenberg, 1997). Though not found to be significantly better, IDyOM did subsume both the proximity and reversal components of the two-factor model (Pearce & Wiggins, 2006). Moving on to more complex stimuli, IDyOM was also used to reanalyse data from Schellenberg (1996, Experiment 1), where participants rated the continuation of tones presented after 8 different English folk songs, four major and four minor and all ending on

implicative intervals. In this analysis, IDyOM explained significantly more of the behavioural data than the two-factor model, explaining 83% of the data, as compared to 75% (Pearce & Wiggins, 2006). Finally, IDyOM was used to analyse data from a betting paradigm from Manzara et al. (Manzara, Witten, & James, 1992), which allows the measurement of expectations as a sequence unfolds. Participants are given a certain amount of capital and they spend it by proportionally placing bets on the pitches they feel are more or less likely to appear next. If they are correct, they gain capital and if they are incorrect, they lose capital. The proportion of bets are a direct measure of relative expectation between potential upcoming pitches. Here, IDyOM outperforms the two-factor model by explaining 63% of the data, compared to 13% (Pearce & Wiggins, 2006), as well as entirely subsuming the proximity and reversal factors. These three data re-analyses with comparison to the two-factor model demonstrate that IDyOM consistently performs better in explaining behavioural data for different levels of musical complexity and different paradigm measurements. It is interesting to note that different viewpoints maximised results in each of the three analyses (details in Table 3.2), presumably reflecting the relative length of the stimuli in each of the behavioural experiments re-analysed.

In Pearce et al. (Pearce, Ruiz, et al., 2010), a new behavioural paradigm for studying

Table 3.2. Summary of viewpoints included in optimal models in the re-analysis of three data sets. Viewpoints with an ‘x’ indicate linked viewpoints.

Experiment	Viewpoints
Experiment 1 – Cuddy & Lunny data (1995)	Interval x Duration; IntFirstPiece; IntervalDuration
Experiment 2 – Schellenberg data (1996)	IntFirstBar; IntFirstPiece; ScaleDegree x Interval; Pitch x IOI
Experiment 3 – Manzara et al. data (2002)	IntFirstPiece; ScaleDegree x DurRatio; ThreadInitPhr

auditory expectation is introduced. In this paradigm, listeners are watching a clock figure whose arrow moves clockwise as the probe tone approaches; when the arrow reaches twelve o'clock, the participant rates the expectedness of that tone. This allows for a continuous collection of expectation ratings that do not interfere with the flow of the sequence, as the betting paradigm does. This paradigm also avoids the closure confound present in any paradigm requiring a rating at the end of a sequence. There was a strong correlation between the probability of notes as predicted by IDyOM, and the perceived degree of unexpectedness as rated by participants.

An electrophysiological study was also undertaken, where participants simply listened to the same stimuli as in the behavioural paradigm above while EEG was being recorded. Results revealed novel, robust ERPs in response to high- and low-probability notes, comparable to the N400 (Miranda & Ullman, 2007), as well as distinct beta oscillation patterns for each type. In conclusion, this model is a powerful, flexible and validated tool for simulating statistical learning based on prior experience of the listener (Hansen, Wallentin, & Vuust, 2013; Hansen et al., 2013; van der Weij et al., 2017), and melodic expectations in human listeners as they are implied by the musical surface (Egermann, Pearce, Wiggins, & McAdams, 2013; Gingras et al., 2016; N. C. Hansen & Pearce, 2014; Marcus T. Pearce, Müllensiefen, et al., 2010).

3.2 Musical corpora

All music corpora, or datasets, used in this thesis are introduced here. All training datasets are monophonic. Where training datasets are predicting polyphonic datasets (Bach Chorale and String Quartet Datasets), each monophonic voice is treated as a separate composition and combined into one dataset.

3.2.1 Montreal Billboard Corpus

The Montreal Billboard Corpus was created to provide a large set of data for the music information retrieval (MIR) community. In summary, the Corpus contains 1365 audio recordings and matching transcriptions taken from the *Billboard* “Hot 100”, a weekly compilation of the most popular music singles in the USA. The parameter of interest here is chord labelling, used to model harmonic progressions in Chapter 5. There are 414 059 labelled beats in the corpus, spread over 638 distinct chords and 99 chord classes, including inversions and chord extensions. These chords are encoded as integers (i.e., C major in root position is 1). The full set of chord progressions (average 11.8 chords per piece) was used to train IDyOM in Chapter 5. Details of the construction, transcriptions and descriptive statistics of the Montreal Billboard Corpus can be found in Burgoyne, Wild, & Fujinaga (2011).

3.2.2 Bach Chorale Dataset

The Bach Chorale dataset was constructed from the 371 Chorales set of kern files downloaded from the Humdrum website¹. This represents all the chorales written by J. S. Bach and published by Breitkopf & Härtel (Breitkopf & Härtel, 1875). The last 20 chorales were removed to be reserved for further validation once primary testing of the melody extraction model was undertaken (Chapter 7). Chorales with a rest in one of the voices were removed so that the final dataset contained 350 4-voice chorales and the validation set contained 19 4-voice chorales.

¹ <http://kern.ccarh.org/cgi-bin/ksbrowse?l=/users/craig/classical/bach/371chorales>

3.2.3 String Quartet Dataset

The String Quartet Dataset is divided into a test set and a training set. The test set is composed of 7 Mozart quartet movements. These were selected from the composers' complete string quartet output based on opening with only four voices (no double stops). Furthermore,

Table 3.3. Catalogue and movement details of the seven Mozart string quartets from the test set of the String Quartet Dataset.

String Quartet	Movement
K428	1
K428	2
K458	3
K464	2
K499	3
K575	2
K590	1

all double stops were removed from test and validation sets (see Chapter 7 for justification). A total of 218 tokens were removed, only 18 of which were suspensions. Kern files were downloaded from the Humdrum website². Table 3.3 contains the quartets' catalogue and movement details.

The training set is composed of 61 Mozart quartet movements and 156 Haydn quartet movements.

3.2.4 Nova Scotia Folk Songs

This is a collection of 152 (monophonic) folk songs from Nova Scotia, Canada, collected by Helen Creighton³.

3.2.5 Essen Folksong Collection Subset

The full Essen Folksong Collection is a collection of over 20 000 folk songs from around the world, encoded by Helmut Schaffrath and now led by Ewa Dahlig-Turek. 8473 have been translated into kern notation, consisting mostly of German and Chinese songs, with fewer from a variety of European and North American, Asian and African countries. In this

² <http://kern.humdrum.org/help/data/>

³ <http://kern.ccarh.org/cgi-bin/ksbrowse?s=nova>

thesis, the *Musikalischer Hausschatz der Deutschen*, or 'fink', subset of the Essen Folksong Collection is used. It is a set of 566 German folk songs with an average of 58.45 events each.

3.2.6 Bach Soprano

This dataset is formed of the soprano lines from the collection of 185 Bach Chorales as encoded in kern format (BWV 253-438)⁴. This dataset is used to train certain IDyOM models for consistency and comparison with previous published research that used this dataset prior to the availability of the full set of 371 chorales described in Section 3.2.2 above.

3.3 Goldsmiths Musical Sophistication Index

Briefly introduced in Section 2.5, the Goldsmiths Musical Sophistication Index, or Gold-MSI is described here in more detail. The Gold-MSI is a recently developed test of musical sophistication that aims to measure musical skill outside of performance and perception such as writing about, analysis of and emotional management using music. This test consists of two music perception tests and a self-report questionnaire and has been validated in the UK (Müllensiefen et al., 2014) and in Germany (Fiedler & Müllensiefen, 2015; Schaal, Bauer, & Müllensiefen, 2014). It has already been translated to German and Danish⁵, with a French version coming soon. A benefit of the Gold-MSI is that it treats musical sophistication as a scale, allowing the user to treat the measurement of musical training as a covariant rather than a two-factor category. The Gold-MSI, specifically the musical training sub-scale (see Figure 3.4), will be used throughout the work presented in this thesis.

⁴ <http://kern.ccarh.org/cgi-bin/ksbrowse?type=collection&l=/musedata/bach/chorales>

⁵ <http://www.gold.ac.uk/music-mind-brain/gold-msi/download/>

3.3.1 Perceptual tests

In the melodic memory test, participants hear three melodies and must identify the odd one out. Difficulty is manipulated in terms of contour (same or different), scale degree (in-key vs. out-of-key) and transposed distance around the circle of fifths. In the beat perception test, participants are to indicate whether or not a series of beeps played alongside a musical excerpt is on or off the beat. The beeps are either shifted in phase or in tempo. Both tests are adaptive, adjusting the difficulty of the task to each participant, allowing for a shorter and more accurate test.

3.3.2 Self-report questionnaire

The self-report portion of the test consists of 7 sub-scales that evaluate a variety of skills, from the emotional regulation with music (i.e. “Pieces of music rarely evoke emotions for me”) to engagement with music (i.e. “I spend a lot of my free time doing music-related activities”) to extent of formal musical training (e.g. “I have had __ years of formal training on a musical instrument (including voice) during my lifetime“). Questions are either positive or negatively phrased, where the scale is inverted in analysis for the negatively phrased questions. Each question can receive a minimum score of 1 point and a maximum score of 7 points, resulting in a minimum of 7 and maximum of 49 for this musical training sub-scale.

Please circle the most appropriate category:

1 – Completely Disagree 2 – Strongly Disagree 3 – Disagree 4 – Neither Agree nor Disagree

5 – Agree 6 – Strongly Agree 7 – Completely Agree

I have never been complimented for my talents as a musical performer. 1 2 3 4 5
6 7

I would not consider myself a musician. 1 2 3 4 5 6 7

Please circle the most appropriate category:

I engaged in regular, daily practice of a musical instrument (including voice) for ___ years.

0 1 2 3 4-5 6-9 10+

At the peak of my interest, I practiced ___ hours per day on my primary instrument.

0 0.5 1 1.5 2 3-4 5+

I have had formal training in music theory for ___ years.

0 0.5 1 2 3 4-6 7+

I have had ___ years of formal training on a musical instrument (including voice) during my lifetime.

0 0.5 1 2 3-5 6-9 10+

I can play ___ musical instruments.

0 1 2 3 4 5 6+

The instrument I play best (including voice) is _____

Age: _____

Gender: Male/Female

Figure 3.4. The musical sophistication sub-scale of the Gold-MSI.

4 An integrated framework for musical auditory streaming

With an understanding of the complexities involved in modelling musical ASA and some tools to study them, this chapter proposes an integrated framework for musical ASA, where all information sources required for modelling musical ASA discussed in Chapter 2 will be addressed using the unifying concept of prediction. However, before it is outlined in detail, an in-depth review of existing approaches to modelling auditory streaming in general and musical streaming in particular will be presented (Sections 4.1-2), where auditory streaming has been explained by models using such varied inspiration as neural firing patterns (Beauvois & Meddis, 1997; McCabe & Denham, 1997), temporal coherence (Elhilali, Ma, Micheyl, Oxenham, & Shamma, 2009), neural oscillatory patterns (von der Malsburg & Schneider, 1986; Wang & Brown, 2006), and predictive processing (Schröger et al., 2014). Musical

auditory streaming in particular has been modelled using corpus analysis (Duane, 2013; Huron, 2001), probability (Temperley, 2009) and perceptual principles (Cambouropoulos, 2008). Sections 4.3-4 will present the framework in more detail, Section 4.5 will describe the type of training data this framework would employ, Section 4.6 will compare the framework to a selection of the models introduced in Sections 4.1-2 and Sections 4.7-8 will discuss limitations of the framework along with the potential it brings for future research in auditory streaming. Finally, Section 4.9 will outline the selection of specific concepts from the framework that will be tested in subsequent chapters.

4.1 General models of auditory streaming

4.1.1 Neural-based models

Beauvois & Meddis (1991) built one of the earliest computer model simulations of auditory stream segregation. It is inspired by neural firing patterns and relies mostly on frequency information, while also incorporating temporal information. Input to peripheral frequency analysis establishes channels (high, medium, low frequency) that are then fed through an inner hair cell and auditory nerve simulation based on physiological information, and divided into pathways based on excitation-level. The channels represent the A frequency, the B frequency, and the harmonic mean of the two, remaining at these frequencies throughout. The dominant channel becomes the foreground while the non-dominant channels are attenuated and become the background. Streaming is detected based on the relative channel amplitude: if the A and B streams have similar amplitude below a given critical threshold, the percept is integrated, if not, the percept is segregated. They later further developed this model to simulate the build-up of segregation over time, the fission and temporal coherence boundaries and the trill threshold in ABAB sequences, as well as the effects of frequency difference between A

and B tones and tempo (Beauvois & Meddis, 1996). McCabe and Denham (1997) developed Beauvois & Meddis' model for multi-channel streaming and considered two stages of processing as originally proposed by Bregman (1990), namely the initial, automatic stage based on bottom-up information (in this case pitch and tempo) and the later stage based on top-down information (here attention). The model organizes input based on competitive interaction between frequency channels and consists of two tonotopically organized arrays of neurons, one for the foreground and one for the background. Activation of one frequency in one array inhibits this frequency in the other, creating a divide between the foreground and background. This model accounts for the effects of frequency difference between A and B tones and tempo, bi-stability and the influence of background organization on the foreground, where *capture* tones were either near or far from the A and B tones, imitating an experiment by Bregman & Rudnick (1975) where capture tones close in pitch to one set of tones but not the other aided segregation.

Neural oscillatory patterns have recently been explored as an explanation for attention (Niebur, Hsiao, & Johnson, 2002; Womelsdorf & Fries, 2007), where neural firing in a region in response to a particular feature(s) increases in synchrony when it is attended to while unattended features' neural firing is desynchronized, a process called active suppression. Similarly, in relation to auditory scene analysis and streaming, a number of models based on neural oscillation have been proposed, where sound objects are represented by units, in turn representing neurons or networks of neurons (Mill, Böhm, Bendixen, Winkler, & Denham, 2013; Pichevar & Rouat, 2007; Rankin, Sussman, & Rinzel, 2015; von der Malsburg & Schneider, 1986; D. Wang & Brown, 2006; D. L. Wang & Brown, 1999; DeLiang Wang & Chang, 2008). Now a potential understanding of the neural mechanisms of attention, von der Malsburg & Schneider (1986) propose a model based on the simple idea that neural firing of

features in the same stream are synchronized while neural firing of features across streams are desynchronized. A more advanced, comprehensive model was later built by Wang and Brown (1999), where both segmentation and grouping are represented sequentially by two layers of two-dimensional, time-frequency grids of oscillators. The first layer forms blocks of synchronized oscillatory activity from similar regions of energy (i.e. harmonics, formants) into segments, which are then grouped together based on compatibility with a fundamental frequency. This model successfully separates vowel pairs but is not applied to tones or musical streams. Wang & Chang's (2008) more recent model is similar, where two-dimensional oscillators are dynamically adjusted by local excitatory and global inhibitory connections to represent pure tones (frequency and time) as integrated or segregated, replicating van Noorden's (1975) integrated, segregated and ambiguous regions. While this model does not simulate perceptual multi-stability, Mill et al.'s (2013) and Rankin et al.'s (2015) models were explicitly designed to do so. Mill et al.'s (2013) model relies on prediction applied to *objects*, a series of linked sound events which once formed is represented by a coupled excitatory and inhibitory population of neurons that interacts with other objects. Objects are cyclically repeating patterns of events, where links between events form probabilistically and competition between objects is affected by the rate of successful predictions the object makes, the rate of pattern rediscovery (where faster equals stronger representation), adaptation, self-excitation, noise and inhibition from other objects predicting the same events. Multi-stability and stochastic switching between percepts was simulated with this model, where objects with high states are considered to be present in perception. The main difference in Rankin et al.'s (2015) model is that the number of possible perceptual objects is fixed (3: A tones, B tones, or A and B tones), while in Mill et al.'s (2013) they are discovered in real time, assuming a tonotopic space with three neural units that accept input from the primary auditory cortex (tokenized

ABA- input). The output is segregation or integration, determined by whether or not the activity of the AB object is more than the average activity of the A and B objects.

4.1.2 Temporal coherence models

Another approach has been to give more consideration to the temporal aspect of streaming, where synchronized onsets are more likely to reflect an integrated percept while unsynchronized onsets encourage a streaming percept. Elhilali et al.'s (Elhilali et al., 2009) model sends input through two stages: temporal integration, and coherence analysis. The degree of multiscale (50-500ms) temporal integration is analysed to inform a coherence matrix over time; areas of non-zero activation in this matrix represent a perceived stream. Tested with either synchronous or alternating A and B tones at frequency separations of .25, .5 and 1 octaves, the model predicts an integrated percept for all synchronous conditions and for the smallest frequency separated asynchronous condition and a segregated percept for the larger two frequency separated asynchronous conditions. These predictions match an experiment where participants identified temporal changes in the B tone sequence in the same sequence conditions as the model: participants performed worse when they perceived two streams, matching the model's predicted segregated conditions. Furthermore, the authors also claim that the model predicts transition points from one to two streams and back, but this aspect has not been explicitly validated by human behavioural data. A model by Ma (2011) also segregates sound based on feature correlation, with features including frequency, scale (frequency component spacing, in cycles per octave), pitch and location. Correlation matrices are calculated for each feature, resulting in a measurement of temporal coherence. Nonlinear principal component analysis is applied to an enlarged correlation matrix to group features into a number of pre-specified units. The model outputs two *masks*, which when applied to the original sound source, reconstruct the identified segregated (or not) source(s). Krishnan et al.

(Krishnan, Elhilali, & Shamma, 2014) have built a similar model, where features are time, frequency, scale and rate (temporal spacing, in Hz) for one function and time and pitch for another. This model was tested on tone sequences and mixed speech streams, simulating van Noorden's (1975) perceptual regions, the bouncing effect (Tougas & Bregman, 1990; tones group together by frequency rather than continuity, similar to the scale illusion (Deutsch, 1975)), and successfully segregating a new tone into a new stream.

4.1.3 Prediction-based models

Some more recent models of auditory scene analysis have been founded on the concept of expectation and predictive regularity (Barniv & Nelken, 2015; Nix & Hohmann, 2007; Schröger et al., 2014). These first two models use Bayesian inference to break down the sound environment into state vectors estimated from input. Barniv & Nelken's (2015) model, which successfully simulates behavioural streaming experiment results, accepts simplified tokens (tones represented by frequency and timing) and assigns these to a 'class', or stream. The assignment decision is made as evidence for one class or another is accumulated over time (dynamic priors); if no existing class is appropriate, a new class is created. Nix & Hohmann's (2007) model relies on sound spectrum direction rather than frequency and timing for state vectors of sound representation, though they only train their model on spectral information. The goal of the model is to determine the state vector's posterior probability, which is adjusted by filtering out vectors that are unlikely to match the input. This model has been successfully applied to voice separation. Another recent example is the Auditory Event Representation System (AERS) by Schröger and colleagues (2014), which brings together research on auditory violation detection and auditory scene analysis. The AERS model has four major components: the formation of auditory stimulus representations, the formation of regularity representations that predict subsequent input, the comparison of prediction and the input at multiple anatomical

and temporal levels and finally the evaluation of the relevance of the relationship between the input and the context. It accepts monophonic, non-overlapping sound sequences, such as ABA sequences and outputs a percept: integration or segregation. CHAINS, the implementation of the AERS framework, uses prediction to assign new objects to existing streams (chains) by detecting and preferring regularity, implementing competition between streams through excitation and inhibition and modelling auditory multi-stability. This model is grounded in the anatomy of the auditory system and all stages are testable experimentally, making it one of the most concrete explanatory models of auditory scene analysis in the literature so far.

4.1.4 Hybrid models

Denham & Winkler's (2006) auditory streaming model combines neural and prediction-based approaches and is inspired by perception as a generative process (Friston, 2005) as well as evidence that auditory bi-stability is similar to visual bi-stability (Pressnitzer & Hupé, 2006). With models of visual bi-stability typically including principles of exclusivity (competition between interpretations), stochasticity (allowing the interpretation to change) and adaptation (interpretations vary in stability), this auditory streaming model applies all three and more, for a model that includes segregation, predictive modelling, competition, stochasticity and adaptation (these last two are subsumed into adaptation).

The model begins by segregating incoming sounds based on evidence of neural functioning from the primary auditory cortex, where prolonged exposure to an ABAB sequence will eventually cause suppression of B tones in the A stream and A tones in the B stream, resulting in two independently encoded streams of tones. Streaming itself depends on frequency difference between A and B tones and the rate of presentation of the sequence, characteristics well established by van Noorden (1975). Each stream then generates predictions about what is coming next, a process here proposed to be the same process that generates the

well-known mismatch negativity (MMN), a neural marker indexing departure from regularity. Once streams are formed and the model has made predictions about what is coming next, competition between alternative predictions is introduced and predictions that best match the incoming stimulus are selected as the dominant interpretation. The model takes into account both local and global predictions as the sequence unfolds over time, where local predictions are more likely to support integration while global predictions can induce segregation over time as evidence for separate streams builds. Finally, an adaptation mechanism is implemented in the model to account for the observation that bi-stability is present in both auditory and visual situations (Leopold & Logothetis, 1999; Pressnitzer & Hupé, 2006). This mechanism simply allows interpretations in the foreground at any given time to weaken over time and allow another interpretation to ‘take over’. The speed of the initial change from integration to segregation depends on stimuli parameters, as does the overall rate of change between percepts.

To summarize, existing models of auditory streaming vary in approaches, goals and implementations. Some models target multi-stability, competition and adaptability (Denham & Winkler, 2006; McCabe & Dehnam, 2006; Mill et al., 2013; Rankin et al., 2015), others focus on replicating van Noorden’s (1975) or their own behavioural data (Barniv & Nelken, 2015; Krishnan et al., 2014; Wang & Chang, 2008). Some models allow the creation of new streams (Barniv & Nelken, 2015; Krishnan et al., 2014; Rankin et al., 2015), while most are fixed. However, with only one exception (Nix & Hohmann, 2007), they all rely heavily, if not exclusively, on frequency and timing information to inform the model. This is most likely heavily influenced by the early work of van Noorden (1975) and Bregman (1990). While these seem to function well by successfully modelling perceptual data, the stimuli in these experiments are reduced to two alternating frequencies, or in a few cases two voices (Nix &

Hohmann, 2007; Wang & Brown, 2006), far simpler than the complex sounds humans are exposed to in daily life, and also far simpler than music.

In the next section, musical auditory streaming models will be presented. While still generally focusing on frequency and timing, the level of abstraction of the information processed by the models is higher. While for example Beauvois & Meddis (1991) model neural firing patterns, Duane (2013) models pitch distance and harmonic relationships between more than two simultaneous sound sources (i.e. a string quartet).

4.2 Musical auditory streaming models

Before presenting existing musical auditory streaming models, it is useful to note the difference between voice separation and true auditory scene analysis. While the former is tied to the score, the latter is a perceptual phenomenon and cannot be so easily defined. The majority of existing models of musical ASA deal with voice separation, often seen as an engineering problem (Chew & Wu, 2004; Jordanous, 2008; Kilian & Hoos, 2002; Kirilin & Utgoff, 2005; Madsen & Widmer, 2006), while relatively few approach the problem concerned with perceptual accuracy (Cambouropoulos, 2008; Temperley, 2009). However, these two phenomena are closely related and, as will be introduced in Section 8.3, the proposed model performs both voice separation and stream segregation, using the former to inform the latter.

4.2.1 Perceptual principles

Many musical ASA models have been built by codifying a number of perceptual principles (Cambouropoulos, 2008; Chew & Wu, 2004; Kilian & Hoos, 2002; Madsen & Widmer, 2006; Marsden, 1992; Szeto & Wong, 2003), where the most common principles are the temporal continuity and pitch proximity principles. Together, these state that within a perceptual stream, notes are closer in time and closer in pitch than across streams. An early

model aimed at improving music transcription is presented by Kilian & Hoos (2002). In this model, polyphonic music is divided into slices where each note in a slice is assigned to a voice. The assignment is done via a cost function that reflects both the relationship between notes within the slice and between slices. The added consideration of groupings within slices allows chords to be assigned to a voice, making this algorithm's output perceptually accurate, where chordal accompaniment, in pop music for example, is typically perceived as one musical voice rather than multiple voices in unison.

Chew & Wu (2004) developed a model based on units called *contigs*, a collection of vertically overlapping notes where the number of voices is constant within a contig. Once a piece of music is divided into these contigs, the notes within each contig are separated into voices using the principle of pitch proximity. Finally, voices from each maximal contig (contigs with the maximum amount of voices) is connected to the appropriate voice from the contig before and after it using a cost function that penalizes large leaps, continuing outwards until the extremities of the piece. It also allows the algorithm to intuitively connect notes that had been broken up between contigs. Though this algorithm performs very well on J.S. Bach's Well-Tempered Clavier and Two- and Three-part Inventions, achieving voice separation with 88.98% accuracy overall (84.39%, 99.29% and 93.35% respectively), it cannot be performed in real time and therefore is not a perceptually accurate reflection of musical ASA. An addition to this approach was proposed by Ishigaki, Matsubara, & Saito (2011), where contig connections are prioritized by number of voices, where contig transitions that increase in voice count are connected first.

Madsen & Widmer (2006) combine the perceptual principle of pitch proximity and real time processing to create an algorithm that performs highly accurate (best 97.58%) assignment of note transitions, termed *soundness*, (two notes in the same voice in the ground truth remain

in the same voice) and moderately accurate (best 73.76%) assignment of notes to the correct voice, termed *completeness*, for the same music by J.S. Bach as Chew and Wu's (2004) evaluation, referred to as the 'chewBach' dataset. Running a few sets of experiments testing various values for their cost function, they also implemented a pattern matching heuristic in one experiment, where if the last five intervals were present in the piece previously, that this pattern not be broken up into multiple voices. This improved the algorithm's performance in some cases, but not in all as the competition between the pattern matching and pitch proximity introduces other types of errors which are not described.

Following in the footsteps of Killian & Hoos (2002), Cambouropoulos' voice integration/segregation algorithm, or VISA (Cambouropoulos, 2008; Karydis, Nanopoulos, Papadopoulos, Cambouropoulos, & Manolopoulos, 2007; Makris, Karydis, & Cambouropoulos, 2016), also allows more than one note per voice, naming this the synchronous note principle. This implementation differs from Killian & Hoos (2002), aiming to model vertical and horizontal grouping simultaneously and progressively as the music unfolds (though here knowledge of future events at any particular time is still necessary). VISA accepts quantized MIDI data, 'sweeping' through a given work by processing vertical slices of music at each possible onset. If a slice contains more than one note, the context in a window around that slice is examined: if co-sounding notes have different onsets and/or offsets, polyphony is assumed and the slice is divided into different voices; if co-sounding notes are synchronous, homophony is assumed and the slice is considered to be one voice. Thus each slice is divided into multiple single-note clusters, or remains one multi-note cluster. Each cluster is then linked to each voice from the set of previously detected voices and a cost function encoding the principles of temporal continuity and pitch proximity is calculated so that clusters closer in time and pitch to a particular voice are assigned to it. The concept of pitch co-

modulation is also incorporated into the latest version of the model (Makris, Karydis & Cambouropoulos, 2016), where synchronous streams moving in non-parallel motion are segregated. Two additional constraints are added: voice crossings should be avoided, even if this results in a sub-optimal cost value, and the top voice (the assumed melody) should be minimally fragmented, where clusters can be further divided if necessary to fulfil this requirement.

4.2.2 Score-based models

Consulting musical scores is a less common, but informative way to gather insight about streaming. In voice leading and orchestration, certain rules and conventions (more or less explicit) aid or deter the streaming percept, based on the goal of the composer. Huron (2001) and Duane (2013) adopted this approach to explore streaming, albeit somewhat differently. Huron (2001) used psychological constructs to derive existing and new voice leading principles, all supported by empirical research, either through the collection of participant data or score analysis. The psychological constructs he explored are toneness (clarity of pitch), temporal continuity (preference for continuous over intermittent sounds), minimum masking (necessity for more space between simultaneous pitches as pitch descends), tonal fusion (consonance encourages fusion), pitch proximity (nearby pitches group together), pitch comodulation (simultaneous pitches moving together group together), onset synchrony (synchronous onsets promote integration), limited density (more than three voices cannot be perceived accurately), timbral differentiation (larger differences encourage segregation) and source location (larger distances encourage segregation). The voice leading principles are the registral compass rule (voice leading is best between pitches F_2 and G_5), the textural density rule (there should be three or more parts; typically four), the chord spacing rule (wider intervals in lower voices), the avoiding unisons rule (voices shouldn't share the same pitch), the common

tone rule (if a pitch can be repeated in a voice it should), the nearest chord tone rule (voices move to the nearest chord tone), the conjunct motion rule (if a pitch must change, a diatonic step is best), the avoiding leaps rule (smaller movement promotes integration within a voice), the part-crossing rule (parts should not cross), the part overlap rule (parts should not overlap), the parallel unisons, fifths and octaves rule (they should be avoided) and the consecutive unisons, fifth and octaves rule (they should be avoided). All of these textbook voice leading rules are derived as realizations of empirically demonstrated psychological principles to promote integration within voices and segregation between voices. Duane (2013) on the other hand analysed string quartets to find features in the music that encourage streaming. Onset synchrony (percent synchronous onsets), offset synchrony (percent synchronous offsets), pitch comodulation (two lines moving in the same distance and the same direction) and harmonic overlap (measure of shared overtones in a given vertical interval) were calculated for pairs of measures coming either from different quartets or from the same quartet but always from different instruments. For each pair, multiple regression using the four measures as predictors revealed that pitch onset and offset were the most important factors in predicting the percept of streaming, according to author and other participants' annotations, followed by pitch comodulation and finally harmonic overlap. Duane also made a distinction between textural streams and musical streams, where textural refers to voices within a work (e.g. violins doubled at the third form a stream) and musical refers to a whole piece of music (e.g. two marches form separate streams in some music by Charles Ives). Harmonic overlap was not an important factor in forming textural streams but was for musical streams. This is likely because harmony is usually a whole-work concept: all voices are involved in harmony at any point in time and so would not play a large role in forming textural streams within pieces.

4.2.3 Data-driven models

Here data-driven models will be discussed, where algorithms are trained on music, learning real patterns rather than following rules imposed by the user.

Kirlin & Utgoff (2005) created VoiSe, a voice separation algorithm in two steps: first a predicate identifies whether two notes are in the same voice or not, and then notes are numbered based on this predicate, where notes in the same voice receive the same number label and notes in different voices receive different number labels. Quite simply, notes with the same number label belong to the same voice. The predicate is based on both pitch and time information, learned from a set of training data. The soundness and completeness evaluation metrics varied widely in accuracy, depending on the training and test data: just above 40% at its worst and 100% at its best. Surprisingly, the algorithm was only evaluated on three 3-measure segments of J.S. Bach's Ciaccona.

Another data-driven approach is suggested by Jordanous (2008), who modifies Chew & Wu's (2004) contig approach by replacing user-defined rules with probabilities acquired through training. The training learns the likelihood of occurrence for each pitch in each voice, and the likelihood for each transition between pitches. Once separated into contigs, the algorithm begins with *marker* contigs that have all possible voices present and whose voices are maximally far apart. Voices are then connected between pairs of these contigs, working towards each other until connection is made. The beginning and end of the piece are then processed from the first and last identified marker contig. This approach performs better on music by J.S. Bach than on Beethoven string quartets, a new style of music for evaluating streaming models thus far, and comparably to other existing models such as Kirlin & Utgoff's (2005), Chew & Wu's (2004), and Madsen & Widmer's (2006).

Temperley (Temperley, 2009) has worked extensively with probabilistic modelling of music (Temperley, 2007, 2008, 2013, 2014), specifically using a Bayesian approach to analyze the musical surface. In his 2009 paper, Temperley introduced a unified probabilistic model of polyphonic music analysis which included metrical analysis, harmonic analysis and stream segregation. The rationale is that these three processes are intimately related, both intuitively and with support from prior research. Harmonies will almost exclusively change on a metrical beat and most often on a strong beat (based on an analysis of the Kostka-Payne corpus), which completely links harmony and meter, while streams tend to begin and end at metrically appropriate places such as a strong beat or the end of a metrical cycle respectively, linking streaming structure to meter, and harmony by extension. The model contains two overarching, chronological processes: a generative process and an analytical process.

The generative process begins with the meter, generating first a tactus level, followed by one beat structure level above and two below, defining the meter as simple or complex, and duple or triple. Events (onsets, offsets, harmonic changes, beats, stream start and end) are restricted to *pips*, discrete timeline points 50ms apart. Note onsets are generated before note offsets. The harmonic structure is then generated, labelling segments of the piece with a chord root. Harmonic changes are only allowed on a tactus beat, an acknowledged simplification. Movement by fifths is highly preferred, if a change is preferred at all. Finally, a stream structure is generated: at each tactus beat, a decision is made as to whether an existing stream will continue or not, or whether a new stream should be generated. Once all these structures are generated, a pattern of notes is generated to fit into these structures. The model first decides whether a note will occur on a particular beat or not, determined by a combination of the probability of a note occurring based on the metrical level and a concept Temperley calls metrical anchoring, where a pitch is more likely to occur if there is already another pitch before

it. The pitches are then generated to fit the harmonic structure determined. Parameters for model generation can be set manually or by training on a corpus of monophonic music.

The goal of the analytical process is to find the most probable metrical, harmonic and streaming structures for a given note pattern. Since it is intractable to search the full problem space, the process is broken down into steps, similarly to the generative process. The analytical process begins where the generation process ended: with streaming structure. The model searches for a streaming structure where the number of streams is fairly small, there are not many rests within streams and pitch intervals within streams are small. Additionally, two different streams cannot cross in pitch nor occupy the same *pip*. Metrical and harmonic analysis are then undertaken together, making use of the concept of the *tactus-root combination* (TRC), the combination of a tactus interval (between two tactus beats) and a chord root, where the probability of a given TRC depends only on the previous TRC and the probability of beats and notes in a given TRC depend only on that TRC itself.

As no other directly comparable model exists, this unified probabilistic model was evaluated by its component parts. The model's metrical and harmonic analysis components compare favourably to Melisma (Temperley & Sleator, 2001) when evaluated on quantized and performed musical corpora. The streaming analysis was not evaluated due to a lack of clarity as to what constitutes a correct response and no annotated corpora addressing this issue exist. Melisma is a similar, module-based model of more general music perception, based not on probability but on a set of rules, or heuristics: it contains modules for meter extraction, segmentation, streaming, harmony analysis and key detection.

4.2.4 Summary

The models summarized above present a range of strengths and weaknesses. First, general models of ASA (Section 4.1) are typically applied to very simple auditory scenes and

though it may be possible and eventually interesting to break down music to the neural firing patterns it creates, this has not yet been done, nor is it likely to happen soon. Therefore, models of musical ASA are needed to approach the problem from a higher level, in this case symbolic musical events that can be identified as standalone perceptual entities. Second, as previously mentioned, musical ASA and voice separation, though related, are different tasks and existing musical ASA models focus on the latter, musicological, problem rather than the psychological process of ASA. Separating music into voices as they are in a score, while useful for engineering problems such as automatic transcription, does not tell us about how a listener parses that same music into meaningful streams such as melody and accompaniment. Third, a systematic comparison of existing streaming models is non-existent due to the variety of stimuli tested, ranging widely in complexity. However, the framework introduced below will combine the strengths of these models to synthesize an integrated framework for musical ASA and address some of these weaknesses. This framework will allow the comparison of all influences to perceptual auditory streaming, as presented in Chapter 2, with each other as well as comparing different types of stimuli complexity. To accomplish this, it is important to analyse music in real time without knowledge of the future; Cambouropoulos' analysis by slice approach will be taken here. Second, Temperley & Sleator's module-based approach will be taken to integrate the many sources of information formulating musical ASA. Finally, these modules will use probabilistic methods applied sequentially and simultaneously to tackle ASA as a perceptual construct applicable to a variety of stimuli.

4.3 A new framework for musical ASA

This thesis proposes a novel framework for an integrated, prediction-based streaming model, fusing top-down and bottom-up information as well as combining vertical and horizontal analysis to imitate human auditory scene analysis of polyphonic music. This

information includes: auditory features, musical features, musical training, attention and expectation. This framework makes use of predictive coding (Section 2.6), where predictions are generated from frequencies present in training data. These predictions will be used to separate symbolic musical input into perceptual streams. Predictions for different types of streaming information will be calculated in modules, which can be combined to produce a single streaming decision, with an identified melody stream and accompaniment stream(s).

This framework will extend the prediction-based approach of the IDyOM model, for three reasons: 1) IDyOM has been behaviourally validated as a cognitive model of musical expectation; 2) IDyOM already employs multiple viewpoints, thus lending itself to a module-based approach; and 3) IDyOM functions in real time, using only past and current information. The basic functioning of the model will be presented in the context of one viewpoint before being discussed in the context of the full range of information required for musical auditory streaming.

4.3.1 Basic framework function

This framework, to be implemented as a predictive model, will perform its analysis in real time following the example of Cambouropoulos' VISA model, where music is divided into vertical slices in time, where each new event onset begins a new slice. Events that span multiple slices are tracked so that they are linked together throughout the analysis. For each slice, the model breaks these slices up into all possible stream structure combinations, with the restriction that non-adjacent voices cannot be a part of the same stream. For example, in an SATB context, one might perceive all four voices as one perceptual stream, SATB, or each as its own stream, S-A-T-B, or the top voice accompanied by the bottom three, S-ATB, vice versa, SAT-B, and so on, but an ST-AB or STB-A combination are not possible. The most probable continuation, given the preceding context (n-1 slices), is added to the context and the analysis

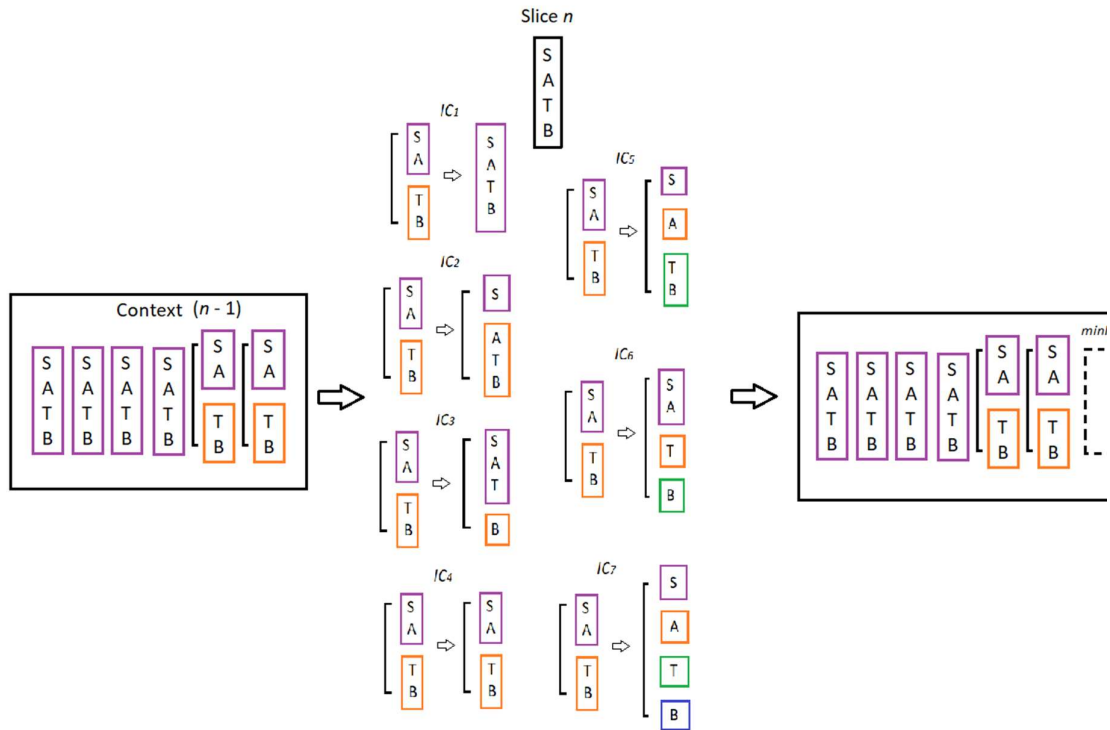


Figure 4.1. Illustration of a predictive module's process, where predictions are generated for each potential streaming structure continuation. These relationships are complex, therefore predictions will be focused on a single viewpoint per module, for example the interval structure of a slice.

continues throughout the piece until all slices are organized into perceptual streams with corresponding average predictability in the form of information content. Figure 4.1 illustrates this process.

The first slice is treated slightly differently due to a lack of contextual information. Presumably like a human listener waiting to hear a piece of music for the first time, expectations would be biased towards the most common type of streaming structure heard previously. Thus, the first slice will be divided into this most common streaming structure, as determined by the model through training. This initial bias, as well as the evaluation of any output by this model will require data annotated with perceptual streaming information, which can be collected from listeners of various musical backgrounds. Finally, to identify the melody

and accompaniment streams, the stream with the highest information content – in other words, the most interesting line – is labelled as melody (Chapter 8).

Outlined above is the overall process of the framework. This iterative analysis process occurs simultaneously in many modules, where each module takes symbolically encoded music as input, here MIDI, kern or text as per IDyOM’s current implementation and outputs information content based on the viewpoint information it models. At each streaming assignment decision, the predictions are linearly combined across modules based on the relative perceptual salience (described in Section 4.4.2) of the viewpoints involved to produce a final

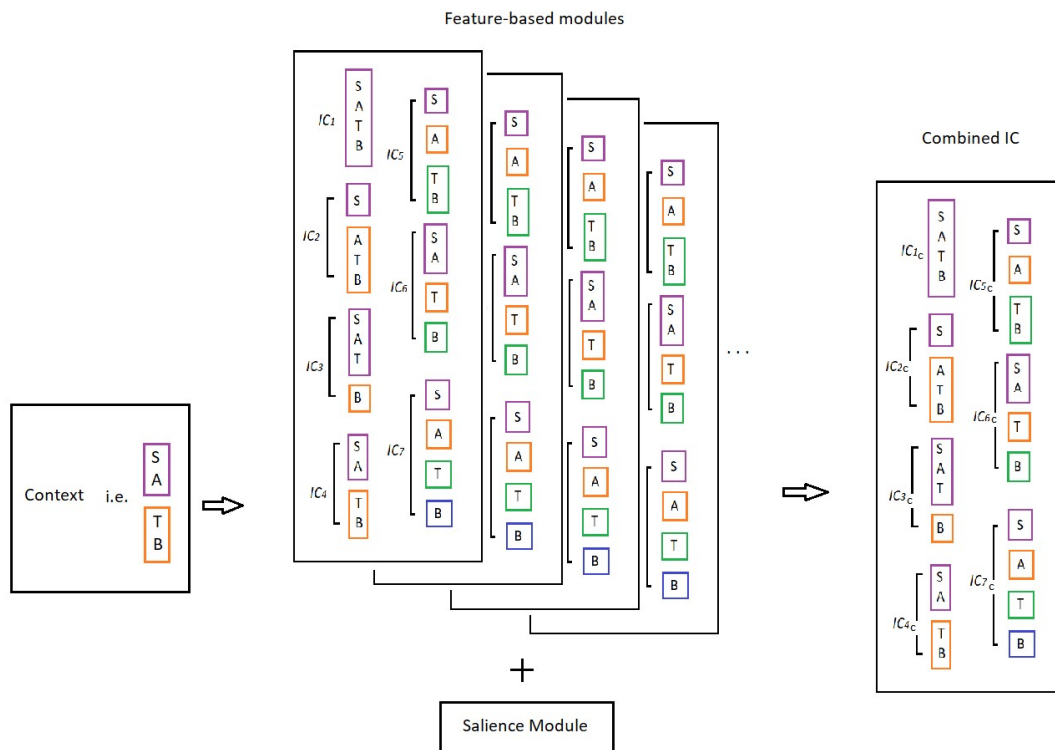


Figure 4.2. An illustration of the proposed streaming model’s work flow. At any point in the analysis, the model will consider the possible streaming structures of the next slice, for example the seven combinations of a four-voice work. Each module will process a feature as described in Section 4.4.1, for example *pitch interval*. The output of all feature-based modules are simply added together, with each module given a weight designated by the salience module, described in Section 4.4.2. The streaming structure with the lowest combined IC, in other words the most likely continuation, is added to the existing context (see Figure 4.1).

stream assignment decision (Figure 4.2). Therefore, it is important that each module is independent and does not rely on the output of another module to create its own.

The remainder of this chapter presents the proposed implementation of all sources of streaming information (Section 4.4) followed by a description of the type of data needed (Section 4.5) and a discussion of potential future research based on the framework (Section 4.6), and its limitations. In Section 4.4.1, acoustic and musical sources of streaming information will be discussed in the context of feature-based modules. In Sections 4.4.2-3, the particular challenges of attention and timbre respectively will be discussed as well as their proposed implementations. Musical training and individual listener's musical knowledge can be modelled using IDyOM's long-term memory implementation. This will be discussed in Section 4.4.4. Finally, in Chapter 2, expectation was also identified as a source of streaming information. As it is prediction-based, expectation is inherently included in the framework. Sections 4.7 and 4.8 will compare this framework to a selection of models reviewed in Sections 4.1-2 and proposes ways in which the framework might be used by researchers from differing backgrounds respectively.

4.4 Streaming information sources in the framework

4.4.1. Feature-based modules

Auditory modules. Based on existing research, modules can be created to process pitch distance, loudness difference, onset synchrony, spatial differences (if relevant) and timbre distance to name a few. In each case, a relevant and meaningful viewpoint must be developed in order for IDyOM to learn its patterns in music and make predictions. These already exist for pitch, interval, scale degree, inter-onset interval and duration, to name a few of the most commonly used in existing research (for a full list, see Chapter 3). Harmonic, timbral and

spatial viewpoints are yet to be implemented, though work on harmonic viewpoints is ongoing in the Music Cognition Lab.

Musical modules. As presented in Chapter 2, parsing musical scenes is influenced by a number of features that play a particular role in the musical auditory scene: harmony, phrase boundaries, repetition and similarity. Harmony strengthens integration in a piece of music overall, where more consonant harmonies (Conklin & Bergeron, 2010) result in the strongest integration. A harmony-processing module can be implemented to reflect this, where an appropriate harmonic viewpoint is developed to learn and predict harmonic patterns.

Phrase boundaries are also a helpful guideline for predicting streaming structure as phrases are typically contained to a single streaming structure throughout. Some models use this to their advantage, easily preventing voice-crossing within segments by analysing segments rather than individual pitches (Chew & Wu, 2004; Ishigaki et al., 2011; Madsen & Widmer, 2006; Rizo et al., 2006). IDyOM already performs real-time phrase boundary detection, therefore this information could be harnessed in its own module to inform streaming decisions, for example by biasing the decision towards inertia – in other words keeping the same streaming structure – between phrase boundaries and allowing more flexibility to change streaming structure at phrase boundaries. This could be implemented by outputting a lower IC for the streaming structure that matches the immediately previous context between phrase boundaries, and outputting a uniform IC across all potential streaming structures at phrase boundaries.

Repetition and similarity are closely related (see Section 2.2), and inform auditory streaming in a similar way: things that are repeated or similar are more likely to originate from the same source. Repetition is inherently modelled in IDyOM through the STM model, where repeated sequences result in low IC as they become increasingly predictable. However, as we

have seen in the context of music, the source can be the overall piece of music performed by either a soloist or an ensemble, or the melody in a piece of music, played by one instrument, or a group of instruments. This complicates repetition and similarity in music substantially, as comparisons must be made between the similarity of phrase variations played by the same instrument and exact phrase repetitions played by two different instruments: which is most similar? Is an exact phrase repetition played by another instrument part of the same perceptual stream, or is it considered a separate, perhaps temporary stream? Further research is required to answer these questions and guide the implementation of modules incorporating repetition and similarity information into this integrated streaming model.

4.4.2 Including attention

Though most auditory and musical aspects of auditory streaming described above can be relatively easily modelled, modelling attention remains elusive and timbre remains a challenging aspect of music to understand. The difficulties of studying and modelling attention have previously been summarized in Section 2.3. An attention module in an integrated framework for auditory streaming would modulate the relative importance, or weight, of all other modules informing the final streaming structure decision for any given musical phrase. This is equivalent to determining the relative salience of all features informing auditory streaming. This would be once again done using information content, where the relative mean IC proportions of feature categories (i.e. pitch, time, harmony) for each stream built up in the model so far is translated into a corresponding weight accorded to the output of feature-based modules. For example, if the mean IC of *pitch interval* for all existing streams is 1.2 times the mean IC of *inter-onset interval* for all existing streams, the IC output of all pitch-related feature-based modules – auditory and musical – making predictions about the streaming structure of the next slice will be weighted 1.2 higher than time-related feature-based modules

for those same streaming structures. The work presented in Chapter 6 of this thesis will explore this hypothesis.

4.4.3 Timbre

Commonly defined as the aspects of sound that are not pitch, duration or loudness, a straightforward definition of timbre still eludes researchers. Attempts to break the concept down result in either many components with limited psychological interpretation in the case of MIR or two main temporal and spectral components that still fail to fully explain timbre in the case of music cognition (Alluri & Toiviainen, 2010; Caclin, McAdams, Smith, & Winsberg, 2005; Lakatos, 2000; McAdams et al., 1995). While determining the distance between pairs of timbres in a two-dimensional space and associating this with an integration or segregation threshold is possible (Sauvé, Stewart, & Pearce, 2014), musical timbre presents two particular challenges: 1) timbre changes as a function of pitch; and 2) music contains timbral blends, as a result of different instruments playing simultaneously in the case of ensemble music. The first challenge is particularly relevant to solo instrumental music, where timbral differences between voices (i.e. piano) or across a piece of music (i.e. flute) are much more subtle than in ensemble music. How important is timbre as a streaming cue in these situations? Furthermore, timbre changes differently as a function of pitch for different instruments, making the creation of an accurate model of solo timbre very complex to begin with; what about when instruments combine? The challenge of instrumental blend has begun to be investigated for pairs of instruments (Kendall & Carterette, 1991; Sandell, 1995). While these studies provide a much-needed start into understanding timbral blend, this line of research is only still in its infancy, and much work is needed to develop timbral understanding for the blend of not only two instruments, but many, in order to model timbral perception in large ensembles such as orchestras.

It is clear that many questions still remain to be answered in order to better understand timbre perception, and I would argue that this issue is the most important to address in order to better model musical ASA. Existing research has established timbre as a relevant streaming cue (Albert S. Bregman & Pinker, 1978; Handel & Erickson, 2004; Iverson, 1995; Marozeau et al., 2013; Sauv e, Stewart, & Pearce, 2014; Singh & Bregman, 1997) and the art of orchestration is concerned with choosing the most appropriate instruments for achieving the desired auditory scene. The same is true in composition, where timbre is one of the most important cues for segregating the melody from the accompaniment, or creating one unified sound mass. Therefore, if timbre perception can be appropriately modelled, so can musical ASA and the challenge in the context of this framework is to create a meaningful viewpoint that can do so.

4.4.4 Including musical training

A further challenge of modelling musical auditory scene analysis is the influence of individual differences on perception. In some cases, these even exceed musical training group definitions (Dean et al., 2014a). Research in musical training suggests that it is a prolonged and focused exposure to music that causes differences in perception between musicians and non-musicians (Fujioka et al., 2006, 2004; Habib & Besson, 2009; Micheyl et al., 2006), but only for domain-specific (i.e., musical) tasks (Bigand & Poulin-Charronnat, 2006; Carey et al., 2015). Those who receive Western musical training are also more likely to be exposed to and understand Western classical music, where non-musicians may have less exposure to this genre (though exposure is unlikely to be zero) and more to popular genres such as rock, pop or dance. Musicians will likely have been exposed to popular genres as well; it is difficult to avoid music in everyday life and individuals have probably been exposed to more genres than they realize.

Therefore, it is the relative contribution of each genre’s knowledge that will help define a listener’s background.

IDyOM has been shown to differentiate between jazz and classical music when trained on each of these specific genres of music (Hansen et al., 2016) by generating lower information content for music in the same genre it was trained on (see also van der Weij, Pearce, & Honing, 2017). This can theoretically be extended to simulate individual or groups of listeners’ musical backgrounds, including different cultures, where IDyOM’s long-term model can be trained on any chosen (currently restricted to monophonic) pieces of music. The choice of training material then becomes crucial as it influences the information content output of the core model, which informs melody selection. As melody selection is based on the highest average information content relative to other voices, the absolute IC value does not matter (i.e., higher IC for a lesser known genre will be higher for the entire piece of music, not just one voice); however, it is possible that highly idiomatic instrument-specific figures (e.g., violin arpeggiations) become well-learned and receive low information content, potentially low enough to push a melody into a non-melodic position according to the melody extraction module. This possible limitation should be explored in future research.

4.5 Training data

One of the biggest obstacles to studying perceptual stream segregation is that there is no existing annotated corpus of data identifying perceptual streams. This makes streaming pattern learning and evaluation of any model output currently impossible. However, as soon as a corpus is created, such models can begin to be tested. This particular framework requires the labelling of perceptual streams in its symbolic music input, and might look something like the following: for each slice, the relationship between each voice is labelled as integrated (0) or segregated (1), so that a four-voice slice may have a completely integrated streaming

structure (SATB) – [(0, 0, 0) (0, 0) (0)] – or a melody and accompaniment – [(1, 1, 1) (0, 0) (0)]. In this notation, each group of integers represents the relationship between pairs of voices, i.e. [(SA, ST, SB) (AT, AB) (TB)]. In this way, the model can learn which features result in vertical integration and which features result in vertical segregation in combination with sequential patterns. This vastly expands the number of transitions being calculated as the model considers such probabilities as the transition between a three-voiced stream with a particular set of pitches to a two-voiced stream with another particular set of pitches or any other combination of voices. Therefore, it is likely that basic viewpoints, such as *chromatic pitch*, become irrelevant as they would be much too situationally specific, and the model learns best by detecting interval patterns, such as parallel movement (i.e. parallel third and sixths are common but not fifths or octaves) between voices in a same stream, or diverging movement that leads to voices splitting into different perceptual streams.

This type of data is also crucial to model evaluation, where a ground truth is needed. While perceptual stream segregation is subjective, a representative ground truth can be generated from a wide range of listeners. It would also be possible to collect ground truths for different types of listeners (i.e. classical musicians, rock musicians, piano players, non-musicians) and compare this ground truth to model outputs trained on corpora approximating these listener profiles.

While the collection of such a corpus is a significant undertaking, it would add significant value to the field of musical auditory scene analysis research. However, as this is a theoretical framework at this stage, this collection was not done as part of this thesis.

4.6 Model comparisons

In this section, the proposed framework just presented will be compared to a selection of existing models of auditory stream analysis, identifying potential areas of improvement

while demonstrating good generalizability. Comparisons will be made with the CHAINS model by Schröger et al., (2014), Denham & Winkler's (2006) predictive model of auditory streaming and Cambouropoulos' VISA model (Cambouropoulos 2008; Karydis, Nanopoulos, Papadopoulos, Cambouropoulos, & Manolopoulos, 2007; Makris, Karydis, & Cambouropoulos, 2016). These were selected to be representative of high- and low-level streaming analysis models applied to both musical and non-musical stimuli.

CHAINS (Schröger et al., 2014), is an implementation of the Auditory Event Representation System (AERS; Section 4.3.1). While a preference for regularity and competition between stream organizations are implemented in the proposed framework, multi-stability as it is understood in CHAINS – switching between one and two streams – is not. It is possible to change streaming structure in the proposed framework, but these changes include more than the two options considered in CHAINS. The phenomenon of switching between integrated and segregated percepts has only been studied in any depth with very simple stimuli (Pressnitzer & Hupé, 2006; Pressnitzer et al., 2011). Presumably, musical stimuli are more stable, where a change in percept might be accompanied by an important change in texture and percepts would remain stable for the duration of a musical phrase at least, as implemented in the framework. Furthermore, the proposed framework, like CHAINS, builds streams using predictive processes. Therefore, the proposed framework subsumes the CHAINS model.

Similar to CHAINS, Denham & Winkler's (2006) auditory streaming model is low-level and deals with simple, monophonic tone sequences. Inspired by a combination of neurophysiological and behavioural evidence from auditory research and visual models of bi-stability, this model incorporates the concepts of segregation, predictive modelling, competition and adaptation. While segregation is established by relative suppression of neural activity between different tones, streams are formed by regularity detection, where each stream

becomes its own predictive model. While this is an interesting approach to using predictive processes to perform auditory streaming analysis, this currently only applies to monophonic sequences, while the proposed framework extends to polyphonic contexts. Finally, adaptation is included in the proposed framework as it can flexibly change percepts over time, with each musical slice labelled with its own streaming structure.

Though the proposed framework conceptually subsumes these two low-level models of auditory streaming, it is worth explicitly considering whether the proposed framework is capable of processing similar simple sequences, having been primarily designed for polyphonic musical input. In its current formulation, monophonic input would lead to only one voice being created by the model and therefore to a single integrated stream, as there would be no other alternative. It would be necessary to allow the segregation of sequential events into separate streams, where perhaps sharp increases in information content might signal the presence of a new stream.

Finally, the VISA model (Cambouropoulos, 2008) performs both vertical and horizontal integration/segregation analysis, in that order, on a range of musical textures by using the perceptual principles of temporal continuity, pitch proximity and pitch co-modulation. VISA's strengths include that it does not a priori establish a fixed number of streams, a strength shared by the proposed framework, and that it can be adjusted to function in real time. The proposed framework incorporates all of VISA's perceptual principles and adds information, thus subsuming VISA.

To summarize, while the proposed framework is comparable to and theoretically subsumes a number of existing auditory streaming models spanning high- and low-level processes and simple and complex stimuli, addressing the majority of the same issues as these models, some discrepancies can be identified, targeting differences between the types of stimuli

handled. Primarily, the proposed framework is restricted to polyphony, where monophonic input can only produce a single integrated streaming percept as output. Importantly, the data needed to implement the proposed framework does not yet exist, keeping it squarely in the realm of theory until such data can be produced. These comparisons have demonstrated the scope and limits of the generalizability of the proposed framework, where generally concepts associated with auditory streaming are well covered while details of implementation for varying inputs are limited.

4.7 Sample use cases

To demonstrate the framework's flexibility and collaborative potential, here are two examples of how researchers from differing disciplines might use the proposed framework.

Musicology. Musicological analysis is score-based (e.g., Duane, 2013; Huron, 2001), where analysis is typically concerned with identifying aspects of the musical surface that inform perception. A musicologist might be interested in investigating particular auditory features such as the relative explanatory power of pitch and loudness, as given by notation and dynamics respectively on collected streaming annotations. These two modules can be used together or in isolation, where the relative weight of each can be systematically manipulated. Thus, the relative weight combination of the two modules that best explain the variance in the streaming annotation data provide new information about how the brain uses information that is specifically encoded on a musical score.

Neuroscience. Neuroscience, on the other hand, is concerned with the neural mechanisms behind perception and cognition, modelling neural activity at a very low level. While such low-level analysis is not currently implemented in the framework, it is possible to add such low-level processing modules, where these might operate on pairs of voices, a simpler stimulus input for which most neural-based streaming models are designed. These low-level

modules could similarly be designed to interpret neural firing patterns induced by the analysed voice pair to produce a streaming decision.

This type of comparison between approaches provides a valuable opportunity to obtain converging evidence for a concept, for example that small pitch distances promote integration, and the potential for increased collaboration across disciplines, where a musicological module can be compared to a neural module, and ideas from one discipline can be tested by another through a different implementation, all within the same overarching framework. This variation in analysis levels also allows the exploration of concepts at various time scales. For example, this may confirm the primacy of pitch in segregation at the neural and cognitive stages, while harmonic movement may only be relevant in higher-level cognitive processing.

4.8 Limitations

A current practical limitation of this framework, based on known use of IDyOM, is long processing time. It is possible that the brain, if indeed making use of mechanisms such as statistical learning, simply has more computing power than current computers, resulting in real-time, millisecond processing of acoustic information. However, it is also possible that the brain is more parsimonious than previously thought and thus exploring parsimony with respect to the information used when making streaming decisions is a useful place to begin in order to reduce processing time. This will be discussed further when the framework is re-evaluated in Chapter 9.

An important theoretical limitation is that the primary assumption made by the model, that statistical learning approximates predictive processes in the brain, is based on a single broad theoretical framework, where if this is ever falsified, the entire framework becomes invalid. Despite this, it remains that assumptions must be established in order to investigate a

theory in new and interesting ways, and to allow it to be disproven. Furthermore, there is a highly established precedent in the literature for the use of this theoretical framework.

This framework also assumes a linear combination of module output, where all feature-based module outputs are simply added. While their relative contribution to the final information content value is modelled by relative salience, the model does not account for the possible discrepancies in time scales at which various parameters operate, treating all equally. For example, while it is possible that pitch is a strong influence in onset to onset streaming perception, harmony is likely to play an important role over a longer period of time. While this approach captures harmonic information, the scale may not be accurate. It will be important to take this limitation into account in investigating auditory streaming using this framework, and investigate potential solutions to overcome it.

Finally, as a high-level, cognitive framework, this proposal contains no direct physiological model. While this was a motivation for the creation of this framework, it is important to recognize its restriction to higher-level concepts, thus limiting compatibility and comparability with low-level, neural implementation models of auditory streaming.

Aside from these limitations, the proposed framework described in this chapter relies on current theories and empirical evidence to create a tool for musical auditory scene analysis research that is flexible, integrative and collaborative by design. With further formal specification, quantification and implementation, the framework can potentially allow researchers to flexibly investigate subsets of auditory scene analysis, working together to improve understanding of the human auditory system, and the brain.

4.9 Conclusion

Thus far, this thesis has presented a theoretical integrated framework for musical ASA, focused on using prediction as a unifying concept. The remaining chapters will be dedicated

to better understanding five aspects of music perception selected from this proposed framework: expectation generated by rhythm, musical emotion, complexity, relative salience of musical parameters and melody extraction. Initially, the effects of timbre, attention and musical training on auditory streaming will be explored in Chapter 5. Then, with IDyOM validation previously focused on pitch viewpoints, Chapter 6 will validate the temporal viewpoints *onset* and *inter-onset-interval*, further supporting IDyOM as a cognitively valid predictive model of musical expectation. Next, musical expectations as generated by pitch and timing aspects of music will be linked to real-time emotional response in listeners, as measured by arousal and valence. In Chapter 7, the hypothesized link between information content and perceived complexity, and complexity and salience, will be tested. Chapter 8 will present an extension of IDyOM that uses two prediction-based hypotheses to extract melody from a symbolic polyphonic context: melodies are internally predictable, and they are the most interesting (unpredictable) voice in a polyphonic work. This model will be evaluated on a selection of string quartets by W.A. Mozart and chorales by J.S. Bach. The final chapter will re-evaluate the theorized framework presented above in the context of the results provided from testing these five concepts.

5 Attention but not musical training affects auditory grouping

The role of timbre as an auditory cue for auditory streaming was briefly introduced in Chapter 2, more specifically Section 2.1 focusing on basic auditory features. This chapter will investigate timbre as a streaming cue, exploring the potential influences of listener background, attentional set and prior expectation on the streaming percept. Furthermore, while timbre is an established streaming cue, stimuli are always simplified for finer control; here, the use of ecological stimuli confirms timbre as such a cue in an increasingly more natural context, using

both the previously introduced ABA paradigm (Section 5.2) and a more ecological paradigm, the interleaved melody paradigm (Section 5.4).

5.1 The study of timbre perception

Timbre is a complex auditory parameter and timbral perception has been investigated in detail using both synthesized tones and real instrumental sounds (Alluri & Toiviainen, 2010; Caclin, McAdams, Smith, & Winsberg, 2005; McAdams et al., 1995). The most common method of investigating timbre has been multidimensional scaling, or MDS. Based on dissimilarity ratings between pairs of timbres, sounds are mapped into a multi-dimensional space representing perceptual distance. In research to date, three dimensions seems to provide an optimal representation of perceptual timbre space; though the first two are fairly stable across experiments, the third is less well established. The first two represent log rise time (the attack), and spectral centroid while the third dimension that emerges is usually a spectro-temporal feature such as spectral flux or spectral irregularity. One of the biggest issues with this research however is that in most cases the rated sounds are synthesized (though see Kendall & Carterette, 1991; Lakatos, 2000 for examples of MDS using natural stimuli). Besides this, our perceptual system is not used to hearing synthetic sounds such as these and may process them differently than natural sounds (Gillard & Schutz, 2012). Therefore, it is important to complement studies using controlled synthesized tones with investigations using natural sounds.

The role of musical training has been extensively studied in the context of auditory skills, including auditory streaming (François et al., 2014; Zendel & Alain, 2009). As a result of training, musicians are more sensitive to changes in auditory stimuli based on pitch, time and loudness for example (Marozeau et al., 2013; Marozeau, Innes-Brown, Grayden, Burkitt, & Blamey, 2010), with discrimination thresholds being lower in musicians than in non-musicians. One problem with treating musicians as a single category is that differences between

instrumentalists may be missed (Tervaniemi, 2009). Pantev and colleagues (Pantev et al., 2001) found that certain instrumentalists were more sensitive to the timbre of their own instrument than to others, as measured by auditory evoked fields (AEF). Violinists and trumpet players were presented with trumpet, violin and sine tones while MEG was recorded. Both instrumentalists presented stronger AEFs for complex over sine tones, and stronger AEFs still for their own instrument. In a similar study (Shahin et al., 2008), professional violinists and amateur pianists as well as young piano students and young non-musicians were presented with piano, violin and sine tones while reading or watching a movie and EEG was recorded. Gamma band activity (GBA) was more robust in professional musicians for their own instruments and young musicians showed more robust GBA to piano tones after their one year of musical training. Furthermore, Drost, Rieger, & Prinz, (2007) found that pianists and guitarists' performance on a performance task was negatively affected by auditory interference, but only if it was their own instrument. Taking a step further and using more ecological stimuli, Margulis, Mlsna, Uppunda, Parrish, & Wong, (2009) explored neural expertise networks in violinists and flautists as they listened to excerpts from partitas for violin and flute by J. S. Bach. Increased sensitivity to syntax, timbre and sound-motor interactions were seen for musicians when listening to their own instrument.

More recently, pianists, violinists and non-musicians listened to music during fMRI scanning (Burunat et al., 2015). The authors investigated the effects of musical training on callosal anatomy and interhemispheric functional symmetry and found that symmetry was increased in musicians, and particularly in pianists, in visual and motor networks. They concluded that motor training, including differences between instrumentalists, affects music perception as well as production. Other research has investigated differences between types of musical training. For example, one study used EEG to show that conductors have improved

spatial perception, when compared to non-musicians and pianists (Nager et al., 2003). Another line of research investigates pianists' formation of action-effect mappings due to the design of their instrument (Baumann et al., 2007, p. 200; Drost, Rieger, Brass, Gunter, & Prinz, 2005; Repp & Knoblich, 2009; Stewart, Verdonschot, Nasralla, & Lanipekun, 2013).

However, such specific effects of instrumental training have not yet been observed in auditory streaming, where an effect would be seen by a change in streaming threshold. Two studies, presented in Sections 5.2 and 5.4, test the hypothesis that due to increased sensitivity to a particular timbre, it would take less time to detect two separate auditory objects when one of these objects is one's own instrument. These use the ABA streaming paradigm in the first instance, and a more ecologically valid paradigm called the interleaved melody paradigm in the second. Another third study, presented in Section 5.3, is a control study examining the effects of prior expectation on streaming perception.

5.2 Study 1: Timbre as a streaming cue

The ABA_ paradigm (van Noorden, 1975) is used here and timbre is manipulated instead of pitch. While the timbre of a *standard sequence* remains static throughout a given trial, a *target sequence* morphs from one timbre to another, creating a qualitative change from a galloping ABA_ rhythm to the perception of two simultaneous, isochronous A_A_A and B__B__B patterns as the standard and target sequences' timbres become more and more different, or vice versa as the timbres become more similar. The point of change in rhythmic perception reflects the detection of a new sound object, or, in the other direction, the merging together of two sound objects. The sound objects (standard and target streams) are defined solely by their timbre, as pitch, length and loudness are controlled. Based on previous work (Sauvé et al., 2014), detection of a sound object defined by one's own instrumental timbre is predicted to occur sooner than for other instrumental timbres, when the participants' instrument

is the target (i.e. it is ‘new to the mix’ and captures attention) and later than for other instrumental timbres when the instrumentalists’ timbre is the standard (i.e. it already holds attention and delays perception of the arrival of a new sound object). This previous study compared seven different instrumental timbres in the same ABA_ paradigm, while additionally exploring the effect of attention on streaming by manipulating participants’ attentional focus. Results guided the design of the current study by providing target effect sizes, refining the test timbres and allowing the elimination of the attention manipulation, as it was confirmed to have a significant impact on the perception of auditory streams.

5.2.1 Participants

Participants were 20 musicians (13 females, average age 34.45; SD = 7.59; range 21-69) recruited from universities and the community. Their average Gold-MSI score (Müllensiefen et al., 2014) for the musical training subscale was 40.15 (SD = 4.23); 5 were violinists, 6 were cellists, 5 were trumpet players and 4 were trombone players. Ethical approval was obtained from the Queen Mary Ethics Committee, QMERC1333.

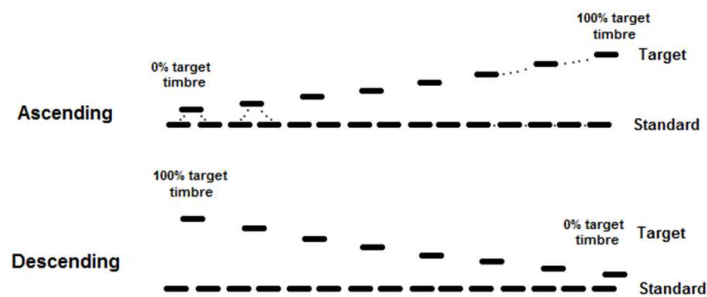


Figure 5.1. Illustration of ABA_ paradigm, ascending and descending, modifying timbre only.

5.2.2 Stimuli

All four timbral sounds (violin, trumpet, trombone, cello) were chosen from the MUMS library (*McGill University master samples collection on DVD*, 2006) with pitches spanning an octave (all 12 pitches between A220 to G#415.30). The files were adjusted to equal perceptual length of 100ms and equal loudness, based on the softest sound. A 10ms fade out was applied to each timbral sound. All editing was done in Audacity and the final product was exported as a CD quality wav file (44,100 Hz, 16 bit). See Appendix A for full details.

Using a metronome in Max/MSP 6, the *standard sequence* was presented by playing a selected timbre with an inter-onset interval of 220ms. The *target sequence* was presented using another metronome at a rate of onset of 440ms, beginning 110ms after the standard sequence to create the well-known galloping ABA_ pattern (van Noorden, 1975). The target sequence was a series of 100ms sound files representing a 30s morph between the standard timbre and the target timbre, achieved using a slightly modified Max/MSP patch entitled ‘convolution-workshop’. This patch is distributed by Cycling ‘74 with Max/MSP. The target sequence morphed from standard to target timbre in the ascending condition, creating a galloping to even rhythm change, and from target to standard timbre in the descending condition, creating an even to galloping rhythm change (see Figure 5.1). Each trial ended when the participant indicated a change in perception or after 30s if participants did not reach a change in perception.

5.2.3 Procedure

The experiment was coded and run in Max/MSP 6, with output presented through headphones and input taken from mouse clicks. Participants were first presented with a practice block with instructions and an opportunity to listen to each timbre and rhythm separately. Up to four practice trials were included in the block and questions were welcomed. Participants then began the first of two experimental blocks.

For each trial, participants indicated by clicking a button on the screen at which point the galloping sequence became perceived as two separate streams of standard and target tones, or the opposite for descending presentation. This point was recorded as the percent of time passed in the trial, which equates to the percent of morphing at that time. Each trial lasted a maximum of 30s, at which point the trial ended automatically and a value of '-1' was recorded, indicating that the participant had not reached a change in perception on that trial. Trials were presented in two blocks, and participants were instructed to indicate a change in rhythm as soon as it was perceived for the ascending block and to hold on to the original rhythm as long as possible for the descending block. Together, this gives two measures of the fission boundary (van Noorden, 1975). The fission boundary was measured instead of the temporal coherence boundary due to its higher sensitivity for detecting timbral effects in perception, and due to confirmation that the fission and temporal coherence boundaries are separate phenomena that can be manipulated by instruction (Sauvé et al., 2014). For every block, every timbre modulated to every other timbre once for a total of 12 trials (4 timbres each modulating to the 3 other timbres), each separated by 4s and each at a different pitch, to reduce trial to trial expectancy and habituation. Participants were randomly assigned to one of two different orders to control for any order effects.

Once both blocks were completed, participants filled out the musical training sub-scale of the Gold-MSI (Müllensiefen et al., 2014).

5.2.4 Analysis

Effect sizes and confidence intervals were used in the analysis of Experiments 1a and 1b, in addition to traditional methods. These methods are based on Cumming (2012; 2013), who advocates wider use of effect sizes and confidence intervals in the research community to increase integrity, accuracy and the use of replication. According to Cumming, the low

occurrence of null results in the literature and a pressure towards new studies and away from replication translates into misrepresentation and inhibition of scientific knowledge. Cumming advocates the use of effect sizes, confidence intervals, and meta-analysis in place of null hypothesis significance testing (NHST). This method is preferred because confidence intervals give more information both about the current effect size, and about potential future replications by offering a range of potential values for a measure, rather than one indicator of significance or non-significance. For more information about effect size and confidence interval methods, see Cummings' book, *The New Statistics* (2012) or the corresponding article for a shorter summary (2013).

5.2.5 Results

Percentage of time passed (degree of morphing) is the dependent variable analysed; for descending trials the percentage was subtracted from 100 so that ascending and descending conditions can be compared directly. A low percentage indicates early streaming in the ascending condition and late integration in the descending condition while a high percentage indicates late streaming in the ascending condition and early integration in the descending condition. Furthermore, trials in the ascending condition where the percentage exceeded 100 were replaced with 100 and trials in the descending condition where the percentage was negative were replaced with 0. These are all cases where the participant listened to the trial for more than 30 seconds and still did not hear a change in rhythm. Five participants' data were removed because they did not hear a change in rhythm in more than half of the trials, in either or both blocks (two violinists, two cellists and a trumpet player). The difference between mean percentage for ascending and descending conditions was 1.2, 95% CI [-4.4, 6.8]. As the CIs include zero, the difference was not significant. However, mean percentage of time passed was significantly higher for the first block of trials than the second, with a difference of 10.6 [2.6,

18.6] for the ascending and 10.7 [3.1, 18.1] for the descending conditions. As both CIs do not include zero, the difference is significant.

Effects of specific instrumental training were investigated next. Data were grouped by instrumentalist and then sub-grouped by standard timbre. For violinists, mean percent time passed when violin was the standard timbre was 56.5 [50.7, 62.3], mean percent for cello was 59.8 [53.5, 66.1], mean percent for trumpet was 65.8 [51.8, 79.8] and mean percent for trombone was 64.6 [50.8, 78.4]. See Table 5.1 for details of all instrumentalists. Data were then sub-grouped by target timbre. When violin was the target timbre, mean percent for violinists was 62.1 [50.0, 74.2], mean percent for cellists was 48.2 [36.9, 59.5], mean percent for trumpeters was 54.5 [44.5, 64.5] and mean percent for trombonists was 48.4 [38.1, 58.7]. See Table 5.1 for details of all target timbres. Figure 5.2 displays results graphically.

Thresholds for an instrumentalists' own timbre were hypothesised to be lower when

Table 5.1. Mean percent of trial duration by standard and target timbre, and by instrumentalist, with 95% confidence interval margins of error (MOE).

		Mean Duration ± MOE			
		Violin	Cello	Trumpet	Trombone
Standard Timbre	Violinist	56.5 ± 5.8	59.8 ± 6.3	65.8 ± 14.0	64.6 ± 13.8
	Cellist	50.7 ± 12.0	46.3 ± 12.1	55.7 ± 11.9	47.8 ± 11.0
	Trumpeter	53.7 ± 11.8	50.8 ± 6.6	50.0 ± 10.3	59.6 ± 12.1
	Trombonist	49.4 ± 13.9	57.0 ± 9.1	48.0 ± 9.9	39.5 ± 9.5
Target Timbre	Violinist	62.1 ± 12.1	62.1 ± 13.4	62.9 ± 9.1	62.0 ± 9.8
	Cellist	48.2 ± 11.3	51.3 ± 11.4	46.8 ± 13.9	53.2 ± 11.6
	Trumpeter	54.5 ± 10.0	47.3 ± 9.0	67.1 ± 11.4	49.2 ± 9.8
	Trombonist	48.4 ± 10.3	45.3 ± 12.3	45.4 ± 11.4	53.4 ± 7.4

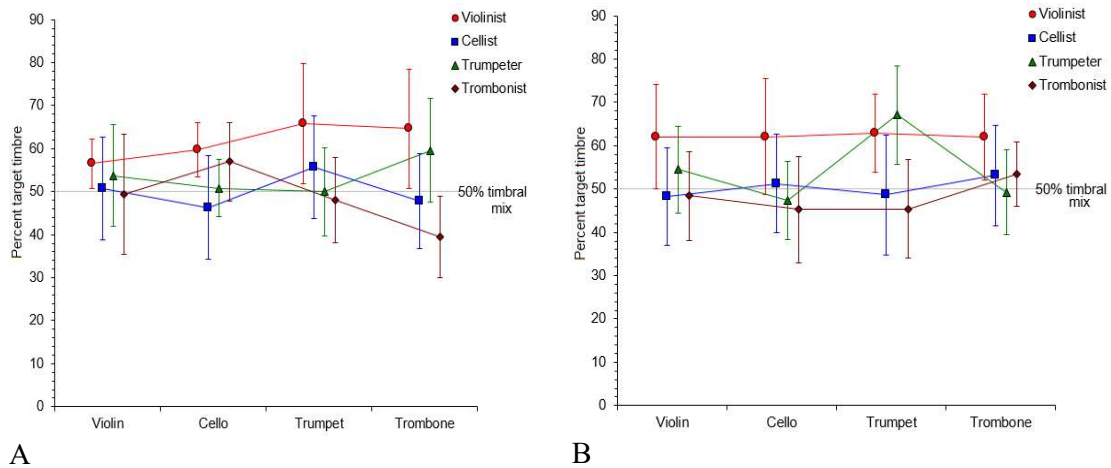


Figure 5.2. Percent target timbre contained in the morphing stream at the point of a change in percept as a function of instrumentalist, and standard (A) and target (B) timbres. Error bars represent 95% CIs.

their own instrument was the target and higher when it was the standard. However, interpreting the CIs above does not reveal any reliable pattern of results. If more than half the margins of error (MOE), which is one half of the CI, overlap when comparing between subject groups, the difference is not considered significant. While two comparisons attain significance (trombone players have a lower threshold than trumpet players for the trombone sound as standard, and trombone players have a lower threshold than trumpet players for the trumpet sound as target), this is not enough to establish a pattern. Comparison of confidence intervals cannot be done so easily for within-subject measures, therefore a mixed effects linear regression model was applied to predict threshold value, where instrument played and standard, or target timbre were fixed effects and participant number was a random effect on intercepts. The instrument played had no effect on perceptual threshold, $\chi^2(3) = 3.83, p = .28$ and $\chi^2(3) = 3.82, p = .28$ for standard and target models respectively.

Effects of instrumental family were also investigated. Performance by instrumental group was analysed for string pair and brass pair trials (i.e. where the standard and target

timbres were both string or both brass instruments). String players performed with a mean percentage of 54.5 [46.0, 63.0] on string pairs and 62.4 [52.2, 72.6] on brass pairs. Brass performed with a mean percentage of 56.0 [46.7, 65.3] on string pairs and 58.4 [48.2, 78.6] on brass pairs (see Figure 5.3). Interpreting the CIs indicates that there was no difference between string players and brass players; however, a mixed effects model to investigate within group differences found that instrument played had an effect on threshold, $\chi^2(1) = 3.54$, $p = .05$, where string players had a lower discrimination threshold for string instruments than for brass instruments.

Trials where participants did not hear a change in rhythm were examined separately. Most participants only had a few trials where this happened, if at all. As noted above, for five participants, this case was more prominent and their data were removed (it is interesting to note that the mean age for these five participants is 51.8 (SD = 12.7) and every participant was at or

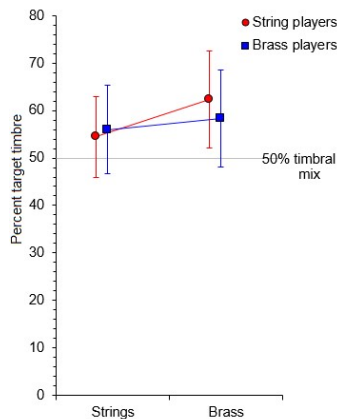


Figure 5.3. Percent target timbre of the morphed stream at the point of a change in percept for brass and string instrumental family groups. Error bars represent 95% CIs.

above the average age for all participants). Every type of instrumentalist was represented in this group of trials; all for the ascending block and all but cellists for the descending blocks. The frequency of each of the standard and target timbres was different within each direction by timbre type condition (i.e. the number of times a trial had cello as the standard or target timbre, versus the other instruments, in the ascending or descending block), but no single timbre was consistently more or less represented. When

looking at pairs of timbres, the cello-trumpet and trombone-trumpet pairs were mostly commonly still perceived as an even percept by the end of a trial in the descending condition and the violin-trombone pair was the most commonly still perceived as a galloping percept by the end of a trial in the ascending condition.

5.2.6 Discussion

This study was designed to corroborate neuroscientific measures showing that instrumentalists are more sensitive to their own instrument's timbre than to others (Pantev et al., 2001). Accordingly, in the ABA_ paradigm, it was hypothesised that a lower timbre discrimination threshold would be found for instrumentalists hearing their own instrument when their instrument is the target timbre, and a higher discrimination threshold when their instrument is the standard timbre.

Results show no reliable effect of instrument played on the perception of timbral stream segregation when looking at individual target instruments. Though thresholds for an instrumentalists' timbre were slightly lower than for other timbres when looking at standards, contrary to the hypothesis, none of these differences were significant. Similarly for target timbres, no threshold differences were significant, though the largest effect was seen in trumpet players, where the threshold when trumpet was the target was higher than for other instruments. There was a small effect of instrument played when comparing performance on instrumental families: string players detected the difference between two brass instruments later than for two string instruments. However, they were no better than brass players at detecting the difference between two string instruments, nor did brass players show an advantage for brass instruments. Thresholds for string instruments were overall lower than for brass instruments. Perhaps the two string instruments were more different than the two brass instruments, thus making them overall easier to distinguish (this is supported by timbre dissimilarity ratings

collected in the second study, reported in Section 5.3). The effect of order is unexpected and could be the result of a familiarization with the task that led to greater sensitivity in the second block.

How can such results be explained when the literature reviewed, particularly Pantev's work (2001), suggests an effect of instrumental training on perception? Let the question first be placed in a more generalized context. Imagining a trained musician listening to an orchestral work, it can be assumed that they clearly hear the melody. What if they were asked to listen to the bass line? Or another instrument? If instrumentalists are more sensitive to their own instrument's timbre, then it would be expected that they could more easily and more accurately pick out (and perhaps transcribe, for potential experimental purposes) their own instrument than any other. However, according to the present results, they could also pick any instrument out of the auditory scene and transcribe it just as well. This would suggest that ability to pick out and transcribe a particular line in a polyphonic work is not related to the instrument one plays, but rather to general musical training, and to where attention is directed. It would be interesting to conduct a transcription experiment along these lines in the future. However, a reasonable explanation of the present results is that listeners simply heard what they paid attention to, though it is only a proposition here and cannot be supported or countered with the current data. The possibility of attention directing perception will be further explored in Sections 5.3 and 5.4.

One of the basic claims of auditory streaming is that coherence is the default percept (Bregman, 1978; Bregman, 1990; Rogers & Bregman, 1998). However, if this were the case, then initial segregation in the descending condition of this experiment would not be possible. The fact that participants were told what they would be hearing (even to galloping for descending blocks and galloping to even for ascending blocks) could have influenced their

perception of the stimuli by setting up a specific expectation. Therefore, a study to control for this was designed and is reported next.

5.3 Study 2: Expectancy control

This study was designed to control for the possible expectation effect of the instructions given in the study presented in Section 5.2 above. Participants were presented with 10s of ABA_ pattern where the timbres are unchanging and maximally different (the same as the beginning of a descending block trial in Study 1) and were asked to report whether they heard an even or a galloping pattern. If participants tend to hear these stimuli as even, then there is cause to revisit the default coherence concept; alternatively, if participants tended to hear the stimuli as galloping, then the instructions given in Study 1 likely set up an expectation which strongly influenced perception, enough to hear an even pattern at first hearing. Participants were also asked to indicate which of the two timbres was most salient. If the standard timbre (the faster stream) is chosen most often then timing tends to attract attention more than timbre; if the standard and target timbres are chosen approximately equally often, then it is the timbre itself that is most salient in capturing focus.

5.3.1 Participants

Data was collected in two groups: first, undergraduate and graduate musicians and, second, individuals with various backgrounds recruited from universities in London and the community. The first group of participants were the same 20 participants as in Study 1 (the same five participants' data was excluded here); they completed both paradigms. The second group was tested separately and included a wider range of backgrounds to control for effects of musical training in the first group. This second group consisted of 20 individuals (7 males, mean age 22.5 years; SD = 4.33; range = 18-32; mean Gold-MSI score = 23.3, SD = 11.9,

range = 7-46) recruited through volunteer email lists, credit scheme and acquaintances. Participants in the first group were entered in a draw for an Amazon voucher while participants in the second group were either entered in a draw for an Amazon voucher or given course credit as part of a university credit scheme.

5.3.2 Stimuli

The stimuli were the same as Study 1, except that there were seven timbres (piano, violin, cello, trumpet, trombone, clarinet, bassoon) and there was no morphing. One timbre was presented at 220ms and the other at 440s with a 110ms offset and the total length of one trial was 10s.

5.3.3 Procedure

This paradigm was also presented in Max/MSP 6. After reading the information sheet and giving written consent, instructions were presented on the screen along with examples of the even and galloping patterns, each accompanied by an illustration to help clearly distinguish the two rhythms. Five practice trials were provided and were compulsory, giving a chance for questions and clarification before beginning the data collection.

When ready to begin, for each trial participants indicated as they were listening which percept they heard first using the keyboard, pressing 'H' (horse) for the galloping pattern and 'M' (morse) for the even pattern (terminology from Thompson, Carlyon, & Cusack, 2011). At the end of the trial, they clicked on the timbre that was most salient to them (the appropriate two were displayed at each trial). Every possible timbre pair was explored, for a total of 21 trials.

Participants then completed the musical training sub-scale of the Gold-MSI (Müllensiefen et al., 2014).

Timbre dissimilarity ratings. Timbre dissimilarity ratings were collected separately using Max/MSP 6. 15 listeners of varying backgrounds, none of which participated in the reported studies, rated the similarity of pairs of timbres on a 7-point Likert scale where 1 was the least similar timbre pair and 7 was the most similar timbre pair, with other pairs rated between these numbers. The participants could listen to the seven musical tones at any time. These were the same as in Sauvé et al. (2014) (piano, violin, cello, trumpet, trombone, clarinet, bassoon). Participants clicked a button to begin a trial: two timbres were presented for comparison and participants rated the similarity between the sounds. There was no time limit and participants submitted each rating on their own time, completing the trial. Pairs of timbres were presented randomly. Results are shown in Figure 5.4.

5.3.4 Results

A comparison of the two groups revealed no significant difference between the initial percept for musicians and for non-musicians, where the difference in proportions was .03 [- .04, .10]. Therefore the remaining analysis was performed on the two participant groups' aggregated data.

The mean of the initial percept, where even was coded as 0 and galloping was coded as 1, was .35 [.32, .39]. Interpreting the CIs in Figure 5.5 indicates that this is significantly different from chance (.5). Because the mean of the initial percept is closer to zero than it is to one, the initial percept is dominantly even. To investigate the relative salience of timbre and timing, a 'matching' variable was created, where if the timbre identified as salient matched the standard timbre, a value of 1 was assigned and if it did not, a value of 0 was assigned. The mean of the matching variable was .69 [.65, .72]. Once again, interpreting the CIs in Figure 5.5 indicates that this is significantly different from chance (.5), confirmed by an exact binomial test, $p < .01$. Therefore, the most salient timbre is most often the standard timbre, which is also

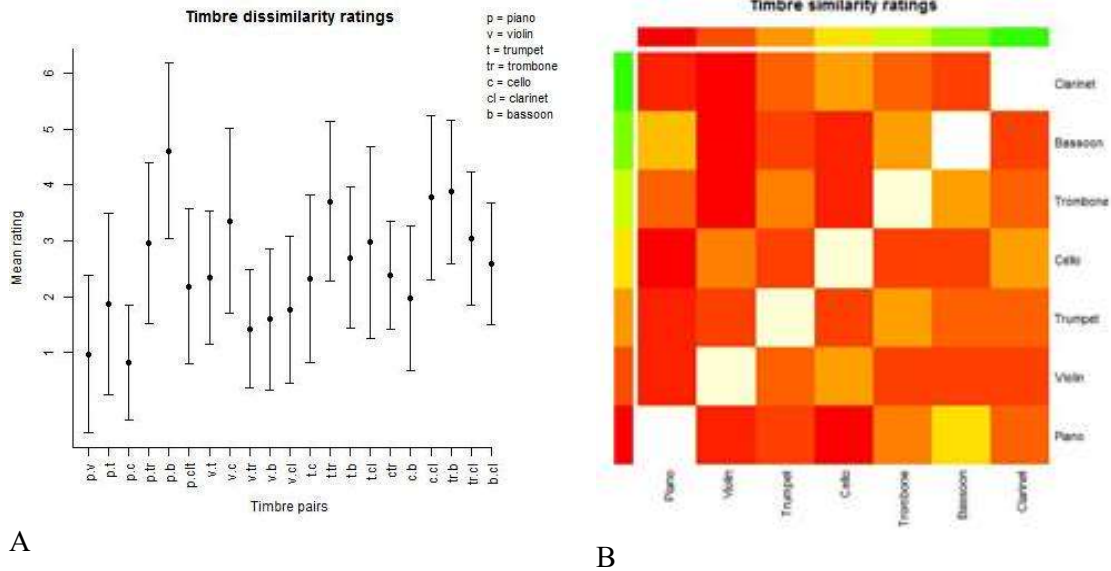


Figure 5.4. A) Timbre dissimilarity ratings (1-7 Likert scale; 1 is very dissimilar, 7 is very similar). When the initial percept is even, timbres are less similar (2.30 [2.22, 2.38]) and when the initial percept is galloping, timbres are more similar (2.90 [2.80, 3.00]). B) Timbre dissimilarity ratings presented in a heat map, where red is most dissimilar and green is most similar.

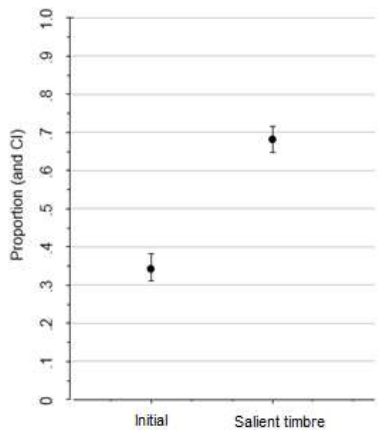


Figure 5.5. Mean of initial (left) and matching (right) variables, both significantly different from chance.

the fastest presented stream. The influence of timbre on initial percept was investigated using timbral dissimilarity ratings to assess whether more similar timbral pairs would encourage integration while less similar pairs would encourage segregation. This pattern was indeed observed in the data. The average dissimilarity rating over all trials where segregation was the initial percept was lower, 2.30 [2.22, 2.38], than when integration was the initial percept, 2.90 [2.80, 3.00], confirmed by a t-test, $t(499) = -8.11, p < .01$.

5.3.5 Discussion

This study was designed to investigate whether the instructions in Study 1 enabled the possibility of initial segregation in the descending blocks by setting up the expectation for segregation, as according to streaming theory, integration is always the default percept until enough evidence is gathered for the existence of two separate streams (Bregman, 1990).

Results indicate that the even percept is the most common initial percept, which is contrary to the streaming theory discussed above. However, this experiment does not rule out the possibility that the build-up of evidence for two streams simply happened very quickly. A reliable neural streaming marker is needed to investigate this question at the millisecond level. While some such markers have been suggested (Alain et al., 2001; Fujioka, Trainor, & Ross, 2008; Sussman, Ritter, & Vaughan, 1999), none of them constitute direct measures of streaming and further research is needed to identify such markers.

Furthermore, the initial percept was depending on similarity between pairs of timbres. Presumably, it takes longer for the brain to find evidence for two streams if the sources are more similar. A similar pattern for pitch was found by Deike et al. (2012), where participants were presented with ABAB sequences and asked to indicate as quickly as possible whether they heard one or two streams. The separation between A and B tones varied from 2 to 14 semitones. Results showed that the larger the pitch separation between A and B tones, the more likely participants were to hear the sequence as segregated in the first place. Predictability was also found to influence degree of segregation (Bendixen et al., 2014): when degree of predictability between two interleaved sequences was high, an integrated percept was supported, while when the predictability within each interleaved sequence alone was high, a predominantly segregated percept was induced. This is contrary to the integration-by-default concept proposed by Bregman (1990). However, auditory scene analysis is complex and the

role of context has yet to be considered, where previous work has shown that context can speed or slow the buildup of evidence for perceptual segregation (Sussman-Fort & Sussman, 2014).

Attentional mechanisms were probed in this study by asking participants which timbre was most salient. Results show that the standard timbre was most often the most salient timbre. In feedback, some participants described it as more driving and therefore more attention-drawing. This suggests that rhythm is a more salient feature than timbre, adding interesting evidence to discussions about the relative salience of different features in the perception of polyphonic music (Chapter 6; Esber & Haselgrove, 2011; Prince, Thompson, & Schmuckler, 2009; Uhlig et al., 2013).

In terms of the influence of instructions in Study 1, it seems that they did influence participants' perception; otherwise, initial segregation on trials with similar pairs of timbres would not be possible. It is already known that attention influences perception in this paradigm (Sauvé et al., 2014), and this experiment suggests that prior expectation about the number of streams also has an impact.

5.4 Study 3: Ecological extension

Study 1 aimed to behaviourally test the hypothesis that instrumentalists are more sensitive to their own instrument's timbre than to others. Study 2 was designed to control for the effect of expectation. However, neither of these paradigms are particularly ecologically valid; the ABA_ pattern is especially synthetic and though the sounds are recorded and not synthesized, the way they are combined is not reminiscent of actual music. This study was designed with the same goal as Study 1 and to allow results to be extended towards more ecological musical listening. The interleaved melody paradigm introduced by Dowling (1973) was selected to achieve this goal. In this paradigm, the notes of two melodies are presented in an alternation, such that melody 'ABCDEF' and melody 'abcdef' become 'AaBbCcDdEeFf'.

Dowling found that as pitch overlap decreased, participants were more easily able to detect, or segregate, each individual melody. Trained musicians could tolerate more pitch overlap than non-musicians. The concept can similarly be applied with many other parameters including loudness and timbre (Hartmann & Johnson, 1991), where it is easier to track a melody if the two interleaved melodies are of different loudness, or played by different instruments.

Here, this paradigm was adapted to timbre perception and the task was to detect one or multiple mistunings, as intonation is a developed skill in many instrumentalists. As in the original paradigm, it is expected that the segregation of the two melodies will be easier with minimal pitch overlap. Additionally, it is hypothesized that instrumentalists should identify mistunings more accurately for their own instrument overall, where the melodies are played by different instruments.

5.4.1 Participants

Participants were 15 musicians, 8 flautists and 7 violinists, recruited from music schools and conservatoires in London and in Canada. If desired, they were entered in a draw for one of two Amazon vouchers.

5.4.2 Stimuli

Melodies were two excerpts from compositions by J. S. Bach: BWV 772-786 Invention 1, mm13 and BWV 772-786, Invention 9, mm14-15.1 (only the first beat of mm15). They are in different meters (4/4 and 3/4 respectively) and different keys (A minor and F minor respectively), but have similar ranges (perfect 12th - octave + perfect 5th - and diminished 12th - octave + diminished 5th - respectively) and similar median pitches (C#4 and B4 respectively). The 4/4 melody was played on a violin and the 3/4 melody on a flute.

A violinist and a flautist were recorded using a Shure SM57 microphone, recorded into Logic Pro 10 and exported as CD quality audio files. These original recordings were verified

by a separate violinist and flautist for good tuning and corrections to tuning were made using Melodyne Editor by Celemony. Melodies were recorded at notated pitch and for every necessary transposition to create each overlap condition, as tuning in a solo instrument changes slightly as a function of key, especially in Baroque music (just intonation).

Using Melodyne Editor, 50 cent sharp mistunings were inserted. Each trial contained either zero, one or two mistunings that could be in either one or both the melodies. The location of each mistuning is presented in Table 5.2. Though it is recognized that sharp or flat tuning may be perceived differently and is dependent on context (Fujioka et al., 2005), only one direction was used here for simplicity. The tempo and note length of the melodies were quantized, and the melodies interleaved, so that the onset of the first note of the second melody fell exactly between the onsets of the first and second notes of the first melody, the second

Table 5.2. Experimental design: details of metrical and instrument location of mistunings (where there are two mistunings, these are separated by a backslash), the higher melody, where attention was directed and the amount of pitch overlap between the mean pitch of the two melodies for each trial, including practice and control trials.

Trial	Location (metrical)	Location (instrument)	Highest melody	Attentional focus	Pitch overlap
Practice 1	4.1	Violin	Violin	Violin	5 th
Practice 2	1.3 / 2.4	Flute / Flute	Flute	Flute	5 th
1	1.3 / 3.3	Flute / Violin	Flute	Flute	2 nd
2	2.4 / 4.2	Violin / Violin	Flute	Violin	2 nd
3	None	None	Violin	Flute	2 nd
4	3.1	Violin	Violin	Flute	2 nd
5	2.2	Flute	Flute	Flute	3 rd
6	1.4 / 2.1	Flute / Violin	Flute	Both	3 rd
7	1.2 / 3.4	Violin / Flute	Violin	Flute	3 rd
8	3.2 / 4.3	Flute / Flute	Violin	Violin	3 rd
9	None	None	Flute	Violin	5 th
10	2.4 / 4.1	Flute / Flute	Flute	Flute	5 th
11	2.3 / 4.1	Violin / Violin	Violin	Violin	5 th
12	3.2 / 4.2	Violin / Flute	Violin	Both	5 th
Control 1	2.3	Violin	-	-	-
Control 2	3.1	Flute	-	-	-

between the second and the third, and so on.

Twelve experimental trials were created, along with two practice trials and two control trials. Five variables were manipulated: *metrical mistuning location*, *instrumental mistuning location*, *highest melody*, *attentional focus* and *pitch overlap*. The mistunings were located either on strong or weak beats, where location is indicated by beat (first number) and subdivision (second number) i.e. 4.2 = beat 4, second subdivision (where each beat is divided into four sixteenth notes). The mistunings were either in the violin or the flute melody, the highest (also the first tone heard) melody was either the violin or the flute melody and the participants' focus was directed at either the violin melody, the flute melody, or both. Pitch overlap was either a 2nd, a 3rd or a 5th, where the distance between the central (in terms of range) pitches of each melody matched these intervals. The instrumental mistuning location, highest melody and attentional focus were manipulated so that they sometimes match and sometimes do not (i.e. the mistuning may not be in the same melody to which the participant is asked to attend). This was intended to assess whether a mistuning in the non-attended melody influences identification of mistunings in the attended melody.

The control trials were single melodies, designed to ensure that participants were able to detect mistunings in a simpler listening situation. In a pilot study, a 50 cent mistuning in a single melody was always detected.

5.4.3 Procedure

This experiment was carried out online, using the survey tool Qualtrics. Once presented with the information sheet and detailed instructions, participants could give informed consent. The two original melodies (with no mistunings) were both presented for participants via SoundCloud to familiarize themselves with the tunes, and in every subsequent trial in case participants wanted to refresh their memory. Each page of the survey contained the two

original melodies, the current trial (also via SoundCloud) and a click track. Participants clicked on the beats where they heard a mistuning; this was set up using Qualtrics' hot spot tool. There was one click track for trials where focus was on one instrument and two, stacked vertically and labelled with the corresponding instrument, when participants were instructed to listen to both (see Figure 5.6). The word 'none' under the click track was also a selection option if participants detected no mistuning.

Participants started with two practice trials, always in the same order. Then, trial 11 was always presented first because it was one of the trials with the least amount of overlap (and, therefore, presumably easier) and all other trials followed in random presentation. Finally, the two control trials were presented, always in the same order. Participants finally selected their primary instrument, either violin or flute, and had the option to submit their email address for the Amazon voucher draw.

5.4.4 Results

Initial inspection of the data showed a high rate of false alarms. Participants were first screened by performance on the control trials; only participants who had correctly identified the mistunings in both control trials, without false alarms, were included in analysis. This left 12 participants; 6 violinists and 6 flautists. A mixed effects binomial logistic regression was performed, with *musicianship* (violinist or flautist), *metrical mistuning location*, *instrumental*

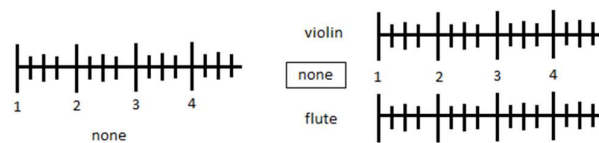


Figure 5.6. Single (left) and double (right) click tracks presented to participants alongside the relevant audio files.

Table 5.3. Details of the mixed effects binomial logistic regression, where accuracy is predicted by fixed effects as described in the text and participant number as random effects on intercepts.

Predictor	Estimate	p-value
Intercept	-2.43	< .01
Musicianship	0.23	.32
Metrical mistuning location	0.30	.07
Instrumental mistuning location	0.47	.03
Highest melody	0.87	< .01
Attentional focus	-0.94	< .01
Pitch Overlap	0.04	.61
Random Intercepts		Variance
Participant	0.02	

mistuning location, *highest melody*, *attentional focus* and *pitch overlap* as fixed effect predictors for accuracy (1 for correctly identified mistuning and 0 for an unidentified mistuning), and participant number as a random intercepts effect. Where there were two mistunings in a trial, each mistuning was considered as its own data point.

Only attentional focus and highest melody were strong significant predictors, $z(1) = -4.85$ and $z(1) = 3.65$ respectively, both $p < 0.01$ while instrumental mistuning location was moderately significant, $z(1) = 2.15$, $p = .03$. There were no significant interactions (see Table 5.3 for details).

When averaged across relevant trials, accuracy when attention was directed to the violin line was highest, at .28 [.22, .35], to the flute line was .25 [.20, .32] and to both was lowest, at .10 [.06, .17], where accuracy of 1.00 would reflect the correct identification of all mistunings. Accuracy when the violin line was on top was lower than when the flute was on top, at .17 [.13,

.22] and .29 [.24, .34] respectively. Accuracy when the mistuning was in the violin line was .38 [.31, .46] and in the flute line was .20 [.16, .25].

5.4.5 Discussion

In this study, the interleaved melody paradigm was used to examine whether musical training on a particular instrument increases timbral sensitivity to that instrument, using mistuning detection in real melodies rather than rhythm judgements for artificial tone sequences, as in Study 1. Contrary to the hypothesis, results converge with Studies 1 and 2 above: musical training does not have an influence on timbre sensitivity, and support the alternate hypothesis proposed in Study 1: attention influences perception. Similarly to the hypothetical orchestral line transcription described before, in this paradigm detection of mistunings, which first requires the separation of the melody from its context, did not depend on the instrument in which the mistuning appeared, but rather which line the listener's attention was directed to. The idea that attention influences perception is certainly not new (Carlyon et al., 2001; W. Jay Dowling, 1990; Snyder, Gregg, Weintraub, & Alain, 2012; Spielmann, Schröger, Kotz, & Bendixen, 2014) and the above results suggest that attentional focus is more important than specific musical training in driving auditory stream segregation, leading to the lack of effect of specific instrumental musical training.

5.5 General Discussion

Though previous literature would suggest that instrumentalists are more sensitive to their own instrument's timbre (Margulis et al., 2009; Pantev et al., 2001), behavioural evidence for this claim was not found in the studies reported in this chapter. Instead, results suggest that behaviour is more strongly guided by attention than musical training, consistent with literature exploring the effects of attention on auditory scene analysis (Andrews & Dowling, 1991;

Bigand et al., 2000; Jones, Alford, Bridges, Tremblay, & Macken, 1999; Macken et al., 2003). This interpretation was supported in both Studies 2 and 3. In Study 2, trials where rhythm captured attention more often also resulted in a segregated percept. In Study 3, attentional focus was a predictor of response accuracy for identifying mistunings, where accuracy depends on participants successfully streaming the relevant melody. Furthermore, performance when participants were asked to identify mistunings in both lines at once was particularly poor, highlighting the importance of attentional focus for successful task completion and the difficulties of divided attention (Bigand et al., 2000).

It is interesting to consider why the present results diverge from those found in cognitive-neuroscientific studies which have found instrument-specific effects of musical training. It may be that methods such as EEG, MEG and fMRI provide more sensitive measures that are capable of picking up on small effects of instrument-specific training which are not expressed in behavioural measures such as those used here. Greater sensitivity of neural over behavioural measures has been observed in research on processing dissonant and mistuned chords (Brattico et al., 2009) and harmonic intervals varying in dissonance (Schöön, Regnault, Ystad, & Besson, 2005). Alternatively, it may be that the instrument-specific effects observed in previous research were actually driven by greater attention to an instrumentalists' own instrument. Further research is required to disentangle these alternative accounts.

Turning to the ABA- paradigm for a closer look, the results of Studies 1 and 2 highlight some interesting observations. Despite listeners most often initially perceiving maximally different timbres as segregated, there were still a fairly large proportion of trials heard as integrated. This was explained above by timbre similarity, but it may not be the only factor: based on personal listening, and participant feedback, the stimuli are clearly bistable, suggesting that timbre alone may not be enough to fully segregate two sounds played with the

same pitch, loudness and length. In a musical sense, this is very useful and is often employed by composers wanting to create instrumental chimerae or even simply writing passages involving the entire sections of the orchestra playing the same line. This suggests that timbre is a less important feature in this simple listening context, with pitch, rhythm and loudness taking precedence. Relative importance of these four parameters for auditory streaming could be evaluated by combining parameters to see which causes streaming first. Some questions concerning salience and combining musical parameters in a streaming paradigm have been investigated (Dibben, 1999; Prince et al., 2009; van Noorden, 1975) but a clear map of relationships between parameters has not yet been established, largely due to the complexity of polyphonic music. Chapter 7 of this thesis uses predictive processes to begin investigating this question of relative salience in the context of polyphonic music, a more complex context than the ABA_ paradigm used here.

To summarize, this chapter confirmed timbre as an auditory streaming cue, notably using ecological stimuli. Together, the three presented studies also addressed the influence of attention, expectation and listening background on streaming perception. While attention and expectation were found to affect streaming, where attention guides perception more strongly than listening background and expectation affected listeners' initial percepts, the instrumental differences between musician participants were not found to have any effects on streaming perception. Furthermore, results suggest that in this simple context, timbre is less salient than pitch, rhythm and loudness. The relationship between different musical parameters and their relative salience will be investigated in more detail in Chapter 7; first, Chapter 6 validates one of the measures necessary for this investigation.

6 Modelling temporal expectations alongside pitch expectations

An understanding of pitch perception through the lens of predictability has been well covered in the research community (from Meyer, 1956 to Eerola, 2016), including behavioural validation of various proposed models of pitch expectation specifically (Carlsen, 1981; Huron, 2006; Krumhansl, Louhivuori, Toiviainen, Järvinen, & Eerola, 1999; Marcus T. Pearce, Ruiz, et al., 2010; Schellenberg, 1996, 1997; Trainor & Trehub, 1992). In the time domain, focus has tended to be on the higher-level temporal structure called meter, the hierarchical organization of beats in music (Desain & Honing, 1999; Dixon, 2001; Temperley, 2010; Volk, 2008). In this chapter, predictability will be used to model low-level pitch and rhythm perception and musical emotion, presenting a study validating two of IDyOM's time-domain viewpoints – onset and inter-onset-interval (IOI) – while jointly investigating the potential role

of predictability and the influence of musical training in eliciting musical emotion (Huron, 2006; Juslin, Liljeström, Västfjäll, & Lundqvist, 2011; Juslin & Västfjäll, 2008; L. Meyer, 1956).

6.1 Rhythm and Predictability

Metrical patterns, similarly to pitch, can be learned through exposure (Cirelli, Spinelli, Nozaradan, & Trainor, 2016; Hannon, Soley, & Ullal, 2012; Hannon & Trehub, 2005a; Hannon & Trehub, 2005), where Western music is dominated by beat patterns in divisions of two, and to a lesser extent, divisions of three. As a result of this dominance, Western listeners develop models of musical rhythm biased towards beat divisions of two and three and are surprised by violations of expectations generated by these models. It is important to keep in mind that as these models and biases are built up through learning, different cultures may exhibit different biases and therefore experience different patterns as surprising. For this reason, focus will remain on Western music and Western listeners only. Proposed computational implementations for such human rhythmic models of music have been suggested using predictive coding (Vuust et al., 2009; Vuust & Witek, 2014) and statistical learning (Pearce, 2005).

Predictive coding, introduced in Section 2.6, has recently been applied to explain music perception as a whole (Gebauer, Kringelbach, & Vuust, 2012; Vuust et al., 2009). For example, Gebauer et al. (2012) present a framework in which dopamine is released during anticipation and an increase or decrease (depending on prediction error) in its release during the evaluation of a given prediction modulates pleasure. This framework explains pleasure in music on a Wundt inverted U-curve (Wundt, 1874), with prediction error magnitude on the x-axis and pleasure on the y-axis. In cases with very low or very high prediction error, music is unpleasant because it is either uninteresting or, conversely, too complex for the brain to predict. In the

range between these two extremes lies musical pleasure, explaining why some expectancy violations can be pleasurable (i.e., musical chills). Predictive coding has also been more specifically applied to rhythm perception as a framework that can explain brain responses to metrical deviation (Vuust et al., 2009) as well as the special cases of syncopation, polyrhythms and groove (Vuust & Witek, 2014). In the case of brain responses, the MMNm and P3am (the magnetic equivalents of the MMN and P3a, classically markers of pattern violation and expectancy respectively (Jongsma, Desain, & Honing, 2004; Näätänen, Gaillard, & Mäntysalo, 1978)) are interpreted as the encoding of prediction error, and subsequent integration of that error higher in the processing hierarchy, respectively. This is supported by a study where jazz musicians and non-musicians listened to drum sequences with three levels of metrical violation: none, syncopation, and missing beat (Vuust et al., 2009). An MMNm was consistently elicited for both musicians and non-musicians, though stronger in musicians, and was consistently localized near the auditory cortex. The P3am on the other hand was elicited more often in musicians and was not consistently localized, spanning the parietal and frontal cortices depending on the participant. This suggests that the P3am is not specific to the auditory cortex and supports its role as integrator of the signal. Syncopation is explained by the predictive coding framework as presenting a tolerable amount of prediction error to the brain, where the metrical violation is perceived and error is generated but the metrical model is maintained. Polyrhythms, the simultaneous occurrence of two different meters (3 against 2 for example) is presented as an example of auditory bi-stability, directly comparable to visual bi-stability (Hohwy, Roepstorff, & Friston, 2008). In the auditory version, since divisions of two or three do not occur simultaneously (similarly to a house and a face not occurring on the same scale at the same time) in the vast majority of Western music, listeners only have models developed for one or the other; while one is being perfectly modelled, the other is receiving

error signals and the listener oscillates between perceiving 3 against 2 and 2 against 3, for example. Finally, groove is a case of continuous syncopation, where the constant pull between the meter and the rhythm falls in the peak of the inverted U, causing a pleasurable reaction.

Statistical learning has much in common with predictive coding: a model of the given environment (here musical features) is built based on exposure, where commonly occurring events or patterns have a high probability of re-occurring and rare events or patterns have a low probability of re-occurring. Those events and patterns that are highly probable are unsurprising and have low information content, while the improbable events and patterns are surprising and have high information content. As described in Section 3.2, IDyOM implements such a model for music over various musical parameters called viewpoints. Use of IDyOM so far has almost exclusively been restricted to the pitch domain (Egermann, Pearce, Wiggins, & McAdams, 2013; Gingras et al., 2016; Hansen & Pearce, 2014; Pearce, Müllensiefen, & Wiggins, 2010; Pearce, Ruiz, Kapasi, Wiggins, & Bhattacharya, 2010), with only a recent expansion into the time domain (Pearce & Müllensiefen, 2017; van der Weij, Pearce, & Honing, 2017), despite the wide range of viewpoints implemented. In this chapter both pitch and timing viewpoints will be used, extending the use of IDyOM and behaviourally validating a subset of time-based viewpoints with respect to perceived expectation.

6.2 Emotion and Predictability

As it was alluded to in Section 6.1, it has been suggested that musical emotion is induced by the fulfilment and violation of expectations (Gebauer et al., 2012; Huron, 2006; Meyer, 1956). While it is not the only possible mechanism for emotional induction (Juslin et al., 2011; Juslin & Västfjäll, 2008), Gebauer et al. (2012) suggest it is the most fundamental mechanism as expectation is a direct internal connection between music and existing psychological mechanisms, while other suggested mechanisms such as evaluative

conditioning, emotional contagion, visual imagery, episodic memory and cognitive appraisal rely on extramusical associations and can be seen as additional sources of emotional meaning on top of expectation (Juslin et al., 2011; Juslin & Västfjäll, 2008). These additional mechanisms are defined as follows: (1) brain stem reflexes refer to changes in arousal caused by sudden psychoacoustic signals (i.e., loudness, dissonance); (2) evaluative conditioning creates a positive or negative emotional reaction when a piece has been repetitively paired with a positive, or negative, situation; (3) emotional contagion is the induction of emotion through mimicry of behavioural or vocal expression of emotion, and is reflected in musical structure; for example shorter durations and ascending pitch contours tend to reflect happiness while longer durations and descending pitch contours communicate sadness; (4) visual imagery refers to the mental imagery evoked by the music, which can have positive or negative affect; (5) episodic memory refers to the pairing between a sound and a past event, triggering the emotion related to that event when the sound is subsequently heard; (6) rhythmic entrainment refers to the induction of emotion through the proprioceptive feedback of internal body entrainment (i.e., heart rate) to the music and; (8) cognitive appraisal refers to the evaluation of music in the context of goals or plans of the listener.

At its simplest, the expectation mechanism of musical emotion proposes that unexpected events are surprising and associated with an increase in tension while expected events are associated with resolution of tension (e.g. Gingras et al., 2016). According to this account, surprising events generally evoke high arousal and low valence (Egermann et al., 2013; Koelsch, Fritz, & Schlaug, 2008; Steinbeis, Koelsch, & Sloboda, 2006). However, there are two cases where these reactions may differ: first, listeners familiar with a piece of music can come to appreciate an event that has low expectancy through an appraisal mechanism, resulting in a high valence response (Huron, 2006); second, as mentioned in Section 6.1, a

certain amount of surprisal (prediction error) can be pleasing and lack thereof uninteresting, leading to high arousal and high valence for moderate levels of surprisal and low arousal and high valence for low levels of surprisal. To avoid the first possibility, only unfamiliar music will be used. The second possibility offers an alternative hypothesis to the simplest interpretation of the expectation mechanism, to be tested in the study described in sections 6.3-6.6. It is worth remembering that there are different sources of influence on musical expectation (Huron, 2006): schematic, veridical and dynamic musical expectations were introduced in Chapter 2 (Section 2.4). Both schematic and dynamic expectations can be simulated as a process of statistical learning and probabilistic generation of expectations in IDyOM (Pearce, 2005) via its LTM and STM models. Furthermore, these may be different for musicians and non-musicians due to extensive exposure and training in a particular style (Juslin & Vastfjall, 2008).

Musical expectancy as a mechanism for the induction of emotion in listeners has previously been studied in an ecological setting: Egermann et al. (2013) asked 50 participants to attend a live concert, during which 6 flute pieces were played. These pieces spanned various musical styles and levels of pitch expectancy. Three types of measure were taken: subjective responses (i.e. arousal levels or ratings of musical expectancy, both of which changed continuously throughout the piece), expressive responses (using video and facial EMG) and peripheral arousal measured by skin conductance, heart rate, respiration and blood volume pulse. IDyOM (Pearce, 2005) was used to analyse pitch patterns of the music and predict where listeners would experience and report low expectancy. Results suggested that expectancy had a modest influence on emotional responses, where high IC segments led to higher arousal and lower valence ratings as well as increases in skin conductance and decreases in heart rate as compared to low IC segments while no event-related changes were found in respiration rate or

facial electromyography (EMG) measures; however, this study was conducted in an ecologically valid, thus non-controlled environment where participants could have focused on something other than the music. For example, visual aspects of performance are highly important to emotional engagement in live music settings (Thompson, Graham, & Russo, 2005; Vines, Krumhansl, Wanderley, & Levitin, 2006). Furthermore, other potential emotion inducing mechanisms, as proposed by Juslin & Vastfjall (2008) were not explicitly controlled for and effects of temporal expectancy on emotional responses were not considered.

6.2.1 Current study

Since pitch and temporal structures generate distinct expectancies, the influence of each as a potential emotional inducer is explored using both correlational and causal methods while also allowing for the possibility of interactions between pitch and timing. The current study is designed to validate pitch and temporal musical expectancy as predicted by IDyOM in a restricted experimental environment, and in the context of musical emotion, controlling for other potential emotional mechanisms (Juslin et al., 2011). To validate IDyOM's implementation of temporal musical expectancy, real-time expectancy ratings are collected. To investigate expectancy as a mechanism of emotional induction, the other mechanisms are controlled as follows: (1) brain stem reflexes are controlled for by maintaining equal tempo, intensity and timbre across all musical excerpts; (2) evaluative conditioning and episodic memory are controlled for by presenting unfamiliar musical excerpts, so that expectation ratings and emotional reactions are not confounded by previous experience with the music; (3) potential effects of emotional contagion are controlled for in the analysis by including pitch and inter-onset-interval (IOI) as predictors of subjective ratings in addition to pitch and IOI predictability (i.e. higher mean pitch and shorter IOI could result in higher arousal and valence ratings, regardless of expectancy); (4) episodic memory is avoided by using unfamiliar music;

(5) the absence of a strong, driving beat and the relatively short duration of the musical excerpts makes deep, emotion-inducing rhythmic entrainment highly unlikely; (6) all participants are listening to these musical excerpts in the context of an experiment, with any other goal or motive being highly unlikely, thus minimising the relevance of the cognitive appraisal mechanism; and (7) irrelevant visual imagery cannot be categorically avoided but the rating tasks are expected to require enough cognitive load to render it unlikely. Furthermore, to the extent that visual imagery is variable between individuals, averaging across participants should remove its influence.

This research aims to address three questions. First, do the predictability of pitch and timing (as simulated by IDyOM) have an effect on listeners' expectations and emotional state, and can explicit manipulation of the stimuli causally influence this effect? It is hypothesized that the degree of musical expectancy for pitch (based on pitch interval and scale degree) and temporal (based on IOI) structures, as predicted objectively by information content provided by IDyOM, will predict listeners' expectation ratings. and have an effect on emotion as measured by the arousal-valence model (Russell, 2003). According to Russell (2003), unexpected events will invoke negative valence and cause an increase in arousal and expected events will invoke positive valence and decreased arousal. Appraisal was not expected to affect this initial reaction as ratings were collected in real time. It is also hypothesized that when both pitch and timing are either expected or unexpected, the emotional response will be more extreme than in conditions of mixed expectedness. Furthermore, direct manipulation of pitch expectancy while keeping temporal expectancy and all other musical features constant is expected to produce the predicted changes in ratings (i.e. transforming unexpected pitches to expected pitches will decrease unexpectedness and arousal, and increase valence ratings).

Second, how do pitch and timing predictability combine to influence expectation and emotion? Though the combination of pitch and timing in music perception has been a research interest for decades (Boltz, 1999; Duane, 2013; Jones, Boltz, & Kidd, 1982; Palmer & Krumhansl, 1987; Prince et al., 2009), no clear conclusions can be drawn as findings regarding this question have low agreement and seem highly dependent on the choice of stimuli, participants and paradigm. For example, while Prince et al. (2009) suggest that pitch is more salient, results from Duane's (2013) corpus analysis suggest that timing is the most reliable predictor of streaming. While the former study uses monophonic melodies, it could be argued that if salience is linked to complexity (Prince et al., 2009), then for melodies where pitch or timing are highly predictable (low complexity), the predictable feature will be less salient than its unpredictable counterpart because it requires less "processing power", and therefore less attention. For melodies where pitch and timing are relatively equally predictable or unpredictable, their relative importance currently remains unknown.

Finally, is there a difference in the effect of pitch and timing predictability on expectation and emotional responses between musicians and non-musicians? The effect of musical training will be evaluated by comparing the responses of musicians and non-musicians, with the expectation that musicians will have higher expectation ratings and more extreme emotional responses to pitch and timing violations due to training (Hansen & Pearce, 2014; Strait, Kraus, Skoe, & Ashley, 2009), where Western musical patterns are more familiar, resulting in violations of these patterns eliciting stronger responses.

6.3 Method

6.3.1 Participants

Forty participants (22 female, 18 male; age range 14-54) were recruited from universities, secondary school and colleges for this experiment: 20 were musicians (mean 3.6

years of musical training, range 1 – 12 years) and 20 were non-musicians (0 years of musical training). Ethical approval was obtained from the Queen Mary Research Ethics Committee, QMREC1078.

6.3.2 Stimuli

The stimuli consisted of 32 pieces of music in MIDI format rendered to audio using a piano timbre: 16 original melodies and 16 artificially-manipulated melodies. Original melodies were divided into the following four categories of predictability: predictable pitch and predictable onset (PP), predictable pitch and unpredictable onset (PU), unpredictable pitch and predictable onset (UP) and unpredictable pitch and unpredictable onset (UU). The artificial melodies were created by changing the pitch predictability of each melody so that PP became aUP, UU became aPU, PU became aUU and UP became aPP, where *a* denotes artificial. All melodies were presented at the same intensity, which was held constant for the duration of all melodies.

Original melodies. The sixteen original melodies were selected from a group of nine datasets, totalling 1834 melodies (see Table 6.1 for details and Figure 6.1 for some examples), all from Western musical cultures to avoid potential cultural influences on expectancy ratings (Hannon & Trehub, 2005a; Palmer & Krumhansl, 1990). All nine datasets were analysed by IDyOM for target viewpoints pitch and onset with source viewpoints pitch interval and scale degree (linked), and inter-onset-interval (IOI) respectively. Both STM and LTM models were engaged; the LTM model was trained on three datasets of Western music, described in Table 6.2.



Figure 6.1. Excerpts from one melody from each of the four different predictability-based types of experimental stimuli. Patterns or notes of interest are marked with a bracket or an arrow respectively. Melody PP is predictable in both pitch and time, where an exact repetition in both dimensions can be seen, marked by a square bracket. Melody PU is predictable in pitch but unpredictable in time, where long notes in general and the rhythmic switch in the last measure specifically contribute to low predictability. Melody UP is unpredictable in pitch but predictable in time, with large leaps to and from C# (marked by arrows) and regular note durations. Melody UU is unpredictable in both pitch and time, where a leap is surprising after such repetitive unison, and the bracketed rhythmic excerpt is a hemiola (here 3 notes in the time of 2).

The 1834 melodies were divided into four categories based on high or low pitch or onset information content (IC). Melodies were considered predictable if they had a lower IC than the mean IC of all samples and unpredictable if the IC was greater than the mean IC of all samples. Four melodies from each category were selected as the most or least predictable by finding maximum and minimum IC values as appropriate for the category; these are the original sixteen melodies. See Table 6.3 for details of mean pitch and onset IC and mean MIDI pitch and IOI values for PP, PU, UP and UU melodies. Notice that categories with unpredictable onset have higher average IOI values; this potential confound is discussed below (see Table 6.4).

Table 6.1. Details of the datasets used in stimulus selection.

Dataset	Description	Number of melodies	Mean events/composition
1	Chorale soprano melodies harmonized by J.S. Bach	100	46.93
2	Alsatian folk songs from the Essen Folk Song Collection	91	49.40
3	Yugoslavian folk songs from the Essen Folk Song Collection	119	22.61
4	Swiss folk songs from the Essen Folk Song Collection	93	49.31
5	Austrian folk songs from the Essen Folk Song Collection	104	51.01
6	German folk songs from the Essen Folk Song Collection: ballad	687	40.24
7	German folk songs from the Essen Folk Song Collection: kinder	213	39.40
8	British folk song fragments used in the experiments of Schellenberg (1996)	8	18.25
9	Irish folk songs encoded by Daiman Sagrillo	62	78.5

Artificial melodies. The sixteen artificial melodies were created as follows. For each original melody, the notes with the highest (for PP and PU) or lowest (for UP and UU) information content were selected for replacement. The notes were replaced with another note from the same melody which shared the same preceding note as the original note in that melody. If several instances of such a note pair existed, the associated IC values were averaged. If several such replacement notes existed, the one with the lowest (for UP and UU) or highest (for PP and PU) information content was selected to replace the original note.

Table 6.2. Details of the training sets used to train IDyOM.

Dataset	Description	Number of melodies	Mean events/composition
1	Songs & ballads from Nova Scotia, Canada	152	56.26
2	Chorale melodies harmonized by J.S. Bach	185	49.88
3	German folk songs	566	58.46

Table 6.3. Mean pitch IC, mean onset IC, mean MIDI pitch and mean IOI (quarter note equals 24) for all melody types. Note that onset IC and mean IOI does not change between original and artificial melodies.

		PP	PU	UP	UU
Original	Pitch IC	1.37-1.85	2.22-2.43	2.83-5.24	2.61-2.78
	Onset IC	0.80-0.92	2.49-4.34	1.13-1.32	4.20-4.39
	Mean pitch	66.85-70.17	66.05-70.23	68.67-72.76	64.40-71.63
	Mean IOI	12.71-21.28	21.41-69.81	13.84-21.69	21.53-64.00
Artificial	Pitch IC	3.49-5.50	4.20-4.56	4.13-6.59	2.79-3.80
	Onset IC	0.80-0.92	2.49-4.34	1.13-1.32	4.20-4.39
	Mean pitch	64.88-69.80	67.05-73.18	64.05-67.76	66.78-72.89
	Mean IOI	12.71-21.28	21.41-69.81	13.84-21.69	21.53-64.00

Where no such replacement existed, the key of the melody was estimated using the Krumhansl-Schmuckler key-finding algorithm (Krumhansl & Schmuckler, 1986) using key profiles updated by Temperley (1999) and the replacement was selected as the scale degree with highest (for UP and UU) or lowest (for PP and PU) tonal stability. All notes labelled as having extremely high or low IC were replaced by a pitch with a less extreme IC. An example of a melody from each category can be seen in Figure 6.1. See Table 6.3 for details of mean pitch and onset IC and mean MIDI pitch and IOI values for aPP, aPU, aUP and aUU melodies. Mean onset IC and mean raw IOI values were unchanged from the corresponding original stimulus predictability category (e.g. aPP has the same mean IOI IC and IOI values as UP). Figure 6.2 illustrates the mean information content of all 32 melodies.

Table 6.4. Summary of 16 original melodies used in this experiment.

File name	Dataset of origin	Number of events	Average pitch (60 = midC)	Average note duration (24 = quarter)	Mean pitch IC	Mean onset IC	Stimulus Predictability
Kindr138	7	33	67.69	74.18	1.3624	.8962	PP
A162	8	21	70.23	27.42	1.4328	.8955	PP
Kindr151	7	51	66.05	22.82	1.5971	.8114	PP
Kindr162	7	19	68.36	26.52	1.5574	.9254	PP
Deut3480	6	19	72.89	36.94	2.4272	4.4488	PU
Jugos052	5	54	66.22	6.77	2.2543	3.7433	PU
I0511	9	53	66.83	11.67	2.0089	2.4660	PU
Deut3284	6	67	69.71	6.52	2.0913	2.5380	PU
I0533	9	39	67.76	11.79	5.6137	1.1401	UP
A120	8	35	64.05	17.31	5.2750	1.3358	UP
Oestr045	5	30	68.90	36.40	4.7200	1.1290	UP
Oestr046	5	35	64.40	32.22	4.6734	1.1983	UP
Deut3244	6	39	67.64	21.84	3.0216	4.7589	UU
Deut3206	6	52	68.67	22.15	2.9122	4.5098	UU
Deut3437	6	29	71.62	19.86	3.0114	4.3796	UU
Deut3524	6	38	72.76	15.15	2.8472	4.3009	UU

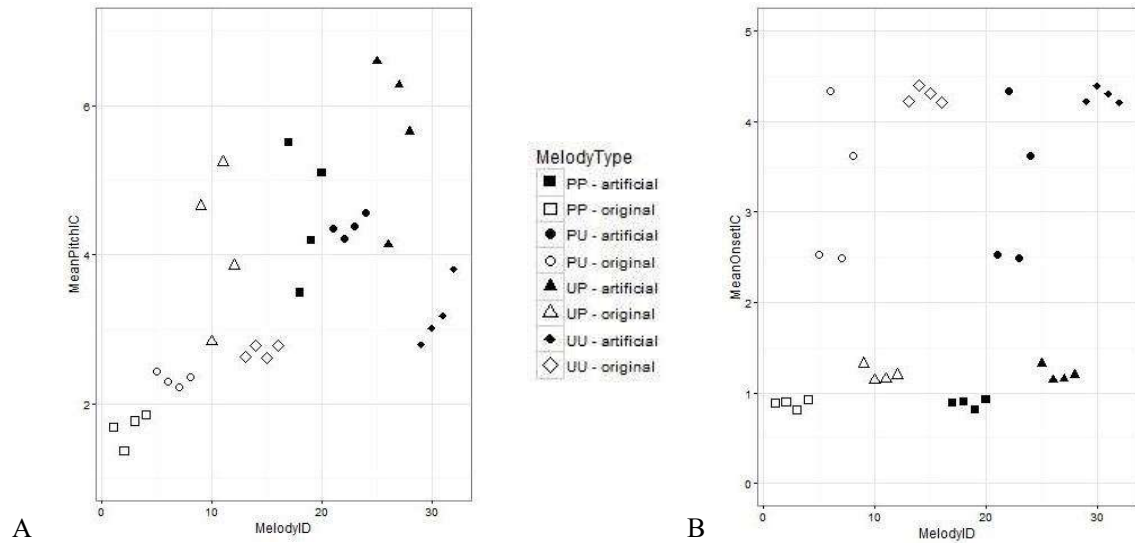


Figure 6.2. Mean (A) pitch IC and (B) onset IC of each melody plotted by stimulus predictability and modification, where original melodies are symbolized by empty symbols and artificial melodies by full symbols.

6.3.3 Procedure

Participants were semi-randomly allocated to one of four (between-subjects) conditions: they were either a musician or a non-musician and, within these groups, randomly assigned to rate either expectancy or emotion (arousal and valence). The experiment was run on software constructed in-house, and on a Samsung Galaxy Ace S5830 (3.5 inches in diameter; running Android 2.3.6). Participants listened through standard Apple headphones and were tested individually in a closed room. The information sheet was presented and informed consent gathered; detailed step-by-step instructions were then presented to participants. Regardless of condition, there was a mandatory practice session: participants heard two melodies and answered the questions appropriate to the condition they were assigned to (either expectancy rating or arousal and valence rating). Participants could also adjust the volume to a comfortable setting during the practice session. Once the practice session was completed, the experimental app was loaded. Participants entered a unique ID number

provided by the experimenter and responded to a short musical background questionnaire. Participants then heard the 32 musical excerpts (mean duration 18.34 s) presented in random order without pause or repeat and performed the appropriate ratings by continuously moving a finger on the screen. Those in the expectancy rating condition reported expectancy on a 7-point integer Likert scale, where 1 was *very expected* and 7 was *very unexpected*. Those in the arousal/valence condition rated induced arousal (vertical) and valence (horizontal) on a two-dimensional arousal/valence illustration (Russell, 2003). Responses, in integers, were collected at a 5Hz sample rate (200ms) (Khalifa, Isabelle, Jean-Pierre, & Manon, 2002). The rating systems used were: Expectancy: 1 – 7 (expected – unexpected); Arousal: 0 – 230 (calm – stimulating); Valence: 0 – 230 (unpleasant – pleasant).

6.3.4 Data collection

Due to the large number of variables included in this analysis, each is described here for clarity. First, the dependent variables are the expectancy, arousal and valence ratings.

The data acquisition software had a sampling rate of 5Hz; for each point in time for each melody and each participant, a *data point*, a value for *pitch*, *IOI*, *musical training*, *stimulus predictability*, *stimulus modification*, *pitch IC* and *onset IC* is assigned, along with *melody ID* and *participant ID*. The first independent variables described are those that were not explicitly manipulated: time, pitch, IOI and musicianship. As mentioned, time is measured in steps of 200ms, the sampling rate of the data acquisition software. Since pitch (interpreted here in MIDI numbers) does not change every 200ms and IOI (in ms) is longer than 200ms in these folk songs (or in Western music in general), their values were interpolated to match the participant ratings' sampling rate of 5Hz so that each point in time has a pitch and IOI value. Finally, the musical training variable had a value of 0 or 1, depending on whether the participant had no musical training or any musical training, respectively.

Next, the following manipulated variables will be described: stimulus predictability, stimulus modification, pitch IC, onset IC. For each data point, the variable stimulus predictability was given a value of 1 if it belonged to the category PP, 2 if it belonged to the category PU, 3 for UP and 4 for UU, regardless of whether these are original or artificial melodies. Similarly, the variable stimulus modification was given a value of 0 if the melody was original or 1 if the melody was artificial. Finally, pitch IC and onset IC, as calculated by IDyOM, were interpolated in the same way as pitch and IOI to match the participant ratings' 5Hz sampling rate. These are the only variables whose values are not integers.

6.3.5 Statistical analysis

For each type of rating (expectancy, arousal, valence) two kinds of analysis were performed: first, a melody-level analysis, in which the time-series for each melody was averaged across participants separately for musicians and non-musicians and temporal position was a discontinuous factor; second, a cross-sectional time-series analysis of the continuous ratings given by each participant throughout each melody. In the melody-level analysis, for each melody, a mean expectancy rating was calculated at every time point across the musician and non-musician groups (10 responses per group). Linear multiple regression modelling was used to evaluate the impact of *time* (point in time at sampling rate of 200ms), *musical training* (musician or non-musician), *stimulus modification* (original or artificial), *stimulus predictability* (predictable/unpredictable pitch/onset), *pitch* and *IOI* on mean expectancy ratings by using a log likelihood test to compare a model with each predictor to a model containing only an intercept. Two additional predictors, *pitch predictability* and *onset predictability*, were derived from stimulus predictability in order to examine the interaction between these two subcomponents: here melodies were coded as having either predictable (1) or unpredictable (0) pitch or onset. While musical training, stimulus modification, pitch

predictability and onset predictability were simple binary factors, stimulus predictability contained four levels, labelled PP, PU, UP and UU. Apart from that between pitch predictability and onset predictability, interactions were not considered due to the difficulty of interpretation for such a complex model. Following these log likelihood comparisons, two *global* linear multiple regression models containing all the above predictors of interest (one containing stimulus predictability and the other containing pitch predictability and onset predictability) plus *time*, *pitch* and *IOI* to parse out any potential effects of time and to analyse potential effects of musical contagion, were evaluated to confirm results.

For the analysis of continuous ratings throughout each melody, cross-sectional time series analysis was employed (CSTSA) similarly to Dean et al. (Dean, Bailes, & Dunsmuir, 2014a, 2014b) to evaluate the predictive impact effects of *pitch IC*, *onset IC*, *stimulus predictability* (predictable/unpredictable), *stimulus modification* (none/artificial), *musical training* and individual differences modelled by random effects on participants' ratings of expectedness, arousal and valence. CSTSA takes account of the autoregressive characteristic of music and the continuous responses of the participants. Pitch IC and onset IC predictors were both scaled to values between 0 and 1 to allow for direct comparison of model coefficients in analysis. A predictor of combined *pitch and onset IC* was also tested, replacing the individual *pitch IC* and *onset IC* predictors. In practice, CSTSA is a mixed-effects model, fitted with maximum random effects as per Barr et al. (Barr, Levy, Scheepers, & Tily, 2013) and fixed effects to account for autocorrelation (lags of endogenous variables, i.e. ratings, denoted by P), and exogenous influence (i.e. information content and its lags, denoted by L). Only optimal models are presented, selected based on BIC, confidence intervals on fixed effect predictors, log likelihood ratio tests between pairs of models, correlation tests between models and the data, and the proportion of data squares fit.

In order to test the relative salience hypothesis put forward in Section 4.2 above, where increased predictability will equate to less salience, four sub-models of each of the three CSTSA models optimised for expectancy, arousal and valence ratings were created, one for each stimulus predictability (PP, PU, UP, UU) in order to compare coefficients between models. Linear multiple regression modelling was used to evaluate the impact of *stimulus predictability*, *lag type* (pitch, onset) and *rating type* (expectancy, arousal, valence) on the coefficients of the sub-models.

6.4 Results

6.4.1 Melody level analysis

In this section, analyses of the mean ratings melody by melody and participant by participant are described: these are discontinuous data, and the experiment manipulated the pitch expectancy of the original melodies to provide a causal test of its influence. Mean ratings are shown in Figure 6.3 and important comparisons are highlighted in Figure 6.4.

Expectancy ratings. There were significant effects of musical training, where musicians rated melody unexpectedness higher (musicians mean = 4.40; non-musicians mean = 4.16; $F(1, 8343) = 73.12, p < .0001$); stimulus modification, where modified melodies, regardless of direction of manipulation (predictable to unpredictable or vice versa), were rated as more unexpected (original melodies mean = 3.92; modified melodies mean = 4.65; $F(1, 8342) = 569.75, p < .0001$); and stimulus predictability, where more predictable melodies were rated with lower unexpectedness than unpredictable melodies (PP melodies mean = 3.48; PU melodies mean = 4.71; UP melodies mean = 3.92; UU melodies mean = 4.66; $F(3, 8340) = 251.58, p < .0001$). Pitch predictability and onset predictability were both significant predictors where mean ratings for melodies with predictable pitch, unpredictable pitch, predictable onset

and unpredictable onset were 4.09, 4.29, 3.70 and 4.68 respectively ($F(1, 8342) = 83.05, p < .0001$ and $F(1, 8342) = 644.31, p < .0001$), and the interaction between the two predictors was also significant, such that there is a more pronounced effect of onset predictability on ratings, $t(3) = -7.36, p < .0001$. We also investigated the effect of stimulus predictability on ratings for original and modified melodies separately, where means for PP, PU, UP and UU melodies were 1.88, 4.47, 3.58 and 5.19 respectively ($F(3, 4223) = 1866.2, p < .0001$) and for aPP, aPU, aUP and aUU melodies were 4.27, 4.16, 5.29 and 4.96, respectively ($F(3, 4112) = 264.36, p < .0001$). The two global models confirmed nearly all the above results, producing two additional findings: pitch ($t(8336) = -3.76, p = .0001$) and IOI ($t(8336) = -3.72, p = .0001$) were significant predictors in both global models and pitch predictability became insignificant in its model ($t(2) = 0.24, p = .80$). In summary, all predictors of interest were significant apart from pitch predictability becoming superseded by pitch and IOI, including the interaction between pitch predictability and onset predictability.

Arousal ratings. There were significant effects of musical training where musicians rate melodies as more arousing overall as compared to non-musicians (musicians mean = 118.16; non-musicians mean = 112.90; $F(1, 8017) = 25.30, p < .0001$) and stimulus predictability where more predictable melodies were rated as more arousing (PP melodies mean = 151.73; PU melodies mean = 109.45; UP melodies mean = 128.86; UU melodies mean = 95.95; $F(3, 8015) = 667.31, p < .0001$). There was no effect of stimulus modification in either direction of manipulation (original melodies mean = 115.83; modified melodies mean = 115.27; $F(1, 8017) = .62, p = .42$). Pitch predictability and onset predictability were both significant predictors where mean ratings for melodies with predictable pitch, unpredictable pitch, predictable onset and unpredictable onset were 125.29, 112.40, 135.29, and 102.7 respectively ($F(1, 8017) = 208.38, p < .0001$ and $F(1, 8017) = 1804.3, p < .0001$), and though

similarly to expectancy ratings, onset predictability had a larger effect on mean ratings than pitch predictability numerically, this effect was not significant, $t(3) = 1.08, p = .28$. Stimulus predictability was also a significant predictor when original and artificial melodies' ratings were investigated separately, with ratings for PP, PU, UP and UU melodies averaging 138.62, 111.14, 121.07 and 100.79 respectively, $F(3, 3956) = 210.16, p < .0001$, and aPP, aPU, aUP and aUU melodies averaging 137.10, 91.56 144.96 and 107.83, respectively, $F(3, 4054) = 556.76, p < .0001$. The two global models confirm all the above results, and add pitch ($t(8011) = -17.72, p < .0001$) and IOI ($t(8011) = 18.58, p < .0001$) as significant predictors. In summary, stimulus modification is the only predictor of interest that did not have a significant effect on arousal ratings, while pitch predictability and onset predictability did not interact significantly.

Valence ratings. There were significant effects of musical training where musicians overall rated melodies as having lower valence (musicians mean = 81.26; non-musicians mean = 84.08; $F(1, 8017) = 5.38, p = .02$); stimulus modification, regardless of direction of manipulation, where original melodies had more positive valence than artificial melodies (original melodies mean = 91.20; artificial melodies mean = 74.33; $F(1, 8017) = 206.84, p < .0001$) and stimulus predictability where more predictable melodies are rated more positively than unpredictable melodies (PP melodies mean = 109.87; PU melodies mean = 74.00; UP melodies mean = 87.00; UU melodies mean = 70.02; $F(3, 8015) = 224.81, p < .0001$). Pitch predictability and onset predictability were both significant predictors where mean ratings for melodies with predictable pitch, unpredictable pitch, predictable onset and unpredictable onset were 91.93, 78.51, 98.43, and 72.01 respectively ($F(1, 8017) = 122.51, p < .0001$ and $F(1, 8017) = 559.04, p < .0001$), and the interaction between the two predictors was significant, where onset predictability again had a larger effect on mean ratings than pitch predictability, $t(3) = 8.40, p < .0001$. Stimulus predictability was also a significant predictor when

investigating original and artificial melodies separately, where PP, PU, UP and UU melodies have mean arousal ratings of 171.90, 77.96, 94.59 and 44.46 respectively, $F(3, 3956) = 1582.6$, $p < .0001$ and aPP, aPU, aUP and aUU melodies have mean ratings of 78.98, 93.21, 45.66 and 70.19 respectively, $F(3, 4054) = 276.84$, $p < .0001$. The two global models include IOI ($t(8011) = 22.07$, $p < .0001$) but not pitch ($t(8011) = -1.48$, $p = .13$) as significant predictors (in both models) and remove pitch predictability ($t(8011) = 0.90$, $p = .36$) from the set of significant predictors found above. In summary, all predictors of interest are significant, including the interaction between pitch predictability and onset predictability, except where pitch predictability was superseded by IOI.

This melody-level analysis has demonstrated that musical training and stimulus predictability predict expectancy, arousal and valence ratings. Furthermore, there is a significant interaction between pitch predictability and onset predictability for expectancy and valence ratings, and a similar pattern for arousal ratings, where onset predictability has a larger effect on ratings than pitch predictability. Stimulus modification is a significant predictor for expectancy and valence ratings only. The results of a cross-sectional time series analysis are presented next.

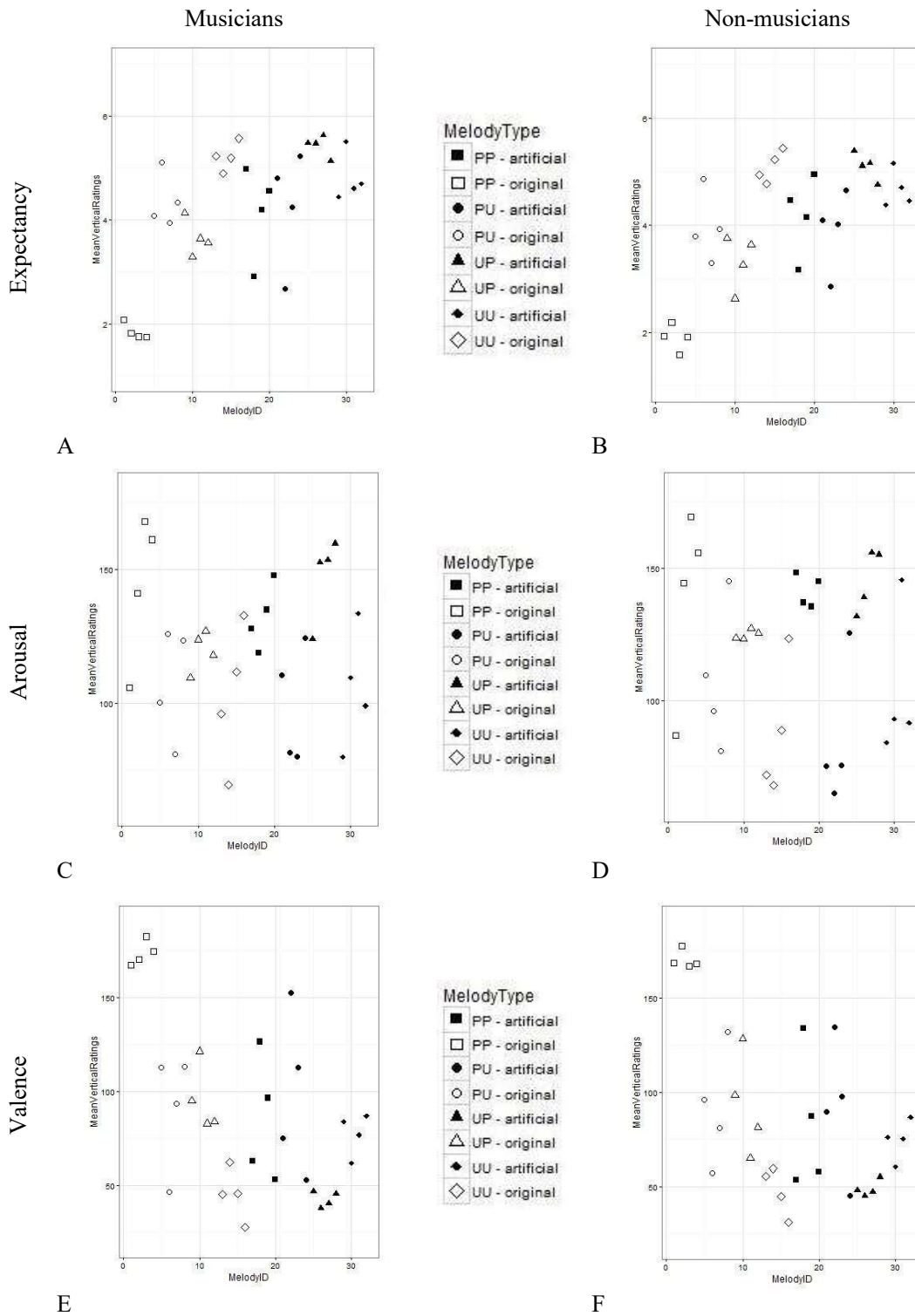


Figure 6.3. Mean expectancy (A, B), arousal (C, D) and valence (E, F) ratings for each melody for musicians (A, C, E) and non-musicians (B, D, F). Hollow shapes illustrate original melodies while filled shapes illustrate artificial melodies.

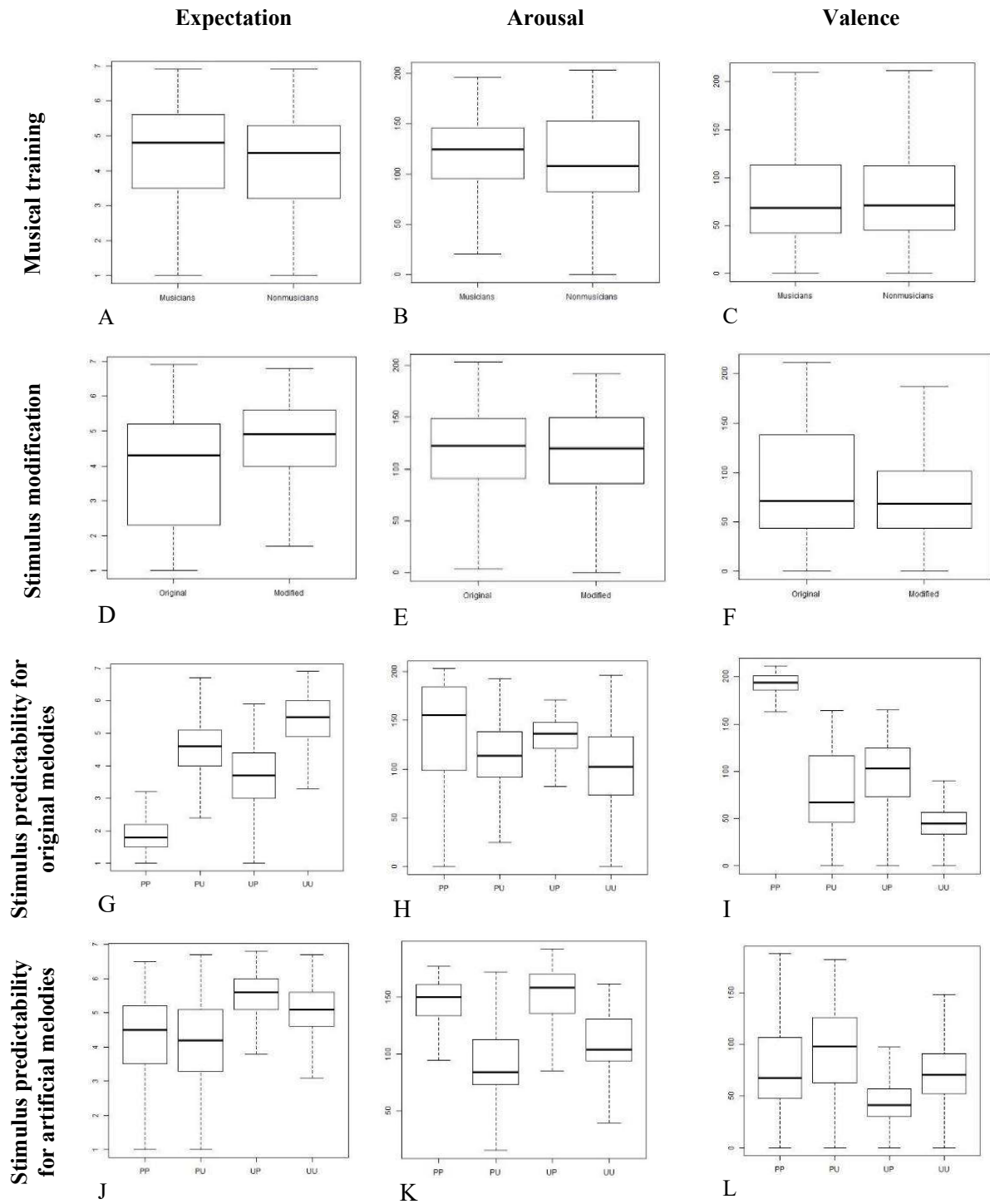


Figure 6.4. Box plots illustrating important mean comparisons between musicians and non-musicians (A, B, C), original and artificial melodies (D, E, F), stimulus predictability categories for original (F, H, I) and artificial (J, K, L) melodies for expectation (A, D, G, J), arousal (B, E, H, K) and valence (C, F, I, J) ratings.

6.4.2 Cross-sectional time series analysis

Here the analyses of the continuous time series data resulting from participants' ongoing responses during listening to the melodies is presented.

Expectancy, arousal and valence ratings were modelled separately using mixed-effects autoregressive models with random intercepts on participant ID and melody ID as well as random slopes on the fixed effect predictor with the largest coefficient before slopes were added. Fixed effects predictors were time, musical training, stimulus predictability, stimulus modification, autoregressive lags of up to 15 (equivalent of 3 seconds) and exogenous lags of pitch and onset information content of up to 15. A combined pitch and onset information predictor was also tested to evaluate whether a combined measure superseded the separate pitch and onset information content predictors. Maximum lags were selected based on previously reported rate of change of emotional responses (Juslin & Västfjäll, 2008) as well as precedent in this type of analysis (Dean et al., 2014a). Pitch and IOI were subsequently added as fixed-effect predictors to investigate the potential confounding effects of musical structure affecting ratings through an emotional contagion mechanism. Figures 6.5 and 6.6 illustrate the variance fitted by random effects, and the fit of the models for a selection of melodies and participants.

Expectancy ratings. The best CSTSA model for expectancy ratings is summarized in Table 6.5. In this model, while autoregression and random effects were duly considered, an effect of musicianship was still clearly observed in addition to pitch IC and onset IC and the optimal selection of their lags. Thus the model included random intercepts and random slopes for L1pitchIC on melody ID and participant ID as well as fixed effects of musicianship, L = 0-1, 7-8 of pitch IC, L = 0-2, 10, 12 of onset IC and P = 1-2, 4-6, 15 of autoregression. All predictors were significant, as Wald 95% confidence intervals did not include zero. The addition of stimulus predictability as a fixed effect did not improve the model, $\chi^2(3) = 1.80$,

$p = .61$ while musicianship and stimulus modification did, $\chi^2(2) = 13.36$, $p = .001$ and $\chi^2(1) = 3.91$, $p = .04$ respectively. The further addition of pitch and IOI significantly improved the model, $\chi^2(2) = 409.33$, $p < .0001$, and removed stimulus modification as a significant predictor. Combined pitch and onset information content with lags of pitch and onset from the best model outlined above was significantly worse, $\chi^2(6) = 972.6$, $p < .0001$.

A correlation test between the data and the model was highly significant, with correlation $.93$, $t(82486) = 783.09$, $p < .0001$. A proportion of data squares fit test was also high, with the model explaining 98% of the data. While this particular model did not converge, a model without random slopes removed did converge where all fixed effects were significant, model fit was equally good (correlation test: $.93$, $t(82486) = 780.53$, $p < .0001$; proportion of data squares fit: 98%) and the inclusion of slopes improved the model significantly; therefore random slopes were reinserted into the best model as per the experimental design (Barr et al., 2013). The final model thus includes design-driven random effects, musicianship, stimulus modification, pitch, IOI, optimal autoregressive lags of expectancy ratings and optimal lags of pitch IC and onset IC.

Arousal ratings. The best CSTSA model for arousal ratings is summarized in Table 6.6. This model revealed stimulus predictability as a significant predictor of arousal ratings in addition to pitch IC and onset IC and a selection of their lags when autoregression and random effects were considered. The model included random intercepts and random slopes for L1onsetIC on melody ID and participant ID as well as fixed effects L = 0-1, 6-8, 10-13, 15 of pitch IC, L = 0-4, 7, 10, 12-15 of onset IC and P = 1, 3, 5-6, 15 of autoregression. All predictors were significant, as Wald 95% confidence intervals did not include zero. The addition of musicianship and stimulus modification as fixed effects did not improve the model, $\chi^2(2) = .60$, $p = .74$ and $\chi^2(2) = 1.72$, $p = .42$ respectively while stimulus predictability did,

Table 6.5. CSTSA modelling of expectancy ratings for all melodies; coefficients for fixed width 95% CI's and variance of random effects.

	Predictor	Coefficient	95% CI	95% CI
Fixed effects	Intercept	0.307	0.251	0.364
	Time	0.001	0.001	0.001
	Musicianship	0.030	0.014	0.004
	Pitch	-0.002	-0.002	-0.001
	IOI	0.003	0.003	0.003
	L1ratings	0.960	0.953	0.967
	L2ratings	-0.065	-0.073	-0.058
	L4ratings	-0.061	-0.069	-0.053
	L5ratings	0.015	0.006	0.025
	L6ratings	0.035	0.023	0.037
	L15ratings	0.015	0.012	0.018
	PitchIC	-0.263	-0.309	-0.217
	L1pitchIC	0.486	0.306	0.666
	L7pitchIC	0.123	0.079	0.167
	L8pitchIC	-0.059	-0.103	-0.016
	OnsetIC	-0.731	-0.794	-0.667
	L1onsetIC	0.845	0.769	0.920
	L2onsetIC	-0.181	-0.240	-0.123
	L10onsetIC	-0.084	-0.129	-0.039
	L12onsetIC	0.138	0.092	0.183
	Predictor	Variance	-	-
Random effects on individuals	Intercept	0.000		
	L1pitchIC	0.000		
Random effects on melody ID	Intercept	0.019		
	L1pitchIC	0.245		
Residual variance		0.421		

$\chi^2 (2) = 14.91, p = .0005$. The further addition of pitch and IOI significantly improved the model, $\chi^2 (2) = 178.89, p < .0001$, where both are significant predictors of arousal ratings. Combined pitch and onset information content with lags of pitch and onset from the best model outlined above was significantly worse, $\chi^2 (13) = 4482.2, p < .0001$.

A correlation test between the data and the model was highly significant, with correlation .96, $t (80183) = 978.48, p < .0001$. A proportion of data squares fit test was also

Table 6.6. CSTSA modelling of arousal ratings for all melodies; coefficients for fixed width 95% CI's and variance of random effects.

	Predictor	Coefficient	95% CI	95% CI
	(Intercept)	1.98	-0.07	4.03
	Time	0.06	0.06	0.07
	Predict2	-3.42	-5.77	-1.06
	Predict3	-0.50	-2.86	1.85
	Predict4	-4.53	-6.88	-2.17
	Pitch	-0.04	-0.05	-0.03
	IOI	-0.03	-0.03	-0.02
	L1ratings	0.95	0.94	0.95
	L3ratings	0.01	0.00	0.01
	L5ratings	-0.05	-0.06	-0.05
	L6ratings	0.03	0.02	0.03
	L15ratings	0.01	0.01	0.01
	PitchIC	-16.6	-17.7	-15.4
	L1pitchIC	16.6	15.5	17.8
	L6pitchIC	2.46	1.31	3.62
Fixed effects	L7pitchIC	2.05	0.70	3.39
	L8pitchIC	-2.14	-3.37	-0.92
	L10pitchIC	1.86	0.63	3.08
	L11pitchIC	-4.43	-5.77	-3.10
	L12pitchIC	4.91	3.57	6.25
	L13pitchIC	-1.95	-3.18	-0.72
	L15pitchIC	2.18	1.23	3.13
	OnsetIC	-11.4	-12.9	-9.83
	L1onsetIC	72.4	48.2	96.6
	L3onsetIC	6.96	5.26	8.66
	L4onsetIC	-8.38	-9.98	-6.77
	L7onsetIC	1.55	.345	2.76
	L10onsetIC	-6.81	-8.12	-5.49
	L12onsetIC	5.43	3.73	7.13
	L13onsetIC	4.47	2.55	6.39
	L14onsetIC	-2.93	-4.83	-1.04
	L15onsetIC	3.09	1.59	4.58
	Predictor	Variance	-	-
Random effects on individuals	Intercept	0.47		
	L1onsetIC	2.94		
Random effects on melody ID	Intercept	13.5		
	L1onsetIC	4815.2		
Residual variance		276.7		

high, with the model explaining 98% of the data. While this particular model did not converge, a model without random slopes removed did converge where all fixed effects were significant, model fit was equally good (correlation test: .95, $t(80183) = 959.73$, $p < .0001$; proportion of data squares fit: 98%) and the inclusion of slopes improved the model significantly, $\chi^2(5) = 335.3$, $p < .0001$; therefore random slopes were reinserted into the best model as per the experimental design (Barr et al., 2013). The final model thus includes design-driven random effects, stimulus predictability, pitch, IOI, optimal autoregressive lags of expectancy ratings and optimal lags of pitch IC and onset IC.

Valence ratings. The best CSTSA model for valence ratings is summarized in Table 6.7. This model revealed significant effects of only pitch IC and onset IC and a selection of their lags when autoregression and random effects were considered. The model included random intercepts and random slopes for L1onsetIC on melody ID and participant ID as well as fixed effects L = 0-1, 5, 8-9, 11-13, 15 of pitch IC, L = 0-1, 3-4, 10, 13 of onset IC and P = 0, 3-7, 9, 15 of autoregression. All predictors were significant, as Wald 95% confidence intervals did not include zero. The addition of musicianship, stimulus predictability and modification as fixed effects did not improve the model, $\chi^2(1) = .29$, $p = .58$, $\chi^2(3) = 4.77$, $p = .18$ and $\chi^2(1) = 3.46$, $p = .06$ respectively. The further addition of pitch and IOI significantly improved the model, $\chi^2(1) = 600.99$, $p < .0001$, where both are significant predictors of arousal ratings. Combined pitch and onset information content with lags of pitch and onset from the best model outlined above was significantly worse, $\chi^2(10) = 194.72$, $p < .0001$.

A correlation test between the data and the model was highly significant, with correlation .94, $t(80183) = 827.83$, $p < .0001$. A proportion of data squares fit test was also high, with the model explaining 98% of the data. While this particular model did not converge, a model without random slopes removed did converge where all fixed effects were significant,

Table 6.7. CSTSA modelling of valence ratings for all melodies; coefficients for fixed width 95% CI's and variance of random effects.

	Predictor	Coefficient	95% CI	95% CI
Fixed effects	(Intercept)	5.38	3.56	7.20
	Time	0.03	0.03	0.04
	Pitch	-0.09	-0.10	-0.08
	IOI	0.16	0.15	0.18
	L1ratings	0.92	0.92	0.93
	L3ratings	-0.02	-0.03	-0.01
	L4ratings	-0.03	-0.04	-0.02
	L5ratings	-0.01	-0.02	-0.00
	L6ratings	0.01	0.00	0.02
	L7ratings	0.01	0.00	0.02
	L9ratings	0.00	0.00	0.01
	L15ratings	0.00	0.00	0.01
	PitchIC	-9.19	-10.6	-7.72
	L1pitchIC	11.2	9.74	12.6
	L5pitchIC	2.62	1.45	3.79
	L8pitchIC	-3.26	-4.72	-1.79
	L9pitchIC	3.29	1.74	4.83
	L11pitchIC	-1.68	-3.22	-0.15
	L12pitchIC	2.91	1.47	4.83
	L15pitchIC	1.28	0.20	2.36
	OnsetIC	-20.0	-22.2	-17.9
	L1onsetIC	48.5	29.7	67.3
	L3onsetIC	4.05	1.92	6.18
	L4onsetIC	-4.02	-5.90	-2.13
	L10onsetIC	-5.35	-6.65	-4.05
	L13onsetIC	3.59	2.32	4.86
		Predictor	Variance	-
Random effects on individuals	(Intercept)	0.11		
	L1onsetIC	0.12		
Random effects on melody ID	(Intercept)	22.2		
	L1onsetIC	2878.7		
Residual variance		439.9		

model fit was equally good (correlation test: .94, $t(80183) = 959.73$, $p < .0001$; proportion of data squares fit: 95%) and the inclusion of slopes improved the model significantly, $\chi^2(4) = 805.25$, $p < .0001$; therefore random slopes were reinserted into the best model as per the

experimental design (Barr et al., 2013). The final model thus includes design-driven random effects, pitch, IOI, optimal autoregressive lags of expectancy ratings and optimal lags of pitch IC and onset IC.

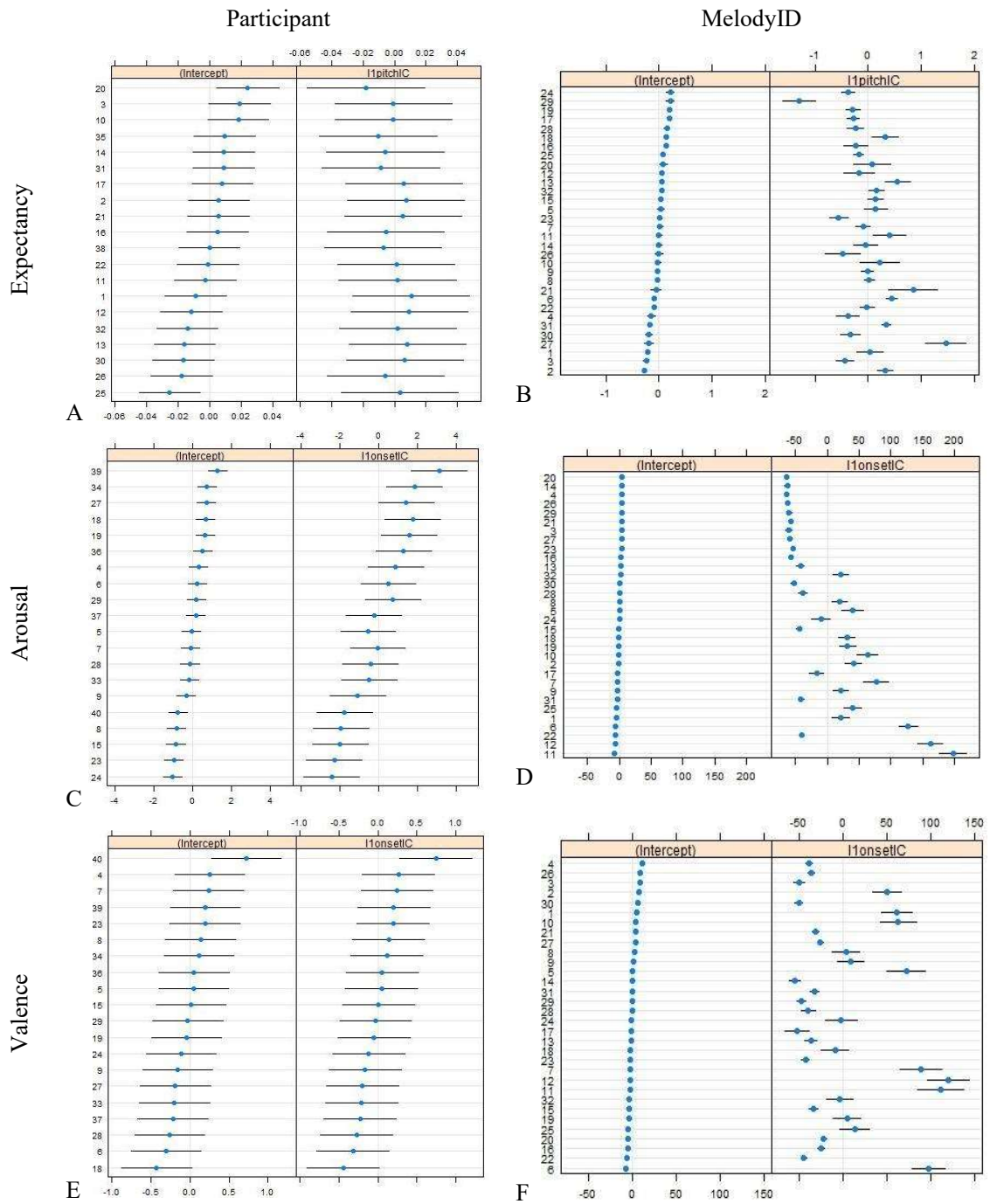


Figure 6.5. Intercept (left) and slope (right; predictors correspond to their respective models) values of random effects on Participant and MelodyID for expectancy, arousal and valence models. These show how each individual participant and melody was modelled and illustrate the variance among participants and melodies.

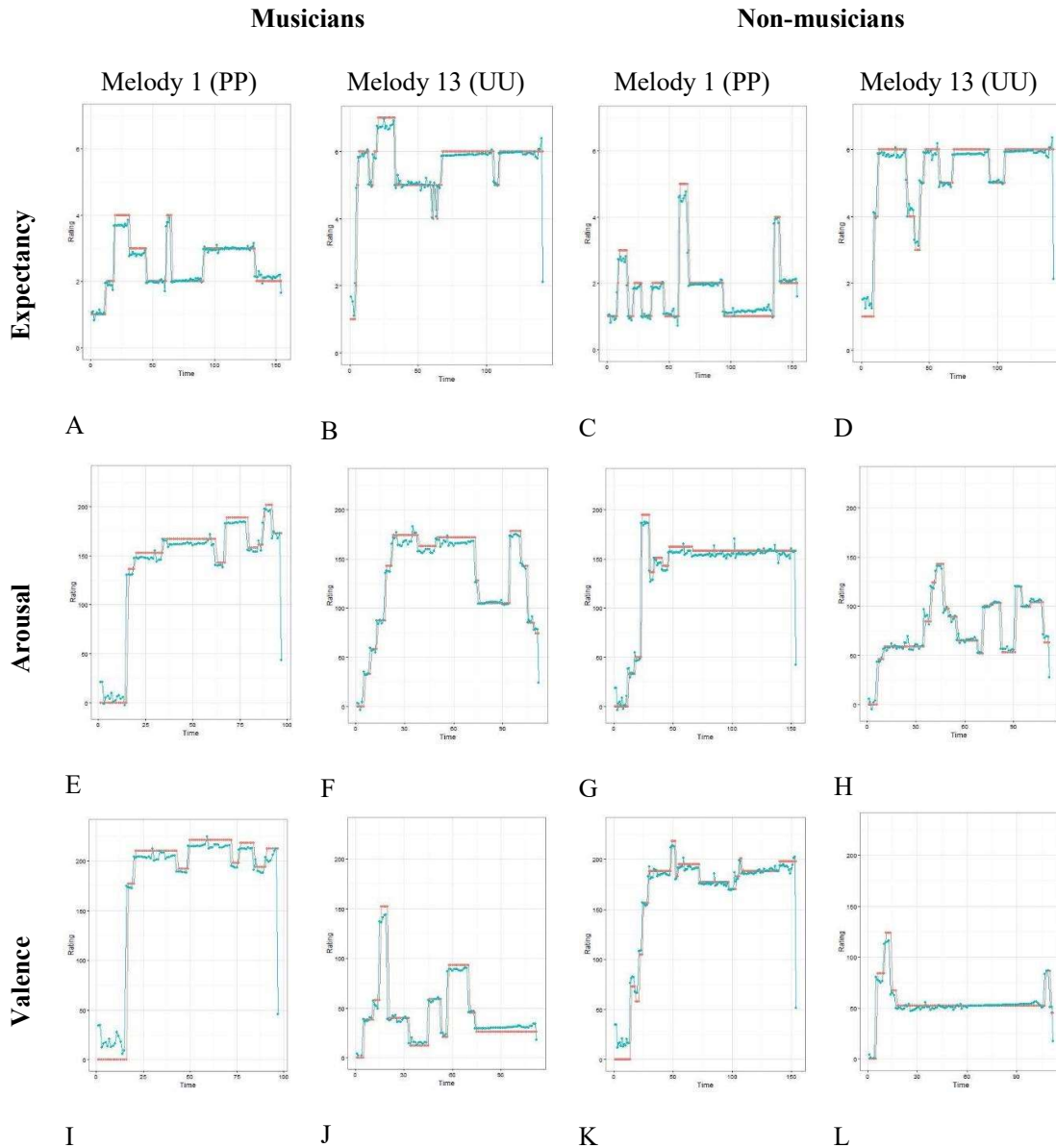


Figure 6.6. Expectancy (A, B, C, D), arousal (E, F, G, H) and valence (I, J, K, L) ratings for single randomly selected participants (6 musicians (A, B, E, F, I, J; participants 14, 35, 34, 18, 27, 7) and 6 non-musicians (C, D, G, H, K, L; participants 1, 10, 8, 33, 5, 37)) are plotted for Melodies 1 (A, C, E, G, I, K) and 13 (B, D, F, H, J, L), examples of PP and UU categories respectively. Ratings predicted by the model (teal) for those melodies for each of those participants only (single extracts) are plotted alongside their actual ratings (pink). Residuals were too small to illustrate on the same plot. These plots illustrate the high explanatory power of our model due to its random effects structure fitted specifically to this data set.

Table 6.8. Coefficients of sub-models for expectancy ratings.

	Coefficient	PP	PU	UP	UU
	Intercept	0.393	0.399	0.235	0.204
	Time	0.002	0.001	0.003	0.001
	Musicianship	0.035	0.031	0.024	0.031
	Pitch	-0.001	0.002	-0.002	-0.002
	IOI	0.001	0.003	0.004	0.005
	L1ratings	0.777	1.00	1.01	1.03
	L2ratings	0.042	-0.100	-0.141	-0.110
	L4ratings	-0.045	-0.075	-0.075	-0.035
Fixed effects	L5ratings	-0.010	-0.031	0.031	0.021
	L6ratings	0.052	-0.007	0.039	0.008
	L15ratings	0.027	0.013	0.019	0.009
	PitchIC	-0.192	-0.197	-0.142	-0.549
	L1pitchIC	0.197	0.461	0.604	0.842
	L7pitchIC	0.242	0.210	-0.021	0.008
	L8pitchIC	-0.005	-0.169	0.030	-0.087
	OnsetIC	-0.756	-1.26	-0.316	-0.769
	L1onsetIC	1.15	1.50	0.687	0.277
	L2onsetIC	-0.258	-0.469	-0.231	0.078
	L10onsetIC	-0.530	-0.169	-0.038	-0.153
	L12onsetIC	0.188	0.034	0.058	0.235
	Participant – Intercept	0.002	0.000	0.002	0.000
Random effects	Participant – l1pitchIC	0.014	0.003	0.000	0.000
	MelodyID – Intercept	0.105	0.005	0.010	0.006
	MelodyID – l1pitchIC	0.078	0.174	0.337	0.178
Residual variance		0.455	0.397	0.536	0.319

6.4.3 Relative salience

Multiple linear regression models were conducted on the coefficients of CSTSA sub-models, one for each stimulus predictability for each rating type; details of these models can be found in Tables 6.8-6.10. Predictors were stimulus predictability, lag type (pitch, onset) and rating type (expectancy, arousal, valence): no predictor was significant, $F(3, 168) = .50, p =$

.67, $F(1, 170) = 2.23$, $p = .13$, $F(2, 169) = .51$, $p = .59$ respectively. There was also no interaction between category and lag type, $F(7, 164) = .79$, $p = .59$.

6.5 Discussion

The results presented above provide answers to all three research questions posed. First, there is evidence that predictability of both pitch and temporal musical structure has an effect on listeners' expectancies and emotional reactions, and that these can be manipulated. This not only agrees with previous behavioural validation of IDyOM pitch-based viewpoints (Pearce et al., 2010) but contributes behavioural validation of IDyOM temporal viewpoints. Second, contrary to a prediction based on complexity, for these stimuli temporal expectancy influences perception more strongly than pitch expectancy. Finally, individual differences generally supersede effects of musical training (Dean et al., 2014a) and inter-melody differences are more substantial than differences between melody predictability groups (PP, UP, PU and UU) or manipulation type, where differences between predictability groups could nevertheless be detected in the discontinuous, melody-level analysis. While the melody-level analysis yielded more significant effects, the CSTSA analysis supplemented it with evidence for a common perceptual mechanism and the importance of individual differences, neither of which could have been detected in the melody-level analysis. Overall, the two types of analysis converge and the implications of these results are discussed further below.

Using IDyOM (Pearce, 2005) to calculate average pitch and onset information content, folk songs were classified into four categories based on overall expectedness, where average pitch expectancy and average onset expectancy could be high or low. Furthermore, pitch expectancy was manipulated to transform expected pitches into unexpected ones, and vice versa. The four melody categories resulted in different subjective ratings of expectancy, arousal and valence, where high pitch and onset information content (UU) resulted in high

unexpectedness ratings, higher arousal and lower valence, low pitch and onset information content (PP) resulted in low unexpectedness ratings, lower arousal and higher valence, and mixed high and low pitch and onset information content (PU and UP) lay somewhere in between, where only the predictable pitch and onset (PP) and unpredictable pitch and predictable onset (UP) categories were not different from each other in arousal ratings. This confirms previous evidence that statistical learning and information content may influence listener expectancies for pitch (Pearce et al., 2010; Pearce & Wiggins, 2006) and arousal and valence ratings of music (Egermann et al., 2013), and provides evidence for statistical learning of temporal information, a hypothesis previously unconfirmed. Additionally, there is a significant interaction between pitch predictability and onset predictability for expectancy and valence ratings, with a similar non-significant pattern for arousal ratings, where onset predictability has a more pronounced effect on ratings than pitch predictability. Cross-sectional time series analysis support these results with excellent models explaining between 93-96% of expectancy, arousal and valence ratings, all including pitch and onset information content, and lags of these of up to 3s (Egermann et al., 2013) as predictors. Additionally, explicit causal manipulation of pitch expectancy – the modification of selected pitches from high to low or from low to high expectancy – results in a change in ratings in the expected direction. For example, melodies transformed into the UP category (filled triangle in Figure 6.3) are rated with higher unexpectedness ratings and lower valence than their original PP counterparts (hollow square in Figure 6.3), yet are also different from the original UP category (hollow triangle in Figure 6.3) melodies. This effect is more pronounced for expectedness and valence ratings than for arousal ratings, which can be explained by the intentionally inexpressive nature of the stimuli. Therefore, the manipulation of pitch expectancy adds causal evidence to

Table 6.9. Coefficients of sub-models for arousal ratings

	Coefficients	PP	PU	UP	UU
Fixed effects	Intercept	-5.66	0.77	-4.68	-3.45
	Time	0.25	0.04	0.24	0.05
	Pitch	-0.02	-0.02	-0.10	-0.00
	IOI	-0.03	0.00	-0.12	-0.00
	L1ratings	0.87	0.96	0.89	0.95
	L3ratings	0.02	0.01	0.01	-0.00
	L5ratings	-0.05	-0.08	-0.03	-0.02
	L6ratings	0.03	0.04	0.04	0.01
	L15ratings	0.05	0.01	0.02	0.01
	PitchIC	-18.3	-12.7	-11.6	-20.5
	L1pitchIC	19.8	8.85	14.5	24.1
	L6pitchIC	9.35	1.42	-1.65	0.32
	L7pitchIC	-2.64	3.32	8.12	2.05
	L8pitchIC	-0.20	-3.68	-3.71	-4.12
	L10pitchIC	6.03	0.66	2.15	0.08
	L11pitchIC	-10.0	-6.78	2.50	-1.29
	L12pitchIC	6.54	8.64	-2.35	1.75
	L13pitchIC	0.08	-7.88	3.12	-0.45
	L15pitchIC	3.69	0.04	2.14	5.71
	OnsetIC	-24.7	-21.1	3.90	-7.01
	L1onsetIC	78.4	91.46	123.5	17.4
	L3onsetIC	10.2	1.84	10.5	4.73
	L4onsetIC	-15.0	-7.82	-7.92	-2.71
	L7onsetIC	-3.79	7.51	0.01	-1.55
	L10onsetIC	-24.2	-1.00	-6.14	-1.82
	L12onsetIC	13.5	-3.32	5.03	6.53
	L13onsetIC	6.51	11.9	1.76	0.47
	L14onsetIC	-3.12	-6.33	-4.64	-0.81
	L15onsetIC	1.30	1.28	5.12	2.40
	Random effects	Participant – Intercept	0.44	1.17	0.12
Participant – l1onsetIC		0.51	0.08	0.24	1.10
MelodyID – Intercept		35.3	8.23	56.4	3.90
MelodyID – l1onsetIC		2443.5	3846.0	11190.0	447.2
Residual variance		368.1	225.3	323.7	186.2

Table 6.10. Coefficients of sub-models for valence ratings.

	Coefficients	PP	PU	UP	UU
	Intercept	-2.42	9.68	1.35	-1.88
	Time	0.16	0.00	0.17	0.05
	Pitch	-0.06	-0.11	-0.13	-0.00
	IOI	0.17	0.14	0.23	0.14
	L1ratings	0.90	0.94	0.88	0.92
	L3ratings	-0.00	-0.01	-0.05	-0.00
	L4ratings	-0.02	-0.01	-0.05	-0.02
	L5ratings	-0.02	-0.04	0.02	-0.01
	L6ratings	0.02	0.02	0.00	0.01
	L7ratings	0.00	0.00	0.03	0.00
	L9ratings	0.01	0.00	0.02	0.00
	L15ratings	0.01	0.00	0.02	0.01
Fixed effects	PitchIC	-8.42	-6.56	-7.56	-14.7
	L1pitchIC	12.9	8.24	6.30	19.05
	L5pitchIC	2.98	-0.13	3.96	3.18
	L8pitchIC	-1.07	-3.18	-3.32	-8.26
	L9pitchIC	3.10	5.09	1.15	5.82
	L11pitchIC	-0.54	-7.44	2.52	-.53
	L12pitchIC	4.25	1.02	2.46	2.62
	L15pitchIC	1.71	-0.69	2.30	2.92
	OnsetIC	-22.7	-24.0	-4.81	-21.2
	L1onsetIC	49.1	80.7	81.4	6.13
	L3onsetIC	17.6	3.28	0.22	3.28
	L4onsetIC	-13.1	-0.27	-1.83	-2.87
	L10onsetIC	-14.3	-4.66	-6.26	0.21
L13onsetIC	7.75	6.75	-1.53	-0.94	
Random effects	Participant – Intercept	0.03	0.72	0.10	0.20
	Participant – l1onsetIC	0.14	0.00	1.37	1.06
	MelodyID – Intercept	46.5	9.41	49.8	4.26
	MelodyID – l1onsetIC	2972.0	4176.0	5024.3	36.0
Residual variance		519.6	375.9	712.1	251.5

previous research by demonstrating a direct link between expectancy manipulation and expectancy, arousal and valence ratings. It would be especially interesting to extend this to include the manipulation of temporal expectancy in the future, additionally allowing better evaluation of the relative contribution of these two dimensions to perceived expectancy, arousal and valence. Furthermore, these patterns support the most basic predictions of an expectancy mechanism while providing no evidence for an inverted U-curve pattern for emotional reactions to music (Gebauer et al., 2012). However, rather than contest the U-curve pattern, it is more likely that the stimuli for this study fall towards extreme ends of the curve.

While melody-level analysis demonstrates an effect of pitch and onset information content, CSTSA assesses the relative contribution of pitch and onset information content to expectancy, arousal and valence ratings in more detail. In the models presented above, onset information content coefficients are almost always approximately 1.1 to 4.3 times larger than pitch information content coefficients for exactly (i.e. L1pitchIC and L1onsetIC) or loosely (i.e. L5pitchIC and L6onsetIC) matching lags. Furthermore, the sum of onset IC lag coefficients is far greater than the sum of pitch IC lag coefficients for arousal and valence rating models, while the sum of pitch IC lag coefficients is greater than onset IC lag coefficients for the expectancy ratings model (though absolute values of individual onset IC coefficients are greater than the pitch IC coefficients). The discrepancy between these results and predictions based on complexity will be discussed further in Section 6.5.1 on Relative Saliency. The sum of lag coefficients is considered rather than the effect of each coefficient individually because the choice of exact combination of lags had minimal effect on the quality of the final model during optimization. This suggests that each lag coefficient does not carry very much interpretable information on its own, nor is this particular combination of lags, with a mix of positive and negative coefficient values, generalizable. Incidentally, every model includes

pitch IC and onset IC lags of 0 and 1, with little overlap beyond this, suggesting that processing time scales for both pitch and onset expectancy are similar soon after a particular note event and diverge after this. This variation in time scales could also explain why a combined pitch and onset IC predictor did not replace the separate pitch IC and onset IC predictors.

Though analysis of mean ratings yielded a main effect of musical training, the amount of variance explained by musical background was superseded by the amount of variance explained by random effects on participant ID for arousal and valence ratings, indicating that though groups can be formed, individual strategies are more important to explain these ratings. Though a large body of literature supports the existence of certain differences between musicians and non-musicians (Brattico et al., 2001; Carey et al., 2015; Fujioka et al., 2004; Granot & Donchin, 2002), similar research by Dean et al. (Dean et al., 2014a; Dean, Bailes, & Dunsmuir, 2014b) has also found that though there were differences between groups, individual differences explain more variance than musical background when rating arousal and valence of electroacoustic and piano music. However, musical background did hold important predictive power for expectancy ratings, where musicians gave slightly higher ratings overall, showing greater unexpectedness. This is in line with the hypothesis presented in Section 6.2, where training is expected to produce stronger expectancies and therefore more extreme reactions to violations (Hansen & Pearce, 2014; Strait et al., 2009). That being said, it is worth noting that the overall difference in ratings between musicians and non-musicians is small, with musicians' ratings being only 0.2 points higher.

Similarly, the differences between individual melodies, as modelled by random intercepts and slopes on Melody ID, outweigh categories of stimulus predictability and stimulus modification in all but two cases: expectancy ratings, where stimulus modification was a significant predictor, and arousal ratings, where stimulus predictability was a significant

predictor, such that $PP > UP > PU > UU$ in terms of arousal ratings. The predictive power of stimulus modification in the context of expectancy ratings can be explained by the overall higher pitch IC in artificial melodies, as shown in Figure 6.3. This is likely due to the fact that the modifications were made by an algorithm and are therefore not as smooth as human-composed changes might have been. As the original melodies already had relatively low IC, it would be difficult to keep mean IC as low or lower with the change of even one note, as this change could also have an effect on the IC of all subsequent notes in a given melody.

As for the importance of stimulus predictability in predicting arousal ratings, which was in the opposite direction to what was expected based on previous empirical (Egermann et al., 2013; Steinbeis et al., 2006) and theoretical (Meyer, 1956; Huron, 2006) research, this could be explained by the potentially confounding effect of duration on ratings. The analysis revealed that note duration did indeed have a significant effect on ratings, where melodies with longer durations, corresponding to low onset expectancy, were rated as more unexpected, less arousing and less pleasant. The pattern of mean arousal ratings by stimulus predictability, with PP and UP (high onset expectancy) rated as more arousing than PU and UU (low onset expectancy) matches this interpretation, which is further supported by previous research establishing a positive correlation between tempo and arousal (Carpentier & Potter, 2007; Husain, Thompson, & Schellenberg, 2002). The significant effect of pitch on ratings is more surprising; a pattern of higher average pitch for PP and UP categories corresponds to lower unexpectedness ratings, higher arousal ratings and higher valence ratings for these categories as compared to PU and UU categories. However, coefficients for pitch and IOI are smaller than almost all other predictors in expectancy, arousal and valence models, suggesting that their overall influence is minimal compared to pitch and onset IC on subjective expectancy and emotion responses.

Also similarly to Dean et al. (2014a), the use of CSTSA evaluates evidence for the presence of a common perceptual mechanism across all pieces of music heard. To do this, predictors encoding melodies by stimulus predictability and modification were added to the basic models, where a null effect of these additional predictors would indicate that the type of melody does not matter and the listeners' ratings depend only on pitch and onset IC in all melodies. In the case of valence ratings, neither stimulus predictability nor stimulus modification were found to provide any additional predictive power to the model, while stimulus modification was a significant predictor for expectancy ratings and stimulus predictability for arousal ratings. However, explanations were proposed for these results (see previous paragraph) and overall the data provides support for a common perceptual mechanism across all melodies.

6.5.1 Relative salience

Having considered the relative importance of pitch and onset IC in the context of models of participant expectancy, arousal and valence ratings, relative salience should also be considered. While the question of relative perceptual weighting between musical parameters such as pitch, timing, structure, and harmony in music cognition will be addressed in more detail in Chapter 7, it is worth mentioning the impact of the above results on this area of research. Studying relative musical salience is challenging and the term itself lacks a unified explanation (Dibben, 1999; Esber & Haselgrove, 2011; Prince et al., 2009; Uhlig, Fairhurst, & Keller, 2013). Generally, pitch or melody is considered the most salient aspect of a piece of music. Prince et al. (2009), for example, argue that there are many more possible pitches than there are rhythmic durations or chords; therefore, pitch takes more attentional resources to process and is more salient. On the other hand, in a corpus analysis of eighteenth- and nineteenth-century string quartets, Duane (2013) found that onset and offset synchrony were

the most important predictors of streaming perception of these string quartets, with pitch explaining half the variance that onset and offset synchrony did, and harmonic overlap explaining an almost insignificant amount. It is also important to consider musical genre when discussing salience, as certain genres are more rhythmically driven (i.e., rap, electronic dance music, African drum music) while others are more melodically driven (i.e., opera). Folk music is more ambivalent and may vary from song to song. Other genres may well produce different results; something which would be worth exploring in the future. The stimuli used in this study best fit Prince et al.'s (2009) description of musical salience, as these melodies contain more different pitches than different rhythmic values. This would imply that the pitch dimension is more complex, and therefore more salient. However, results indicate that onset information content is more salient than pitch information content, though here perception of emotion, as opposed to auditory streaming, was evaluated alongside the subjective experience of expectancy. Interestingly, work in cue salience in the context of associative learning explores the effect of predictability and uncertainty on salience (Esber & Haselgrove, 2011), with one model predicting increased salience for cues with high predictability (Mackintosh, 1975) and another model predicting increased salience for cues with high uncertainty (Pearce & Hall, 1980). Though contradictory, these models have each accumulated significant evidence and have more recently led to the development of both hybrid (Pearce & Mackintosh, 2010) and new unified models of cue salience (Esber & Haselgrove, 2011). The possibility that high and low uncertainty and pitch and onset lag coefficients interacted was considered so that melodies with high pitch predictability (expectancy) and low onset predictability (PU) led to larger pitch IC coefficients than onset IC coefficients, and vice versa. This effect was not found in the data (see Section 6.3.3), so it is concluded that in this particular paradigm, onset is the more salient cue overall.

6.6 Conclusion

In addition to validating IDyOM's implementation of statistical learning for temporal viewpoints, the present study makes a single but significant step towards isolating individual mechanisms for the induction of musical emotion. The study explicitly controlled for six of the eight proposed mechanisms (Juslin et al., 2011) and manipulated one while considering another as a covariate. Brain stem reflexes, evaluative conditioning, episodic memory, visual imagery, rhythmic entrainment and cognitive appraisal were controlled for by presenting novel stimuli with equal tempo, intensity and timbre alongside a rating task. Emotional contagion, information conveyed by musical structure itself, was addressed by including pitch and duration values into the CSTSA models of the expectancy, arousal and valence ratings. Though these were significant predictors, they carried less weight than the lags of information content predictors. Musical expectancy was examined by selecting stimuli with either high or low pitch and onset expectancy and additionally explicitly manipulating pitch expectancy, finding evidence for a consistent effect of pitch and onset expectancy on ratings of arousal and valence by musicians and non-musicians. Additionally, onset was found to be more salient than pitch and musicians gave higher unexpected ratings than non-musicians, where group differences were overridden by individual differences on emotion ratings.

In this chapter, IDyOM's implementation of the temporal viewpoints inter-onset-interval and onset were validated experimentally by collecting expectedness ratings from participants listening to musical excerpts in real time. These excerpts had either high or low objective expectancy along the pitch and temporal dimensions, as modelled by statistical learning of folk and Baroque music. In addition, this chapter provided strong evidence for expectancy as a mechanism for the induction of musical emotion, demonstrating the versatility and potential explanatory power of predictiveness in the context of cognitive mechanisms.

Though applied to a monophonic context, this chapter has addressed some of the basic auditory features relevant to auditory streaming, validating an expectancy approach to these, as well as taking into account musical training. In the next chapter, expectancy of basic auditory, as well as musical features will be applied to a polyphonic context to further explore relative salience between the pitch and temporal domains, as well as the harmonic domain, helping gradually build the knowledge necessary for the design of an integrated, prediction-based framework for auditory streaming.

7 Relative Salience in Polyphonic Music

In Chapter 2, one of the challenges of modelling auditory streaming discussed is that many auditory factors influence how sounds are grouped – from auditory features such as raw frequency and timbre to cognitive mechanisms such as attention and prediction to musical training – with each factor, or combination of factors having a distinct impact on perception

This chapter investigates this issue, evaluating an information content-based solution to relative perceptual salience. A combination of existing conceptual and methodological approaches is employed to investigate the relative perceptual salience of melody, harmony and rhythm. The long-term aim is to predict what feature of a piece of music a listener is most likely to be focused on at any given time – knowledge that is essential to modelling auditory streaming.

7.1 Defining Salience

From the start, salience is a tricky term. In laymen's terms, salience can be defined as something that stands out, but operationalizing it is more elusive, as evidenced by the varying existing approaches in the literature (Collins, Laney, Willis, & Garthwaite, 2011; Dibben, 1999; Lerdahl, 1989; Prince, Thompson, & Schmuckler, 2009). To summarize the state of the salience literature in music cognition, three definitions of salience will be discussed.

First, Collins et al. (2011) relate salience to repetition in a study of pattern importance in Chopin's Mazurkas, where it is proposed that a repeated pattern has some sort of musical importance. They studied 90 patterns from a selection of Chopin Mazurkas that encompassed five types of repetition:

- Exact;
- With interpolation: indicates the presence of additional notes in repetitions, usually between existing pattern notes;
- Transposed real: each pattern note is transposed the same number of semitones;
- Transposed tonal: the pattern is transposed, but some minor changes are allowed to keep the excerpt in its key; and
- Durational: a rhythmic pattern.

Half of the patterns were selected by the first author and the other half by Meredith et al.'s (2002) structural inference algorithm for translational equivalence classes (SIATEC), to cover a range of pattern plausibility. For each of these patterns, 29 features were calculated and used to predict pattern importance as rated by music students on a scale of 1-10 (where 1 is unimportant and 10 is very important). The final model, explaining 71% of their data, included three features: (1) compactness, the ratio of number of notes in the pattern to total number of notes in the texture while the pattern occurs; (2) expected occurrences, the likelihood of

occurrence of a given pattern based on empirical distributions representing the music; formula given by Conklin & Bergeron (2008); and (3) compression ratio, equal to coverage divided by the sum of cardinality and number of occurrences minus 1, where coverage is the number of notes in the analysed piece(s) of music that are included in the given pattern and cardinality is the number of notes in a pattern (Formula 6.1). This model suggests that the more compact and the more often a pattern occurs, the more likely it is to be salient to a listener. This supports the intuition that less information is easier to remember, leading to the conclusion that a short and memorable pattern will be more salient.

$$\frac{\textit{coverage}}{(\textit{cardinality} + \textit{occurrence}) - 1} \quad (6.1)$$

Second, Dibben (1999) and Lerdahl (1989) explore salience in music without tonal hierarchy, arguing that the stability of tonal music subsumes salience where musical structure is concerned. Salience here refers to attention-drawing musical mechanisms that define a piece's structure, where there is no tonality to do so. In other words, if a piece of music is tonal, tonality provides structure; if not, salience provides structure in place of tonality. Several features of the musical surface (as opposed to hierarchical structure) are defined as carriers of salience. Lerdahl develops a rather sophisticated hierarchy of these features that he suggests are inferred by listeners. It is not clear however how this hierarchy was established and is presumably based on Lerdahl's own intuition; it should be tested empirically before it can explain perception in a reliable way. There are more parameters than need to be mentioned here, but here are a few of the most straightforward: in this hierarchy, notes will be more salient if they are in an extreme registral position or parallel to a choice made elsewhere in the piece (parallelism), slightly less if they are relatively loud, long or timbrally prominent and even less if they are in a relatively strong metrical position. Dibben also identifies register and parallelism as carriers of salience, and summarizes other attention-drawing parameters

suggested by Lerdaahl as phenomenal accents (i.e. a loud note). Unlike Collins, this approach to salience deals with individual events and considers events that are different from their context to be salient as opposed to simple, repeated patterns.

Third, Prince et al. (2009) link salience to complexity, arguing that increased feature complexity requires higher processing demands, leading to increased allocation of attention to that feature – in other words, that feature becomes salient. This conclusion is derived from a set of two goodness-of-fit probe tone experiments, and four speeded classification tasks. In these, tonal and metrical hierarchies of two-bar contexts and the tonal and metrical relationships between these contexts and a probe were manipulated both together and independently to evaluate each parameter's impact on goodness-of-fit ratings or classification performance. Goodness of fit ratings ranged from 1 (poor fit) to 7 (good fit) and the classification tasks asked participants whether a probe tone was on or off beat, or in or out of key, depending on the parameter under investigation. Results showed that pitch class of the probe tone better predicted goodness-of-fit ratings than its temporal position, even when listeners were instructed to ignore pitch. In the speeded classification tasks, temporal classification was biased by tonal relationships but pitch classification was not affected by temporal position. To explain this asymmetry, the authors apply their definition of salience as it relates to complexity: in Western music, there are many more commonly used different pitches (not pitch classes) than there are different commonly used durational values, therefore pitch is more complex, requires more attention to process, and is more salient. This of course ignores octave equivalence; if octave equivalence was taken to make two pitches absolutely equal perceptually, then the overall pitch options in Western music are reduced to the 12 pitch classes of the chromatic scale. This would render pitch and rhythm almost equal in complexity. However, two pitches an octave apart are still two different frequencies and, assuming neuro-

typical perception, can be recognized as two different pitches, alongside the recognition of belonging to the same pitch class. Thus, the argument for pitch superiority can be considered convincing, particularly as Western music is heavily melody-based, and even atonal music is pitch-based (for example serialism). Unlike the previous two, this definition of salience considers the contents of a given parameter isolated from the others and over the span of a whole piece of music.

The differences between these definitions have already quickly been highlighted: for Collins et al. (2011), simplicity and repetition is salient, for both Dikken (1999) and Lerdahl (1989), extreme events are salient, and for Prince et al. (2009), complexity is salient. However, an investigation into the broader psychology literature on salience sheds some light on these discrepancies and proposes a definition that supersedes all those above. The school of Gestalt psychology first defined salience as contrast (Wundt, 1874), where the more different an object is from its context, the more salient it will be perceived to be. Spanning research in vision (Fink, Marshall, Halligan, & Dolan, 1998; Hoffman & Singh, 1997; Parkhurst, Law, & Niebur, 2002), attention (Horstmann, Becker, & Ernst, 2016; Summerfield & Eger, 2009) and language acquisition (Ellis, 2006; Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006), the concept of salience was still both varied and ill-defined. A recent special edition of *Frontiers in Psychology* (Blumenthal-Dramé et al., 2017) addresses this issue. A number of papers suggest that different definitions may simply reflect different aspects of salience: top-down salience or bottom-up salience, and that these can be explained as a function of expectation (Blumenthal-Dramé, Hanulíková, & Kortmann, 2017; Horstmann et al., 2016; Jaeger & Weatherholtz, 2016; Schmid & Günther, 2016). In this interpretation, bottom-up salience is a result of surprise in response to an unexpected event while top-down salience is a result of confirmed expectation, where an object, or a word, is salient if one expects that object or word

because it was recently mentioned, imagined, or part of a routine for example. This is supported by Ellis' (2006) explanation of some of the challenges of second language acquisition, particularly grammatical subtleties and link words. Ellis argues that these aspects of language are so common (i.e. the 's' in the third person conjugation) that they become dropped by first language speakers because they are implied. This lack of salience, driven by high frequency, makes second language acquisition more difficult for these aspects of language, where less common vocabulary is spoken more clearly and is therefore more salient and better understood. Also in language processing research, Zarcone et al. (Zarcone, van Schijndel, Vogels, & Demberg, 2016) integrate salience into the predictive coding framework, where predictability and attention at multiple processing levels, already incorporated into the framework, influence perceived salience. For example, a highly predictable event outside of attention will not be salient, while the same type of event inside the focus of attention will be salient. In vision research, salience has been measured by reaction and fixation times using eye tracking, where faster reaction times indicate higher salience and salience is correlated with higher contrast, supporting the bottom-up aspect of salience perception (Fink et al., 1998; Hoffman & Singh, 1997; Parkhurst et al., 2002). Though this unifying, expectancy-based definition is encouraging, it is very new and more research is needed to validate it; this chapter contributes to this need.

Jaeger & Weatherholtz (2016) specifically suggest that computational models of expectation are the way forward in salience research; in the present chapter this idea is applied to a musical context. The expectancy framework, with its two contrasting types of salience, largely resolves discrepancies between the definitions discussed from the music cognition literature. Each of those definitions relies on some type of information content calculated (algorithmically) or extracted (by a human annotator) for a given feature. In the case of Collins

et al. (2011), as a listener (real or algorithmic) experiences a repeated pattern, that pattern will become more predictable and IC for that pattern will decrease. Alternatively, IC can also replace compactness by measuring similarity through compression (Pearce & Müllensiefen, 2017). In the case of Dibben (1999) and Lerdahl (1989), IC for viewpoints such as loudness, timbre and register (though these do not yet exist in IDyOM) would be high for extreme events as these are also unlikely, thus identifying these events as surprising and salient. In the case of Prince et al. (2009), IC is a potential measure of complexity, where high IC reflects unexpectedness and higher unexpectedness equates to more complexity, and therefore salience. This is similar to Dibben and Lerdahl's definitions, where low IC reflects expectedness and lack of complexity, and therefore lack of salience. Again, while these definitions are contrary to Collins', the expectation framework can resolve this: these studies are addressing different aspects of salience, where Collins et al. considers top-down salience, Dibben, Lerdahl and Prince et al. address bottom-up salience. Furthermore, the scale of application of each definition of salience differs, where Collins et al. refers to patterns repeated throughout a piece of music, Prince et al. to parameters spanning a piece of music and Dibben and Lerdahl to individual events in relation immediate context. Finally, the paradigms used in Collins et al. and Prince et al. differ, where Collins et al. investigates the effect of sequential repetition while Prince et al. investigates the simultaneous perception of multiple lines. As the goal of this chapter is to investigate the relative salience of musical parameters in polyphonic music, which involves simultaneous perception and bottom-up salience, the predictions associated with Prince et al. (2009)'s definition of salience will guide its hypotheses.

7.1.1 Operational definitions

Given that information content provides a common thread to the music cognition definitions of salience given above and can represent both types of salience suggested, the following operational definition of salience is proposed:

Salience is proportional to information content (1)

To validate this definition, it is decomposed into two sub-definitions:

Information content is proportional to complexity (2)

Complexity is proportional to salience (3)

This serves two purposes: (1) to methodologically break the statement down into smaller, more concentrated studies and (2) to provide empirical evidence linking complexity to information content as calculated by IDyOM, theorized here and elsewhere (Eerola, 2016; Huron, 2006) but not yet tested.

The next section will present the stimuli that will be used for two studies, one for each sub-definition, described in Sections 7.3 and 7.4 respectively, while Section 7.5 summarizes and discusses the results of both in the context of the wider salience and music perception literature. The first of these studies tests the hypothesized link between information content and complexity by asking participants to rate the perceived complexity of short 3-voice musical excerpts specially composed for these studies and manipulated in terms of melodic, harmonic and rhythmic information content. The second tests the hypothesized link between complexity and salience by asking participants to listen to pairs of these same excerpts and identify whether the middle voice, called *target melody*, is the same or different. Both simple and complex target melodies were created to test two types of relationship between the outer and middle voices: middle voice less complex than outer voices, and middle voice more complex than outer voices (details in Section 7.4).

7.2 Materials

A total of 24 basic stimuli were designed, 8 for each of the three musical parameters investigated: melody, harmony and rhythm. These eight stimuli represent eight objective levels of complexity where information content, as measured by IDyOM, is progressively larger as levels increase. Each two-bar excerpt is written for three voices, where only the two outer voices are manipulated according to information content and the middle voice (target melody) remains mostly static: there are four versions of the simple target and four versions of the complex target, each differing by only one note (see Figures 7.1 and 7.2). These differences are important for the same-different paradigm in which participants are asked to identify whether a pair of stimuli are the same or different (details in Section 7.4). Two of these versions contain all in-key notes and two contain out-of-key notes according to the implied harmony of the target melody alone; however, in-key notes and out-of-key notes could become out-of-key or in-key depending on the harmonic context they are set in. Each of the 24 basic stimuli were created in four versions, each with a different target melody, for each of the simple and complex targets, for a total of 192 different two-bar musical excerpts. Each was rendered to an audio file, with a violin sound applied to the upper voice, a clarinet sound applied to the target melody, and a bassoon sound applied to the lower voice. These instruments were chosen so

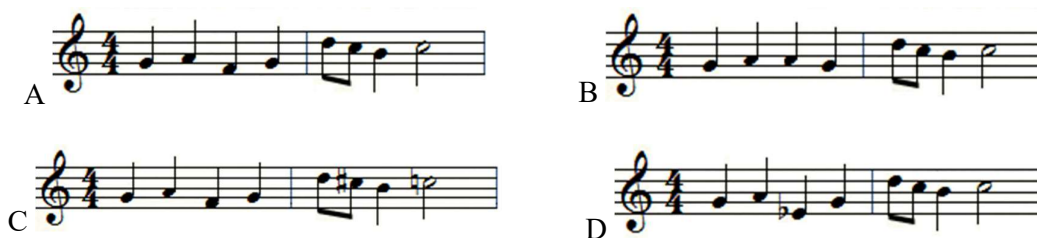


Figure 7.1. Simple target stimuli (A), all in the key of C, with each of the following modifications: A instead of F in the third beat of bar 1 (B), C# instead of C in the second half of the downbeat of bar 2 (C), and E^b instead of F in the third beat of bar 1 (D).

that each outer/inner voice timbral pair would be roughly equally perceptually dissimilar, as found from collecting timbre dissimilarity ratings in a previous study (Chapter 5, Figure 5.4; Sauv e, Stewart, & Pearce, 2014). Other combinations piloted include oboe as the target melody, with violin the upper and bassoon in the lower voices, and piano for all voices, but these combinations were too easy and too difficult, resulting in ceiling and floor performance, respectively. The combination of timbres used in the present studies resulted in a mean pilot performance around 75%.

The construction of these stimuli, including the information content details of the target melodies, and the manipulations of the outer voices in terms of melodic, harmonic and rhythmic information content will now be described.

7.2.1 Simple target stimuli

All melodic information content was measured using IDyOM, using pitch interval and scale degree source viewpoints to predict pitch. IDyOM was trained on soprano lines from 185 Bach chorales, a collection of 152 Nova Scotia folk songs, and the German *fink* sub-collection of the Essen Folk Song Collection, which consists of 566 songs. Both the long- and short-term models were engaged. All rhythmic information content was measured using the inter-onset-interval source viewpoint to predict onset. All other model parameters were held constant. The

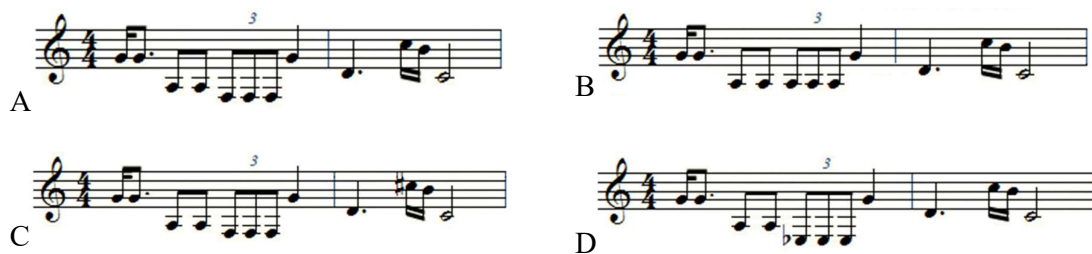


Figure 7.2. Complex target stimuli (A), all in the key of C, with each of the following modifications: A instead of F in the third beat of bar 1 (B), C# instead of C in the second half of the downbeat of bar 2 (C), and E^b instead of F in the third beat of bar 1 (D).

mean melodic information content of simple target melodies A, B, C and D (Figure 7.1) was 3.08, 3.63, 4.52, and 4.45 respectively and the mean rhythmic information content was 1.59 for all versions. Though there is some variation in IC between these four versions of the target, the range and variation are smaller than the possible values of mean melodic IC for the eight levels of complexity presented in Section 7.2.3. Furthermore, any potential effect of this difference in target stimuli IC will be included in the analysis. Mean harmonic information content is not calculated for this monophonic line.

7.2.2 Complex target stimuli

Complex target stimuli contain the same pitch classes as simple stimuli, where some pitches have simply been shifted down one octave. Using the same models, the mean melodic information content of complex target melodies A, B, C and D (Figure 7.2) were 5.61, 6.55, 8.66, and 8.13 respectively and the mean rhythmic information content was 3.70 for all versions.

7.2.3 Melodic complexity levels

The outer voices of these excerpts were designed to vary in mean information content between each of eight levels, so that level 1 had the lowest mean IC and level 8 had the highest. Mean melodic information content for each complexity level ranged from 2.79 to 7.62 (SD = 1.57). Figure 7.3 illustrates the mean information content for all parameters for all 24 basic stimuli, for each of simple and complex target melodies. Increased IC was reflected in larger intervals and more out-of-key pitches, which is logical considering the source viewpoints used. Harmonic complexity varied across levels (range = 4.20-8.87, SD = 1.78) while rhythmic complexity did not, IC = 1.59. See Figure 7.4 for the excerpts.

7.2.4 Harmonic complexity levels

Harmonic complexity was also measured using IDyOM. Chord progressions were four chords long, two beats per chord, with information content of chord transitions determined by training IDyOM on chord progressions from the Montreal Billboard Corpus (see Section 3.2.1 for details), where chords were encoded as integers to simulate MIDI pitch. The long- and short-term models were both engaged. Possible chords included major, minor, seventh, and various extension chords in almost any inversion. Each stimuli's chords were encoded as a series of four integers, one for each chord. For example, the progression I – IV⁶ – V – I was encoded as 1 – 55 – 3 – 1. This type of input simulates the existing cpitch viewpoint in IDyOM. Average harmonic information content for each complexity level ranged from 5.67 to 9.55 (SD = 1.89), where higher information content was reflected in rarer chords and chord transitions. Melodic complexities were fairly equal across levels (range = 2.35-4.04, SD = 0.66) while rhythmic complexity remained fixed, IC = 1.59. The close relationship between melody and harmony makes complete independence impossible, though as illustrated in Figure 7.3, some separation is possible. This relationship will affect interpretation of the results, where a significant effect of harmonic IC implies a related influence of melodic aspects of the music. An effect of melodic IC in this polyphonic context would also imply some influence of harmony. See Figure 7.4 for the stimuli.

7.2.5 Rhythmic complexity levels

IDyOM's inter-onset interval (IOI) viewpoint was used to predict note onset for the outer voices of each stimulus, using the same training set and parameters as the melodic complexity stimuli. Average rhythmic information content for each complexity level ranged from 1.49 to 3.04 (SD = 0.53), where higher information content was reflected in greater diversity of IOI values, or decreased repetition. The melodic and harmonic complexities

remained fairly equal across levels for both melody and harmony, ranging from 3.55 to 4.32 (SD = 0.28) and 6.13 to 9.13 (SD = 1.78) respectively. See Figure 7.3 for an illustration of relative parameter IC and Figure 7.4 for the excerpts.

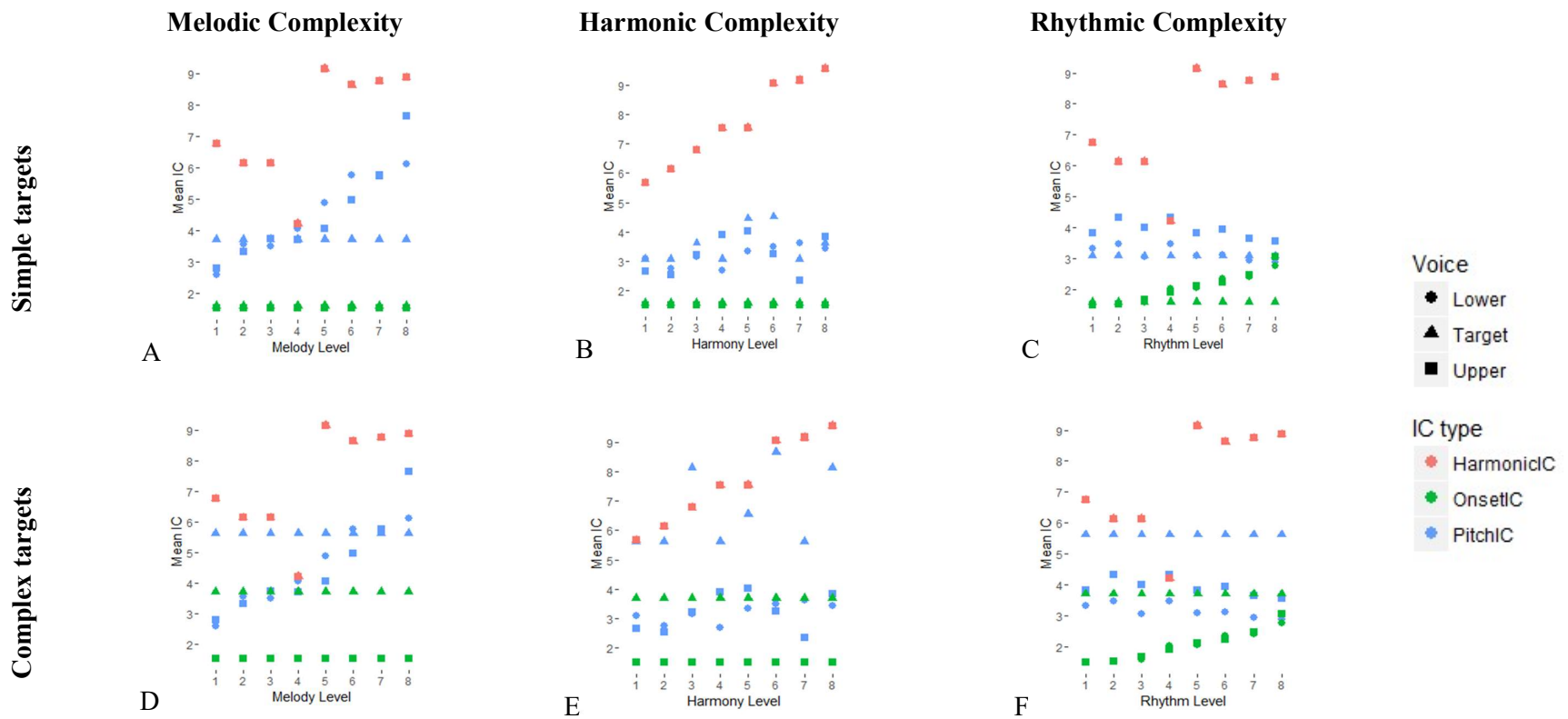


Figure 7.3. The mean IC of each voice plotted for each of the 48 stimuli by type of manipulation and type of target. The manipulation of melodic information content in the outer voices as compared to the target melody is clear in the first column, with more diffuse distributions in last two columns and the manipulation of rhythmic information content is similarly clear in the third, with a flat distribution in the first two columns. The manipulation of harmonic information content is closely entangled with melody and is therefore difficult to manipulate completely independently, though some separation is possible; the outer columns have identical, more sporadic patterns than the central column.











	Melodic Complexity	Harmonic Complexity	Rhythmic Complexity
Level 1			
Level 2			
Level 3			
Level 4			

Figure 7.4. Scores of excerpts for complex target stimuli for Levels 1-4.

	Melodic complexity	Harmonic complexity	Rhythmic complexity
Level 5			
Level 6			
Level 7			
Level 8			

Figure 7.4 con'd. Scores of excerpts for complex target stimuli for Levels 5-8.

7.3 Linking information content to complexity

To test the hypothesis that objectively measured information content approximates subjective perceived complexity, complexity ratings on 96 3-voice stimuli were collected, using all 4 versions of each of the 24 basic stimuli set with a complex target melody. All four versions were used to verify that ratings were stable when only one note in a given excerpt was changed. Ratings ranged from 1 – 7, where 1 was not complex and 7 was very complex. Participant ratings are expected to correlate with information content for melodic, harmonic and rhythmic IC manipulations, where high IC will yield higher complexity ratings.

7.3.1 Participants

Data was collected from 28 participants (12 female), mean age 43.03 (SD = 16.34) and mean Gold-MSI musical training subscale (see Chapter 3, Section 3.3 for details) score 36.60 (SD = 9.31). Participants were recruited through musicology and psychology mailing lists and social media. Ethical approval was obtained from the Queen Mary Research Ethics Committee, QMREC1536a.

7.3.2 Procedure

Data was collected via online survey tool Qualtrics. Participants first read through the information sheet and provided consent before reading the instructions and answering two practice trials to familiarise themselves with the type of stimuli and form an idea of their complexity. Participants were not given any definition of complexity but rather instructed to decide for themselves what it meant to them when listening to these stimuli. This way, if a relationship is found between information content and complexity, it is not because participants were told it existed. They were encouraged to use the full range of the complexity scale and to perform these ratings by judging complexity of an excerpt in relation to the other excerpts

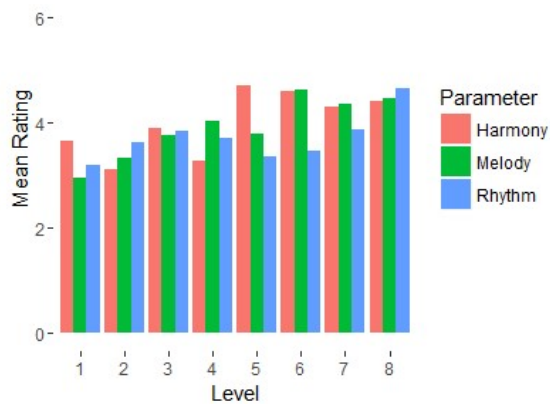


Figure 7.5. Mean complexity ratings for each level by parameter.

rather than in relation to other music they know. Each participant rated 48 excerpts, where two from each set of four modifications were randomly selected. The 48 excerpts were divided into three blocks of 16 excerpts, one for each of the melodically, harmonically and rhythmically manipulated stimuli.

These 16 excerpts were presented in random order and the presentation order of the blocks was also randomized.

7.3.3 Analysis

Primary analysis was performed in R (3.3.2) on mean complexity ratings, averaged across the four versions and across participants for each of the 24 basic stimuli. Mean complexity ratings are plotted in Figure 7.5. This primary analysis consisted of three multiple linear regression models answering three questions about the relationship between the stimuli and the collected complexity ratings. These are specified in Section 7.3.4 below. These models included fixed effects only and were constructed according to the question being posed. Variance explained by each model is given by R-squared, calculated by a correlation test between the model's predictions and the data. The overall F-statistic of the model is also given. Statistical significance of each predictor is given by the result of a likelihood-ratio test between a null model (intercept only) and a model containing the single evaluated predictor. Statistical significance of each individual factor level for a given predictor is evaluated by the *t* statistic

given by the lme4 package for a multiple linear regression model, and by the 95% confidence intervals, where an interval not including zero indicates a significant predictor.

The influence of musical training on complexity ratings was also evaluated for each analysis question posed. This was done by adding a fixed effect for training, reflected by Gold-MSI scores for each participant, where each trial's response was treated individually (long format data). In this format, and to model with maximal effects in accordance with the experimental design (Barr et al., 2013), random intercepts on participant and stimulus number were added to all models including musical training to create multiple linear mixed effects models. To evaluate the significance of musical training, mixed effects models with and without the predictor were compared using a log likelihood ratio test.

Finally, the multiple linear mixed effects model was evaluated using a correlation test between the model's predictions and the data.

7.3.4 Results

As described in Section 7.3.3 above, multiple linear regression analyses were carried out to answer three questions about the relationship between the stimuli and the complexity ratings. The influence of musical training (fixed effect) was then evaluated for each question.

First, does the objective manipulation of complexity predict subjective ratings? For this model, *objective complexity* (1-8), *parameter* (melody, harmony or rhythm) and *version* (four types of target melody) were included as predictors of mean ratings. For objective complexity, all factor levels were compared to Level 1, for parameter factor levels were compared to rhythm and for version, comparisons were made to Version A. All but the second and fourth levels of objective complexity had a significant impact ($t(83) = -0.19, p = .27$, $t(83) = 0.43, p = .01$, $t(83) = 0.27, p = .12$, $t(83) = 0.54, p = .002$, $t(83) = 0.82, p < .0001$, $t(83) = 0.76, p < .0001$, $t(83), 1.10, p < .0001$ for levels 2 through 8 respectively) along with one factor

Table 7.1. Summary of the fixed effects manipulation model, including coefficients, 95% confidence intervals and R² of each predictor. Each factor of objective complexity is in relation to Level 1; each factor of parameter is in relation to rhythm; and each factor of version is in relation to version A.

Predictor	Coefficient	2.5%	97.5%	R ²
(Intercept)	3.25	2.93	3.57	-
Level 2	-0.19	-0.54	0.15	
Level 3	0.43	0.07	0.78	
Level 4	0.27	-0.07	0.62	
Level 5	0.54	0.19	0.89	.68
Level 6	0.82	0.46	1.17	
Level 7	0.76	0.41	1.12	
Level 8	1.10	0.75	1.45	
Melody	0.19	-0.01	0.41	.03
Harmony	0.27	0.06	0.49	
Version B	-0.13	-0.38	0.11	
Version C	-0.02	-0.27	0.22	.01
Version D	0.07	-0.17	0.32	

of parameter, harmony ($t(83) = 2.55, p = .01$ for harmony and $t(83) = 1.86, p = .07$ for melody); no level of version was significant (all $p > .05$). Overall, objective complexity was a significant predictor, $F(7, 88) = 11.36, p < .0001$, parameter was not, $F(3, 93) = 1.89, p = .15$ and version was not, $F(3, 92) = 0.47, p = .69$. This model has an R² of .72 and $F(12, 83) = 7.78, p < .0001$. The addition of *musical training* to predict ratings in long format marginally improved a model without it, $\chi^2(1) = 3.58, p = .05$, but the model's fit to the data was low, $r^2 = .27, t(2686) = 14.94, p < .0001$. A summary of the fixed effects model predicting mean ratings and of the maximally fitted mixed effects model predicting individual ratings can be found in Tables 7.1 and 7.2 respectively.

The second question is whether complexity is accurately simulated by information content? This model included *mean melodic information content*, *mean harmonic information content* and *mean rhythmic information content*, based on the mean IC of the outer voices

Table 7.2. Summary of the maximally fitted manipulation model, including coefficients, 95% confidence intervals and R^2 of each predictor. Each factor of objective complexity is in relation to Level 1; each factor of parameter is in relation to rhythm; and each factor of version is in relation to version A.

Predictor	Coefficient	2.5%	97.5%	R^2
(Intercept)	2.72	2.22	2.96	-
Level 2	-0.10	-0.48	0.27	
Level 3	0.22	-0.15	0.60	
Level 4	0.11	-0.25	0.49	
Level 5	0.22	-0.15	0.60	-.01
Level 6	0.50	0.02	0.78	
Level 7	0.34	-0.04	0.72	
Level 8	0.54	0.16	0.92	
Melody	0.11	-0.11	0.34	.00
Harmony	0.12	-0.11	0.35	
Version B	0.01	-0.25	0.27	
Version C	-0.05	-0.33	0.21	-.00
Version D	1.16	-0.10	0.43	
Gold-MSI	-0.15	-0.32	0.00	.00
	Predictor	Variance		
Random intercepts	Participant	0.15		.20
	Stimulus	0.06		.08
Residual variance		4.53		

calculated from a selection of pitch and timing viewpoints (see Section 7.2.1 for details). As the range of mean information content for these predictors varies (mean melodic IC range = 2.35 – 7.62; mean harmonic IC range = 4.20 – 9.55; mean rhythmic IC range = 1.27 – 3.04), each predictor was transformed into z-scores (mean = 1, SD = 1) so that the mean melodic IC range became -1.42 – 3.48, mean harmonic IC range became -2.03 – 1.32 and mean rhythmic IC range became -0.55 – 3.20. To confirm its influence, or lack thereof, melodic IC of the target voice was also included as a predictor (rhythmic IC was not included because it does not vary). Melody and harmony IC predictors were significant ($F(1, 94) = 16.18, p = .0001$ and

Table 7.3. Summary of the fixed effects information content model, including coefficients, 95% confidence interval and R² for each predictor.

Predictor	Coefficient	2.5%	97.5%	R ²
(Intercept)	3.86	3.75	3.96	-
Melody IC	0.17	0.06	0.28	.38
Harmony IC	0.22	0.11	0.33	.15
Rhythm IC	-0.02	-0.12	0.08	.00
Target Pitch IC	-0.02	-0.12	0.08	.00

Table 7.4. Summary of the maximally fitted information content model, including coefficients, 95% confidence interval and R² for each predictor.

Predictor	Coefficient	2.5%	97.5%	R ²
(Intercept)	2.93	2.75	3.10	-
Melody IC	0.09	-0.00	0.19	.00
Harmony IC	0.11	0.01	0.21	-.01
Rhythm IC	-0.01	-0.11	0.08	.00
Target Pitch IC	0.03	-0.06	0.12	.00
GoldMSI	-0.15	-0.32	0.00	.00
Predictor		Variance		
Random intercepts	Participant	0.15		.20
	Stimulus	0.06		.08
Residual variance		4.54		

F (1, 94) = 23.05, p < .0001 respectively) but not rhythm IC nor melodic IC for the target voice (F (1, 94) = 0.09, p = .75 and F (1, 94) = 0.11, p = .73 respectively), with an R² of .53 and F (3, 92) = 12.03, p < .0001. The addition of *musical training*, predicting ratings in long format, marginally improved the model, $\chi^2(1) = 3.57$, p = .05, but correlation to the data was low, r² = .27, t(2686) = 14.96, p < .0001. A summary of the fixed effects model, predicting mean ratings,

and of the maximally fitted mixed effects model, predicting individual ratings, can be found in Tables 7.3 and 7.4 respectively.

The third question is: what musical properties might correspond to the variations in complexity ratings? In other words, how can the differences in complexity perception be characterised in musical terms? For this model, the *mean interval size* (in semitones; range 0.41 – 8.41) of the two outer voices of each stimuli, the *mean pitch* (MIDI; range 72.14 – 78.14 for the upper voice and 43.85 – 51 for the lower voice) of each of the two outer voices separately, the *mean note duration* (in milliseconds, where 1 beat = 24ms; range 16 – 27.42) of the two outer voices, the proportion of out-of-key notes in relation to the total number of pitches in the two outer voices, *key proportion* (range 0 – 0.38) and *syncopation score*, where a lower score equates to more syncopation (Lerdahl & Jackendoff, 1983) were calculated. Mean duration and key proportion were significant predictors in this model, though only key proportion was significant when tested against a null model, $F(1, 94) = 50.63, p < .0001$. Mean duration has significance only in the context of the rest of the model, $t(89) = -3.21, p = .001$. The model overall has a total R^2 of .68 and $F(5, 90) = 12.53, p < .0001$. The addition of *musical training* marginally improved the model, $\chi^2(1) = 3.588, p = .05$, but correlation to the data remains low, $r^2 = .27, t(2686) = 14.75, p < .0001$. A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.5 and 7.6 respectively.

Table 7.5. Summary of the fixed effects musical properties model, including coefficients, 95% confidence intervals and R² for each predictor.

Predictor	Coefficient	2.5%	97.5%	R²
(Intercept)	8.83	1.46	16.20	-
Key proportion	3.59	2.65	4.53	.59
Mean Duration	-0.11	-0.18	-0.04	.06
Syncopation Score	-0.06	-0.14	0.006	.01
Mean Interval Size	-0.02	-0.09	0.04	.00
Mean Pitch – Upper voice	-0.02	-0.09	0.05	.00
Mean Pitch – Lower voice	0.06	-0.006	0.13	.02

Table 7.6. Summary of the maximally fitted musical properties model, including coefficients, 95% confidence intervals and R² for each predictor.

Predictor	Coefficient	2.5%	97.5%	R²
(Intercept)	4.84	-2.75	12.43	-
Key proportion	1.77	0.80	2.74	-.01
Mean Duration	-0.05	-0.12	0.01	.00
Syncopation Score	-0.03	-0.10	0.04	.00
Mean Interval Size	-0.01	-0.08	0.05	.00
Mean Pitch – Upper voice	-0.00	-0.08	0.07	.00
Mean Pitch – Lower voice	0.03	-0.03	0.11	.00
GoldMSI	-0.15	-0.32	0.00	.00
	Predictor	Variance		
Random intercepts	Participant	0.15		.20
	Stimulus	0.05		.08
Residual variance		4.53		

7.3.5 Discussion

In order to validate the operational definition stating that information content is proportional to complexity, complexity ratings were collected for a series of stimuli manipulated in terms of melodic, harmonic and rhythmic information content as calculated by IDyOM. The primary analysis results support the definition, as both the experimental manipulations, based on information content, and raw information content successfully predicted mean ratings, explaining 72% and 53% of the variance in the data respectively. When random effects are added to account for individuals and stimulus, the fixed effects lose almost all their predictive power, but do not consistently explain more of the data. Musical training did not significantly improve any models when added as a fixed effect. These additions will be addressed in more detail below. While the information content model demonstrates a correlational relationship between information content and complexity, the explicit manipulation of information content in the experimental design crucially also provides causal evidence. The coefficients and the relative R^2 of these models can be interpreted to draw some interesting conclusions.

First, the intercepts are both close to the centre of the rating scale, indicating a slightly lower than mid-scale baseline rating. Continuing with the manipulations model, there is an imperfect, yet steady increase in the coefficients for every factor level of the *objective complexity* predictor, and positive coefficients for both factor levels of the *parameter* predictor. In the case of objective complexity, this equates to stimuli with higher objective complexity resulting in higher subjective complexity ratings as compared to Level 1, as expected. In the case of parameter, these coefficients indicate that melodic and harmonic complexity stimuli receive higher complexity ratings overall than rhythmic complexity stimuli, with harmony being the highest and significantly different from those stimuli manipulated in terms of rhythm.

This suggests that listeners consider unpredictable harmonic progressions to be more complex than unpredictable melodies (e.g., large leaps or out-of-key notes), and these more complex than rhythmic variety. However, note that the explanatory power of parameter is negligible compared to objective complexity. Finally, *version* was not a significant predictor, supporting the assumption that the four versions of each stimulus would be rated similarly. The information content model provides somewhat conflicting evidence for the interpretation of the above predictor parameter, where melodic IC carries the most explanatory power, followed by harmonic IC and finally rhythmic IC, with the same pattern of magnitude seen in their coefficients. However, as mean information content is a more fine-grained measure than objective complexity level and parameter, melody IC captures some of the features of unpredictable harmonic progressions, such as out-of-key pitches, resolving the discrepancy between the results of the two models. Thus far, rhythmic complexity is rated as simpler, and onset information content negligible when explaining ratings.

Several raw musical properties of the stimuli were also tested for potential predictive power to attempt to locate the musical features the models and the listeners are responding to when assessing complexity. Mean interval size for both outer voices, mean pitch of each outer voice, mean note duration for both outer voices, the proportion of out-of-key notes among both outer voices and the degree of syncopation of both outer voices was calculated. Together, these represent melodic, harmonic and rhythmic dimensions of music. Analysis on mean ratings reveals an effect of note duration and proportion of out-of-key notes. Both are in the expected direction, where higher proportions of out-of-key notes lead to higher complexity ratings, and longer mean durations, corresponding to stimuli in the lower levels, result in slightly lower complexity ratings. The proportion of out-of-key notes not only provides the most explanatory power, its coefficient combined with the intercept add up to a complexity rating of

approximately 6.5, near the top end of the rating scale. Knowing that the range of the key proportion predictor is small (ranging from 0 – 0.38), this interestingly puts all ratings quite high on the scale when ratings are predicted by a model based on musical features. The importance of the out-of-key versus in-key proportion is in line with the two previous models where experimental manipulations and information content are predictors: the composed unexpected harmonic progressions contain more out-of-key pitches, which is accounted for in melodic information content since these will have low probability in context.

There was, perhaps surprisingly, no effect of musical training on ratings, where it might be expected that increased exposure to music would yield better models and therefore lower information content, and lower perceived complexity. Additionally, when random effects for participants were included in the mixed effects models, this explained the majority of the variance. Thus, it can be concluded that individual differences supersede any effects of experimental manipulation, information content, musical features or musical training on ratings, and in the cases of information content, increase the predictive power of the model. However, when results are collapsed across participants and more general patterns are considered, these same experimental manipulations, measures of information content and musical features explain at least half the variance in the data in each case, up to as much as 72% in the case of experimental manipulations.

In summary, we find that the results demonstrate a strong link between information content and perceived complexity and furthermore, that harmonically manipulated stimuli, melodic information content and out-of-key notes – all closely related – have the largest influence on complexity ratings, followed by melodic manipulations and harmonic information content and finally rhythmic manipulations and information content, where other melodic and rhythmic musical features have negligible influence.

7.4 Linking complexity to salience

In the above study, information content was validated as a reasonable proxy for perceived complexity, providing evidence for operational definition 2: information content is proportional to complexity. The current study will test operational definition 3: complexity is proportional to salience. In accordance with Prince et al. (2009), it is hypothesized that more complex stimuli will be more salient, as they demand more attention to process. This relationship will be tested with a modified version of a streaming paradigm used by Marozeau and colleagues (Marozeau et al., 2013). It is worth emphasizing that raw information content is not a perfect proxy for perceived complexity, only explaining half of the variance in the collected complexity ratings, while experimental manipulations of objective complexity (based on information content) for melody, harmony and rhythm explained 72% and musical features explained 68%. Nevertheless, a link between these measures of complexity and salience is worth exploring as there is theoretical (Jaeger & Weatherholtz, 2016; Prince et al., 2009) evidence for such a link as well as empirical evidence for a proxy link (Section 7.3).

In the Marozeau et al. paradigm, a four-note target melody was interleaved with a four-note pseudorandom distractor sequence and participants rated how easy it was to hear the target as the distractor was gradually manipulated over repeated exposure to the pattern. This distractor sequence was manipulated such that it masked the target more or less well based on how closely it matched the target melody on a given parameter. Intensity and spectral and temporal envelopes were manipulated to investigate effects of loudness and timbre on auditory streaming. For example, in the case of intensity, the distractor tones were manipulated where equal intensity between target and distractors made the task very difficult – because the two sequences were integrated – while a difference in intensity eventually allowed identification of

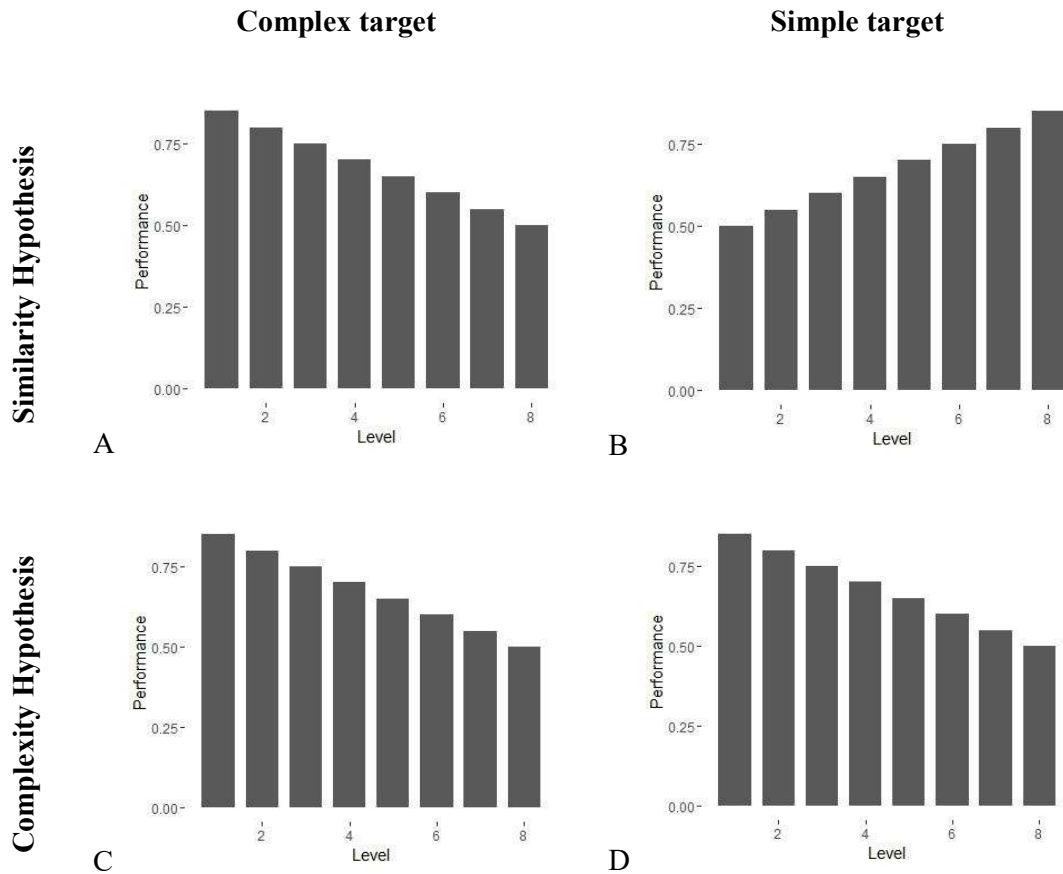


Figure 7.6. Theoretical task performance patterns according to the similarity (A, B) and complexity (C, D) hypotheses for complex (A, C) and simple (B, D) target conditions.

the target – because the sequences were segregated. The parameter distance between the target and distractor – the *threshold* – needed to perceive the target varied with musical training and with the parameter itself. The parameters that required the least amount of difference from the target – those that triggered segregation first – were considered the most salient.

In this study, the parameters are higher-level, but the same principle applies. In 3-voice, two-bar excerpts, melodic and rhythmic information content in the outer voices match or differ from the melodic and rhythmic information content of the target voice. Harmonic information content must be treated differently, as harmony applies to the voices as a unit rather than to individual voices (implied harmony is not considered here); it will be manipulated for the whole

excerpt. The task associated with the paradigm is also modified: instead of participants rating how easy it is to perceive the target melody, participants must decide whether pairs of stimuli are the same or different. In the Marozeau et al. (2013) paradigm, increased similarity between target and distractors made the task more difficult. In the present study, for the condition with a complex target and increasingly complex outer voices, this would result in a decrease in performance as outer voice complexity increases (Figure 7.6A), while for the condition with a simple target and increasingly complex outer voices, performance should increase as outer voice complexity increases (Figure 7.6B). On the other hand, based on the relationship between information content and complexity (Section 7.3), it is also possible that higher melodic and rhythmic information content in the outer voices – equating to higher complexity – will mask the target more effectively and make streaming of the target voice difficult, thus making a same-different judgment task between two excerpts difficult. This is because voices with high information content require more attention to process and become more salient, distracting from the target. If the outer voices are less salient than the target, as in the case of the complex target condition, the target will be easy to identify and performance will decrease with increased outer voice complexity (Figure 7.6C). Similarly, if the outer voices are more salient than the target, as in the case of the simple target condition, the target will be difficult to identify and performance will also decrease with increased outer voice complexity (Figure 7.6D). To disentangle these two contrasting interpretations in terms of similarity and complexity, a two by two factorial design was employed. To locate a threshold where the outer voices mask, or reveal, the target melody (as in Marozeau et al., 2013), eight levels of complexity for each of melodic and rhythmic dimensions were designed for the outer voices along with eight levels of harmonic complexity incorporating all three voices. Furthermore, as is illustrated in Table 7.7, the target melody is either simple or complex while the outer voices vary from simple to

Table 7.7. Illustration of experimental design, with levels of target and outer voice complexity. X's highlight the 2 x 2 factorial portion of the design.

		Outer voices complexity							
		Simple 1	- 2	- 3	- 4	- 5	- 6	- 7	Complex 8
Target melody complexity	Simple	X							X
	Comple x	X							X

complex. This creates situations where either all voices are relatively simple, all voices are relatively complex or a fully factorial comparison of the two. The four marked squares represent the 2x2 factorial part of the design. Figure 7.3 shows the relationship between target melody complexity and melodic, rhythmic and harmonic complexity for the outer voices of the 24 basic stimuli.

In the case of the effect of harmonic sequence information content on task performance, two possibilities exist. In the first case, as above, a simple harmonic sequence will allow the target melody to stand out while a complex harmonic sequence will detract attention. The alternative possibility is that the simple harmonic sequence creates such strong integration that the target is more difficult to pick out, while a complex harmonic sequence is more disjointed and allows the target melody to be perceived through the texture. It is worth noting that the integrating nature of harmony is likely to make the task more difficult overall in comparison to the original paradigm (Marozeau et al., 2013).

While outer voice complexity and manipulated musical parameter remained within-subject variables, target melody complexity was between-subjects, with participants completing the experiment with either the complex target or the simple target.

7.4.1 Participants

Data was collected from 245 participants (152 female, mean age 27.05, SD = 9.80) for the complex target version of the study and 90 participants (62 female, mean age 27.02, SD = 10.72) for the simple target version of the study. All participants only took part in one of the two versions and none had taken part in the complexity rating study (Section 7.3). Complex target participants had mean Gold-MSI musical training sub-scale scores of 24.70 (SD = 9.90) and simple target participants had a mean score of 26.28 (SD = 10.51), a non-significant difference, $t(151) = -1.31$, $p = .18$. Participants were recruited through musicology and psychology mailing lists and social media, as well as Slice the Pie, a crowdsourcing website⁶. Ethical approval was obtained from the Queen Mary Research Ethics Committee, QMREC1536a.

7.4.2 Procedure

Both complex and simple versions of the same-different paradigm were presented using online survey system Qualtrics; the procedure was the same for both. After reading the introduction and providing informed consent, participants had the opportunity to hear each of the four simple target melodies, followed by a short training session in which participants had to correctly identify each target (arbitrarily labelled 1, 2, 3 and 4) twice. Targets were presented in blocks of four, with each version present once in each block. If one or more of the four were labelled incorrectly, a new block was presented until each was correctly identified twice (not necessarily in sequential blocks). This was followed by two practice trials and 48 experimental trials, where in half the cases the two stimuli were the same and in half the cases they were different. Each trial consisted of two stimuli and the participant's task was to identify whether

⁶ <https://www.slicethepie.com/>

the target melody was the same or different. As there are four target melodies, there are three possibilities for the pairs of stimuli to be different; only one was randomly selected for each participant. There were three blocks of 16 trials, each containing the stimuli from a different parameter: melodic complexity, harmonic complexity and rhythmic complexity. The 16 trials were presented in random order within each block, and the presentation order of the three blocks was randomized to avoid order effects. Between blocks, participants could listen to the four target melodies and were offered a short break; they could resume when they felt ready. The study concluded with the eight questions of the Gold-MSI musical training sub-questionnaire and basic demographic questions.

7.4.3 Analysis

Analysis for this study follows a similar methodology to the complexity ratings analysis (see Section 7.3.3), where the effects of experimental manipulations, raw information content and musical features, as well as musical training are evaluated. In addition, complexity ratings for each stimulus from the previous experiment (Section 7.3) were tested as a predictor. This analysis differs in the type of model used: here, logistic regression models are employed, as the dependent variable is binomial, where 0 is an incorrect response and 1 is a correct response. Evaluation criteria for the models are therefore slightly different. While confidence intervals on model coefficients and model comparisons using likelihood ratio tests still apply, R^2 cannot be calculated. Instead, a combination of ΔAIC (with respect to a null model), ΔBIC (with respect to a null model) and residual deviance will be employed, where residual deviance will be evaluated in relation to the χ^2 value for the associated degrees of freedom for that model: the more deviance is larger than the χ^2 value, the better the model. ΔBIC will also be calculated for each predictor, as an equivalent measure to R^2 . Here, the ΔBIC value will be reported

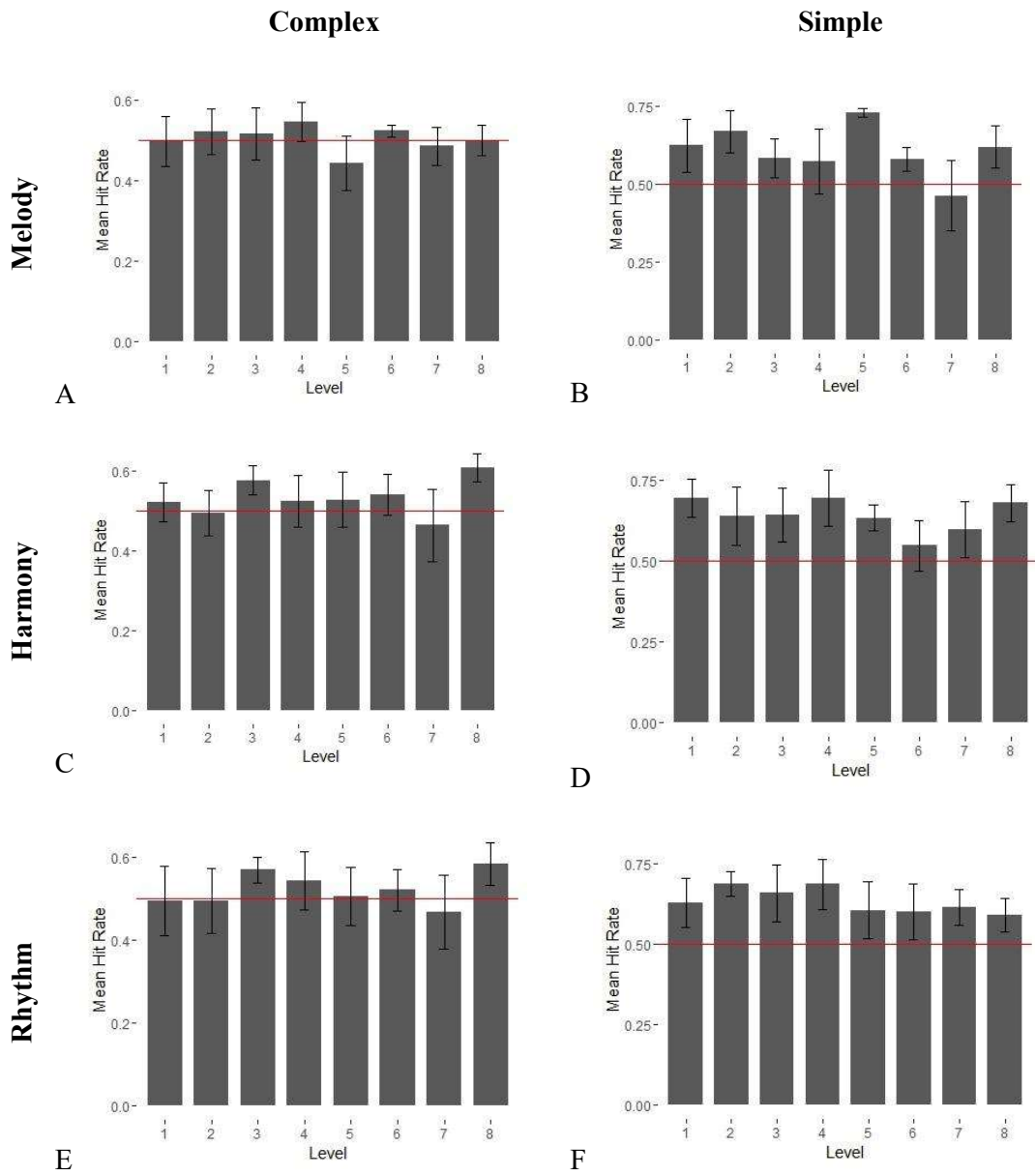


Figure 7.7. Mean hit rate (chance = 50%, red line) by objective complexity level for complex (A, C, E) and simple (B, D, F) targets for melody (A, B), harmony (C, D) and rhythm (D, F). Error bars represent standard error. In comparison to Figure 7.6, the hypotheses, no clear pattern can be seen or matched.

always in relation to the null model, with predictors added sequentially as they are presented in each table, with random effects before fixed effects where applicable.

For illustration purposes, mean hit rates were calculated for each trial (see Figure 7.7) as the number of correct responses divided by the number of participants.

7.4.4 Results

Mean hit rates for each trial (four measurements per trial, one for each version) are plotted in Figure 7.7. Initial visual inspection indicates that few conditions in the complex target condition achieved performance above chance (error bars include 50%) while performance in the simple condition exceeds chance (error bars do not include 50%) most of the time. All further analysis, as described in Section 7.4.3 above, will be carried out on individual responses.

Complex targets. Each of the 96 trials received at least one response, with ‘different’ trials receiving on average 80.73 responses each (SD = 7.08), corresponding to approximately 1/3 of total participants. The experimental manipulation model, with predictors *objective complexity*, *parameter* and *version*, yielded a significant effect of Level 4 complexity ($z(11614) = 2.24, p = .02$) and all versions ($z(11614) = -2.45, p = .01, z(11614) = -7.66, p < .0001$ and $z(11614) = -6.69, p < .0001$ for versions B, C and D respectively), with a ΔAIC of -74, ΔBIC of 15 and residual deviance of 15835 (df = 11614), $p < .0001$. Overall, neither objective complexity nor parameter were significant predictors, $\chi^2(2) = 5.53, p = .06$, and $\chi^2(7) = 10.41, p = .16$ while version was, $\chi^2(3) = 83.69, p < .0001$. The addition of *musical training* improved the model significantly, $\chi^2(1) = 50.08, p < .0001$. This model has $\Delta AIC = -179, \Delta BIC = -84$ and residual deviance of 15728 (df = 11565), $p < .0001$. Finally, the maximally fitted mixed effects model performed best overall with $\Delta AIC = -703, \Delta BIC = -560$ and residual variance of 15232 (df = 11611), $p < .0001$. One level of version and musical

Table 7.8. Summary of the fixed effects manipulation model for complex targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter). Each factor of objective complexity is in relation to Level 1; each factor of parameter is in relation to rhythm; and each factor of version is in relation to version A.

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.39	0.27	0.51	-
Level 2	0.06	-0.08	0.20	
Level 3	0.10	-0.04	0.25	
Level 4	0.17	0.02	0.32	
Level 5	-0.00	-0.15	0.14	55
Level 6	0.05	-0.09	0.20	
Level 7	0.00	-0.14	0.15	
Level 8	0.04	-0.11	0.19	
Melody	-0.06	-0.15	0.02	69
Harmony	0.04	-0.04	0.13	
Version B	-0.12	-0.21	-0.02	
Version C	-0.43	-0.54	-0.32	15
Version D	-0.37	-0.48	-0.26	
GoldMSI	0.13	0.09	0.17	-33

training were significant predictors in this maximal model, $z(11611) = -2.25$, $p = .02$ and $z(11611) = 4.18$, $p < .0001$ respectively. Random effects on *participant* explained more variance than random effects on *trial number* (see Table 7.9). A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.8 and 7.9 respectively.

The information content model, with predictors *mean melodic IC*, *mean harmonic IC*, *mean rhythmic IC* and *target IC* (all scaled), yielded a significant effect of harmonic IC only ($z(11622) = -2.14$, $p = .03$), with Δ AIC = 31, Δ BIC = 24, and a residual variance of 15920 ($df = 11622$), $p < .0001$. The addition of *musical training* improved the model significantly, $\chi^2(1) = 50.13$, $p < .0001$. This model has Δ AIC = -53, Δ BIC = -17 and residual deviance of 15870 ($df = 11621$), $p < .0001$. Finally, the maximally fitted mixed effects model performed best with

Table 7.9. Summary of the maximally fitted manipulation model for complex targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects). Each factor of objective complexity is in relation to Level 1; each factor of parameter is in relation to rhythm; and each factor of version is in relation to version A.

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.17	-0.16	0.50	-
Level 2	0.07	-0.29	0.44	
Level 3	0.15	-0.21	0.53	
Level 4	0.24	-0.13	0.61	
Level 5	-0.00	-0.37	0.37	-621
Level 6	0.20	-0.16	0.58	
Level 7	-0.09	-0.47	0.27	
Level 8	0.15	-0.21	0.52	
Melody	-0.08	-0.31	0.14	-574
Harmony	0.03	-0.19	0.26	
Version B	-0.02	-0.28	0.23	
Version C	-0.30	-0.57	-0.04	-552
Version D	-0.20	-0.47	0.05	
GoldMSI	0.15	0.08	0.23	-552
	Predictor	Variance		
Random intercepts	Participant	0.24		-273
	Trial number	0.17		-652

Δ AIC = -676, Δ BIC = -624 and residual variance of 15243 (df = 11619), $p < .0001$. Only musical training was a significant fixed effect predictor ($z(11619) = 4.18$, $p < .0001$) and random effects on *participant* explained more variance than random effects on *trial number* (see Table 7.11). A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.10 and 7.11 respectively.

Table 7.10. Summary of the fixed effects information content model for complex targets, including coefficients, 95% confidence interval and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.25	0.21	0.29	-
Melody IC	-0.02	-0.06	0.01	4.5
Harmony IC	-0.02	-0.08	-0.00	9.3
Rhythm IC	0.00	-0.03	0.04	18
Target Pitch IC	-0.03	-0.07	0.00	24
GoldMSI	0.13	0.09	0.16	17

Table 7.11. Summary of the maximally fitted information content model for complex targets, including coefficients, 95% confidence interval and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.11	-0.00	0.22	-
Melody IC	0.00	-0.10	0.11	-643
Harmony IC	-0.06	-0.17	0.04	-634
Rhythm IC	0.04	-0.06	0.15	-625
Target Pitch IC	0.00	-0.10	0.10	-616
GoldMSI	0.15	0.08	0.23	-17
	Predictor	Variance		
Random intercepts	Participant	0.24		-273
	Trial number	0.20		-652

The musical features model, with predictors *mean interval*, *mean pitch – upper voice*, *mean pitch – lower voice*, *mean duration*, *key proportion* and *syncopation score*, yielded a significant effect of mean interval, $z(11620) = -2.32$, $p = .02$, had Δ AIC = 2, Δ BIC = 46 and a residual variance of 15923 ($df = 11620$), $p < .0001$. The addition of *musical training* to significantly improved the model, $\chi^2(1) = 50.12$, $p < .0001$. This model has Δ AIC = -46, Δ BIC = 5 and a residual variance of 15873 ($df = 11619$), $p < .0001$. Finally, the maximally fitted mixed effects model is nearly unidentifiable. Previously, when this occurred (Chapter 6), a

model without random effects was tested for soundness and random effects were reintroduced in order to represent the experimental design. Therefore, random effects were kept, though coefficients are not as reliable here. The maximally fitted model had $\Delta\text{AIC} = -673$, $\Delta\text{BIC} = -606$ and residual variance of 15262 ($df = 11617$), $p < .0001$. Only musical training was a significant fixed effect ($z(11617) = 4.18$, $p < .0001$) and random effects on *participant* explained more variance than random effects on *trial number* (see Table 7.13). A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.12 and 7.13 respectively.

Finally, *mean complexity ratings* for each trial, collected from the previous study (Section 7.3) were also tested as a predictor, which was significant, $z(11625) = -2.26$, $p = .02$. The model has $\Delta\text{AIC} = -3$, $\Delta\text{BIC} = 4$, and a residual variance of 15923 ($df = 11625$), $p < .0001$. The addition of musical training improves the model, $\chi^2(1) = 50.02$, $p < .0001$, with $\Delta\text{AIC} = -46$, $\Delta\text{BIC} = -31$ and a residual variance of 15883 ($df = 11624$), $p < .0001$. A maximally fitted model performs best, with significant fixed effect of musical training ($z(11622) = 4.18$, $p < .0001$) random effects on *participant* explaining more variance than *trial number* (see Table 7.15). This model has $\Delta\text{AIC} = -680$, $\Delta\text{BIC} = -650$, and a residual variance of 15245 ($df = 11622$), $p < .0001$. A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.14 and 7.15 respectively.

Overall, performance on this task was poor and rarely above chance. Fixed effects representing experimental manipulation, raw information content, musical features and mean complexity ratings are weak predictors of the data, generally increasing a null model's BIC, while random intercepts on participants explain the most variance in all models in which it is included.

Table 7.12. Summary of the fixed effects musical properties model for complex targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	-2.38	-5.13	0.36	-
Mean Interval Size	-0.04	-0.07	-0.00	3.7
Mean Duration	0.01	-0.00	0.04	13
Key Proportion	0.11	-0.31	0.54	22
Syncopation Score	0.01	-0.00	0.03	29
Mean Pitch – Upper voice	0.01	-0.01	0.04	37
Mean Pitch – Lower voice	0.00	-0.02	0.04	46
GoldMSI	0.13	0.09	0.16	5

Table 7.13. Summary of the maximally fitted musical properties model for complex targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	-3.50	-10.87	3.86	-
Mean Interval Size	-0.04	-0.14	0.04	-643
Mean Duration	0.00	-0.04	0.04	-634*
Key Proportion	0.41	-0.74	1.57	-625*
Syncopation Score	0.01	-0.04	0.07	-616*
Mean Pitch – Upper voice	0.03	-0.04	0.12	-608*
Mean Pitch – Lower voice	0.00	-0.09	0.10	-598*
GoldMSI	0.15	0.08	0.23	-606*
	Predictor	Variance		
Random intercepts	Participant	0.24		-273
	Trial number	0.19		-652

*model does not converge

Table 7.14. Summary of the fixed effects complexity ratings model for complex targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.53	0.28	0.78	-
Complexity Ratings	-0.07	-0.13	-0.00	4

Table 7.15. Summary of the maximally fitted complexity ratings model for complex targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	-0.01	-0.67	0.63	-
Complexity Ratings	0.03	-0.13	0.20	-643
GoldMSI	0.15	0.08	0.23	-650
	Predictor	Variance		
Random intercepts	Participant	0.24		-273
	Trial number	0.20		-652

Simple targets. All 96 trials received at least one response, with the remaining ‘different’ trials receiving on average 29.95 responses each (SD = 4.07), corresponding to approximately 1/3 of total participants. Performance on simple target stimuli was better than on complex target stimuli, $\chi^2(1) = 161.03$, $p < .0001$, where overall success rates were 67% and 56% respectively. The experimental manipulation model, with predictors *objective complexity*, *parameter* and *version*, yielded a significant effect of version C and D only, $z(4298) = -6.11$, $p < .0001$ and $z(4298) = 03.36$, $p = .0007$ respectively, with a Δ AIC of -43, Δ BIC of 34 and residual deviance of 5374 (df = 4298), $p < .0001$. Overall, neither objective complexity nor parameter were significant predictors, $\chi^2(2) = 3.44$, $p = .17$, and $\chi^2(7) = 11.83$,

Table 7.16. Summary of the fixed effects manipulation model for simple targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter). Each factor of objective complexity is in relation to Level 1; each factor of parameter is in relation to rhythm; and each factor of version is in relation to version A.

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.96	0.74	1.18	-
Level 2	-0.15	0.10	0.13	
Level 3	-0.23	0.02	0.13	
Level 4	0.03	-0.23	0.30	
Level 5	-0.11	-0.37	0.13	47
Level 6	-0.12	-0.38	0.12	
Level 7	-0.15	-0.41	0.09	
Level 8	-0.15	-0.41	0.10	
Melody	-0.08	-0.23	0.29	60
Harmony	0.06	-0.09	0.22	
Version B	-0.01	-0.18	0.15	
Version C	-0.58	-0.77	-0.39	34
Version D	-0.33	-0.52	-0.13	
GoldMSI	0.28	0.22	0.35	-35

$p = .10$ while version was, $\chi^2(3) = 54.38$, $p < .0001$. The addition of *musical training* as a predictor improved model significantly, $\chi^2(1) = 77.59$, $p < .0001$. This model has Δ AIC = -118, Δ BIC = -35 and residual deviance of 5297 (df = 4297), $p < .0001$. The previously significant effects of versions C and D remain significant. Finally, the maximally fitted mixed effects model does not converge, with Δ AIC = -584, Δ BIC = -478 and residual variance of 4837 (df = 4295), $p < .0001$. Only one level of version remained significant, version C with $z(4295) = -2.93$, $p = .003$ along with musical training, $z(4295) = 3.46$, $p < .0001$ and random effects on *participant* explained more variance than random effects on *trial number* (see Table 7.17). A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.16 and 7.17 respectively.

Table 7.17. Summary of the maximally fitted manipulation model for simple targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor. Factors are the same as in Table 7.16 above.

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	1.11	0.59	1.63	-
Level 2	0.09	-0.43	0.62	
Level 3	0.00	-0.52	0.53	
Level 4	-0.02	-0.54	0.50	
Level 5	-0.10	-0.63	0.41	-502
Level 6	-0.16	-0.69	0.36	
Level 7	-0.41	-0.93	0.11	
Level 8	-0.18	-0.71	0.33	
Melody	-0.14	-0.46	0.18	-487
Harmony	0.05	-0.27	0.37	
Version B	0.03	-0.32	0.40	
Version C	-0.60	-0.98	-0.23	-505*
Version D	-0.29	-0.67	0.09	
GoldMSI	0.39	0.17	0.62	-478*
Predictor	Variance			
Random intercepts	Participant	1.01		-339
	Trial number	0.29		-557

*model does not converge

The information content model, with predictors *mean melodic IC*, *mean harmonic IC*, *mean rhythmic IC* and *target IC* (all scaled), yielded a significant effect of melody IC only ($z(4306) = -1.97$, $p = .04$), with Δ AIC = -5, Δ BIC = 20, and a residual variance of 5428 ($df = 4306$), $p < .0001$. The addition of *musical training* improved the model significantly, $\chi^2(1) = 74.88$, $p < .0001$. This mixed model has Δ AIC = -78, Δ BIC = -46 and residual deviance of 5353 ($df = 4305$), $p < .0001$. Melody IC remained a significant fixed effect, $z(4305) = -1.99$, $p = .04$. Finally, the maximally fitted mixed effects model performed best with Δ AIC = -575, Δ BIC = -530 and residual variance of 4852 ($df = 4303$), $p < .0001$. Musical training was the only significant fixed effect predictor, $z(4303) = 3.45$, $p = .0005$ and random effects on

Table 7.18. Summary of the fixed effects information content model for simple targets, including coefficients, 95% confidence interval and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.74	0.67	0.80	-
Melody IC	-0.07	-0.14	-0.00	3.0
Harmony IC	-0.06	-0.13	0.01	6.1
Rhythm IC	-0.04	-0.11	0.02	12
Target Pitch IC	0.01	-0.04	0.08	20
GoldMSI	0.28	0.21	0.34	-46

Table 7.19. Summary of the maximally fitted information content model for simple targets, including coefficients, 95% confidence interval and Δ BIC for each predictor.

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.77	0.51	1.02	-
Melody IC	-0.09	-0.25	0.07	-550
Harmony IC	-0.06	-0.23	0.09	-543
Rhythm IC	-0.05	-0.21	0.10	-535
Target Pitch IC	0.04	-0.10	0.20	-527
GoldMSI	0.39	0.17	0.61	-530
	Predictor	Variance		
Random intercepts	Participant	1.01		-399
	Trial number	0.37		-557

participants explained more variance than random effects on *trial number* (see Table 7.19). A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.18 and 7.19 respectively.

The musical features model, with predictors *mean interval*, *mean pitch – upper voice*, *mean pitch – lower voice*, *mean duration*, *key proportion* and *syncopation score*, yielded no significant effects, with Δ AIC = 3, Δ BIC = 42 and a residual variance of 5432 (df = 4304), $p < .0001$. The addition of *musical training* significantly improved the model, $\chi^2(1) = 74.74$, $p < .0001$, with Δ AIC = -70, Δ BIC = -25 and a residual variance of 5358 (df = 4303), $p < .0001$.

Table 7.20. Summary of the fixed effects musical properties model for simple targets, including coefficients, 95% confidence intervals and Δ BIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.20	-4.63	5.06	-
Mean Interval Size	-0.02	-0.08	0.04	5.8
Mean Duration	0.02	-0.00	0.05	13
Key proportion	-0.56	-1.32	0.18	17
Syncopation Score	0.01	-0.02	0.05	25
Mean Pitch – Upper voice	-0.00	-0.06	0.05	33
Mean Pitch – Lower voice	-0.00	-0.07	0.06	42
GoldMSI	0.28	0.21	0.34	-25

Table 7.21. Summary of the maximally fitted musical properties model for simple targets, including coefficients and Δ BIC for each predictor. Confidence intervals were not computable for this unidentifiable model.

Predictor	Coefficient	2.5%	97.5%	Δ BIC
(Intercept)	0.18	-	-	-
Mean Interval Size	-0.04	-	-	-549
Mean Duration	0.03	-	-	-541
Key proportion	-0.59	-	-	-534*
Syncopation Score	0.02	-	-	-526*
Mean Pitch – Upper voice	-0.03	-	-	-518*
Mean Pitch – Lower voice	0.02	-	-	-509*
GoldMSI	0.39	-	-	-512*
	Predictor	Variance		
Random intercepts	Participant	1.01		-399
	Trial number	0.37		-557

*model does not converge

Finally, the maximally fitted mixed effects model did not converge, with $\Delta AIC = -570$, $\Delta BIC = -512$ and residual variance of 4853 ($df = 4301$), $p < .0001$. Musical training was a significant fixed effect, $z(4301) = 3.45$, $p = .0005$, and random effects on *participant* explained more variance than random effects on *trial number* (see Table 7.21). A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.20 and 7.21 respectively.

Finally, *mean complexity ratings* for each trial, collected from the previous study (Section 7.3) were also tested as a predictor, which was not significant, $z(4309) = -1.67$, $p = .09$. The model has $\Delta AIC = -1$, $\Delta BIC = 6$, and a residual variance of 5438 ($df = 4309$), $p < .0001$. The addition of musical training improves the model, $\chi^2(1) = 74.50$, $p < .0001$, with

Table 7.22. Summary of the fixed effects complexity ratings model for simple targets, including coefficients, 95% confidence intervals and ΔBIC for each predictor (change in BIC associated with the addition of each parameter, beginning with random effects).

Predictor	Coefficient	2.5%	97.5%	ΔBIC
(Intercept)	1.07	0.67	1.48	-
Complexity Ratings	-0.08	-0.19	0.01	-11
GoldMSI	0.28	0.21	0.34	-61

Table 7.23. Summary of the maximally fitted complexity ratings model for simple targets, including coefficients, 95% confidence intervals and ΔBIC for each predictor.

Predictor	Coefficient	2.5%	97.5%	ΔBIC
(Intercept)	1.16	0.17	2.15	-
Complexity Ratings	-0.10	-0.35	0.14	-549
GoldMSI	0.39	0.16	0.65	-552
	Predictor	Variance		
Random intercepts	Participant	1.01		-399
	Trial number	0.38		-557

$\Delta AIC = -74$, $\Delta BIC = -61$ and a residual deviance of 5363 ($df = 4308$), $p < .0001$. A maximally fitted model performs best, with random effects on *participant* explaining more variance than *trial number* (see Table 7.22). This model has $\Delta AIC = -578$, $\Delta BIC = -552$, and a residual variance of 4855 ($df = 4306$), $p < .0001$. A summary of the fixed effects model and of the maximally fitted mixed effects model can be found in Tables 7.22 and 7.23 respectively.

Overall, performance on this task was generally above chance and below 75%. Fixed effects representing experimental manipulation, raw information content, musical features and subjective complexity ratings are weak predictors of the data, generally increasing a null model's BIC, while musical training is consistently a good predictor, where increased training is proportional to increased performance, and random intercepts on participants explain the most variance in all models in which it is included.

7.4.5 Discussion

In this study, the aim was to validate operational definition number three, stating that complexity is proportional to salience. To do so, complexity of the outer voices of short 3-voice musical excerpts was systematically manipulated in terms of melody, harmony and rhythm and participants were asked to identify whether the middle voice – the target – was the same or different when presented with pairs of these stimuli. While a similarity hypothesis (Marozeau et al., 2013) predicted that in the complex condition, performance would decrease with increasing complexity and in the simple condition, performance would increase with increasing complexity, a complexity hypothesis (Prince et al., 2009) predicted that performance would decrease with increasing complexity for both complex and simple target conditions (see Figure 7.7).

Though results do not provide strong statistical support for a link between complexity and salience, nor for either the similarity or complexity hypotheses, several patterns are worth

discussing. Firstly, despite piloting with diverse listeners achieving performance ranging from chance to ceiling, the complex target condition proved to be very difficult and participants rarely achieved performance above chance. The simple target condition was easier and provides a wider range of performance, mostly above chance.

Second, random effects on participants consistently explain the majority of variance in all conditions, though musical training remains significant throughout. This suggests that listeners' streaming success was due to a combination of musical training and individual differences and not due to information content or various musical features. Perhaps different information content, or complexity between voices, is not a sufficiently strong segregator and musicians performed better due to their training improving their auditory streaming skills (François et al., 2014). They presumably used cues other than complexity to successfully segregate the target from the outer voices. Alternatively, it is possible that individual differences masked any variance in salience due to complexity. The adaptation of existing paradigms investigating relative salience in the general psychology literature to music would be an interesting avenue to pursue, perhaps disentangling these two possibilities. For example, the well-known Stroop task (Stroop, 1935) might be adapted to music by presenting congruent or incongruent visual and auditory stimuli in terms of pitch, harmony or time in order to compare the interference levels of each on the other parameters. Another task asks participants to rate the similarity of two objects that were either explored visually or haptically (Lakatos & Marks, 1999) and found that haptic exploration resulted in higher salience of local object features (i.e., grooves) as compared to global object features (i.e., shape) for short handling periods (1s) but not for longer handling periods (16s). Musical excerpts could be manipulated to differ in varying parameters as was done in the present study, where the task is instead to

rate the similarity of pairs of excerpts, where small variations to the most salient parameter will result in the lowest similarity ratings.

Third, for models without random effects, fixed effect predictors (other than musical training) overwhelmingly worsened the AIC and BIC values with respect to a null model, though residual variance did decrease somewhat. Only three of these fixed effects were significant predictors of performance, where Δ BIC had a smaller positive value than for other predictors: target versions C and D, as well as increased melodic IC led to worse performance (as indicated by coefficients; Tables 7.8, 7.9, 7.16 and 7.17). This pattern of decreasing performance with increased melodic complexity (higher IC) corresponds to predictions made by the complexity hypothesis (Figure 7.6); however, this effect is weak and random effects explain more variance than complexity ratings when included. It is interesting that versions C and D also led to worse performance in all analyses. These versions contain out-of-key notes in the context of the target itself but are not necessarily out of key in the context of the overall 3-voice stimulus due to varying harmonic progressions. A decrease in performance, in this paradigm caused by the inability to segregate the target from its polyphonic context, suggests increased integration for these stimuli. This in turn suggests that unexpected chord progressions (chords containing out-of-key target notes tend not to be ‘regular’ chord progressions; see Figure 7.3) are either particularly strong integrating factors or strong distractors, taking attention away from the target. While the latter possibility is in line with the complexity hypothesis, the integration hypothesis cannot be supported or rejected by these results. Nevertheless, both possibilities are partially supported by the influence of harmonic IC on performance in the complex target condition, where an increase in IC leads to a decrease in performance (as indicated by coefficients; Tables 7.10 and 7.11). The same pattern exists in the simple target condition, but does not reach significance (see Tables 7.18 and 7.19). Instead,

melodic IC is a significant predictor of performance in the simple target condition, where once again increased IC leads to a decrease in performance (as indicated by coefficients; Tables 7.18 and 7.19). This provides some support for the complexity hypothesis, where increased IC in the outer voices distracts from the simple target, making discrimination between a pair of simple targets more difficult.

In summary, though results provide only weak statistical support for a link between complexity and salience, some interesting patterns emerged that support the complexity hypothesis, but not the similarity hypothesis, of salience. These results will next be discussed together with results from Section 7.3 and Section 6.4, bringing these studies together in the context of relative perceptual salience.

7.5 General Discussion

Thus far, this chapter has presented an operational definition of salience – salience is proportional to information content – and has broken this statement down into two parts for individual testing, beginning with the lowest-level concept: (1) information content is proportional to complexity; and (2) complexity is proportional to salience.

The first reported study (Section 7.3) provided evidence supporting a relationship between information content and complexity by collecting complexity ratings on musical excerpts that were manipulated in terms of melodic, harmonic and rhythmic information content. Excerpts with higher information content resulted in higher complexity ratings. Though predictors approximating all three examined musical parameters had some influence on ratings, pitch-based predictors carried the most predictive power, and substantially more than time-based predictors. While predictors such as rhythmic information content, mean duration and rhythmic complexity levels had little impact on ratings, melodic information

content, the proportion of out-of-key pitches in the outer voices and harmonic complexity levels – all closely related – explained the majority of the variance in the data in each of their models. While different types of objective measurements exist (Eerola, 2016; Narmour, 1992; Vuust & Witek, 2014), this is the first time that a relationship between information content as produced by IDyOM and perceived complexity has been empirically tested and serves as a very useful link to current research in complexity in music by providing a valid, objective proxy.

The second reported study did not find evidence for a relationship between complexity and salience. A same-different paradigm was employed asking participants to identify whether the target melodies in a pair of musical excerpts (the same as for the complexity rating task) were the same or different. Performance was rarely above chance for the complex target condition and no higher than 75% for the simple target condition. Despite this, existing patterns weakly support a complexity-based definition of salience over a similarity-based definition, where performance matches the former's predictions more closely than the latter (Figure 7.7). Effects of melodic and harmonic features of the music are found here as in the first study, where stimuli with higher information content in these parameters and larger intervals (rated as more complex) lead to worse performance on the same-difference task.

It is interesting that melodic and harmonic aspects of the stimuli are the most important in this pair of studies, as the study presented in Chapter 6 rather provided evidence for rhythm being more salient than pitch. It is possible that this discrepancy is due to the difference in overall complexity of both the stimuli and the task involved in the two sets of studies. While the study in Chapter 6 used longer stimuli and some had high mean IC, these were monophonic. The polyphonic stimuli for the studies reported here were designed to interfere with the target to varying degrees. Furthermore, the three voices were harmonically related, increasing the chances of integration of these voices; this was observed particularly for complex harmonic

progressions. It is worth remembering here that Duane (2013) found evidence that onset and offset synchronization were the most important streaming factor in 18th and 19th century string quartets, while these excerpts are also harmonically integrated. In the current study's stimuli, onset synchrony between all voices was 0.88 for simple target stimuli and 0.46 for complex target stimuli, reflecting higher integration in simple target stimuli; however, performance was better for these simple target stimuli overall and rhythmic aspects of the stimuli had little to no effect on performance, indicating that something was counteracting this integration and performance here is influenced by other sources of information. An important difference between Duane's stimuli and the present stimuli is that the latter, though following Western classical common practice as much as possible, are synthetic in comparison to existing string quartets. However, it is important to approach this issue with both natural and manipulated stimuli to examine it from different angles. Another possibility is that the paradigm task also affected perception, where in Chapter 6, listeners were asked to make ratings in real time while these studies require single decisions after hearing the stimuli. While all music occurs over time, it is possible that a time-dependent task such as providing continuous ratings may have drawn attention to temporal aspects of the music.

Clearly, more research is needed to explore when and why different musical parameters appear to be more or less salient in different contexts. It would be interesting to pursue this line of research in the context of expectation more explicitly (Jaeger & Weatherholtz, 2016), as this approach is hypothesized to explain bottom-up salience as well as top-down salience. Furthermore, this universality and grounding in prediction is in line with the predictive coding framework (Zarcone et al., 2016), offering yet another avenue of research into the applications of this potential grand unifying theory. Some recent work fits into this research agenda, investigating whether predictable, rather than unpredictable sequences are salient (Southwell

et al., 2017). Southwell et al. used either regular or random sequences as irrelevant distractors from a task where participants were asked to identify a change in an auditory scene. The task scene and distractors were presented to different ears. If predictability was salient and captured attention, worse performance on the task would be expected when a regular distractor was presented; however, a regular distractor resulted in better performance than a random distractor, supporting a negative relationship between information content and saliency, the top-down side of the salience question.

Before concluding, one major question must be addressed: if there is such a strong link between information content and complexity, and increasing evidence and discussion in the literature (Blumenthal-Dramé et al., 2017; Ellis, 2006; Horstmann et al., 2016; Jaeger & Weatherholtz, 2016; Prince et al., 2009; Schmid & Günther, 2016; Southwell et al., 2017) for a link between information content and salience and complexity and salience, why was this link so weakly observed here? As previously suggested (Section 7.4.5), it is possible that information content, or complexity, are not strong enough segregation cues to allow successful performance on the same-difference task, which relies on successfully streaming the target voice from the outer voices. This is supported by the positive effect of musical training on performance, where musical training improves auditory streaming (François et al., 2014). Perhaps if the three voices were less strongly integrated and the overall perception was ambiguous, then relative complexity of the voices would have a stronger effect. For example, each outer voice could have been another octave away from the target voice. Furthermore, it seems that harmony, which appears to have a particularly strong integrative role, should not be considered equal to melody and rhythm in its perception and thus not directly compared, as was the case here. These are interesting considerations for future research.

7.6 Conclusion

Overall, this chapter validates information content as a proxy for perceived complexity. While not a perfect proxy, the majority of variance in the behavioural data is explained by experimental design and information content. Furthermore, pitch-based musical features were found to be the best indicators of perceived complexity and melodic complexity seems to be the most salient parameter for complex auditory scenes. Despite these preliminary conclusions, more research is needed to understand more specifically when and in what contexts various musical parameters are most salient. Finally, the results presented in this chapter are relevant to the use of relative salience in the proposed auditory streaming framework, where it is suggested that modules modelling parameters with higher average information content should carry more weight to make a streaming decision. However, the relationship between information content and salience was not supported by the results of these studies, so an alternative approach should be explored. The implications of these results will be discussed further in Chapter 9, where the proposed integrated framework for auditory streaming will be re-evaluated in the context of all results presented in this thesis.

8 Prediction-based melody extraction

How do we know a melody is a melody? This is probably not a question often contemplated when listening to music, as listeners *just know* a melody when they hear one. It is however a very relevant question when considering musical ASA, as it can inform the process of melody extraction, or in other words, identifying the foreground of a musical scene. This task, which became very popular in the past few decades in the music information retrieval (MIR) community, consists of identifying the melody in a polyphonic music context, either from audio or symbolic data. As a core part of the proposed integrated framework for auditory scene analysis presented in Chapter 4, a predictive approach to melody extraction is investigated in this chapter. However, due to some constraints, the implementation of this task will be different from how it is described in Chapter 4 while keeping to the same principles. First, as IDyOM does not yet have implemented harmonic viewpoints, each stream will be

monophonic. Secondly, as there is yet to be any annotated data for perceptual streams, this task will identify score-based voices, evaluated on chorales by J.S. Bach and string quartets by W.A. Mozart. This approach is based on two hypotheses: melodies are internally coherent and melodies are the most interesting stream in a given musical work.

To begin, melody will be defined and work in MIR melody extraction will be summarized, including common performance metrics and datasets (Section 8.1). In Section 8.2, the proposed prediction-based model and its implementation, an extension of the current IDyOM system, will be presented. Section 8.3 will present the model evaluation and results, and finally, Section 8.4 will discuss the model’s output and potential improvements for its future development.

8.1 Melody: definition and literature

When discussed in an informal setting, people usually understand each other as to what is meant by *melody*. It’s one of those things that everyone understands without needing to put specifically into words. However, for the purposes of empirical research, a definition is needed. Drawn from music theory (Toch, 1923), melody can be defined as *a succession of different pitch sounds brightened up by the rhythm*. A more recent music dictionary, the New Grove Dictionary of Music and Musicians (Rycroft & Sadie, 1983), similarly defines melody as *a combination of a pitch series and a rhythm having a clearly defined shape*. These are quite generic; the most commonly employed definition of melody in audio MIR is that it is *the sequence of monophonic pitches that a listener might sing or hum when asked to reproduce a polyphonic piece of music, and encompasses the core identity of the piece* (Salamon, Gomez, Ellis, & Richard, 2014). While still generic, this last definition allows for a “correct interpretation” through the identification of melody by a listener. Beyond this general

definition, melody has been broken down further into different types, as described by Selfridge-Field (1998):

- *compound melodies* describe melodies where some pitches are melodic and some are either another melody or an accompaniment; this is also called pseudopolyphony and is most common in solo string music
- *self-accompanying melodies* are melodies where some pitches act as both main theme and harmonic support, also another form of pseudopolyphony
- *submerged melodies* are melodies in inner voices of a polyphonic work
- *roving melodies* are melodies that move from part to part, or instrument to instrument, in an ensemble
- *distributed melodies* are melodies spread across various instruments and the theme cannot be represented by one part alone

Overall, these definitions are heavily biased towards Western ideas of melody in that it is assumed that there is only one such dominant line, characterized by pitch (as opposed to rhythm or timbre), that can be sung to represent a piece of music and that this line is monophonic (though doublings aren't especially rare in Western music, a melody is generally thought of as monophonic). One caveat to keep in mind is that it is not guaranteed that every listener will sing back the same line; currently, when there is disagreement, the most common interpretation is considered correct. Another typical assumption in the MIR field is that the melody cannot change instruments throughout the piece, which is appropriate and performs well for pop music but performs substantially worse for Western classical music, where in instrumental ensembles it is common for the melody to change instruments or rove, as defined by Selfridge-Field. Another related challenge is to identify whether there is a melody present at all, a problem called *voicing* (Salamon et al., 2014). For this chapter, melody will be defined

by the MIR definition presented above, considering the melody to be a monophonic sequence of pitches and that the majority of listeners agree to be labelled “melody”. However, a few differences will be allowed: 1) the melody is allowed to change instrumental lines (the sub-definition roving melodies), and 2) voicing will not be considered. While not every moment in Western classical music contains a voice that could be called melody, thinking particularly of textural passages and transitions between themes, here any non-melodic notes identified as melody will simply be considered false positives when compared to a ground truth. Though this would have a negative influence on performance, this will happen in a small enough proportion of the music analysed to have negligible impact (J. J. Bosch, Marxer, & Gómez, 2016).

Now that melody is defined, current methods for melody extraction in MIR from both audio and symbolic data formats will be summarized. First tackled in (1999) by Matasaka Goto, approaches to the melody extraction problem have multiplied since then, presenting both improvements and new challenges (Salamon et al., 2014). Algorithms that extract melody from audio files can generally be divided into two broad categories: source separation based approaches and salience based approaches, while algorithms using symbolic data tend to use variations of probability-based strategies.

With such a variety of approaches, it is worth first defining some common performance measures and thinking about challenges in evaluating performance. These measures are either related to pitch - whether the correct pitch was extracted (audio) or assigned (symbolic); or voicing - whether the melody was currently identified as present.

- *Raw pitch accuracy* is defined as the proportion of melody frames (vertical time divisions of the music) of the ground truth where pitch is considered correct (within half a semitone for audio-based approaches).

- *Raw chroma accuracy* is similar to raw pitch accuracy, but pitch is mapped to a single octave, allowing forgiveness for octave errors in audio-based approaches.
- *Voicing recall rate* is the proportion of frames labelled as melody in the ground truth that are also labelled as melody by the algorithm.
- *Voicing false alarm rate* is the proportion of frames labelled as non-melody in the ground truth that are labelled as melody by the algorithm.
- *Overall accuracy* combines pitch and voicing measures and is defined by the proportion of all correctly labelled frames (whether labelled melody or non-melody) and for correctly labelled melody frames, the pitch is also correctly identified (within half a semitone of the ground truth for audio-based approaches).

The evaluation of this chapter’s model will employ raw pitch accuracy alone, as voicing is not considered.

The evaluation datasets for these approaches vary widely, with all datasets being relatively small in MIR terms. This is due to the need for a ground truth – a definitive labelling of melody pitches for a piece of music – which is a perceptual construct and, as mentioned, may differ between listeners and therefore cannot simply be lifted from a score. Datasets containing annotations indicating melody in a polyphonic context are few and far between due to the extensive time commitment involved in building such a dataset, though over the past few years a number of datasets have been built, the majority for audio-based melody extraction. Tables 8.1 and 8.2 summarize known existing datasets for audio and symbolic research respectively. Through the Music Information Retrieval Evaluation eXchange (MIREX) competition, there are a few datasets on which many approaches have been tested; however, these are quite small and inaccessible due to the way the competition is designed and so many

Table 8.1. Details of audio datasets used for melody extraction evaluation; partially reproduced from (Bittner et al., 2014).

Name	# Songs	Song Duration	Total Duration	% Vocal Songs	Genres	Content
ADC2004	20	~20 s	369 s	60%	Pop, jazz, opera	Real recordings, synthesized voice and MIDI
MIREX05	25	~10–40 s	686 s	64%	Rock, R&B, pop, jazz, solo classical piano	Real recordings, synthesized MIDI
INDIAN08	8	~60 s	501 s	100%	North Indian classical music	Real recordings
MIREX09	374	~20–40 s	10020 s	100%	Chinese pop	Recorded singing with karaoke accompaniment
MIR1K	1000	~10 s	7980 s	100%	Chinese Pop	Recorded singing with karaoke accompaniment
RWC	100	~240 s	24403 s	100%	Japanese Pop, American Pop	Real recordings
MedleyDB	108	~20–600 s	26831 s	57%	Rock, pop, classical, jazz, rock, pop, fusion, world, musical theater, singer-songwriter	Real recordings
ORCHSET	64	~10-32 s	1379 s	0%	Classical, Romantic and 20 th century period symphonic works	Real recordings

Table 8.2. Details of symbolic datasets used in melody extraction evaluation, including the String Quartet dataset, previously described in Chapter 3.

Name	# Pieces	Genres	Content
CL200	200	Classical	MIDI files
JZ200	200	Jazz	MIDI files
KR200	200	Popular music	MIDI files
String Quartets (test & validation)	41	Classical	Kern files

evaluation datasets are built by authors to suit their needs. More recently, datasets such as MedleyDB (Bittner et al., 2014) and ORCHSET (Bosch & Gómez, 2014) have been developed with the aim of providing annotated audio datasets for MIR research in general. Symbolic datasets are also difficult to compile for this particular problem as existing symbolic data repositories lack melody annotations. Ponce de León et al. (Ponce de León Amador, Iñesta Quereda, & Rizo Valero, 2008) created three annotated datasets of 200 MIDI files, each in a different genre: classical, jazz and karaoke. For each file, none, one or more tracks in each file were hand labelled as melody by a musician, and the other tracks were labelled non-melody.

8.1.1 Melody extraction from audio

Saliency-based. Emilia Gómez is one of the most well-known leaders in the field of audio-based melody extraction and has mentored many in developing this line of research (Bittner, Salamon, Essid, & Bello, 2015; J. Bosch & Gómez, 2015; Gómez, Klapuri, & Meudic, 2003; Salamon & Gomez, 2012; Salamon et al., 2014). With Salamon, they have created one of the best saliency-based melody extraction algorithms yet (Salamon & Gomez, 2012), in 2009 achieving the best overall accuracy seen in the MIREX competition at the time. Since then, improvements seem to be stagnating (Salamon et al., 2014). This particular approach is saliency-based, where *saliency* refers to the relative presence of a particular frequency throughout at slice of time, called a *frame*. A frame is a slice in time encompassing all

frequencies present in the signal at that moment. Saliency algorithms extract *peaks*, particular frequencies that are salient relative to other frequencies in the frame. There may be more than one peak in any given frame if more than one frequency (translating to note) is relatively salient; thresholds are fine-tuned for each algorithm to strike a balance between removing non-melody notes, while keeping all potential melody notes.

There are a number of innovations that make Salamon and Gómez's approach particularly successful. First, their approach targets *pitch contours*, extracted with some initial audio processing and a saliency function, rather than defined notes with fixed onsets and offsets. Note offsets are particularly difficult to identify in audio, and this approach simply bypasses the problem by dealing with a continuous pitch estimation over time. This is a sensible thing to do, as the beginning and end of each individual note isn't necessarily helpful in identifying a melody against an accompaniment. Second, rather than selecting melody notes (peaks) directly, non-melody notes are filtered out. Once pitch contours, which can be anywhere from the equivalent of one note to a short phrase consisting of multiple joined salient peaks are formed, some straightforward characteristics are computed: pitch mean, pitch deviation, contour mean saliency, contour total saliency, contour saliency deviation, length and vibrato presence (true or false, based on Herrera & Bonada, 1998). This information is used to filter out non-melody notes. Most prominently, contour mean saliency is useful as melody and non-melody notes have separate distributions in a pitch contour, where non-melody notes are less salient, meaning that a threshold can be determined where most non-melody notes can be cut out while only minimally affecting potential melody notes. Furthermore, notes with vibrato are automatically included, as the probability of a non-melody note performed with vibrato is less than 5%. This keeps potentially less salient melodic notes for consideration. The use of pitch contours also minimizes octave errors by allowing computation of contour trajectories

over time. If two contours are an octave apart, then clearly there is a case of an octave doubling. The correct octave is chosen through a combination of salience (the most salient probably being the melody), and context, where large leaps are avoided. Finally, outliers can be removed by removing anything outside an octave of the mean pitch of the remaining contour.

Source separation. Durrieu et al. (Durrieu, Richard, David, & Fevotte, 2010) employ an approach akin to a source/filter model (Fant, 1971), where pitch and timbre characteristics of the audio are extracted separately. In this particular approach, the original audio source is modelled by a combination of two source/filter models, one for the melody and one for the accompaniment, where the melody model retains the melody and filters out the accompaniment and the accompaniment model does the opposite. This is also carried out over frames, as is standard in MIR.

Combined approaches. More recently, Bosch has combined the two main approaches described above and fine-tuned melody extraction for orchestral music (Bosch & Gómez, 2015; Bosch, Bittner, Salamon, & Gómez, 2016). More specifically, this algorithm combines Salamon & Gómez's and Durrieu et al.'s methods to use both aspects of salience-based extraction and source separation, namely source-filter models (Durrieu et al., 2010) and pitch contour extraction (Salamon & Gomez, 2012), including some basic auditory streaming cues to minimize octave errors and outliers, as in Salamon & Gómez (2012). Furthermore, by testing aspects of several approaches, Bosch et al. (2016) found that a good salience function is the most important aspect of a successful melody extraction model and the combination of harmonic summation and pitch contour selection give the best results for orchestral music, which is particularly dense and rich in doublings.

While research in audio-based melody extraction continues to develop, the remaining review of the literature will focus on symbolic melody extraction, the method employed by this

chapter's proposed extraction model. Symbolic representation is chosen because this model is concerned with the cognitive processing of music once relevant features (such as pitch, rhythm, harmony) have already been abstracted.

8.1.2 Melody extraction from symbolic data

Approaches to melody extraction from symbolic data tend to rely on the use of the statistical properties of certain musical features, typically chosen for their perceptual relevance. Some approaches consider full tracks or voices, and select one as the melody, while others separate the polyphonic material into a set of monophonic voices and subsequently select the melody from these extracted voices.

One of the earliest attempts at melody extraction relies on a very simple heuristic: choose the highest sounding pitch at any given time. Known as the skyline algorithm (Uitdenbogerd & Zobel, 1998), this works well for pop music or for a style like Bach chorales, but breaks down for instrumental music. Friberg & Ahlbäck (Friberg & Ahlbäck, 2009) give us an example of entirely perception-inspired melody extraction, where they use a number of basic features, some derived from Huron's principles of voice-leading (Huron, 2001) to identify melody. These are:

- *pitch* is derived from the *toneness principle*, which states that pitch perception is more accurate around a central pitch of D4
- *articulation*, defined using duration/IOI, so that a legato note is near or exactly 1 and a staccato note closer to 0, is derived from the *temporal continuity principle*, where an auditory stream is more stable if it is continuous

- *interval size* is derived from the *pitch proximity principle*, where smaller intervals make groups of notes fuse more easily (the Gestalt principle of good continuation also applies here)
- *IOI* is derived from the authors' *temporal sensitivity principle*, where music is best perceived when IOI is an average of around 250ms – not too slow so as not to lose continuity and not too fast so as not to miss events
- *timbre*, labelled as 0, 0.5 or 1 (nonharmonic, percussive harmonic or sustained harmonic), based on the idea that harmonic sounds are overwhelmingly more likely to also be melodic
- *sound level*, where the loudest voice is often the melody
- *total duration*, where the melody would be expected to be long but not as long as the whole piece, with the accompaniment occupying more time than the melody
- *polyphony*, measured as the number of simultaneous notes in each voice; the melody is assumed to most often be monophonic
- *narrative*, a measure introduced by the authors that is designed to convey the melody as an interesting story, where the more new material is encountered in a voice or track over the course of the piece, the more likely it is to be the melody

Using multiple regression on 242 polyphonic ringtones in MIDI format whose tracks were manually labelled melody or accompaniment (half were used for training and half for evaluation), it was found that all features but articulation were correlated with melody, in the expected direction based on the definitions given above. While this model correctly identified the melody track in 90% of the evaluation set, annotator agreement (the two authors) was only 81%. Overall, the authors conclude that melody is voice-like, with pitch and duration close to speech height and speed, overwhelmingly monophonic, louder than the accompaniment and

more original (narrative). They also recognize that ideal feature selection is an ongoing exploration and that these selected features in particular are not all ideal for symbolic representation, as in particular articulation, timbre and loudness were not systematically encoded.

Guided less explicitly by perceptual features, Rizo et al. (Ponce de León Amador et al., 2008; Rizo, De León, Pérez-Sancho, Pertusa, & Quereda, 2006) select the most melodic track of a given MIDI file by using pattern recognition on pitch, interval, duration and rhythmic information. More specifically, a set of 19 descriptors, based on the four labels just listed or track information, is computed for each track. Then, some tracks are used to train a random forest classifier (Breiman, 2001) and others are used for evaluation, where the random forest has learned the statistical properties of the 19 descriptors and assigned a likelihood of ‘melodiness’ to each track. The track with the highest melody likelihood is selected as the melody. This method was tested on three sets of 200 MIDI pieces, one classical, one jazz and one pop (karaoke) (see Table 8.2), where overall success exceeded 96% in all three datasets.

Madsen & Widmer (2007) present the approach that is the most similar to the approach presented in Section 8.2 below, computing various measures of complexity and information content to select the melody amongst given (whether from the score or separated by algorithm) voices in a polyphonic piece of music. They cite the link between complexity and attention, observed in music cognition research as motivation for this type of approach. As seen in Chapter 7, increased complexity demands more processing power and therefore draws more attention. Madsen and Widmer selected *Shannon entropy* (Shannon, 1948) as a measure of complexity. Shannon entropy is a measure of uncertainty; in this case, uncertainty about what pitch will come next in a given sequence. However, an important difference between Madsen

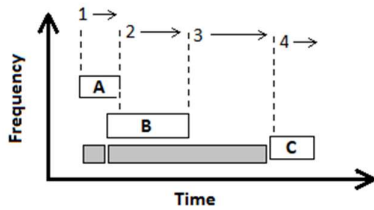


Figure 8.1. Illustration of window formation in Madsen & Widmer's (2007) streaming algorithm. Lettered boxes represent pitches and windows are numbered. Prediction periods include all pitches between the start of two windows and are represented by gray boxes (note that these are different from IOI). The first prediction period includes only window 1 while the second includes windows 1 and 2, and therefore events A and B.

& Widmer's model and the present work is that the former uses zeroth-order models while the latter uses variable-order models, through IDyOM.

Being very certain of what comes next equates to low entropy, while being very uncertain to high entropy. For example, before a piece of music begins, entropy is very high because there is no indication as to what the first pitch or pitches could be. By the end of the piece, in certain styles (such as most Western classical music) the last pitch is predictable with nearly absolute certainty, so the prediction has very low entropy. In this approach, a given piece of polyphonic MIDI music is processed from beginning to end in a series of *windows*, so that each window is different but every transition between notes in the music is included: a new window will begin either at the first offset of a pitch in the current window or at the first onset of a pitch beginning after the current window. *Prediction periods* are created, including all notes between the start of two windows (see Figure 8.1 for an illustration of this). Entropy for each voice in a prediction period is calculated from the frequency tabulations of three features and three combinations of these in each window included in the prediction period: *pitch*, *pitch interval* and *duration*, as well as the combination of pitch and duration, pitch interval and duration, and a weighted combination of all three. In each prediction period, the voice with the

highest average entropy is labelled as melody and the rest as non-melody. In this way, the melody can move from voice to voice. In terms of cognitive relevance, a drawback of this method is that it includes pitches beyond the current prediction period in the entropy calculation, which equates to using information about pitches that has not yet been heard by a listener. This happens when there are overlapping windows: the prediction period ends when the second window begins, but the whole window is included in the entropy calculation.

This model was evaluated on two melody-annotated pieces: String Quartet No 58, Op. 54, No. 2 in C major, first movement by Franz Joseph Haydn and Symphony No 40 in G minor (KV550), first movement by Wolfgang Amadeus Mozart. Window sizes of 1, 2, 3, and 4 seconds were tested for each of the six different entropy measures. The combination of pitch and interval gave the best precision, recall and F-measure for all window sizes, ranging from .81 to .87, except 4s for the string quartet, where pitch alone had slightly better recall, and therefore also F-measure. Performance was much better for the string quartet than for the symphony, which is not surprising as one has four voices and the other, ten. As compared to a skyline algorithm baseline, this approach performed much better in the symphony (mean .51 as opposed to .36), but worse for the string quartet (mean .85 as opposed to .93). This is not surprising, as in a string quartet the first violin most often plays the melody, while in a symphony it is not only more variable, but woodwinds' accompaniment, particularly in the case of the flute, often has higher pitch than the melody, even if the latter is played by a violin. Despite its drawbacks, this latest approach, which uses a measure of complexity, is a successful model of symbolic melody extraction when tested on music by Haydn and Mozart.

8.2 Melody extraction: a prediction-based approach

As seen in Section 8.1, while many use probability-based techniques, existing approaches to symbolic melody extraction vary widely in their implementations. The model proposed here is similarly grounded in prediction and is based on two rules: a melody is internally coherent and is the most interesting stream. These rules are both intuitive and empirically supported. The first hypothesis, that a melody is internally coherent, is supported by the Gestalt principle of good continuity and the perceptual principle of pitch proximity, where pitches that are closer together form coherent streams (Huron, 2001; Wundt, 1874). In this implementation, this rule translates to high probability, or low information content: melodies tend to contain steps and small leaps, therefore these patterns will be more probable and preferred to model melody. The second hypothesis, that a melody is the most interesting stream in a piece of music, is supported by previous work demonstrating this pattern using measures of information content (Duane, 2012; Friberg & Ahlbäck, 2009). In this implementation, this rule translates to melody being identified as the voice with the highest information content. Voice is used here instead of stream because auditory streaming has not been performed; rather, for reasons outlined in the opening of this chapter, this model performs voice separation. Details of the model's implementation are presented in Section 8.2.1 below.

Highly influenced by Madsen & Widmer's (2007) approach described above, the current model nevertheless differs in a number of respects. First, the model's measure of information content comes from IDyOM, a cognitively valid method of simulating prediction of music listening (Pearce et al., 2010). Unlike the zeroth-order distributions used by Madsen & Widmer (2007), IDyOM combines predictions from multiple time scales, simulates short- and long-term memory and can automatically select the best feature (optimized by lowest information content) if desired. Second, as a result of using IDyOM, only information that

would also be known to the listener is used (i.e. all in the past), creating a non-causal process. Third, the analysis windows are much smaller, only considering the current sounding pitches. Music is, after all, processed both vertically and horizontally and as context is included in the analysis of the current window by default, a larger window size is not necessary. Fourth, said context consists of one voice at any given time so that each currently sounding pitch is considered as a potential continuation for each existing voice. Furthermore, voice information is unknown and therefore voice separation is performed, an additional task not present in Madsen & Widmer's (2007) model. Finally, as a result of this build-up of voices over time, the melody is selected at the end of the piece rather than at each prediction period as the voice with the highest information content: the most complex.

While it is theoretically possible to calculate information content for any viewpoint implemented in IDyOM, this method will focus on pitch-based viewpoints. While rhythm and timing play a role in melodic identity, pitch can be considered the most important aspect of Western music (Prince et al., 2013). Furthermore, rhythm as a predictor of melody is irrelevant for Bach chorales since these are homophonic; all voices share the same rhythmic patterns, which would not help separate the melody from its context based on predictability. While this is not the case for string quartets, focus will remain on pitch for simplicity and comparison.

In summary, the approach taken in this chapter takes symbolic source material, breaks it down into multiple monophonic voices based on the lowest information content option and selects the melody from these voices based on the highest average information content. The implementation of this approach is presented next.

8.3.1 Model implementation

The model proposed here was implemented in three versions, each an improvement on the last. The overall approach is identical and the versions differ in small details, highlighted

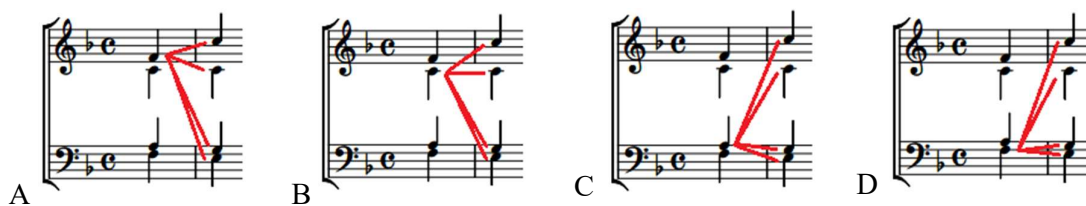


Figure 8.2. Illustration of the iteration process of the first and second model implementations. For each slice, each voice (A-D) is processed in turn and each potential following note is considered; the most likely continuation is selected and added to that voice.

throughout the description that follows. The models all process music in vertical slices, chronologically over time, similarly to Cambouropoulos’ (2008) streaming model. A *slice* is created at each note onset in the polyphonic texture and contains all notes sounding at the time,

Table 8.3. Tabulation of all transitions between bars 1 and 2 of Mozart’s String Quartet No. 16, second movement. The current voice corresponds to the model’s voice numbers and the source voice corresponds to the score voice, where 1 corresponds to the first violin and 4 to the cello.

Current voice	Source voice	Pitch (MIDI)	IC
1	1	63	0.73
1	2	60	5.37
1	3	56	6.86
1	4	44	9.47
2	1	63	3.96
2	2	60	3.55
2	3	56	5.52
2	4	44	6.91
3	1	63	6.00
3	2	60	3.77
3	3	56	3.20
3	4	44	13.85
4	1	63	8.11
4	2	60	12.75
4	3	56	11.86
4	4	44	0.93

whether that represents a new onset or a continuation of a previous note. The model initiates the analysis of a piece of music by creating one *voice* for each note in the first vertical *slice*, which also makes up the first *interpretation*. This is organized so that the soprano is “on top”, or voice 1, and the bass is “on the bottom”, or voice 4. Remaining in this first slice, the first version of the model then iterates through each voice, beginning with the highest pitch. The model generates predictions using both IDyOM’s LTM and STM models. At each voice, the information content of each potential next pitch (the *next*

slice) is determined by using the current voice as context, where this context is specifically modelled by the STM in addition to the LTM, and the pitch with the lowest information content is selected to be added to that stream. The model then moves to the next slice, and iterates through its voices and so on until the end of the work (Figure 8.2). The voices with the highest average information content (Duane, 2012) is then returned as the melody. In the first version, LTM predictions for each voice are drawn from an LTM trained on the equivalent chorale voice (i.e. soprano predictions are generated by an LTM trained on soprano voice only). This was to ensure that each voice would be modelled as accurately as possible. In the second version, only one LTM is employed, trained on all four chorale voices. This is the preferred implementation, as IDyOM's STM should be capable of adjusting predictions to each voice according to context. The biggest drawback of these first two versions is that they do not handle fluctuations in the number of voices. The third version of the model addresses this drawback and is highly similar to the first two, differing in its iteration procedure. Instead of iterating through each slice and each voice in order, the model tabulates the information content of all possible *transitions* between streams and iteratively selects the transition with the lowest information content (see Table 8.3). At each iteration, all potential transitions that include (1) the voice with a newly assigned pitch, (2) that pitch, and/or (3) the voice that pitch came from (score-based) are removed from the possible transitions. This modification was implemented to handle the frequent addition and subtraction of voices in string quartets, so that all potential pitches are assigned to a voice first before voices left without new pitches are assigned *rest objects*, created with pitch -1 as 'placeholders' to maintain vertical alignment until a pitch is assigned in that voice again. These rest objects are ignored when information content is calculated as this model is only currently dealing with pitch-based viewpoints. See Figure 8.3 for an illustration of this process.

Table 8.4. Tabulation of all transitions between bars 5 and 6 of Mozart’s String Quartet No. 16, first movement. This transition table corresponds to the illustrated example given in Figure 8.3, where transitions are marked on the score.

Current voice	Source voice	Pitch (MIDI)	IC	Transition
1	1	65	3.16	A-B
1	2	60	7.89	A-D
2	1	65	3.16	C-B
2	2	60	7.89	C-D
3	1	65	3.76	E-B
3	2	60	5.46	E-D
4	1	65	3.76	F-B
4	2	60	5.46	F-D

Mozart K.428, movement 1, mm 4-5

Voice 1	68	67	#1 65
Voice 2	68	67	#3 -1
Voice 3	56	55	#2 60
Voice 4	56	55	#4 -1

Figure 8.3. Illustration of the iteration process of the third model implementation, where for each slice the IC of each possible transition between voices is calculated (Table 8.4) and the lowest IC is assigned first. Focusing on the transition over the barline, the lowest IC value is 3.16 and represents the transition from G to F (labels A – B and C – B); this value exists for two transitions because the previous context is exactly the same for Voices 1 and 2. The pitch is assigned to Voice 1 simply because it occurs first in iteration as everything else is equal. Once all transitions containing the F are removed from the table, the lowest IC left is 5.46, for the transition between Voices 3 or 4 and middle C (E – D and F – D). It is assigned to Voice 3. As there are voices left with unassigned pitches, these are assigned rest objects, with a place-holder pitch value of -1.

Though a straightforward concept, a few heuristics were added to avoid certain issues such as crossing streams and dealing with equidistant or equally probable pitches. First, any pitches that are a continuation of an already sounding pitch are assigned to the voice containing that pitch before any transitions are considered based on IC; this currently assumes that there will only be one such pitch match. Second, as IDyOM relies on local context as well as a training corpus for generating its predictions, this model initially chooses the next pitch for each voice by selecting the closest pitch for s slices before engaging the information-content based selection method. The default value of s is 5 and can be adjusted by the user. This provides the model with a context in order to generate more accurate predictions for each voice early in the analysis. Third, if two potential pitches are equidistant from the current pitch or are equally likely to follow the current pitch, in the first two versions of the model the first option is assigned to the current interpretation and a new interpretation is created for each alternate possibility, retaining all previous context from all voices until that point, and continuing with the slice and voice iteration as usual. In the case where there are multiple possible interpretations for the organization of pitches into voices, the interpretation with the least amount of similarity between its voices is selected, based on the assumption that different voices in polyphonic music do not share many pitches. Similarity is simply measured by the number of unique pitches in each slice of each interpretation, where more unique pitches equates to less similarity. Melody is then selected from the streams of this selected interpretation. In the third version of the model, no alternative interpretations were necessary as the instrumentation in string quartets results in a larger pitch spread and extremely low probability of equidistant or equally probable pitches for any given voice. In the rare instance where the situation does occur, in the case of equal distance, either the pitch that has not been

previously assigned to a voice, or the highest pitch is assigned for simplicity and in the case of equiprobable transitions, the transition with the smallest distance between pitches is assigned.

8.3 Model evaluation

The model was evaluated on two types of polyphonic music: chorales by J. S. Bach and a selection of string quartet movements by W. A. Mozart. In the first instance, Bach chorales were chosen because of the straightforward ground truth: these chorales are by definition harmonized melodies and so the melody simply equates to the soprano voice. String quartets were chosen both for comparison to previous work identifying melody as the most complex voice (Duane, 2012) and to evaluate the model in a musical situation where voice does not equate to stream and the melody may move between voices.

8.3.1 Bach chorales

The Bach chorale test and validation datasets are described in Chapter 3, Section 3.2.2. All 350 test set chorales were analysed using tenfold cross-validation, where for each validation set, in this case a subset of 35 chorales were analysed while the remaining chorales were used as a training set. As IDyOM only learns from monophonic sequences, two training sets were constructed: the first is a combination of all four monophonic chorale voices for all training chorales, considering information from the full chorales; the second contains only the soprano voice, training the model on melodic information only. The first two versions of the model were tested using the first training set along with a variety of source viewpoints: *chromatic pitch*, *pitch interval* and *scale degree*, and every combination of these three. Only Version 2 of the model was trained on a melody-only training set, as having only one voice to train on would render testing Version 1 redundant. A further caveat is that training on one voice restricts the viewpoints that can be tested to derived viewpoints only. Since the model is not exposed

to alto, tenor and bass voice pitches in training, it cannot process them when they appear in the music analysed; therefore, only *pitch interval*, *scale degree*, and their combination can be tested as these higher-level patterns are present in all voices. Three performance metrics were calculated: 1) percent match of the extracted melody to each score voice; 2) percent overlap between voices extracted by the model and 3) percent match between each extracted voice and its corresponding score voice (based on the first pitch of the extracted voice). While the first metric explicitly evaluates melody extraction, the second verifies whether the model effectively separates the voices as they are written in the score and the third verifies the quality of this voice separation. Percent match is calculated by comparing the extracted melody pitch to each score voice for each slice in each chorale. Percent overlap is calculated by averaging the percentage of unique pitches for every slice, where the minimum is 25% (1/4) and the maximum is 100% (4/4). As this represents the number of unique pitches, minimal overlap equates to a mean percentage of 100% and maximal overlap of 25%. This is counterintuitive; therefore, values are subtracted from 100 so that a low percentage reflects low overlap and a high percentage, high overlap and the new minimum is 0% and the new maximum is 75%. This type of metric is only possible because of the nature of chorales, where each voice is independent and has a straightforward ground truth. Percent match between extracted voices and corresponding score voices is calculated the same as percent match between the extracted melody and score voices, where extracted voices and score voices are compared based on the first pitch of each extracted voice. Table 8.5 presents the results of these tests alongside results returned by the first version of the model, with a different LTM for each voice, using IDyOM's LTM modelling only, without considering local context. This additional test allows the evaluation of the influence of context on results.

Subsequent validation was performed on the 19 remaining chorales from the original 370 that had never before been seen by the model, whether during model development or initial evaluation. Training sets were derived from the 350 chorales of the Bach chorale test dataset. Table 8.6 presents the results of this validation test.

Finally, as it has been used in melody extraction elsewhere (Madsen & Widmer, 2006; Manzara et al., 1992), entropy will also be tested as an alternative measure to information content for the sake of comparison. While closely related, information content and entropy measure different aspects of prediction: information content measures how expected an event is, while entropy measure how certain the model is of its expectedness (see Section 2.3). There is no change in the model implementation, only in the metric that the decisions are based on; entropy is already implemented in IDyOM (Pearce, 2005). Results of this analysis can be found in Tables 8.7 and 8.8.

To evaluate the impact of test type and viewpoints used, a multiple linear regression model was constructed, with test type (version 1, 2, version 1 LTM only or version 2 melody-training), data type (information content or entropy) and viewpoint as fixed effects predicting voice match performance. Neither predictor was significant when compared to a model containing only an intercept using a likelihood ratio test, $F(3, 188) = 0.01, p = .99$, $F(1, 190) = 0.005, p < .94$ and $F(6, 185) = 0.001, p = 1$. While these values are extreme, the residual sum of squares was equal for both models compared, indicating that neither test type nor viewpoint had any influence on model performance. However, values fluctuate according to voice, and this was indeed found to be a significant predictor of performance, $F(3, 188) = 16.12, p < .0001$. Voice also significantly interacted with data type, where the voice match values for the bass and tenor voices are lower, and voice match values for the soprano and alto voices are higher when entropy is used as a prediction metric, $F(7, 184) = 29.43, p < .0001$.

Results for the validation set show no significant predictors of version, data type or viewpoint, $F(3, 188) = 0.07, p = .97$, $F(1, 190) = 0.02, p = .88$ and $F(6, 185) = 0.05, p = .99$ respectively, where voice was a significant predictor, $F(3, 188) = 9.99, p < .0001$.

Table 8.5. Results of the evaluation of two model versions on the Bach chorale test dataset as well as an LTM-only Version 1 and Version 2 training on melodic information only, using seven different viewpoint combinations of chromatic pitch (CP), pitch interval (PI) and scale degree (SD). All values are percentages. *Voice match* describes the percent match between the extracted melody line and each of the SATB chorale voices. *Voice overlap* describes the percent overlap between the four voices extracted by the algorithm. *Voice:Ground* describes the percent match between each extracted voice and its ground truth (i.e., the SATB voices respectively). The best performing condition is highlighted in bold in each column.

		Version 1			Version 2			Version 1 LTM only			Version 2 melody training		
		Voice match	Voice overlap	Voice:Ground	Voice match	Voice overlap	Voice:Ground	Voice match	Voice overlap	Voice:Ground	Voice match	Voice overlap	Voice:Ground
CP	S	0.7		28.4	10.3		45.5	0.7		59.6			
	A	2.7	35.8	45.8	22.7	42.8	58.3	5.1	28.8	58.1	-	-	-
	T	18.2		51.3	29.9		59.8	28.3		61.7			
	B	82.0		36.2	40.4		46.6	69.3		69.7			
S	2.4	38.4		12.9	46.0		2.2	48.5		21.7	33.8		
PI	A	9.9	32.7	34.8	23.6	42.8	55.0	12.6	37.1	47.8	17.2	29.4	
	T	29.7		40.3	29.7		55.5	41.8		42.5	20.3	62.2	30.0
	B	61.5		46.6	37.1		50.1	46.8		47.9	38.9	33.5	
	S	1.9		42.9	15.1		44.0	2.0		55.0	16.9	32.7	
SD	A	12.3	33.8	39.0	24.0	40.0	47.0	13.6	35.5	39.2	24.7	37.5	
	T	37.5		41.3	26.6		50.0	43.7		39.3	27.5	50.9	42.7
	B	51.9		43.2	37.2		51.1	43.8		43.9	32.5	47.2	
	S	2.4		78.0	3.6		75.5	0.3		66.7			
CP PI	A	17.4		69.8	9.6		73.6	3.0		59.8			
	T	47.5	21.9	70.0	23.6	23.6	71.3	27.3	27.8	62.3	-	-	-
	B	36.7		79.9	66.8		60.2	72.8		73.0			

CP SD	S	2.5		27.0	11.1		45.5	1.1		71.3			
	A	9.6		45.7	24.8		58.9	2.2		61.9	-	-	-
	T	29.9	37.2	51.3	31.3	45.0	57.5	21.2	24.5	66.0			
	B	61.6		34.4	36.1		41.6	78.7		80.2			
PI SD	S	1.1		41.5	3.2		75.9	2.0		53.4	16.3		20.9
	A	3.3		36.8	9.1		72.3	12.8		49.8	23.7		21.4
	T	18.7	32.7	41.9	23.2	23.3	71.2	41.9	38.3	45.4	28.6	47.9	19.4
	B	80.6		49.0	68.1		71.9	46.4		47.0	32.6		14.4
CD PI SD	S	1.2		38.0	8.0		51.8	1.4		71.1			
	A	6.1		41.5	17.9		60.2	7.7		59.6	-	-	-
	T	24.5	34.3	47.2	28.6	38.0	60.9	32.5	29.2	60.5			
	B	71.8		37.5	49.0		57.5	62.6		62.5			

Table 8.6. Results of the evaluation of two model versions and their variations on the Bach chorale validation dataset, using seven different viewpoint combinations.

		Version 1			Version 2			Version 1 LTM only			Version 2 melody training		
		Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground
CP	S	0.1		64.5	10.2		46.3	0.1		68.9			
	A	1.6		60.4	26.4		55.0	1.5		57.7	-	-	-
	T	21.7	26.5	63.6	29.0	44.4	53.9	20.0	25.6	66.7			
	B	80.1		80.1	37.1		46.5	81.7		81.7			
PI	S	1.1		51.5	13.2		42.0	1.1		54.3	10.7		44.3
	A	6.6	36.6	52.4	28.1	46.0	53.4	6.2	36.2	48.2	27.6	45.8	54.2

	T	44.3		46.2	29.8		47.1	42.1	45.6	31.0		49.7
	B	51.4		51.4	32.1		46.0	53.9	53.9	33.8		45.5
SD	S	1.6		56.1	17.1		32.8	1.0	43.6	22.9		33.4
	A	10.5	35.2	39.9	24.1	49.1	35.0	9.2	39.7	19.6	51.2	30.7
	T	46.3		39.6	28.0		41.7	47.0	34.6	36.8		42.8
	B	44.9		44.9	33.7		45.0	46.2		46.2	29.5	45.1
	S	0.2		65.2	10.8		47.8	0.2		66.1		
CP PI	A	3.0	28.3	54.0	26.1	44.0	54.4	2.4	54.8	-	-	-
	T	23.7		63.7	28.0		52.6	23.3	27.5	65.2		
	B	76.4		76.4	38.0		45.6	77.3		77.3		
CP SD	S	0.1		66.7	12.1		40.1	0.2		71.3		
	A	1.9	25.1	62.7	28.2		57.4	1.5		62.1	-	-
	T	21.2		67.2	30.5	46.9	56.7	18.8	23.3	69.9		
	B	80.5		80.5	31.8		41.2	83.2		83.2		
PI SD	S	1.2		58.6	12.5		49.2	1.2	59.3	39.2		51.0
	A	7.2	36.4	50.9	20.0	43.7	50.9	7.5	50.2	31.5	44.9	48.7
	T	42.2		47.2	28.3		55.2	41.8	37.0	49.0	21.8	53.4
	B	52.6		52.6	42.1		48.9	52.8		52.8	9.8	47.7
CD PI SD	S	0.5		69.4	10.2		38.5	0.6	71.0			
	A	5.3	28.2	58.0	23.0	47.3	49.3	6.4	58.2	-	-	-
	T	29.9		62.9	28.3		53.3	31.5	28.5	61.0		
	B	67.8		67.8	41.4		51.4	65.0		65.0		

Table 8.7. Results of the evaluation of entropy models on the Bach chorale training dataset, using seven different viewpoint combinations.

		Version 1			Version 2			Version 1 LTM only			Version 2 melody training		
		Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground
CP	S	40.4		59.7	31.3		62.1	40.4		10.6	-		-
	A	35.4	71.1	58.1	27.8	52.0	42.6	39.9	70.9	22.0	-		-
	T	14.9		61.7	22.3		30.3	13.5		49.9	-	-	-
	B	14.1		69.8	20.7		27.1	10.9		10.6	-		-
S	41.8	48.5		29.3	40.7		41.5	70.1		25.4		43.1	
PI	A	13.0	62.7	47.8	23.6	50.6	39.6	12.9	61.1	15.7	31.8	44.9	49.2
	T	13.2		45.2	24.7		38.6	14.1		21.2	28.6		49.0
	B	37.0		48.0	21.8		34.4	36.8		79.3	17.2		49.9
	S	16.4		55.0	20.6		22.4	16.9		25.3	20.2		32.5
SD	A	25.2	64.3	39.2	23.9	51.0	37.2	25.2	55.4	28.4	22.9	49.9	37.5
	T	30.3		39.3	25.4		40.2	30.3		31.0	29.8		42.7
	B	33.2		43.9	32.3		45.4	30.1		35.7	30.1		47.2
	S	55.8		66.7	30.2		60.6	56.2		8.6	-		-
CP PI	A	29.5	71.9	59.8	28.0	52.7	42.7	26.1	72.6	20.4	-	-	-
	T	8.9		62.3	22.1		30.6	8.9		46.0	-		-
	B	13.7		73.0	22.0		26.9	13.0		10.0	-		-
	S	40.2		71.2	29.3		62.7	40.7		10.4	-		-
CP SD	A	34.7	75.3	61.9	26.9	52.8	44.1	38.7	65.6	30.1	-	-	-
	T	15.7		65.9	23.0		30.9	14.1		40.9	-		-
	B	14.4		80.2	22.9		28.9	11.4		10.9	-		-
	S	34.5		61.5	53.5		26.1	54.1		39.3	34.8		64.2

PI SD	A	16.3		49.8	27.0		40.7	16.4		19.6	30.8	47.6
	T	18.3		45.4	23.2		38.5	18.3		24.6	29.8	49.0
	B	36.4		47.0	25.2		36.2	35.9		59.2	19.9	52.1
CD PI SD	S	53.4		70.8	28.8		59.9	50.3		8.5	-	-
	A	28.9	70.6	59.2	27.1	53.5	44.9	32.1	69.9	25.8	-	-
	T	10.0		60.7	23.5		31.7	10.7		39.0	-	-
	B	12.4		62.5	22.9		28.2	11.9		10.4	-	-

Table 8.8. Results of the evaluation of entropy models on the Bach chorale validation dataset, using seven different viewpoint combinations.

		Version 1			Version 2			Version 1 LTM only			Version 2 melody training		
		Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground	Voice match	Voice overlap	Voice: Ground
CP	S	48.2		10.2	34.9		66.6	46.2		10.5	-	-	
	A	32.0		22.2	28.9		41.6	39.2		21.2	-	-	
	T	8.3	28.8	49.1	19.9	53.0	23.9	8.4	28.7	50.0	-	-	
	B	15.4		11.8	18.2		27.0	10.4		10.4	-	-	
PI	S	37.3		73.8	32.5		41.9	37.7		73.6	30.6	45.1	
	A	12.7		14.2	23.3		36.4	12.6		14.1	33.0	46.1	
	T	12.9	39.6	21.1	22.8	50.4	36.5	12.9	39.8	21.0	24.8	45.7	
	B	42.6		82.7	23.6		36.2	42.3		82.5	15.0	48.2	
SD	S	18.2		27.3	23.9		34.6	20.0		27.4	18.4	31.0	
	A	27.1		29.1	20.2		33.2	26.2		27.1	21.9	35.6	
	T	29.4	44.0	30.2	27.2	51.7	36.7	29.9	44.7	29.8	28.9	51.1	
	B	30.8		35.8	31.6		47.7	29.3		34.8	32.9	45.5	

CP PI	S	62.1		8.3	32.7		64.0	63.3	8.0	-		-	
	A	25.6		21.2	30.9		42.3	24.9	22.1	-		-	
	T	5.4	25.5	43.3	19.4	53.8	24.4	5.5	25.4	44.7	-	-	
	B	10.5		10.5	18.8		26.6	9.7	9.7	-		-	
CP SD	S	45.1		8.8	28.0		63.7	42.7	8.1	-		-	
	A	36.4		30.4	30.0		44.0	39.2	29.3	-		-	
	T	11.0	32.2	37.4	20.5	55.2	26.2	11.8	32.5	39.5	-	-	
	B	11.7		11.7	23.8		29.6	11.2	11.2	-		-	
PI SD	S	27.7		49.6	24.6		36.9	30.6	49.5	23.8		40.6	
	A	13.2		16.6	26.6		39.1	13.8	16.3	33.8		48.7	
	T	21.4	36.1	26.3	27.1	54.7	34.6	21.3	36.3	25.1	28.0	46.0	45.4
	B	43.4		62.7	23.4		35.6	40.2	62.6	17.4		53.9	
CD PI SD	S	56.1		8.4	28.1		60.7	53.5	7.9	-		-	
	A	30.1		29.0	28.2		44.3	33.3	28.6	-		-	
	T	6.7	29.4	36.3	19.8	54.8	28.1	7.0	29.9	36.3	-	-	
	B	11.0		11.0	26.0		29.4	10.5	10.5	-		-	

8.3.2 String quartets

The String quartet dataset is described in Chapter 3, Section 3.2.2. Cross-validation was not used here; rather, the test, training and validation datasets are described in Section 3.2.2. This is due to the monumental task of obtaining a ground truth for this large amount of data. While there is a bias towards the first violin, or the highest pitch (Duane, 2012), it is not guaranteed that the melody will be perceived as the quartet’s ‘soprano’ voice. Therefore, a ground truth was built by annotation, where each quartet in the test set was annotated by three different listeners, the author and two others. All listeners studied music formally at the university level. Listeners were instructed to highlight (or otherwise mark) the melody on a score as they listened to the piece. It was specified that in this situation melody was restricted to monophony and only one pitch could be highlighted at a time (i.e., if harmonized, pick one note). They were also instructed to not leave any marks if they felt there was no melody at any point (i.e., textural passages). Rather than generating one common ground truth from the three annotations, analysis was performed for each annotation and results presented as an average of these three. This was done to maximise the information given by each annotation, as a ground truth containing parts of each annotation is not representative of any single one and valuable information may be lost. As results demonstrate, the deviation in performance of the model between annotations is low, indicating high agreement overall.

An important adjustment to the string quartets analysed is that all double stops were removed. This is due to the definition of melody being employed, which is restricted to monophony. While it would be possible to assign extra vertical sonorities to a new voice, this would be perceptually inaccurate the majority of the time, where double stops tend to provide harmonization and texture. Of the 218 tokens removed from the kern files, only 18 made up musical relevant voices (suspensions) for one or two measures at a time. Though it renders the

resulting analysis partially incomplete, this omission represents a negligible fraction of the overall analysis.

Though not a problem in Bach chorales, voicing becomes more relevant in string quartets as it is not guaranteed that a melody will always be present. It is however expected that unvoiced segments form only a negligible percentage of the music analysed. To verify this, the number of unvoiced slices, averaged according to the three annotators, were counted and represent 2.53% of the total number of slices (489 unvoiced slices of 19258).

As the voices of a string quartet do not directly correspond to an equivalent instrument in the score, voice overlap and comparisons between individual voices and their score equivalents as were carried out for the J.S. Bach chorales could not be done here. Instead, the percent match between the extracted melody and the annotated melody was calculated for each quartet, using the same viewpoints and combinations thereof as for the chorale analysis. In order to effectively evaluate model performance, two extremely simple comparison models were created. The first is a model that randomly selects one pitch or rest from the options present in the next slice as melody was also run. This model did not consider continuations and could therefore select a pitch whose onset belonged in a previous slice. This comparison is necessary because chance performance in string quartets does not equate to 25% due to rests and common thin textures. Output from this random model was only considered a hit when the selected pitch corresponded to the onset of a pitch but not when the selected pitch was already sounding. The second is a model that implements the skyline rule, selecting the highest sounding pitch at any given time. This comparison allows evaluation of the model against a simple heuristic. Results can be found in Table 8.9.

To effectively compare these values, a multiple linear mixed effects model was run, with random intercepts for each piece (7) and each annotation (3) and viewpoint (incl. random)

as a fixed effect. Viewpoint was a significant predictor when compared to a model containing only an intercept, $F(7, 160) = 212.76$, $p < .0001$, and remained significant when random intercepts were contained in the model, where scale degree, linked pitch interval and scale degree, and skyline performed significantly better than the rest, $t(161) = 4.06$, $p < .0001$, $t(161) = 2.58$, $p < .01$ and $t(161) = 34.96$, $p < .0001$ respectively, and a random model performed significantly worse, $t(161) = -2.24$, $p < .02$. The random intercepts improved a basic model,

Table 8.9. Mean (standard deviation) percent match for the Mozart string quartets between extracted melody and annotated melody for all three annotations for all combinations of *chromatic pitch* (CP), *pitch interval* (PI) and *scale degree* (SD) viewpoints, as well as a random model and skyline analysis. Performance for each string quartet is given separately as well as an overall mean across quartets. The best performing viewpoint (excluding random and skyline analyses) is bolded in each column.

	K428-1	K428-2	K458-3	K464-2	K499-3	K575-2	K590-1	Overall Mean
RAND	11.26 (0.55)	7.84 (2.82)	11.43 (2.10)	12.69 (2.90)	10.72 (1.07)	15.05 (0.88)	16.77 (0.25)	12.25 (3.17)
SKY	87.30 (2.92)	89.64 (3.85)	84.59 (1.63)	79.48 (8.26)	83.31 (8.46)	83.37 (5.18)	79.62 (2.71)	83.90 (5.70)
CP	27.18 (0.37)	13.35 (1.36)	16.79 (0.98)	16.30 (0.81)	12.81 (2.18)	8.74 (1.27)	10.57 (0.97)	15.10 (5.83)
PI	26.55 (0.74)	31.84 (2.29)	25.81 (4.23)	13.91 (0.56)	17.48 (1.72)	5.21 (0.91)	11.85 (0.45)	18.95 (9.13)
SD	25.79 (1.00)	19.40 (1.40)	25.73 (0.45)	16.48 (0.36)	19.63 (0.72)	22.14 (2.28)	41.59 (0.51)	24.39 (7.94)
CP	27.87	12.60	14.17	13.63	13.68	10.72	9.37	14.58
PI	(0.25)	(0.83)	(0.36)	(2.60)	(0.36)	(1.49)	(1.22)	(5.90)
CP	26.72	18.09	13.53	12.05	13.43	6.34	9.30	14.21
SD	(1.07)	(1.07)	(2.47)	(1.09)	(1.58)	(0.87)	(0.45)	(6.38)
PI	26.01	32.00	28.30	9.62	35.41	6.14	13.11	21.53
SD	(0.41)	(2.22)	(1.22)	(0.74)	(1.26)	(0.44)	(0.55)	(11.12)
CP	25.28	29.12	13.54	12.43	17.14	6.74	11.44	16.57
PI	(0.42)	(3.23)	(0.09)	(1.73)	(1.05)	(0.88)	(1.45)	(7.69)
SD								

with BIC reduced from -296.55 to -288.10, with random intercepts on piece explaining more variance than random intercepts on annotation, which had an associated variance of zero.

8.4 Discussion

While prediction-based, symbolic voice extraction is not new, the model presented above was created to offer a cognitively valid approach to the problem. In the first instance, it is an extension of an existing, validated model of cognitive expectancy in music, processing music in a non-causal manner and integrating long and short-term memory while taking past context into account. Second, its guiding hypotheses, that a melody is internally coherent and that it is the most interesting voice, are based on perceptual principles (Huron, 2001) and human and computational corpus analysis respectively (Duane, 2012). This new model was tested on chorales by J.S. Bach, which are by definition harmonized melodies, offering a straightforward ground truth where the melody is the soprano voice, and a selection of string quartet movements by W.A. Mozart, where ground truth was established using human annotation.

8.4.1 Bach chorales

Three metrics were used to evaluate the quality of the voice extraction algorithm for Bach chorales: 1) percent match between extracted melody and each score-based voice; 2) percent overlap between extracted voices; and 3) percent match between each extracted voice and its corresponding score-based voice. Several pitch viewpoint combinations were tested, using *chromatic pitch*, *pitch interval* and *scale degree*, as well as two test sets, using either all voices or only the soprano voice. Furthermore, two versions of the model were tested on the Bach chorales: one with an LTM especially trained for each voice, where for example predictions for the soprano line only had knowledge of other soprano lines and one with only one LTM model, with knowledge of all chorale voices to make all predictions. The first version

of the model was also tested using the LTM portion of IDyOM only in order to evaluate the impact of context on melodic extraction.

Results for Version 1 of the model (Table 8.5) suggest an advantage of chromatic pitch, with an 82% match between the extracted melody and the score's bass line. However, the lowest amount of overlap and best match between each extracted voice and corresponding score voice was achieved by a linked viewpoint between chromatic pitch and pitch interval. This pattern suggests that pitch is the best viewpoint for melody extraction, while a combination of pitch and pitch interval are better at source separation. However, results from multiple linear regression analyses indicate that viewpoint had no statistical effect on performance overall. Scale degree performs particularly badly on its own, though performs almost equally to chromatic pitch when combined with pitch interval. These results are somewhat surprising, particularly when thinking about voice separation, as it might be expected that pitch performs best on all metrics due to each extracted voice learning only from its corresponding score voice. Intuitively, these voices are defined by pitch much more than interval or scale degree patterns, where for example the soprano and bass lines may both utilise similar scale degrees but have no pitch overlap. The influence of local context in this version of the model is important as a model using LTM alone performed worse for most viewpoints. The exceptions to this are two pairs of linked viewpoints: chromatic pitch and pitch interval, and chromatic pitch and scale degree. The latter combination is the best performing viewpoint on all metrics, the most uniform result of all model tests undertaken. This pattern suggests that interval and scale degree, and particularly scale degree, can define chorale voices when large amounts of data are considered; however, when only considering a single chorale, these large-scale patterns are lost in the influence of the preceding context, which seems to blur the distinction between voices



Figure 8.4. First measures of Bach Chorale No. 6, where the closest next pitch in the soprano voice is in fact the same pitch, that occurs in the alto voice.

based on interval and scale degree in favour of pitch alone. Once again, despite these observed patterns, no statistical effect of viewpoint was observed.

When moving to a single LTM, best performance decreases overall, where the best viewpoints for melody extraction and

the majority of source separation metrics are linked pitch interval and scale degree. Linked pitch and pitch interval performs best for some of the source separation metrics but only by a very narrow margin. Similarly, where the interval and scale degree linked viewpoint performs best, the pitch and pitch interval linked viewpoint is only marginally worse. It is surprising that given these combinations, the combination of all three viewpoints does not perform particularly well in any metric with either model implementation. A further test using a combination of the two pairs of linked viewpoints chromatic pitch and pitch interval and chromatic pitch and scale degree does not perform better than any single linked viewpoints on their own. This indicates that an increase in musical information available to the model is in fact detrimental to melody extraction and voice separation, where focus on one or a single pair of viewpoints separates voices and selects the most interesting voice most effectively. Training on melody alone did not produce better results. On the contrary, though the extracted voice was most often matched to the score's soprano voice, all metrics demonstrated worse performance. While this experiment tested whether learning from melody alone could better identify the melody in context, voice separation comes first and thus results show that training on melody alone does not perform good voice separation.

In terms of voice overlap, where 75% is the maximum possible overlap with this metric, an overlap in half the slices equates to 37.5%. While no condition using Version 1 of the model reaches this, some come very close. This is due to convergence in extracted voices, where if an ‘incorrect’ pitch is assigned to a voice once, it may continue in the voice where this ‘incorrect’ pitch originated from, so that the final analysis contains two voices that are partially identical. Voice overlap is worse overall when training on melody alone, once again due to poor performance on voice separation with a restricted training set. The $s=5$ heuristic where the algorithm initially selects the closest pitch rather than the most probable is also sometimes a nuisance rather than a benefit and can lead to voice overlap from the first transition between slices. For example, in Chorale No. 6 (Figure 8.4), while the soprano voice moves from an F to an A, the closest pitch is in fact another F, sung by the alto voice. In this case, the algorithm is incorrect from the start. An s value of 5 was selected in prior piloting on a smaller dataset; it would be interesting to investigate the effect of context using this heuristic in the future. Despite these occasional errors, voice separation is generally successful for this test set, as evidenced by most extracted voices matching their respective score voices by more than 50%.

Entropy was also tested as a comparison measure of predictive processes in melody extraction. Results show that entropy performs more poorly in terms of voice separation, including some cases of particularly high, almost maximal, voice overlap. This leads to the following major difference seen in the voice match metric: while a model based on information content predominantly and clearly matches the score’s bass voice, a model based on entropy predominantly but less distinctly matches the score’s soprano voice. This pattern is supported by the significant interaction between voice and data type. This presents an interesting pattern of results, particularly when comparing to previous research: entropy has been used to extract melody from scores that were already separated into their respective voices (Madsen &

Widmer, 2007), whereas here entropy is used to perform both voice separation and melody extraction. Perhaps these processes are better served by different types of predictive information: information content is best at voice separation, using predictability to track voices, while entropy is best at melody extraction, using uncertainty to measure the ‘interest’ of a voice (Friberg & Ahlbäck, 2009). This combination could be tested in future research. In terms of viewpoints, best performance across metrics was mostly shared by pitch interval, scale degree and linked chromatic pitch and scale degree. However, there is no statistical difference in performance between these, and ‘best’ performance is often marginally better than the next best value.

Results of testing on the 19-chorale validation set are similar on all points, where the linked viewpoint chromatic pitch and scale degree performs well overall and in these chorales even outperforms chromatic pitch alone for tests using Version 1 of the model, though not by very much. Entropy still performs poorer voice separation. The similarities and small differences between the test and validation datasets highlight the vital importance of evaluation on varied datasets of various sizes, where the influence of training set size as well as musical style is currently under investigation in the Music Cognition Lab and preliminary results indicate an influence on predictions generated by IDyOM as a function of these factors.

The results presented in Section 8.3.1 also demonstrate a clear disagreement with one of the central assumptions of this study, where the soprano voice was defined as melody. Rather, the model has identified the bass line as the most interesting line. An analysis of mean information content for each voice in all chorales from the evaluation set, as calculated by the default IDyOM model with chromatic pitch as a source viewpoint, reveals that the bass voice is the most ‘interesting’ voice, as defined by information content. The soprano voice has an average IC of 2.07, the alto voice of 2.51, the tenor voice of 2.51 and the bass voice of 2.68.

This may vary based on the viewpoint selected, therefore the same analysis was performed using pitch interval, with average IC 2.11, 2.74, 1.80 and 2.77 for soprano, alto, tenor and bass voices respectively, where here the bass voice is the most complex and using scale degree, with average IC 2.89, 3.39, 3.33 and 3.57 for each voice respectively, where the bass voice is again the most complex. This simple analysis demonstrates that the algorithm successfully accomplished what it was designed to do by extracting the most interesting voice in the polyphonic context; therefore, the definition of melody was inappropriate for application to Bach chorales, and the second hypothesis of this chapter is unsupported. The same analysis using entropy as a metric yielded mean entropies of 2.58, 2.78, 2.74 and 3.27 respectively for the soprano, alto, tenor and bass voices when chromatic pitch was the source viewpoint; 2.88, 2.98, 3.00 and 2.99 when pitch interval was the source viewpoint; and 4.16, 4.20, 4.22 and 4.61 when scale degree was the source viewpoint. The mean entropy of the soprano voice is consistently the lowest, which also disproves the chapter's second hypothesis and counters the suggestion above that while information content is best for voice separation, entropy may be a better metric for melody selection. Given that the Bach chorale melodies' mean entropies are lower than all other voices, it is likely that the higher prevalence of melodic notes extracted by the model when using entropy are a confound of poorer voice separation. This is supported by the greater presence of all voices in the extracted melody (see Table 8.7, Voice Match).

It is worth noting that the percentages for the first metric, voice match, do not add up to 100%. This is because of voice overlap in the score, which occurs from time to time in Bach chorales. In these cases, a match will be recorded for more than one voice, thus exceeding 100%. Similarly, there is an apparent discrepancy between the first and third metrics, however, these are not directly related. The voice identified as melody may not be the bass voice as determined by the first pitch; for example, what began as the tenor voice may have been

assigned a higher proportion of bass notes, and due to occasional shared pitches between voices in the ground truth and overlap between extracted voices, the third metric is not directly comparable to the first. It is interesting that this discrepancy only appears in the test dataset and not in the validation set, where it is clear that the bass voice was maintained and selected as the melody. This is likely due to the scale of the dataset, where the bass voice was selected as melody in all 19 validation chorales but not in all 350 test chorales.

8.4.2 String Quartets

While Bach chorales are harmonized melodies, string quartets present new challenges to melody extraction: the potential for roving melodies, the presence of rests and fluctuating voice numbers, and the presence of non-melodic segments. Where existing models of melody extraction cannot handle roving melodies, the melody extraction model described in Section 8.3.1 can do so as it selects the most likely continuation for each voice without knowledge of the corresponding score-based voice. Versions 1 and 2 of the model require homophonic textures, whereas Version 3 was created to handle rests and fluctuating voice numbers, where sounding notes are assigned to voices first. It was therefore this version that was used to extract the melody in seven string quartet movements by W.A. Mozart (see Table 3.3 for details). These movements were randomly selected from all Mozart quartet movements that contain exactly four notes in the opening slice. This choice was made for simplicity and to restrict the otherwise large amount of data available for analysis. Of these seven, two are first movements, three are second movements and two are third movements. This in turn reflects differing styles, as string quartet movements are designed to be contrasting. Some of these are *adagio* movements, characterized by many scalar and decorative passages, and one is a *minuet*, a particularly sparse and repetitive style. These differences have a visible effect on results, as will be discussed in more detail below.

Due to the potential for the melody to rove, straightforward comparison between model output and score is not possible here, therefore melody annotations were collected from three different listeners and the percent match between each annotation and model output was calculated, where the mean and standard deviation of these three values are presented in Table 8.9. The model performed melody extraction on these seven quartet movements using three pitch-based viewpoints and all their combinations: *chromatic pitch*, *pitch interval* and *scale degree*. Two additional models were run for comparison: a random model and a skyline model. Results suggest the best overall approach to be the skyline model (see Table 8.9), however, when considering only the prediction-based model viewpoints, the scale degree viewpoint performed best, either alone (three movements) or in combination with pitch interval (three movements). The remaining movement is best represented by the linked chromatic pitch and pitch interval viewpoint. These results highlight scale degree as a particularly important viewpoint in this task.

However, it is worth noting that in some cases the best performing viewpoint was only marginally better than the next best, as in the cases of movements K428-1, K428-2 and K646-2. In the case of K428-1, all viewpoint combinations perform similarly, with performance at least double that of the random model. In the case of K428-2, linked pitch interval and scale degree only slightly outperforms pitch interval alone, which is also slightly better than all three linked viewpoints. Finally, in the case of K464-2, chromatic pitch is the runner up to scale degree. While in the first two cases the best performing viewpoint combinations share common viewpoints, for K464-2 they do not, and it is interesting that the combination of chromatic pitch and scale degree performs worse than the two individually. However, unlike other movements, no viewpoint or viewpoint combination for K464-2 performs better than 4% above random-selection model performance, where the sparse texture of the movement is an issue for the

prediction-based model. While a linear mixed modelling analysis identifies viewpoint as a significant predictor when it is a fixed effect alone, this effect is overshadowed by differences in performance between movements, evident in Table 8.9 and highlighted in the discussion below.

Overall performance on this task is also much lower than analysis performed on Bach chorales. To effectively evaluate the quality of this model's performance, a model that simply selects one random pitch (or rest) from the next slice of music as melody, and a skyline analysis were also run. As can be seen in Table 8.9, the prediction-based model generally performed better than the random model while performing worse than the skyline approach. Exceptions to improved performance in comparison to the random model include K464-2, K575-2 and K590-1, where for some viewpoint systems the random model outperforms the prediction-based model, though the best-performing viewpoint system outperforms the random model in all cases. These three particular movements are sparse, containing many slices with less than four voices, where in these sections the voice chosen as melody by the model overwhelmingly contains rests instead of any notes. By contrast, a model randomly selecting pitches is likely to select a pitch at some point, where the longest sparse section lasts 20 measures (K590, movement 1, mm 77-98) and there are 2-3 voices sounding for the majority of the section.

Despite this, K590-1 is the movement with the highest model performance at 41.59% when using the scale degree viewpoint alone. This is presumably due to the scalar nature of the piece and the fact that the opening theme of the movement (Figure 8.5) was picked up by the model immediately, either in the first violin or in the cello, and recognized regularly throughout, though not every instance was present in the extracted melody voice; it is possible that these were assigned to other voices instead. However, it is worth noting that this opening excerpt would be perceived as one voice and one melody, rendering any 'selection' of a melody

perceptually incorrect. A model beginning with stream formation, such as the one proposed in Chapter 4, is necessary to possibly achieve this perceptual accuracy. Since the implementation in this chapter is score voice based, this issue must be conceded for the time being.

Patterns. Despite performance remaining below 50%, it is worth noting that the prediction-based melody extraction model is generally good at pattern detection, with pitch interval and scale degree, and combinations thereof being the best viewpoints to identify these. However, not all patterns detected are perceptually correct, which leads to a maintained low performance as once these are identified, they tend to be maintained throughout the movement. Some examples of successful and unsuccessful pattern detection will be described here for each movement.

Figure 8.6 illustrates examples from the first quartet movement, K428-01, where patterns A and D were detected more successfully and more reliably than patterns B and C,



Figure 8.5. Opening three measures of Mozart K590, movement 1. This pattern was detected by all viewpoints, though all but the scale degree viewpoint selects the pattern in the cello octave while scale degree selects the pattern in the Violin I octave.

K428-1
 A mm1-4
 B mm5-6
 C mm75
 D mm77-8

Figure 8.6. Examples of pattern detection in Mozart string quartet movement K428-1; clefs same throughout. The opening pattern (A) was picked up by most viewpoints, with the exception of the first pitch, which tended to be the lower octave. The following pattern (B) was sometimes picked out successfully (blue) by the CP and CP-PI viewpoints and sometimes unsuccessfully but consistently (red, illustrating the extraction by the pitch interval viewpoint). Pattern C, always annotated as melodic, was only picked up by the model in its full 6-note pattern once each by the CP-SD, PI-SD and CP-PI-SD viewpoints, out of 14 occurrences in the piece. Finally, pattern D (arpeggiated triplets), was also always annotated as melodic, and was picked up by every viewpoint except scale degree.

which were either sometimes incorrectly, yet reliably, or incompletely, yet reliably, identified respectively. Red boxes here represent patterns identified by annotation, where in the case of pattern C, it was not successfully detected by the model, and in the case of pattern D, it was successfully detected using all but one viewpoint configuration. Figure 8.7 illustrates examples from K428-02, whose extracted melody always began in the cello (red box), sometimes staying there for several measures and sometimes moving to other instruments. The annotated melody, in contrast, was always in the first violin. The scalar pattern found in mm32 (pattern B, red box) is best detected by the pitch interval viewpoint, matching melodic annotations. Figure 8.8

Figure 8.7. Examples of pattern detection in Mozart string quartet movement K428-2. Pattern A (mm1), though never annotated as melody, is reliably detected by all viewpoints in extraction. Pattern B (mm32), always annotated as melody, is best picked up by viewpoints including pitch interval information.

illustrates examples from K458-3, an adagio movement featuring many scalar and decorative passages. These are best detected by the pitch interval (e.g., pattern B) and scale degree (e.g., pattern A) viewpoints, where the two illustrated patterns are always or most often annotated as melodic

Figure 8.8. Examples of pattern detection in Mozart string quartet movement K458-3. Pattern A (mm11), often but not always annotated as melodic, is only successfully picked up by viewpoints containing scale degree information while pattern B (mm15-6), always annotated as melodic, is picked up by most viewpoints, though the pattern is not always complete. Viewpoints containing pitch interval information perform best on pattern B.

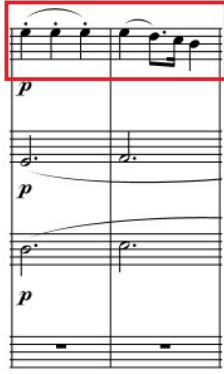


Figure 8.9. The principle melodic pattern of Mozart's movement K464-2 (mm5-6) is poorly identified by all viewpoints.

respectively. Figure 8.9 illustrates examples from K464-2, the minuet movement whose sparsity has been challenging for the prediction-based model. Accordingly, pattern detection is poor, with the movement's principle melodic pattern detected only once by the PI-SD linked viewpoint and twice by the CD-PI-SD linked viewpoint, out of 24 instances of the pattern. However, it is worth recognizing that this pattern sometimes occurs in canon (overlaps) in this

movement, which restricts successful detection as the model is confined to monophony. In K499-3, another adagio movement, scales and octaves are particularly prevalent. Figure 8.10B illustrates an example of successful scalar pattern extraction by the PI-SD linked viewpoint, where the model matches melody annotations. While octaves are rarely annotated as melody, they are a pattern well picked up by the CP-SD linked viewpoint, commonly found in the accompanying voices. Another notable moment in this movement is the partial identification of the first violin line as melody by the PI-SD linked viewpoint (Figure 8.10 A). This opening pattern is also partially picked up when it recurs in mm54, though different notes are selected. Figure 8.11 illustrates examples from K575's sparse second movement, where pattern A, the principle melodic pattern of the movement, is only sometimes picked out by scale degree. Similarly, secondary pattern B is only partially picked up by all viewpoints. Quartet movement K590-1 is the movement with the best overall performance, with a mean 41.59% match between the extracted melody and the three melody annotations. As mentioned previously, the opening pattern (Figure 8.5) was

Figure 8.10 consists of two musical excerpts, A and B. Excerpt A shows the opening measures of a string quartet movement, with four staves (Violin I, Violin II, Cello, and Double Bass). Red asterisks are placed above notes in the first and second violin staves, and blue asterisks are placed below notes in the cello staff. The dynamic marking *p* and the instruction *cresc.* are visible. Excerpt B shows a scalar passage in the first violin (top staff, red box) and an octave passage in the cello (bottom staff, red box). The dynamic marking *p* is visible.

Figure 8.10. Examples of pattern detection in Mozart string quartet movement K499-3. The opening measures (A) commonly result in the selection of the cello voice (blue), while only the PI-SD combined viewpoint selects some notes from the first and second violins. Featuring many scalar passages (B, mm35, violin I) and octaves (B, cello), viewpoints containing pitch interval and scale degree perform best at identifying these patterns, where scalar passages are more often annotated as melodic (B red box).

identified by most viewpoints, though sometimes identified in the lower octave (cello). A particularly successful excerpt of melody extraction is illustrated in Figure 8.12A, while the measure in Figure 8.12B is representative of the long stretches of sparsity in the movement where melody extraction by prediction fails.

Figure 8.11 consists of two musical excerpts, A and B. Excerpt A shows a sparse passage with four staves. A red box highlights a pattern in the first violin staff. Excerpt B shows a sparse passage with four staves. A red box highlights a pattern in the cello staff.

Figure 8.11. Examples of pattern detection in Mozart string quartet movement K575-2. Being a sparse movement, pattern detection is poor, with annotated patterns A (mm5-6) only sometimes detected by the scale degree viewpoint and B (mm21-3) only partially detected.

Though most patterns are straightforward, there are some instances where it will be impossible for this prediction-based model to successfully match the melody annotations provided by listeners. Though these cases are few, it is worth mentioning when these situations arise in order to address these better in the future. The first, already briefly mentioned, is the possibility of canonical excerpts, where a melody is performed in a round, with a second repetition of the melody beginning before the first is finished. In this situation, the listener would likely be able to perceive that the melody is repeated in an overlapping manner and would hold the melody perceptually intact throughout. Such excerpts are present in quartet K464, movement 2, where the canonized pattern is shown in Figure 8.9. Another, similar example is illustrated in Figure 8.13, where a pattern moves from one voice to another while the other voices sustain notes. Melodic annotation of this excerpt tended to follow the moving pattern rather than the sustained notes. However, due to the model's treatment of note continuations, where a continuation must be assigned to the voice containing its onset, the complete pattern can never be captured by the model in its current implementation. However, this is currently a relatively rare problem faced by the model.

The detailed look at model performance on all quartet movements above highlight some of the more salient issues faced by the current prediction-based model of melody extraction. The first, evident by poor performance for quartets K464-2, K575-2 and K590-1, is the difficulty it has in sparse musical conditions, where slices contain rests in one or more voices. It seems that pitches are assigned to other voices, which have lower IC overall, while rests are assigned to the voice with the highest mean IC. One possible explanation for this pattern is the design of the model itself: pitches are to be assigned according to the lowest IC. This hypothesis clearly does not hold here as it does for Bach chorales. The second

A

Figure 8.12. Examples of pattern detection in Mozart’s string quartet movement K590-1. The segment above (A; mm105-8) is part of a 12-measure long portion of highly successful melody extraction by the best performing viewpoint for this movement, scale degree. Here highlighted pitches (red) represent the pitches from the first violin that were not present in the extracted melody; no accompanying voices were included in the extracted melody in this segment. The measure on the right (B; mm21) is representative of the sparse texture in the movement where melody extraction fails, where the voice returned as the melody contains rests throughout these sections.

B

main issue, illustrated in Figure 8.10A, is the high frequency of movement between score voices as the model frequently selects pitches separated by large intervals, resulting in melodic patterns being broken into incomplete pieces. While the model should already prefer smaller intervals to larger ones through training, this pattern was not seen in the model’s performance. Perhaps the viewpoint should be calculating compound intervals (the interval between pitch classes rather than the interval between pitches) instead of basic intervals, concentrating the model’s learning to within an octave. It is interesting that this solution would strongly imply the use of pitch interval for melody extraction, resulting in the assumption that this viewpoint is salient to listeners. These adjustments would be interesting to implement and test in future research.

While voice separation cannot be appropriately evaluated in relation to perceptual validity, a few tests can be conducted in order to better understand the model's successes and shortcomings. With each voice differing in length and containing varying numbers of rests in varying locations, it is difficult to automatically calculate overlap between the model's extracted voices. Instead, observation of the data indicates very high overlap between the two upper and two lower voices and high instances of large leaps. As this prediction-based model's melody extraction performance depends on successful initial voice separation, it is possible that poor voice separation is the cause of poor performance and that the assumption that melody is the most complex voice remains valid (Duane, 2013; Madsen & Widmer, 2007). To test this assumption, mean IC of each score-based voice was



Figure 8.13. Mozart string quartet K428, movement 1, mm34-8. In this excerpt, the melody is typically annotated as moving from one voice to another, ending on a long, sustained note before changing voices. However, due to constraints in treating continuation events, the current model is incapable of matching such an annotation.

calculated for *chromatic pitch*, *pitch interval* and *scale degree* viewpoints (Table 8.10). While these do not necessarily correspond to perceived voices, the success of the skyline approach, where the first violin most often plays the melody in these quartets, allows the approximation to be an interesting guideline. In the case of chromatic pitch, the first violin line narrowly has the highest mean IC, 3.58, while other voices have slightly lower mean IC: 3.45, 3.58 and 3.27 for the second violin, viola and cello respectively. However, for pitch interval viola is the most complex voice,

Table 8.10. Mean information content for each score-based voice and annotated melody (mean of the three annotations) for seven Mozart string quartets.

	Chromatic Pitch	Pitch Interval	Scale Degree
Violin 1	3.58	3.05	4.61
Violin 2	3.45	3.16	4.27
Viola	3.58	3.27	4.34
Cello	3.27	2.90	4.63
Annotations (3)	3.33	2.87	4.74

Table 8.11. Mean information content of the extracted melody voice for each tested viewpoint system, including chromatic pitch (CP), pitch interval (PI) and scale degree (SD).

	CP	PI	SD	CP-PI	CP-SD	PI-SD	CD-PI-SD
Mean IC	6.22	6.27	6.15	6.35	6.52	6.23	6.45

where mean IC values are 3.05, 3.16, 3.37 and 2.90 for first and second violins, viola and cello respectively and for scale degree the cello is the most complex voice, where mean IC values are 4.61, 4.27, 4.34 and 4.63 for the first and second violins, viola and cello respectively. For comparison, mean IC for the same viewpoints were calculated for the annotated melodies, with mean IC 3.33, 2.87 and 4.74 for chromatic pitch, pitch interval and scale degree respectively. While falling in similar ranges, only the scale degree viewpoint might have resulted in the melody being selected as the most complex voice, assuming that other annotated voices have lower mean IC. The variation in highest mean IC may be due to the fact that pitch interval and scale degree detect patterns more successfully, reducing the mean information content as repeated patterns are predictable and therefore have low IC. Perhaps something other than information content characterises melody. For example, a compression-based analysis would be interesting to conduct as an alternative to information content, as both melodic and accompanying voices contain patterns, where their differences typically lie in length and complexity. While a melody would be expected to be the least compressible voice, with longer

and more complex patterns, such an analysis may reveal new and interesting information and would make an interesting comparison to the current approach.

While mean IC values for score-based voices and the annotated melodies are comparable within each viewpoint, mean IC for these same score-based voices are not comparable to the mean IC of extracted melodies, which are much higher (see Table 8.11). This supports the observed high frequency of large leaps, which increases the mean IC. While it is possible that the melody may be in a different extracted voice and that the frequency of large leaps inflated the IC of the selected voice, observation indicates that voice separation is poor across all voices. Once again, this indicates that improving voice separation should be the focus of further model development.

8.4.3 General Discussion

In this chapter, a prediction-based model of melody extraction was presented, guided by two hypotheses: a melody is internally coherent, and a melody is the most interesting stream. The model was evaluated on a total of 369 chorales by J.S. Bach and 7 string quartet movements by W.A. Mozart and results support the first hypothesis, but not the second.

A melody's hypothesized internal coherence is based on the pitch proximity principle (Huron, 2001) and is implemented in the current melody extraction model through the selection of the most likely pitch continuation for a given voice, with pitches within the range of a small leap (< perfect fifth) hypothesized to be the most likely pitches. Testing the model using chorales by J.S. Bach allows the evaluation of the quality of melodic coherence due to the compositional nature of the chorales, where each voice can be considered an independent melody and these do not cross voices. Results demonstrate good voice separation, where voice overlap is a result of any mis-attribution of a pitch to its correct voice having a long-term effect

on the remaining analysis, where the voice receiving the incorrect pitch will likely continue in that voice. Results also demonstrate a good match between extracted voices and their respective ground truths, with percent match ranging from 61.9% to 80.2% for linked chromatic pitch and pitch interval, chromatic pitch and scale degree, and pitch interval and scale degree viewpoints, depending on the model version tested. The same evaluation cannot be conducted for string quartets, as there is no such direct correspondence between perceived voices and score. However, observations of exceptionally well extracted excerpts (see Figure 8.12A) support the hypothesis that melodies are internally coherent and that this melodic characteristic can be extracted by relying on information content. On the other hand, high occurrence of large leaps in other areas, such as the model identifies in Figure 8.10A, suggests that the most likely continuation for a voice is not necessarily the closest pitch. However, the mis-attribution of rests may be the cause of inflated information content (Table 8.11). This causes a higher tolerance for large leaps, where pitches following a rest belong to a new phrase and thus are unexpected with respect to the immediate context (Pearce, Müllensiefen, et al., 2010). While this may be the case, melodic annotations of the string quartets indicate that pitch proximity is preferred to large leaps in this music. Therefore, future implementations of this model may integrate temporal information in order to more effectively handle rests in voices.

Melody has previously been defined as the most interesting voice in a piece of music (Duane, 2012; Madsen & Widmer, 2007). However, while Bach chorales are by definition harmonized melodies, the melodic voice extracted by the current model corresponds most strongly to the bass voice. As discussed in Section 8.4.1, mean IC of each score-based voice is indeed highest for the bass, with the soprano voice often having the lowest mean IC. This suggests a completely opposite definition for melody: the melody is the most predictable voice. Though melodic pitches were more often included in melodies extracted using entropy rather

than information content, so were other voices and melodic extraction results are no better overall, with score-based mean entropy being highest in the bass voice as well. However, with existing evidence suggesting otherwise (Duane, 2012; Madsen & Widmer, 2007), it is worth highlighting the differences in musical style from which these definitions are derived. As previously mentioned, chorales, composed in the style of 18th century counterpoint, are composed of four independent voices, each complex in its own right. In contrast, string quartets and symphonic works (Duane, 2012; Madsen & Widmer, 2007) are typically composed of a melody and an accompaniment, where both may rove and where the accompaniment is much simpler and more repetitive than the melody, often simply providing harmonic support and context for the melody. Therefore, the ideal melody extraction model should be able to recognize its context and apply the appropriate melodic definition to the situation.

Another important consideration in melody extraction is the role of high-frequency salience. It is well known that the highest pitch in a polyphonic context is typically the most salient to a listener, a phenomenon known as high voice superiority (Marie, Fujioka, Herrington, & Trainor, 2012; Marie & Trainor, 2012; Marie & Trainor, 2014; Trainor, Marie, Bruce, & Bidelman, 2014). This is reflected in the creation of the *skyline* algorithm (Uitdenbogerd & Zobel, 1998), which simply selects the highest pitch in the polyphonic context as the melody. This works well for pop music and would perform perfectly for Bach chorales, but breaks down for symphonic music, where melody tends to rove and where accompaniment is commonly higher in pitch than the melody (i.e., flutes). String quartets represent a middle ground between these two styles, where with only four instruments, roving is minimized and timbral similarity results in the minimization of the masking of melodies played by lower pitched instruments by higher pitched instruments. In the quartets analysed above, melodic annotations from three listeners greatly favoured the highest pitch as evidenced by the high

performance in melody extraction by the skyline approach over the prediction-based approach, suggesting that this would be a useful bias to incorporate into a melody extraction model for application to music for small ensembles in particular. As the prediction-based model succeeds where the skyline approach fails (e.g. Figure 8.8A and B, where the prediction-based model identifies the melodic pattern even when it is not the highest pitch), it would be interesting to combine the skyline and prediction-based approaches in a hybrid model that includes a measure of certainty to the prediction-based decision. In this hybrid model, the skyline portion would always select the highest pitch with 100% certainty, while the prediction-based model would select a pitch with less certainty, directly corresponding to the entropy associated with the prediction. If the entropy is low, the pitch selected by the prediction-based model would be selected. Conversely, if entropy is high, the model would default to the skyline choice. Such a model would likely favour the highest pitch, while still allowing the possibility of roving across instruments.

Returning to the current results, there is a marked decrease in performance between melody extraction for Bach chorales and Mozart string quartets. From observation of the extracted melodies, the issue, previously highlighted in Section 8.4.2, is related to voicing, where chorales contain four voices throughout while the number of voices in string quartets fluctuates. It is likely that no other extracted string quartet voice is a better match to the annotated melodies, as voice separation for these quartet movements is only partially successful (e.g. Figure 8.12A) while mostly poor (e.g., Figure 8.10A). Therefore, voice separation must first be improved to improve melody extraction more generally. As mentioned previously (Section 8.4.2), this could be solved by estimating melody at each slice rather than once at the end of the piece, preferring the assignment of pitches over rests to the melody at any given slice.

While this evaluation allows the testing of the two hypotheses of this chapter, performance of this model cannot be directly compared to other models of melody extraction due to differences in evaluation datasets. Furthermore, while most melody extraction and voice separation models focus on score-based voice separation where melody tends to be confined to a single voice, this model is performing perceptual melody extraction, where the melody is identified by common practice or expert listeners and can change voices through the piece of music. Therefore, comparison to a random model is the most effective evaluation possible for this model at the present time, assessing the extent to which the model performs better than chance. Additional comparison to the skyline approach on the other hand identifies the gaps in the model's performance and gives a minimum performance to aim for.

Finally, it is worth clarifying the extent of the cognitive validity of this model. Though based on IDyOM, which has been evaluated for cognitive validity (Hansen & Pearce, 2014; Pearce & Müllensiefen, 2017; Pearce, Müllensiefen, et al., 2010, 2010; van der Weij et al., 2017), these evaluations are based on monophonic stimuli. Therefore, while it may be true that listeners learn patterns in music as IDyOM models them, listeners are unlikely to learn about chorale or string quartet structures by listening to each individual instrument, as the melody extraction model was trained here. Harmonic viewpoint implementations and polyphonic learning need to be tested to fully implement cognitive validity for polyphonic music perception in IDyOM.

8.5 Conclusion

In this chapter, a prediction-based model of melody extraction was presented and evaluated on Bach chorales and Mozart string quartet movements, testing two hypotheses: a melody is internally coherent and a melody is the most interesting voice in a polyphonic piece

of music. While the first hypothesis is largely supported by current results and by existing literature (Huron, 2001), the second hypothesis was not supported, with results suggesting that the melody is instead the least interesting voice. Consideration of the conflicting evidence highlights the importance of musical style in this problem, where Bach chorales are composed of four independent melodies while string quartets are built in melody-accompaniment style. While voice separation performs well for Bach chorales, the model struggles with fluctuating voice numbers, where this is an issue to target in future research.

9 Conclusion

This thesis set out to do two things: 1) propose an integrated framework for auditory streaming using existing literature to produce a powerful, flexible and potentially collaborative research tool for future research, and 2) investigate a range of aspects of auditory streaming using probabilistic approaches. Chapter 4 presented an integrated theoretical model of auditory streaming in music perception and Chapters 5-8 each investigated a distinct, specific subset of the auditory streaming problem for music, all together spanning the five categories of source information necessary to inform streaming. Identified in detail in Chapter 2, these are: 1) auditory features; 2) musical features; 3) attention; 4) expectation; and 5) listener background, or musical training. In this final chapter, Section 9.1 will summarize the content of this thesis, followed by a presentation of its outcomes and limitations, in Sections 9.2-3. Section 9.4 will

present a systematic evaluation of the integrated framework for auditory streaming presented in Chapter 4 and Section 9.5 will wrap up with suggestions for future research.

9.1 Summary

First, Chapter 4 presented an integrated framework for musical ASA composed of a combination of modules, each containing a predictive model. This framework, when implemented, would be the first to incorporate information from all sources relevant to musical ASA, namely auditory features, musical features, attention, expectation and listener background, on complex stimuli and combine these to produce a perceptually motivated analysis of auditory streaming, where existing streaming models perform score-based voice separation. Furthermore, this framework would be an extremely rich research tool available to researchers, allowing the experimental exploration of musical ASA in parts and as a whole with the potential to develop even further as research in the field continues due to its modules' common modelling approach: prediction.

Chapter 5 investigated the role of timbre as a streaming cue and asked whether the specific instrument a musician played would bias their streaming perception, where it was hypothesized that listeners would be more sensitive to their own instruments' timbre. This effect was not found; rather, results suggest that directed attentional set and expectations about what they were about to hear influenced listeners' perception more strongly than their listening background. Additionally, the success of the paradigm, never before attempted with ecologically valid timbral sounds, further validates timbre as a relevant streaming cue. Finally, the bistable percept reported by participants in the ABA paradigm suggests that in this simple streaming context, timbre is a less salient cue than pitch, timing or loudness. Thus, this chapter

addressed a combination of auditory features, attention, expectation and listener background aspects of the integrated framework for auditory streaming.

By asking listeners to rate their expectancies, or arousal and valence in real time, the study in Chapter 6 experimentally validated a subset of IDyOM's temporal viewpoints and applied these along with pitch viewpoints to the prediction of expectation and perceived emotion in monophonic musical excerpts, where both pitch and timing expectancy were significant predictors of rated expectancy, arousal and valence for monophonic folk music. This both extends the application of predictive processes in the context of music perception, providing evidence for a prediction-based mechanism of musical emotion (Egermann, Pearce, Wiggins, & McAdams, 2013; Juslin, Liljeström, Västfjäll, & Lundqvist, 2011; Meyer, 1956), and sets the foundation for the studies of Chapter 7, which make use of the same temporal viewpoints.

The studies in Chapter 7 address a particularly important aspect of musical auditory streaming: the relative perceptual salience of musical parameters, where the most salient parameter will dictate the organization of the auditory scene. As proposed in Chapter 2, these studies investigate the potential link between predictability and salience, where it is hypothesized that less predictable, and therefore more complex, parameters are more salient due to higher cognitive processing demands. While predictability was strongly linked to complexity, its link to salience in the chapter's second study was tenuous. Rather, results indicated that perceived ratings of complexity in relation to polyphonic stimuli manipulated in melodic, rhythmic and harmonic predictability were better predictors of detection of differences between pairs of stimuli than measures of predictability.

Chapter 8 presented an extension of IDyOM that is also a simplified implementation of the proposed integrated framework for musical ASA from Chapter 4, investigating the use of

prediction as a method for voice segregation and melody extraction. In this simplified implementation, IDyOM is applied to a polyphonic context by analysing each vertical slice of music iteratively, assigning the most likely continuation for each existing voice, the number of which is determined by the number of pitches in the first slice. Once all voices have been constructed, the melody is selected as the voice with the highest average information content, in other words, the most interesting voice. This algorithm was tested on chorales by J.S. Bach and string quartet movements by W.A. Mozart. While voice separation was successful for chorales, the presence of rests and the fluctuation of the number of voices hindered performance for string quartets. Furthermore, it was found that the bass voice rather than the melody was the most interesting voice in Bach chorales. Where successful melody extraction here depends on successful voice separation, performance on string quartets was poor. Compared to a chance model and a skyline approach, results from the prediction-based model performed better than chance but worse than the skyline approach, resulting in the consideration of a hybrid model combining prediction with the rule-based skyline decision. This will be discussed further in Section 9.4.5 below.

9.2 Outcomes

This thesis has produced a number of useful outcomes to the scientific community. First, in Chapter 5, ecologically valid timbral sounds were confirmed as valid streaming cues, extending current knowledge of timbre as a basic auditory feature cue. Second, where IDyOM is fast becoming a recognized cognitive model of musical expectation, its temporal viewpoints have only recently been used in research (M. Pearce & Müllensiefen, 2017; Marcus T. Pearce, Müllensiefen, et al., 2010; van der Weij et al., 2017) and only confirmed as valid models of perceived expectation here (Chapter 6; Sauv e, Sayed, Dean & Pearce, in review). Third,

Chapter 7 established a strong link between information content, a measure of predictability derived using IDyOM, and perceived complexity, a link previously theorized but not empirically tested (Eerola, 2016; Huron, 2006). Fourth, Chapter 8 presented an extension of the IDyOM model for melody extraction that will be published in a subsequent release of the software. Fifth, Chapters 5-7 provided new evidence concerning the link between musical training and perception, where most of this evidence supports the initial assumption (Section 2.5) that previous training is an important aspect to take into account for any study of auditory streaming, influencing expectations (Chapter 5) and accounting for individual differences in perception (Chapters 6 and 7). However, it is important to note that in Chapter 5, attentional set and expectancies outweighed effects of listening background. This discrepancy contributes to the rich findings of the current literature (Section 2.5) and highlights the need for further research. Finally, Chapter 4 presented the synthesis of a large body of research, proposing a powerful framework for musical auditory streaming.

9.3 Limitations

This thesis also has its limitations. Chapter 4's most glaring limitation is that the integrated framework it proposes is theoretical and is not yet implemented, nor empirically validated as a whole. Other model limitations were discussed in Section 4.8. The small sample size in Chapter 5 is that chapter's most important limitation, where larger power would produce more reliable statistical results. In Chapter 6, pitch and temporal expectations successfully predicted listeners' perceived expectations and emotional reactions to the music; however, this music was selected specifically for its extreme expectancy features. Furthermore, these melodies are monophonic as opposed to polyphonic and while the models produced have excellent fit to the data, their ability to predict new data remains untested. These three tests of

generalizability would strengthen the findings presented in this chapter considerably. In Chapter 7, neither experimental manipulations, information content nor musical features could fully explain perceived complexity ratings or, in particular, same/different task performance; alternative predictors should be explored to better characterize perceived complexity and salience and until then, results should be considered encouraging but tentative. In Chapter 8, it became clear that melody is not universally the most interesting line in a polyphonic piece of music, as exemplified by chorales by J.S. Bach, where the bass line was instead selected, thus disproving the chapter's primary hypothesis. This highlights the need for context to be considered in melody extraction. Furthermore, melody was constrained to monophony, which is not always accurate, for example in the case of a canon. Future development of the model should allow more than one voice to be melodic, though there should be a strong prior for monophony so that only strong evidence for multiple melodies will result in such an analysis. Finally, the analysis was only conducted on quartets opening with exactly four pitches and double stops were not considered. Future editions of this model should be able to analyse all string quartets in their original form.

9.4 Framework evaluation

In this section, the integrated framework for auditory streaming proposed in Chapter 4 will be systematically evaluated in relation to the evidence provided by the studies presented in Chapters 5-8.

9.4.1 Timbre, musical training, attention and expectation

The pair of studies in Chapter 5 covered four of the five sources of auditory streaming information, namely auditory features, musical training, attention and expectation. In the proposed integrated framework for auditory streaming, only an idealized method of timbre

modelling could be presented as there is no existing model of polyphonic timbre. While the results presented in Chapter 5 bring us no further to such a model, they do confirm timbre as a valid streaming cue. Results additionally demonstrate differences in perception for more or less similar timbres (based on listener similarity ratings), where more similar timbres are more often perceived as integrated.

Musical training does not affect streaming perception as a function of timbre in these studies, but attention and expectations do, supporting their inclusion in an integrated framework for auditory streaming. However, the lack of influence of musical training should not be taken to suggest that this source of information is irrelevant but that perhaps such minor distinctions between instrumentalists are too subtle to have an effect on auditory streaming perception.

9.4.2 Temporal viewpoints

The proposed integrated framework for auditory streaming presented in Chapter 4 assumed the inclusion of temporal viewpoints, among others, while these had yet to be validated in the IDyOM system. Results from Chapter 6 validate the use of at least one temporal viewpoint, demonstrating that inter-onset interval predictability matches listener predictability ratings in short monophonic melodies. While more of IDyOM's viewpoints, temporal in particular, require validation for use in the proposed integrated framework, this study provides the first validation outside of monophonic pitch viewpoints.

9.4.3 Relative salience

In Chapter 7, two studies were designed to connect information content to perceived complexity, and in turn to relative salience. While a link between information content and perceived complexity was supported by the data, a link between complexity and salience, and therefore information content and salience, was not. This affects the theoretical framework

proposed in Chapter 4, as the implementation of attentional bias relied on the modelling of relative salience using information content. In the absence of evidence for such a link, corroborating evidence should be sought in replications and simplified experiments, where only pairs of parameters are compared, as well as an alternative implementation. For example, perhaps the model would make the default assumption that melody is the most salient aspect of a piece of music and would prioritize another parameter, such as rhythm, only when onset synchrony exceeds a particular threshold, or harmony, only when the information content of a harmonic progression suddenly exceeds a particular threshold.

However, while a strong link between information content and salience was not established, a few patterns can be extracted from results of Chapters 6-8. First, Chapter 7 was focused specifically on the relative perceptual salience of melody, harmony and rhythm in polyphonic music, where salience was defined by information content. Results from this chapter support previous evidence that the pitch dimension is more salient than the rhythmic dimension (Palmer & Krumhansl, 1987; Prince et al., 2009), where predictability and salience are correlated, while the relative salience of harmony with respect to melody and rhythm is unclear and would be an interesting research avenue to pursue. On the other hand, results from Chapter 6 suggest that the rhythmic dimension is more salient, with onset information content predicting both expectancy and arousal and valence ratings of monophonic folk melodies with more weight than pitch information content. However, the musical contexts of these studies vary greatly in complexity, where in Chapter 7 the stimuli are short, dense 2-bar, 3-voice excerpts while in Chapter 6 the stimuli are an average of 30s long, monophonic folk songs, a generally simple musical style. These differences suggest that timing information may be more salient, or more easily accessible in simpler musical contexts while pitch information is needed to process more complex scenes. In Chapter 8, the hypothesis that melody was the most

interesting (or unpredictable) voice in a polyphonic piece of music was tested, where results indicated that this was not the case in Bach chorales. Written in the style of 18th century counterpoint, these chorales are composed of four independent, interesting voices and the bass line was in fact the most interesting, with the highest mean information content and that the soprano line, in chorales equating to the melody, was the least interesting, with the lowest mean information content. This is in complete contradiction to existing evidence suggesting that the melody is the most complex voice (Duane, 2012; Madsen & Widmer, 2007). However, evidence for the latter case comes from string quartet and symphonic works, styles commonly built on a melody and accompaniment structure where the accompaniment is typically quite repetitive, reducing mean information content. Therefore, context is once again crucial to defining the relationship between predictability and salience, where this evidence supports the important role of musical style in melody perception. Finally, in Chapter 5, while predictability was not a factor, the bistability of the ABA_ sequences reported by listeners even when timbre was maximally different between the two tones suggests that timbre alone is not a sufficient cue for complete segregation. With existing evidence that parameters like pitch and tempo are sufficient (van Noorden, 1975), this in turn suggests that pitch and time are more salient parameters than timbre, when all other parameters are held constant. This evidence is provided in the context of extremely simple stimuli, where orchestration practice suggests it is likely that timbre plays a much more significant role in musical auditory scene analysis in more complex contexts such as symphonic works. However, it is difficult to come up with examples of such works where pitch, rhythm and loudness are equally salient and where timbre can therefore become a defining characteristic of the music. Perhaps the single best example of such a musical situation is Boléro by Maurice Ravel; in this situation, the only thing left to manipulate is timbre, and this defines the work.

To conclude, it seems that the salience of an event depends entirely on perceptual context, including the preconceptions of the perceiving agent; the role of context will be discussed further in Section 9.4.5 below.

9.4.4 Melody extraction

In chapter 8, the concept of prediction was applied to the common engineering problem of melody extraction, where melodies were hypothesized to be internally coherent and the most interesting line in a piece of music. While the implementation in this chapter is a simplification of the proposed framework from Chapter 4, the concepts and hypotheses with relation to melody identification are the same. While the first hypothesis was supported by tests on chorales by J.S. Bach, where voices were well separated into score-based voices using prediction alone, the second was not, where the voice with the highest average information content – or highest average entropy – was not the melody. Neither hypothesis was supported when the implemented melody extraction module was tested on Mozart string quartet movements; however, overall performance on this task was poor due to the difficulties the model has dealing with rests, where future work is needed to address this implementation issue. While these results disagree with the literature inspiring the high information content melody hypothesis (Duane, 2012; Madsen & Widmer, 2006), it is important to point out that these experiments were carried out on different musical styles. Bach chorales consist of four independent voices, while string quartets and symphonies are structured differently. Once again, though this thesis did include a test on string quartets, the voice separation was too poor to perform meaningful melody extraction.

Therefore, once again, the influence of context is inescapable, and thus will be discussed in more depth forthwith.

9.4.5 Considering Context

As it has become clear in the past few pages, the role of context deserves some discussion; as the existing literature is compared and new results are obtained, it seems that particular approaches only function best in particular contexts and do not generalize to all music. For example, our prediction-based melody extraction model performs better for homophonic music by J.S. Bach than for polyphonic music by W.A. Mozart. While pointing out context seems a default explanation, humans can easily identify the melody in both contexts – the challenge remains to model it. The obvious initial solution is to present the model with all sorts of contexts to learn from, so that it may in time recognize patterns as belonging to one context or another and therefore apply the appropriate response. This would require substantial amounts of encoded musical corpora, which certainly do not exist in the form necessary for the proposed integrated framework of Chapter 4. In the meantime, perhaps each musical context requires its own hypothesis, though the issue then becomes one of categorization: what music belongs to what context label, and how is melody defined in each?

Another possibility might be to complement the predictive approach with a different type of modelling, one based on heuristics, here reflecting the use of less information to make decisions. This approach, and its potential as a complement to the n-gram approach of prediction used in this thesis and as the basis of an integrated framework for auditory streaming, is introduced and discussed next.

9.4.6 Considering Heuristics

As powerful as the predictive coding framework is, ignoring information can be a more efficient cognitive process. These are heuristics, or rules about the environment that reduce processing load and are recently being shown to be more accurate than models designed to take

into account all available information in a given environment (Gigerenzer & Brighton, 2009; Hutchinson & Gigerenzer, 2005). This is in contrast to the mentality developed towards the end of the 20th century that heuristics were second-best, the result of cognitive limitations, and that more information is always better. Gigerenzer & Brighton (2009) give an excellent review of the state of the systematic study of heuristics, which will be briefly summarized before being tied into the predictive coding framework.

The idea that “less is more” came to light in the 1970s, where it was demonstrated using two methods: tallying, and take-the-best. In tallying, models with equal weights (coefficients) on each cue predicted data better than multiple linear regression, where weights were fitted to the data. Multiple linear regression fit the data better, but had poorer predictive power due to over-fitting (Czerlinski, Gigerenzer, & Goldstein, 1999). While this works best for small sample sizes in relation to available cues, dependency between cues and low predictability of these cues, it is pointed out that this is often the situation in natural environments. Rarely is all the possible information available and new situations arise regularly where quick, accurate decisions are needed. Take-the-best is one of the one-good-reason class of heuristics, which finds the first cue that allows a decision to be made and ignores all other information (Gigerenzer & Goldstein, 1996, 1999). Cues are considered independent and compared only to the object of the decision. Again, this works best when information is sparse.

Heuristics are typically associated with bias, which generally carries a negative connotation because it is considered irrational. Rational processes consider all the available information to make a decision; heuristics do not and are therefore irrational. However, humans rarely have access to all the relevant information and must use what is available to make decisions. This is why heuristics are considered such a powerful model of human cognition. This is well understood in machine learning, where total prediction error is a combination of

bias, variance and noise (Friedman, Hastie, & Tibshirani, 2001). A model with zero bias reflects the underlying process perfectly while some bias is only an approximation of this process. A model with zero variance is insensitive to differences between samples from which the underlying process is being extracted; this is practically impossible. Studies have shown that higher variance is more harmful to accurate prediction than higher bias, as high variance is a symptom of an over-generalized model that contains many adjustable predictors and fits particular samples too well (Gigerenzer & Brighton, 2009). Therefore, better fit leads to poorer predictions; yet, the current status quo in psychology is to search for the best fitting model without considering its performance in predicting new or other data (Roberts & Pashler, 2000).

The breakdown of prediction error into bias, variation and noise has the potential to seamlessly fit into the predictive coding framework, as there is not yet any specific detailing of how its prediction error is encoded. If predictive coding is at its core the creation of models of the world continuously updated through error processing, why could these models not be, for example, take-the-best models, with high bias and low variance? Heuristics are adaptable, updated through new information and error (Gigerenzer & Brighton, 2009), exactly as learning is proposed to function according to predictive coding (see Section 2.6). The integrated framework for musical auditory streaming proposed in Chapter 4 could therefore potentially contain heuristic predictive models as part of its framework, leading to fast decisions based on prioritized cues chosen as a result of the immediate context. While heuristics are typically interpreted and implemented as rules, such as in some of the streaming models presented in Sections 4.1 and 4.2, this conception of heuristics as models explaining prediction error with bias, variation and noise is different and offers a new way of testing the predictive coding framework, as well as thinking about human cognition in general. This approach could potentially help solve the context issue discussed in Section 9.4.5 above and speed up the

overall processing time of the proposed framework, a current limitation mentioned in Section 4.8. This alternative implementation method also offers the opportunity to compare the n-gram approach to predictive coding with a heuristic approach to predictive coding, offering interesting insights into human cognition.

9.5 Future directions

Despite the limitations presented in Section 9.3, this thesis contributes new insight into the relationship between prediction and the perception of musical auditory scenes. Beginning with an ambitious proposed integrated framework for auditory streaming, this thesis breaks the problem down into smaller chunks investigated separately before returning to evaluate the proposed framework in the context of this new data. Clearly, much work is still needed to develop the implementation of this proposed integrated framework in order to evaluate it. First, adapting IDyOM for polyphonic music is key, diversifying the number of viewpoints, especially harmonic, that it can learn and predict. There is also a need to cognitively validate these; both tasks are currently being tackled in the Music Cognition Lab. Second, the ‘fusion’ stage of the framework, where the output of modules are collected to merge and inform a streaming decision must be implemented. Third, data with annotated streaming perception should be collected. This step is arguably valuable for all music streaming research, where perceptual streaming can begin to be investigated in more depth once the appropriate data is made available. Finally, the framework should be made accessible, where only high-level coding is necessary to develop a new module, thus making collaboration easier on a larger scale.

More specifically, the results (both positive and negative) of the work presented in this thesis can guide the development of interesting new research questions based on each chapter.

From Chapter 4, an endless supply of research questions exploring musical ASA can be posed and investigated using the proposed framework as a powerful, flexible, collaborative research tool. First, auditory and musical modules can be included selectively, allowing the comparison of performance using only pitch interval, or only timbre, or the combination of both, or any combination of auditory and musical modules. Direct comparisons such as these within a single framework will help unify the disjointed evidence in the current literature, where different model approaches evaluate different information sources differently. Second, the relative weights of these modules can be manipulated, investigating relative salience. Furthermore, alternatives to linear combinations of these modules can be explored, such as the product or exponential combination of outputs. Third, the framework can be trained to specialize in different musical styles and to reflect different areas of expertise (i.e. jazz, raga, flautist, etc.), simulating differences in perception between these listening groups. How significant these differences might be, as well as where these differences lie provide interesting questions for future research. Fourth, can the framework be more efficient with the inclusion of predictive heuristics, and to what extent? Should all decisions be made by predictive heuristic models or only the most high- or low-level (i.e. parameter or viewpoint) ones? These are only some of the research questions that can be explored with this new framework, once implemented. From Chapter 5 we may ask whether directed attention influences perception more than listener background using an alternative task, such as transcription from a polyphonic context. From Chapter 6 we may ask whether the manipulation of temporal expectancy affects listeners' expectations and emotional reactions similarly to manipulations of pitch expectancy. From Chapter 7, other potential predictors of perceived complexity such as entropy or change in complexity should be explored, along with using a wider range of stimuli (i.e. monophony, pairs of voices, chorales, short and long excerpts, etc.). This in turn may better inform a

definition of perceptual salience and its relationship to predictability and its local context. From Chapter 8, we may ask whether an increased use of heuristics, and better pattern recognition, may achieve better voice separation and therefore better melody extraction. One heuristic may be to rely on one viewpoint alone, in a take-the-best model approach, or to ignore pitches outside of a learned range. Overall, the contents of this thesis make valuable contributions to the research community, helping us to move forward with new tools, new knowledge and new questions.

Appendix A – Details of the stimuli for Chapter 4, Study 1

Timbre (original file name)	Pitch	Length (ms)	Peak Amplitude (dB)	Fadeout (ms)
Cello (CelA3_3.84sec)	A3			
Cello (CelC#4_2.44sec)	C#4			
Cello (CelD4_2.77sec)	D4	114	-16	10
Cello (CelE4_2.67sec)	E4			
Cello (CelF4_2.56sec)	F4			
Cello (CelF#4_2.12sec)	F#4			
Trombone (TTbnG3_2.17sec)	G3		-12	
Trombone (TTbnG#3_2.22sec)	G#3			
Trombone (TTbnB3_2.54sec)	B3	113	-15	10
Trombone (TTbnD4_2.81sec)	D4			
Trombone (TTbnD#4_3.54sec)	D#4		-16	
Trombone (TTbnF4_3.01sec)	F4			
Trumpet (CTptG#3_6.06sec)	G#3		-12.5	
Trumpet (CTptA#3_2.75sec)	A#3		-16	
Trumpet (CTptC4_7.44sec)	C4	111	-13	10
Trumpet (CTptD#4_3.54sec)	D#4		-12	
Trumpet (CTptE4_7.42sec)	E4		-15	

Trumpet (CTptF#4_6.55sec)	F#4		-14	
Violin (VlnG3_8.79sec)	G3		-16	
Violin (VlnA3_8.98sec)	A3		-15	
Violin (VlnA#3_8.58sec)	A#3		-16	
Violin (VlnB3_9.67sec)	B3	114	-12	10
Violin (VlnC4_7.69sec)	C4		-14	
Violin (VlnC#4_7.12sec)	C#4		-16	

Bibliography

- Agres, K., Abdallah, S., & Pearce, M. (2017). Information-Theoretic Properties of Auditory Sequences Dynamically Influence Expectation and Memory. *Cognitive Science*. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/cogs.12477/full>
- Aksentijevic, A., Smith, A., & Elliott, M. A. (2014). Rate-specific Entrainment of Harmonic Pitch: Effects of Music Training. *Music Perception: An Interdisciplinary Journal*, 31(4), 316–322. <https://doi.org/10.1525/mp.2014.31.4.316>
- Alain, C., Arnott, S. R., & Picton, T. W. (2001). Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. *Journal of Experimental Psychology: Human Perception and Performance*, 27(5), 1072–1089. <https://doi.org/10.1037/0096-1523.27.5.1072>
- Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30(8), 2960–2966.
- Alluri, V., & Toiviainen, P. (2010). Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre. *Music Perception: An Interdisciplinary Journal*, 27(3), 223–242. <https://doi.org/10.1525/mp.2010.27.3.223>
- Andreou, L.-V., Kashino, M., & Chait, M. (2011). The role of temporal regularity in auditory segregation. *Hearing Research*, 280(1–2), 228–235. <https://doi.org/10.1016/j.heares.2011.06.001>
- Andrews, M. W., & Dowling, W. J. (1991). The development of perception of interleaved melodies and control of auditory attention. *Music Perception*, 8(4), 349–368.
- Anvari, S. H., Trainor, L. J., Woodside, J., & Levy, B. A. (2002). Relations among musical skills, phonological processing, and early reading ability in preschool children. *Journal of Experimental Child Psychology*, 83(2), 111–130. [https://doi.org/10.1016/S0022-0965\(02\)00124-8](https://doi.org/10.1016/S0022-0965(02)00124-8)
- Aslin, R. N., Saffran, J. R., & Newport, E. L. (1999). Statistical learning in linguistic and nonlinguistic domains. In B. MacWhinney & B. (Ed) MacWhinney (Eds.), *The emergence of language*. (pp. 359–380). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Aucouturier, J.-J., Pachet, F., & others. (2002). Music similarity measures: What's the use? In *ISMIR* (pp. 13–17). Retrieved from <http://www.music.mcgill.ca/~ich/classes/mumt614/similarity/AucouturierSimilar.pdf>
- Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology*, 41(3), 254–311. <https://doi.org/10.1006/cogp.2000.0738>
- Barniv, D., & Nelken, I. (2015). Auditory Streaming as an Online Classification Process with Evidence Accumulation. *PLOS ONE*, 10(12), e0144788. <https://doi.org/10.1371/journal.pone.0144788>

- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). <https://doi.org/10.1016/j.jml.2012.11.001>
- Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: a key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance*, *6*(3), 501.
- Baumann, S., Koeneke, S., Schmidt, C. F., Meyer, M., Lutz, K., & Jancke, L. (2007). A network for audio-motor coordination in skilled pianists and non-musicians. *Brain Research*, *1161*, 65–78. <https://doi.org/10.1016/j.brainres.2007.05.045>
- Beauvois, M. W., & Meddis, R. (1991). A computer model of auditory stream segregation. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *43A*(3), 517–541.
- Beauvois, M. W., & Meddis, R. (1996). Computer simulation of auditory stream segregation in alternating-tone sequences. *Journal of the Acoustical Society of America*, *99*(4, Pt 1), 2270–2280. <https://doi.org/10.1121/1.415414>
- Beauvois, M. W., & Meddis, R. (1997). Time delay of auditory stream biasing. *Perception & Psychophysics*, *59*(1), 81–86. <https://doi.org/10.3758/BF03206850>
- Bendixen, A., Denham, S. L., Gyimesi, K., & Winkler, I. (2010). Regular patterns stabilize auditory streams. *The Journal of the Acoustical Society of America*, *128*(6), 3658–3666. <https://doi.org/10.1121/1.3500695>
- Bendixen, A., Denham, S. L., & Winkler, I. (2014). Feature Predictability Flexibly Supports Auditory Stream Segregation or Integration. *Acta Acustica United with Acustica*, *100*(5), 888–899. <https://doi.org/10.3813/AAA.918768>
- Bendixen, A., Prinz, W., Horváth, J., Trujillo-Barreto, N. J., & Schröger, E. (2008). Rapid extraction of auditory feature contingencies. *Neuroimage*, *41*(3), 1111–1119.
- Bendixen, A., Roeber, U., & Schröger, E. (2007). Regularity extraction and application in dynamic auditory stimulus sequences. *Journal of Cognitive Neuroscience*, *19*(10), 1664–1677.
- Bendixen, A., SanMiguel, I., & Schröger, E. (2012). Early electrophysiological indicators for predictive processing in audition: A review. *International Journal of Psychophysiology*, *83*(2), 120–131. <https://doi.org/10.1016/j.ijpsycho.2011.08.003>
- Bendixen, A., & Schröger, E. (2008). Memory trace formation for abstract auditory features and its consequences in different attentional contexts. *Biological Psychology*, *78*(3), 231–241.
- Benetos, E. (2015, July). *ASyMMuS Workshop on Audio-Symbolic Music Similarity Modelling*. Workshop, London, UK.
- Berenzweig, A., Logan, B., Ellis, D. P., & Whitman, B. (2004). A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, *28*(2), 63–76.
- Bigand, E., McAdams, S., & Forêt, S. (2000). Divided attention in music. *International Journal of Psychology*, *35*(6), 270–278. <https://doi.org/10.1080/002075900750047987>

- Bigand, E., & Poulin-Charronnat, B. (2006). Are we “experienced listeners”? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130. <https://doi.org/10.1016/j.cognition.2005.11.007>
- Bigand, Emmanuel, Parncutt, R., & Lerdahl, F. (1996). Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, *58*(1), 125–141. <https://doi.org/10.3758/BF03205482>
- Bilhartz, T. D., Bruhn, R. A., & Olson, J. E. (1999). The Effect of Early Music Training on Child Cognitive Development. *Journal of Applied Developmental Psychology*, *20*(4), 615–636. [https://doi.org/10.1016/S0193-3973\(99\)00033-7](https://doi.org/10.1016/S0193-3973(99)00033-7)
- Bittner, R. M., Salamon, J., Essid, S., & Bello, J. P. (2015). Melody Extraction by Contour Classification. In *ISMIR* (pp. 500–506). Retrieved from <http://ai2-s2-pdfs.s3.amazonaws.com/03e6/8704aea0929e217e0f99449f458ba85d7c44.pdf>
- Bittner, R. M., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014). MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *ISMIR* (Vol. 14, pp. 155–160). Retrieved from http://matthiasmauch.de/_pdf/bittner2014medleydb.pdf
- Blumenthal-Dramé, A., Hanulíková, A., & Kortmann, B. (2017). Perceptual linguistic salience: Modeling causes and consequences. *Frontiers in Psychology*, *8*. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5360726/>
- Bod, R. (2002). Memory-Based Models of Melodic Analysis: Challenging the Gestalt Principles. *Journal of New Music Research*, *31*(1), 27–36. <https://doi.org/10.1076/jnmr.31.1.27.8106>
- Bogdanov, D., Serrà, J., Wack, N., Herrera, P., & Serra, X. (2011). Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, *13*(4), 687–701.
- Boltz, M. G. (1999). The processing of melodic and temporal information: independent or unified dimensions? *Journal of New Music Research*, *28*(1), 67–79.
- Bosch, J., & Gómez, E. (2014). Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms. In *Proc. 9th Conference on Interdisciplinary Musicology–CIM14, Berlin, Germany*. Retrieved from http://phenicx.upf.edu/system/files/publications/cim14_submission_114_ready.pdf
- Bosch, J., & Gómez, E. (2015). Melody extraction by means of a source-filter model and pitch contour characterization (MIREX 2015). *11th Music Information Retrieval Evaluation EXchange (MIREX), Extended Abstract, Málaga, Spain*. Retrieved from <http://www.mtg.upf.edu/system/files/publications/Bosch%20MIREX%202015.pdf>
- Bosch, J. J., Bittner, R. M., Salamon, J., & Gómez, E. (2016). A comparison of melody extraction methods based on source-filter modelling. *Proc. ISMIR, New York*. Retrieved from http://m.mr-pc.org/ismir16/website/articles/256_Paper.pdf

- Bosch, J. J., Marxer, R., & Gómez, E. (2016). Evaluation and combination of pitch estimation methods for melody extraction in symphonic classical music. *Journal of New Music Research*, *45*(2), 101–117.
- Brattico, E., Näätänen, R., & Tervaniemi, M. (2001). Context effects on pitch perception in musicians and nonmusicians: evidence from event-related-potential recordings. *Music Perception: An Interdisciplinary Journal*, *19*(2), 199–222. <https://doi.org/10.1525/mp.2001.19.2.199>
- Brattico, E., Pallesen, K. J., Varyagina, O., Bailey, C., Anourova, I., Järvenpää, M., ... Tervaniemi, M. (2009). Neural discrimination of nonprototypical chords in music experts and laymen: an MEG study. *Journal of Cognitive Neuroscience*, *21*(11), 2230–2244. <https://doi.org/10.1162/jocn.2008.21144>
- Bregman. (1978). Auditory Streaming is Cumulative. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(3), 380–387.
- Bregman, A. S., & Rudnicki, A. I. (1975). Auditory segregation: stream or streams? *Journal of Experimental Psychology. Human Perception and Performance*, *1*(3), 263–267.
- Bregman, Albert S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press.
- Bregman, Albert S., & Pinker, S. (1978). Auditory streaming and the building of timbre. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *32*(1), 19–31. <https://doi.org/10.1037/h0081664>
- Breitkopf, & Härtel. (1875). *371 vierstimmige Choralgesänge von Johann Sebastian Bach*. (4th edition. Plate Number: V.A.10. Rtypeset c. 1915 as Edition Breitkopf 10. Reprinted by Associated Music Publishers, Inc., New York [c. 1940].). Leipzig: Alfred Dörrffel.
- Burgoyne, J. A., Wild, J., & Fujinaga, I. (2011). An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis. In *ISMIR* (Vol. 11, pp. 633–638). Retrieved from <http://ismir2011.ismir.net/papers/OS8-1.pdf>
- Burunat, I., Brattico, E., Puoliväli, T., Ristaniemi, T., Sams, M., & Toiviainen, P. (2015). Action in perception: prominent visuo-motor functional symmetry in musicians during music listening. *PLoS ONE*, *10*(9), e0138238. <https://doi.org/10.1371/journal.pone.0138238>
- Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: a confirmatory study using synthetic tones. *The Journal of the Acoustical Society of America*, *118*(1), 471–482.
- Cambouropoulos, E. (2008). Voice and stream: Perceptual and computational modeling of voice separation. *Music Perception*, *26*(1), 75–94. <https://doi.org/10.1525/mp.2008.26.1.75>
- Cambouropoulos, E. (2009). How similar is similar? *Musicae Scientiae, Discussion Forum 4B*, 7–24. <https://doi.org/10.1177/102986490901300102>
- Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81–105. <https://doi.org/10.1016/j.cognition.2014.12.005>

- Carlsen, J. C. (1981). Some factors which influence melodic expectancy. *Psychomusicology: A Journal of Research in Music Cognition*, 1(1), 12–29. <https://doi.org/10.1037/h0094276>
- Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, 27(1), 115–127. <https://doi.org/10.1037/0096-1523.27.1.115>
- Carpentier, F. R. D., & Potter, R. F. (2007). Effects of Music on Physiological Arousal: Explorations into Tempo and Genre. *Media Psychology*, 10(3), 339–363. <https://doi.org/10.1080/15213260701533045>
- Cavnar, W. B., Trenkle, J. M., & others. (1994). N-gram-based text categorization. *Ann Arbor MI*, 48113(2), 161–175.
- Chan, A. S., Ho, Y.-C., & Cheung, M.-C. (1998). Music training improves verbal memory. *Nature*, 396(6707), 128–128. <https://doi.org/10.1038/24075>
- Chew, E., & Wu, X. (2004). Separating Voices in Polyphonic Music: A Contig Mapping Approach. In U. K. Wiil (Ed.), *Computer Music Modeling and Retrieval* (pp. 1–20). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-31807-1_1
- Cirelli, L. K., Spinelli, C., Nozaradan, S., & Trainor, L. J. (2016). Measuring Neural Entrainment to Beat and Meter in Infants: Effects of Music Background. *Auditory Cognitive Neuroscience*, 229. <https://doi.org/10.3389/fnins.2016.00229>
- Clark, A. (2013a). The many faces of precision (Replies to commentaries on “Whatever next? Neural prediction, situated agents, and the future of cognitive science”). *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00270>
- Clark, A. (2013b). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, 36(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Cleary, J. G., Teahan, W. J., & Witten, I. H. (1995). Unbounded length contexts for PPM. In *Data Compression Conference, 1995. DCC '95. Proceedings* (pp. 52–61). <https://doi.org/10.1109/DCC.1995.515495>
- Conklin, D. (2002). Representation and discovery of vertical patterns in music. In *Music and artificial intelligence* (pp. 32–42). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-45722-4_5
- Conklin, D., & Bergeron, M. (2010). Discovery of Contrapuntal Patterns. In *ISMIR* (Vol. 2010, p. 11th). Retrieved from <http://www.ehu.es/cs-ikerbasque/conklin/papers/ismir2010-36.pdf>
- Corrigall, K. A., & Trainor, L. J. (2011). Associations Between Length of Music Training and Reading Skills in Children. *Music Perception: An Interdisciplinary Journal*, 29(2), 147–155. <https://doi.org/10.1525/mp.2011.29.2.147>
- Critchley, H., & Seth, A. (2012). Will studies of macaque insula reveal the neural mechanisms of self-awareness? *Neuron*, 74(3), 423–426.

- Cuddy, L. L., & Lunney, C. A. (1995). Expectancies generated by melodic intervals: perceptual judgments of melodic continuity. *Perception & Psychophysics*, 57(4), 451–462.
- Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Routledge.
- Cumming, G. (2013). The New Statistics Why and How. *Psychological Science*, 0956797613504966. <https://doi.org/10.1177/0956797613504966>
- Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? Retrieved from <http://psycnet.apa.org/psycinfo/1999-04366-005>
- Dean, R. (2016). Algorithmically-generated corpora that use serial compositional principles can contribute to the modeling of sequential pitch structure in non-tonal music. *Empirical Musicology Review*, 11(1), 27–46.
- Dean, R. T., Bailes, F., & Dunsmuir, W. T. M. (2014a). Shared and distinct mechanisms of individual and expertise-group perception of expressed arousal in four works. *Journal of Mathematics and Music*, 8(3), 207–223. <https://doi.org/10.1080/17459737.2014.928753>
- Dean, R. T., Bailes, F., & Dunsmuir, W. T. M. (2014b). Time series analysis of real-time music perception: approaches to the assessment of individual and expertise differences in perception of expressed affect. *Journal of Mathematics and Music*, 8(3), 183–205. <https://doi.org/10.1080/17459737.2014.928752>
- Deike, S., Heil, P., Böckmann-Barthel, M., & Brechmann, A. (2012). The Build-up of Auditory Stream Segregation: A Different Perspective. *Frontiers in Psychology*, 3, 461. <https://doi.org/10.3389/fpsyg.2012.00461>
- Deliège, I. (2003). Special issue on music similarity. *Musica Scientiae*.
- Deliège, Irène. (2001). *Similarity perception ↔ categorization ↔ cue abstraction*. JSTOR. Retrieved from <http://www.jstor.org/stable/10.1525/mp.2001.18.3.233>
- Deliège, Irène. (2007). Similarity relations in listening to music: How do they come into play? *Musicae Scientiae*, 11(1_suppl), 9–37.
- Demorest, S. M., & Morrison, S. J. (2016). 12 Quantifying Culture: The Cultural Distance Hypothesis of Melodic Expectancy. *The Oxford Handbook of Cultural Neuroscience*, 183.
- Denham, S. L., & Winkler, I. (2006). The role of predictive models in the formation of auditory streams. *Journal of Physiology, Paris*, 100(1–3), 154–170. <https://doi.org/10.1016/j.jphysparis.2006.09.012>
- Denham, Susan L., Gyimesi, K., Stefanics, G., & Winkler, I. (2013). Perceptual bistability in auditory streaming: How much do stimulus features matter? *Learning & Perception*, 5(Supplement 2), 73–100. <https://doi.org/10.1556/LP.5.2013.Suppl2.6>
- Desain, P., & Honing, H. (1999). Computational Models of Beat Induction: The Rule-Based Approach. *Journal of New Music Research*, 28(1), 29–42. <https://doi.org/10.1076/jnmr.28.1.29.3123>
- Deutsch, D. (1975). Two-channel listening to musical scales. *The Journal of the Acoustical Society of America*, 57(5), 1156–1160. <https://doi.org/10.1121/1.380573>

- Dibben, N. (1999). The Perception of Structural Stability in Atonal Music: The Influence of Salience, Stability, Horizontal Motion, Pitch Commonality, and Dissonance. *Music Perception: An Interdisciplinary Journal*, 16(3), 265–294. <https://doi.org/10.2307/40285794>
- Dixon, S. (2001). Automatic Extraction of Tempo and Beat From Expressive Performances. *Journal of New Music Research*, 30(1), 39–58. <https://doi.org/10.1076/jnmr.30.1.39.7119>
- Dowling, W. J. (1973). The perception of interleaved melodies. *Cognitive Psychology*, 5(3), 322–337. [https://doi.org/10.1016/0010-0285\(73\)90040-6](https://doi.org/10.1016/0010-0285(73)90040-6)
- Dowling, W. Jay. (1990). Expectancy and attention in melody perception. *Psychomusicology: A Journal of Research in Music Cognition*, 9(2), 148–160. <https://doi.org/10.1037/h0094150>
- Dowling, W. Jay, & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology: A Journal of Research in Music Cognition*, 1(1), 30.
- Drost, U. C., Rieger, M., Brass, M., Gunter, T. C., & Prinz, W. (2005). When hearing turns into playing: Movement induction by auditory stimuli in pianists. *The Quarterly Journal of Experimental Psychology Section A*, 58(8), 1376–1389. <https://doi.org/10.1080/02724980443000610>
- Drost, U. C., Rieger, M., & Prinz, W. (2007). Instrument specificity in experienced musicians. *The Quarterly Journal of Experimental Psychology*, 60(4), 527–533. <https://doi.org/10.1080/17470210601154388>
- Duane, B. (2012). Agency and information content in eighteenth- and early nineteenth-century string-quartet expositions. *Journal of Music Theory*, 56(1), 87–120.
- Duane, B. (2013). Auditory Streaming Cues in Eighteenth- and Early Nineteenth-Century String Quartets: A Corpus-Based Study. *Music Perception: An Interdisciplinary Journal*, 31(1), 46–58. <https://doi.org/10.1525/mp.2013.31.1.46>
- Durrieu, J. L., Richard, G., David, B., & Fevotte, C. (2010). Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3), 564–575. <https://doi.org/10.1109/TASL.2010.2041114>
- Eerola, T. (2016). Expectancy-Violation and Information-Theoretic Models of Melodic Complexity. *Empirical Musicology Review*, 11(1), 2. <https://doi.org/10.18061/emr.v11i1.4836>
- 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>
- Egner, T., Monti, J. M., & Summerfield, C. (2010). Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience*, 30(49), 16601–16608.

- Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron*, *61*(2), 317–329. <https://doi.org/10.1016/j.neuron.2008.12.005>
- Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.
- Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society of London B: Biological Sciences*, *278*(1718), 2553–2561. <https://doi.org/10.1098/rspb.2011.0836>
- Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (Vol. 2). Walter de Gruyter. Retrieved from <https://books.google.co.uk/books?hl=en&lr=&id=UY0iAAAAQBAJ&oi=fnd&pg=PA5&dq=G.+Fant,+Acoustic+Theory+of+Speech+Production.&ots=4eAVWbeB-U&sig=E3RAvCvFzzqV-7-5L3dlYYhfJlk>
- Fiedler, D., & Müllensiefen, D. (2015). Validation of the Gold-MSI questionnaire to measure musical sophistication of German students at secondary education schools. *Musikpädagogische Forschung / Research in Music Education*, *36*, 199–219.
- Fink, G. R., Marshall, J. C., Halligan, P. W., & Dolan, R. J. (1998). Hemispheric asymmetries in global \ local processing are modulated by perceptual salience. *Neuropsychologia*, *37*(1), 31–40.
- Flexer, A., Schnitzer, D., & Schlüter, J. (2012). A MIREX Meta-analysis of Hubness in Audio Music Similarity. In *ISMIR* (pp. 175–180). Retrieved from http://ismir2012.ismir.net/event/papers/175_ISMIR_2012.pdf
- François, C., Jaillet, F., Takerkart, S., & Schön, D. (2014). Faster Sound Stream Segmentation in Musicians than in Nonmusicians. *PLoS ONE*, *9*(7), e101340. <https://doi.org/10.1371/journal.pone.0101340>
- Franklin, M. S., Moore, K. S., Yip, C.-Y., Jonides, J., Rattray, K., & Moher, J. (2008). The effects of musical training on verbal memory. *Psychology of Music*. <https://doi.org/10.1177/0305735607086044>
- French-St. George, M., & Bregman, A. S. (1989). Role of predictability of sequence in auditory stream segregation. *Perception & Psychophysics*, *46*(4), 384–386. <https://doi.org/10.3758/BF03204992>
- Friberg, A., & Ahlbäck, S. (2009). Recognition of the main melody in a polyphonic symbolic score using perceptual knowledge. *Journal of New Music Research*, *38*(2), 155–169.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York. Retrieved from <http://statweb.stanford.edu/~tibs/book/preface.ps>
- Friston, K. (2003). Learning and inference in the brain. *Neural Networks*, *16*(9), 1325–1352. <https://doi.org/10.1016/j.neunet.2003.06.005>

- Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 360(1456), 815–836. <https://doi.org/10.1098/rstb.2005.1622>
- Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, 13(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, 11(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Frith, C. (2012). Explaining delusions of control: The comparator model 20years on. *Consciousness and Cognition*, 21(1), 52–54.
- Fujioka, T., Ross, B., Kakigi, R., Pantev, C., & Trainor, L. J. (2006). One year of musical training affects development of auditory cortical-evoked fields in young children. *Brain: A Journal of Neurology*, 129(10), 2593–2608. <https://doi.org/10.1093/brain/awl247>
- Fujioka, T., Trainor, L. J., & Ross, B. (2008). Simultaneous pitches are encoded separately in auditory cortex: An MMNm study. *NeuroReport: For Rapid Communication of Neuroscience Research*, 19(3), 361–366. <https://doi.org/10.1097/WNR.0b013e3282f51d91>
- Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., & Pantev, C. (2004). Musical training enhances automatic encoding of melodic contour and interval structure. *Journal of Cognitive Neuroscience*, 16(6), 1010–1021. <https://doi.org/10.1162/0898929041502706>
- Fujioka, T., Trainor, L. J., Ross, B., Kakigi, R., & Pantev, C. (2005). Automatic encoding of polyphonic melodies in musicians and nonmusicians. *Journal of Cognitive Neuroscience*, 17(10), 1578–1592. <https://doi.org/10.1162/089892905774597263>
- Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., & Griffiths, T. D. (2011). Neural prediction of higher-order auditory sequence statistics. *NeuroImage*, 54(3), 2267–2277. <https://doi.org/10.1016/j.neuroimage.2010.10.038>
- Gebauer, L., Kringelbach, M. L., & Vuust, P. (2012). Ever-changing cycles of musical pleasure: The role of dopamine and anticipation. *Psychomusicology: Music, Mind, and Brain*, 22(2), 152.
- Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1), 107–143.
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, 103(4), 650.
- Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In *Simple heuristics that make us smart* (pp. 75–95). Oxford University Press. Retrieved from http://pubman.mpdl.mpg.de/pubman/item/escidoc:2102907/component/escidoc:2102906/GG_Betting_1999.pdf
- Gillard, J., & Schutz, M. (2012). Improving the efficacy of auditory alarms in medical devices by exploring the effect of amplitude envelope on learning and retention. In *Proceedings of the International Conference on Auditory Display* (pp. 240–241). Atlanta, GA.

- Gingras, B., Pearce, M. T., Goodchild, M., Dean, R. T., Wiggins, G., & McAdams, S. (2016). Linking melodic expectation to expressive performance timing and perceived musical tension. *Journal of Experimental Psychology: Human Perception and Performance*, 42(4), 594–609. <https://doi.org/10.1037/xhp0000141>
- Globerson, E., Granot, R., Tal, I., Harpaz, Y., Zeev-Wolf, M., & Golstein, A. (2017). Brain responses to regular and octave-scrambled melodies: A case of predictive-coding? *Journal of Experimental Psychology: Human Perception and Performance*, 43(3), 487.
- Gómez, E., Klapuri, A., & Meudic, B. (2003). Melody description and extraction in the context of music content processing. *Journal of New Music Research*, 32(1), 23–40.
- Goto, M., & Hayamizu, S. (1999). A real-time music scene description system: Detecting melody and bass lines in audio signals. In *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis* (pp. 31–40). Retrieved from https://www.researchgate.net/profile/Satoru_Hayamizu/publication/2455822_A_Real-time_Music_Scene_Description_System_Detecting_Melody_and_Bass_Lines_in_Audio_Signals/links/540146ac0cf23d9765a49199.pdf
- Granot, R., & Donchin, E. (2002). Do Re Mi Fa Sol La Ti—Constraints, Congruity, and Musical Training: An Event-Related Brain Potentials Study of Musical Expectancies. *Music Perception: An Interdisciplinary Journal*, 19(4), 487–528. <https://doi.org/10.1525/mp.2002.19.4.487>
- Gromko, J. E., & Poorman, A. S. (1998). The Effect of Music Training on Preschoolers' Spatial-Temporal Task Performance. *Journal of Research in Music Education*, 46(2), 173–181. <https://doi.org/10.2307/3345621>
- Habib, M., & Besson, M. (2009). What do Music Training and Musical Experience Teach Us About Brain Plasticity? *Music Perception: An Interdisciplinary Journal*, 26(3), 279–285. <https://doi.org/10.1525/mp.2009.26.3.279>
- Habibi, A., Wirantana, V., & Starr, A. (2013). Cortical Activity During Perception of Musical Pitch: Comparing Musicians and Nonmusicians. *Music Perception: An Interdisciplinary Journal*, 30(5), 463–479. <https://doi.org/10.1525/mp.2013.30.5.463>
- Haimson, J., Swain, D., & Winner, E. (2011). Do Mathematicians Have Above Average Musical Skill? *Music Perception: An Interdisciplinary Journal*, 29(2), 203–213. <https://doi.org/10.1525/mp.2011.29.2.203>
- Halpern, A. R., Bartlett, J. C., & Dowling, W. J. (1998). Perception of mode, rhythm, and contour in unfamiliar melodies: Effects of age and experience. *Music Perception: An Interdisciplinary Journal*, 15(4), 335–355.
- Handel, S., & Erickson, M. L. (2004). Sound Source Identification: The Possible Role of Timbre Transformations. *Music Perception: An Interdisciplinary Journal*, 21(4), 587–610. <https://doi.org/10.1525/mp.2004.21.4.587>
- Hannon, E. E., Soley, G., & Ullal, S. (2012). Familiarity Overrides Complexity in Rhythm Perception: A Cross-Cultural Comparison of American and Turkish Listeners. *Journal of Experimental Psychology: Human Perception and Performance*, 38, 543–548.

- Hannon, E. E., & Trehub, S. E. (2005a). Metrical categories in infancy and adulthood. *Psychological Science*, *16*, 48–55.
- Hannon, Erin E., & Trehub, S. E. (2005). Tuning in to musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(35), 12639–12643. <https://doi.org/10.1073/pnas.0504254102>
- Hansen, M., Wallentin, M., & Vuust, P. (2013). Working memory and musical competence of musicians and non-musicians. *Psychology of Music*, *41*(6), 779–793. <https://doi.org/10.1177/0305735612452186>
- 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>
- Hansen, N. C., Vuust, P., & Pearce, M. (2016). “If You Have to Ask, You’ll Never Know”: Effects of Specialised Stylistic Expertise on Predictive Processing of Music. *PloS One*, *11*(10), e0163584.
- Harrison, P. M. C., Müllensiefen, D., & Collins, T. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *In press*.
- Hartmann, W. M., & Johnson, D. (1991). Stream segregation and peripheral channeling. *Music Perception*, *9*(2), 155–183.
- Herrera, P., & Bonada, J. (1998). Vibrato extraction and parameterization in the spectral modeling synthesis framework. In *Proceedings of the Digital Audio Effects Workshop (DAFX98)* (Vol. 99). Retrieved from <http://www.academia.edu/download/34327454/dafx98-perfe.pdf>
- Hoffman, D. D., & Singh, M. (1997). Saliency of visual parts. *Cognition*, *63*(1), 29–78.
- Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108*(3), 687–701. <https://doi.org/10.1016/j.cognition.2008.05.010>
- Honing, H., Ladinig, O., Háden, G. P., & Winkler, I. (2009). Is Beat Induction Innate or Learned? *Annals of the New York Academy of Sciences*, *1169*(1), 93–96.
- Horstmann, G., Becker, S., & Ernst, D. (2016). Perceptual saliency captures the eyes on a surprise trial. *Attention, Perception, & Psychophysics*, *78*(7), 1889–1900.
- Huron, D. (2001). *Tone and Voice: A Derivation of the Rules of Voice-Leading from Perceptual Principles*.
- Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA, US: The MIT Press.
- Hurwitz, I., Wolff, P. H., Bortnick, B. D., & Kokas, K. (1975). Nonmusical Effects of the Kodaly Music Curriculum in Primary Grade Children. *Journal of Learning Disabilities*, *8*(3), 167–174. <https://doi.org/10.1177/002221947500800310>
- Husain, G., Thompson, W. F., & Schellenberg, E. G. (2002). Effects of Musical Tempo and Mode on Arousal, Mood, and Spatial Abilities. *Music Perception: An Interdisciplinary Journal*, *20*(2), 151–171. <https://doi.org/10.1525/mp.2002.20.2.151>

- Hutchinson, J. M., & Gigerenzer, G. (2005). Simple heuristics and rules of thumb: Where psychologists and behavioural biologists might meet. *Behavioural Processes*, *69*(2), 97–124.
- Ishigaki, A., Matsubara, M., & Saito, H. (2011). Prioritized contig combining to segregate voices in polyphonic music. In *Sound and Music Computing Conference (SMC 2011)* (Vol. 119). Retrieved from http://www.smc-conference.org/smc11/papers/smc2011_119.pdf
- Istók, E., Friberg, A., Huottilainen, M., & Tervaniemi, M. (2013). Expressive Timing Facilitates the Neural Processing of Phrase Boundaries in Music: Evidence from Event-Related Potentials. *PLOS ONE*, *8*(1), e55150. <https://doi.org/10.1371/journal.pone.0055150>
- Iverson, P. (1995). Auditory stream segregation by musical timbre: effects of static and dynamic acoustic attributes. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 751–763.
- Jaeger, T. F., & Weatherholtz, K. (2016). What the heck is salience? How predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, *7*. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4971435/>
- Jehee, J. F., & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Comput Biol*, *5*(5), e1000373.
- Jones, D., Alford, D., Bridges, A., Tremblay, S., & Macken, B. (1999). Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 464–473. <https://doi.org/10.1037/0278-7393.25.2.464>
- Jones, Mari R. (1976). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, *83*(5), 323–355. <https://doi.org/10.1037/0033-295X.83.5.323>
- Jones, Mari Riess, Boltz, M., & Kidd, G. (1982). Controlled attending as a function of melodic and temporal context. *Perception & Psychophysics*, *32*(3), 211–218. <https://doi.org/10.3758/BF03206225>
- Jongsma, M. L. A., Desain, P., & Honing, H. (2004). Rhythmic context influences the auditory evoked potentials of musicians and nonmusicians. *Biological Psychology*, *66*(2), 129–152. <https://doi.org/10.1016/j.biopsycho.2003.10.002>
- Jordanous, A. (2008). Voice separation in Polyphonic Music: a Data-Driven Approach. In *ICMC*. Retrieved from <http://www.academia.edu/download/6249561/cr1582.pdf>
- Juslin, P. N., Liljeström, S., Västfjäll, D., & Lundqvist, L.-O. (2011). How does music evoke emotions? Exploring the underlying mechanisms. In P. N. Juslin & J. Sloboda (Eds.), *Handbook of Music and Emotion*. Oxford University Press. Retrieved from <https://philpapers.org/rec/JUSHDM>
- Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: the need to consider underlying mechanisms. *The Behavioral and Brain Sciences*, *31*(5), 559–575; discussion 575-621. <https://doi.org/10.1017/S0140525X08005293>

- Karydis, I., Nanopoulos, A., Papadopoulos, A., Cambouropoulos, E., & Manolopoulos, Y. (2007). Horizontal and vertical integration/segregation in auditory streaming: a voice separation algorithm for symbolic musical data. In *Proceedings 4th Sound and Music Computing Conference (SMC'2007)*. Retrieved from <http://smc-conference.org/smc07/SMC07%20Proceedings/SMC07%20Paper%2050.pdf>
- Kendall, R. A., & Carterette, E. C. (1991). Perceptual Scaling of Simultaneous Wind Instrument Timbres. *Music Perception: An Interdisciplinary Journal*, 8(4), 369–404. <https://doi.org/10.2307/40285519>
- Khalfa, S., Isabelle, P., Jean-Pierre, B., & Manon, R. (2002). Event-related skin conductance responses to musical emotions in humans. *Neuroscience Letters*, 328(2), 145–149.
- Kilian, J., & Hoos, H. H. (2002). Voice Separation-A Local Optimization Approach. In *Proceedings of the Third International Conference on Music Information Retrieval* (pp. 39–46). Paris: IRCAM - Centre Pompidou. Retrieved from <http://www-devel.cs.ubc.ca/~hoos/Publ/KilHoo02.pdf>
- Kirlin, P. B., & Utgoff, P. E. (2005). VOISE: Learning to Segregate Voices in Explicit and Implicit Polyphony. In *ISMIR* (pp. 552–557). Citeseer. Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.74.1474&rep=rep1&type=pdf>
- Koelsch, S., Fritz, T., & Schlaug, G. (2008). Amygdala activity can be modulated by unexpected chord functions during music listening. *Neuroreport*, 19(18), 1815–1819. <https://doi.org/10.1097/WNR.0b013e32831a8722>
- Krishnan, L., Elhilali, M., & Shamma, S. (2014). Segregating complex sound sources through temporal coherence. *PLoS Computational Biology*, 10(12), e1003985.
- Krumhansl, C. L., Louhivuori, J., Toiviainen, P., Järvinen, T., & Eerola, T. (1999). Melodic Expectation in Finnish Spiritual Folk Hymns: Convergence of Statistical, Behavioral, and Computational Approaches. *Music Perception: An Interdisciplinary Journal*, 17(2), 151–195. <https://doi.org/10.2307/40285890>
- Krumhansl, C. L., & Schmuckler, M. (1986). Key-finding in music: An algorithm based on pattern matching to tonal hierarchies. Presented at the 19th Annual Meeting of the Society of Mathematical Psychology, Cambridge, MA.
- Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., Patterson, R. D., ... Griffiths, T. D. (2011). Predictive coding and pitch processing in the auditory cortex. *Journal of Cognitive Neuroscience*, 23(10), 3084–3094.
- Lakatos, S. (2000). A common perceptual space for harmonic and percussive timbres. *Perception & Psychophysics*, 62(7), 1426–1439.
- Lakatos, Stephen, & Marks, L. E. (1999). Haptic form perception: Relative salience of local and global features. *Attention, Perception, & Psychophysics*, 61(5), 895–908.
- Lange, K. (2010). Can a regular context induce temporal orienting to a target sound? *International Journal of Psychophysiology*, 78(3), 231–238.
- Lattner, S., Grachten, M., Agres, K., & Chacón, C. E. C. (2015). Probabilistic Segmentation of Musical Sequences Using Restricted Boltzmann Machines. In T. Collins, D. Meredith,

- & A. Volk (Eds.), *Mathematics and Computation in Music* (pp. 323–334). Springer International Publishing. https://doi.org/10.1007/978-3-319-20603-5_33
- Leopold, D. A., & Logothetis, N. K. (1999). Multistable phenomena: changing views in perception. *Trends in Cognitive Sciences*, 3(7), 254–264.
- Lerdahl, F., & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. MIT Press.
- Li, T., & Ogihara, M. (2004). Content-based music similarity search and emotion detection. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04). IEEE International Conference on* (Vol. 5, pp. V–705). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/1327208/>
- Ma, L. (2011). *Auditory Streaming: Behavior, Physiology, and Modeling*. (doctoral). University of Maryland, College Park, MD.
- Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, 29(1), 43–51. <https://doi.org/10.1037/0096-1523.29.1.43>
- Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4), 276–298. <https://doi.org/10.1037/h0076778>
- Madsen, S. T., & Widmer, G. (2006). Separating voices in MIDI. In *ISMIR* (pp. 57–60). Retrieved from <https://pdfs.semanticscholar.org/0510/f591c6ae52bd412ef65684af5e7a2764bd15.pdf>
- Madsen, S. T., & Widmer, G. (2007). A Complexity-based Approach to Melody Track Identification in MIDI Files. Retrieved from <http://ai2-s2-pdfs.s3.amazonaws.com/6f0d/b229f11dd8596cfbbf31bbd1218a34f3612a.pdf>
- Makris, D., Karydis, I., & Cambouropoulos, E. (2016). VISA3: Refining the voice integration/segregation algorithm. Retrieved from https://www.researchgate.net/profile/Ioannis_Karydis2/publication/307866735_VISA_3_Refining_the_Voice_IntegrationSegregation_Algorithm/links/57cfeceb08ae582e0695cec7.pdf
- Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.
- Manzara, L. C., Witten, I. H., & James, M. (1992). On the Entropy of Music: An Experiment with Bach Chorale Melodies. *Leonardo Music Journal*, 2(1), 81–88. <https://doi.org/10.2307/1513213>
- Mardirossian, A., & Chew, E. (2005). Key distributions as musical fingerprints for similarity assessment. In *Multimedia, Seventh IEEE International Symposium on* (pp. 6–pp). IEEE. Retrieved from <http://ieeexplore.ieee.org/abstract/document/1565887/>
- Mardirossian, A., & Chew, E. (2006). Music Summarization Via Key Distributions: Analyses of Similarity Assessment Across Variations. In *ISMIR* (pp. 234–239). Retrieved from <https://pdfs.semanticscholar.org/79a8/096ec06f2a7178c57890e70e0adff22d5e0d.pdf>

- Margulis, E. H. (2012). Musical repetition detection across multiple exposures. *Music Perception: An Interdisciplinary Journal*, 29(4), 377–385.
- Margulis, E. H. (2013). Aesthetic responses to repetition in unfamiliar music. *Empirical Studies of the Arts*, 31(1), 45–57.
- Margulis, E. H. (2014). *On repeat: How music plays the mind*. Oxford University Press. Retrieved from <https://books.google.co.uk/books?hl=en&lr=&id=4a6cAQAAQBAJ&oi=fnd&pg=PP1&dq=music+repetition+margulis&ots=w4JLN6C7--&sig=5IU3XVRffwdmPMIAXfvfAHnsJkc>
- Margulis, E. H., Mlsna, L. M., Uppunda, A. K., Parrish, T. B., & Wong, P. C. M. (2009). Selective neurophysiologic responses to music in instrumentalists with different listening biographies. *Human Brain Mapping*, 30(1), 267–275. <https://doi.org/10.1002/hbm.20503>
- Marie, Céline, Fujioka, T., Herrington, L., & Trainor, L. J. (2012). The high-voice superiority effect in polyphonic music is influenced by experience: A comparison of musicians who play soprano-range compared with bass-range instruments. *Psychomusicology: Music, Mind, and Brain*, 22(2), 97–104. <https://doi.org/10.1037/a0030858>
- Marie, Celine, & Trainor, L. J. (2012). Development of simultaneous pitch encoding: infants show a high voice superiority effect. *Cerebral Cortex*, 23(3), 660–669.
- Marie, Céline, & Trainor, L. J. (2014). Early development of polyphonic sound encoding and the high voice superiority effect. *Neuropsychologia*, 57, 50–58.
- Marozeau, J., Innes-Brown, H., & Blamey, P. J. (2013). The Effect of Timbre and Loudness on Melody Segregation. *Music Perception: An Interdisciplinary Journal*, 30(3), 259–274. <https://doi.org/10.1525/mp.2012.30.3.259>
- Marozeau, J., Innes-Brown, H., Grayden, D. B., Burkitt, A. N., & Blamey, P. J. (2010). The effect of visual cues on auditory stream segregation in musicians and non-musicians. *PloS One*, 5(6), e11297. <https://doi.org/10.1371/journal.pone.0011297>
- Marsden, A. (1992). Modelling the perception of musical voices: a case study in rule-based systems. *Computer Representations and Models in Music*, 239–263.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3), 177–192. <https://doi.org/10.1007/BF00419633>
- McAuley, J. D., Jones, M. R., Holub, S., Johnston, H. M., & Miller, N. S. (2006). The time of our lives: life span development of timing and event tracking. *Journal of Experimental Psychology: General*, 135(3), 348.
- McCabe, S. L., & Denham, M. J. (1997). A model of auditory streaming. *Journal of the Acoustical Society of America*, 101(3), 1611–1621. <https://doi.org/10.1121/1.418176>
- McGill University master samples collection on DVD. (2006). [Montreal, Quebec, Canada]: McGill [University].

- Menning, H., Roberts, L. E., & Pantev, C. (2000). Plastic changes in the auditory cortex induced by intensive frequency discrimination training. *NeuroReport: For Rapid Communication of Neuroscience Research*, 11(4), 817–822. <https://doi.org/10.1097/00001756-200003200-00032>
- Meredith, D., Lemström, K., & Wiggins, G. A. (2002). Algorithms for discovering repeated patterns in multidimensional representations of polyphonic music. *Journal of New Music Research*, 31(4), 321–345. <https://doi.org/10.1076/jnmr.31.4.321.14162>
- Meyer, L. (1956). *Emotion and Meaning in Music*. University of Chicago Press. Retrieved from <http://www.press.uchicago.edu/ucp/books/book/chicago/E/bo3643659.html>
- Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, 219(1–2), 36–47. <https://doi.org/10.1016/j.heares.2006.05.004>
- Micheyl, C., Kreft, H., Shamma, S., & Oxenham, A. J. (2013). Temporal coherence versus harmonicity in auditory stream formation. *The Journal of the Acoustical Society of America*, 133(3), EL188–EL194. <https://doi.org/10.1121/1.4789866>
- Micheyl, C., & Oxenham, A. J. (2010). Pitch, harmonicity and concurrent sound segregation: Psychoacoustical and neurophysiological findings. *Hearing Research*, 266(1–2), 36–51. <https://doi.org/10.1016/j.heares.2009.09.012>
- Mill, R. W., Böhm, T. M., Bendixen, A., Winkler, I., & Denham, S. L. (2013). Modelling the Emergence and Dynamics of Perceptual Organisation in Auditory Streaming. *PLOS Computational Biology*, 9(3), e1002925. <https://doi.org/10.1371/journal.pcbi.1002925>
- Miller, G. A., & Heise, G. A. (1950). The trill threshold. *Journal of the Acoustical Society of America*, 22, 637–638. <https://doi.org/10.1121/1.1906663>
- Miranda, R. A., & Ullman, M. T. (2007). Double dissociation between rules and memory in music: An event-related potential study. *NeuroImage*, 38(2), 331–345. <https://doi.org/10.1016/j.neuroimage.2007.07.034>
- Moreno, S., Friesen, D., & Bialystok, E. (2011). Effect of Music Training on Promoting Preliteracy Skills: Preliminary Causal Evidence. *Music Perception: An Interdisciplinary Journal*, 29(2), 165–172. <https://doi.org/10.1525/mp.2011.29.2.165>
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PLoS ONE*, 9(2), e89642. <https://doi.org/10.1371/journal.pone.0089642>
- Münste, T. F., Kohlmetz, C., Nager, W., & Altenmüller, E. (2001). Superior auditory spatial tuning in conductors. *Nature*, 409(6820), 580–580. <https://doi.org/10.1038/35054668>
- Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 99(23), 15164–15169.
- Näätänen, R., Gaillard, A. W., & Mäntysalo, S. (1978). Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica*, 42(4), 313–329.

- Nager, W., Kohlmetz, C., Altenmüller, E., Rodríguez-Fornells, A., & Münte, T. F. (2003). The fate of sounds in conductors' brains: an ERP study. *Cognitive Brain Research*, *17*(1), 83–93. [https://doi.org/10.1016/S0926-6410\(03\)00083-1](https://doi.org/10.1016/S0926-6410(03)00083-1)
- Nan, Y., Knösche, T. R., & Friederici, A. D. (2006). The perception of musical phrase structure: A cross-cultural ERP study. *Brain Research*, *1094*(1), 179–191. <https://doi.org/10.1016/j.brainres.2006.03.115>
- Nan, Y., Knösche, T. R., Zysset, S., & Friederici, A. D. (2008). Cross-cultural music phrase processing: An fMRI study. *Human Brain Mapping*, *29*(3), 312–328. <https://doi.org/10.1002/hbm.20390>
- Narmour, E. (1992). *The Analysis and Cognition of Melodic Complexity: The Implication-realization Model*. University of Chicago Press.
- Nettl, B. (2010). *The study of ethnomusicology: Thirty-one issues and concepts*. University of Illinois Press. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=_vlrrG7HvP4C&oi=fnd&pg=PR3&dq=bruno+nettl&ots=jAI17MgskK&sig=7MuMaASA7RwI15sYITI7cVNQkvc
- Neuhaus, C., Knösche, T. R., & Friederici, A. D. (2006). Effects of Musical Expertise and Boundary Markers on Phrase Perception in Music. *Journal of Cognitive Neuroscience*, *18*(3), 472–493. <https://doi.org/10.1162/jocn.2006.18.3.472>
- Niebur, E., Hsiao, S. S., & Johnson, K. O. (2002). Synchrony: A neuronal mechanism for attentional selection? *Current Opinion in Neurobiology*, *12*(2), 190–194. [https://doi.org/10.1016/S0959-4388\(02\)00310-0](https://doi.org/10.1016/S0959-4388(02)00310-0)
- Nix, J., & Hohmann, V. (2007). Combined estimation of spectral envelopes and sound source direction of concurrent voices by multidimensional statistical filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(3), 995–1008.
- Palmer, C., & Krumhansl, C. L. (1987). Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity. *Perception & Psychophysics*, *41*(6), 505–518. <https://doi.org/10.3758/BF03210485>
- Palmer, C., & Krumhansl, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 728–741.
- Pampalk, E., Flexer, A., Widmer, G., & others. (2005). Improvements of Audio-Based Music Similarity and Genre Classification. In *ISMIR* (Vol. 5, pp. 634–637). London, UK. Retrieved from http://www.cp.jku.at/research/papers/pampalk_ismir_2005.pdf
- Pantev, C., Roberts, L. E., Schulz, M., Engelien, A., & Ross, B. (2001). Timbre-specific enhancement of auditory cortical representations in musicians. *NeuroReport: For Rapid Communication of Neuroscience Research*, *12*(1), 169–174. <https://doi.org/10.1097/00001756-200101220-00041>
- Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107–123.
- Paton, B., Skewes, J., Frith, C., & Hohwy, J. (2013). Skull-bound perception and precision optimization through culture. *The Behavioral and Brain Sciences*, *36*(3), 222. <https://doi.org/10.1017/S0140525X12002191>

- Paulus, M. P., & Stein, M. B. (2006). An insular view of anxiety. *Biological Psychiatry*, *60*(4), 383–387.
- Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. <https://doi.org/10.1037/0033-295X.87.6.532>
- Pearce, M., & Müllensiefen, D. (2017). Compression-based Modelling of Musical Similarity Perception. *Journal of New Music Research*. Retrieved from <http://www.tandfonline.com/eprint/XWP6EzEjbrSI2nkQk3Ue/full>
- Pearce, M. T. (2005, December). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (doctoral). City University London. Retrieved from <http://openaccess.city.ac.uk/8459/>
- Pearce, Marcus T., Müllensiefen, D., & Wiggins, G. A. (2010). 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, Marcus T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., & Bhattacharya, J. (2010). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *NeuroImage*, *50*(1), 302–313. <https://doi.org/10.1016/j.neuroimage.2009.12.019>
- Pearce, Marcus T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*(5), 377–405. <https://doi.org/10.1525/mp.2006.23.5.377>
- Phillips, D. (1976). An investigation of the relationship between musicality and intelligence. *Psychology of Music*, *4*(2), 16–31. <https://doi.org/10.1177/030573567642003>
- Pichevar, R., & Rouat, J. (2007). Monophonic sound source separation with an unsupervised network of spiking neurones. *Neurocomputing*, *71*(1), 109–120.
- Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). On Rhythm and General Music Similarity. In *ISMIR* (pp. 525–530). Retrieved from http://www.cp.jku.at/people/schedl/Research/Publications/pdf/pohle_ismir_2009.pdf
- Ponce de León Amador, P. J., Iñesta Quereda, J. M., & Rizo Valero, D. (2008). *Mining digital music score collections: melody extraction and genre recognition*. Intech. Retrieved from <http://rua.ua.es/dspace/handle/10045/16184>
- Pressnitzer, D., & Hupé, J.-M. (2006). Temporal dynamics of auditory and visual bistability reveal common principles of perceptual organization. *Current Biology: CB*, *16*(13), 1351–1357. <https://doi.org/10.1016/j.cub.2006.05.054>
- Pressnitzer, D., Suied, C., & Shamma, S. A. (2011). Auditory scene analysis: The sweet music of ambiguity. *Frontiers in Human Neuroscience*, *5*. <https://doi.org/10.3389/fnhum.2011.00158>
- Prince, J. B., Thompson, W. F., & Schmuckler, M. A. (2009). Pitch and time, tonality and meter: how do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1598–1617. <https://doi.org/10.1037/a0016456>

- Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R. M., & Hennon, E. A. (2006). The Birth of Words: Ten-Month-Olds Learn Words Through Perceptual Salience. *Child Development*, *77*(2), 266–280.
- Rammsayer, T., & Altenmüller, E. (2006). Temporal Information Processing in Musicians and Nonmusicians. *Music Perception: An Interdisciplinary Journal*, *24*(1), 37–48. <https://doi.org/10.1525/mp.2006.24.1.37>
- Rammsayer, T. H., Buttkus, F., & Altenmüller, E. (2012). Musicians Do Better than Nonmusicians in Both Auditory and Visual Timing Tasks. *Music Perception: An Interdisciplinary Journal*, *30*(1), 85–96. <https://doi.org/10.1525/mp.2012.30.1.85>
- Rankin, J., Sussman, E., & Rinzel, J. (2015). Neuromechanistic Model of Auditory Bistability. *PLOS Computational Biology*, *11*(11), e1004555. <https://doi.org/10.1371/journal.pcbi.1004555>
- Rao, R. P., & Sejnowski, T. J. (2002). 16 Predictive Coding, Cortical Feedback, and Spike-Timing Dependent Plasticity. *Probabilistic Models of the Brain*, 297.
- Rauscher, F. H., & Hinton, S. C. (2011). Music Instruction and its Diverse Extra-Musical Benefits. *Music Perception: An Interdisciplinary Journal*, *29*(2), 215–226. <https://doi.org/10.1525/mp.2011.29.2.215>
- Repp, B. H., & Doggett, R. (2007). Tapping to a Very Slow Beat: A Comparison of Musicians and Nonmusicians. *Music Perception: An Interdisciplinary Journal*, *24*(4), 367–376. <https://doi.org/10.1525/mp.2007.24.4.367>
- Repp, B. H., & Knoblich, G. (2009). Performed or observed keyboard actions affect pianists' judgements of relative pitch. *The Quarterly Journal of Experimental Psychology*, *62*(11), 2156–2170. <https://doi.org/10.1080/17470210902745009>
- Rizo, D., De León, P. J. P., Pérez-Sancho, C., Pertusa, A., & Querada, J. M. I. (2006). A Pattern Recognition Approach for Melody Track Selection in MIDI Files. In *ISMIR* (pp. 61–66). Retrieved from <http://www.academia.edu/download/32238952/ismir2006.pdf>
- Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358.
- Rogers, W. L., & Bregman, A. S. (1993). An experimental evaluation of three theories of auditory stream segregation. *Perception & Psychophysics*, *53*(2), 179–189. <https://doi.org/10.3758/BF03211728>
- Rogers, W. L., & Bregman, A. S. (1998). Cumulation of the tendency to segregate auditory streams: Resetting by changes in location and loudness. *Perception & Psychophysics*, *60*(7), 1216–1227. <https://doi.org/10.3758/BF03206171>
- Rolland, P.-Y. (1999). Discovering patterns in musical sequences. *Journal of New Music Research*, *28*(4), 334–350.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172.
- Rycroft, D. K., & Sadie, S. (1983). *The New Grove Dictionary of Music and Musicians*. JSTOR. Retrieved from <http://www.jstor.org/stable/30249775>

- Salamon, J., & Gomez, E. (2012). Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(6), 1759–1770. <https://doi.org/10.1109/TASL.2012.2188515>
- Salamon, J., Gomez, E., Ellis, D. P. W., & Richard, G. (2014). Melody Extraction from Polyphonic Music Signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, *31*(2), 118–134. <https://doi.org/10.1109/MSP.2013.2271648>
- Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining “blended” instrument pairings in orchestration. *Music Perception*, *13*(2), 209–246.
- Sauvé, S., Stewart, L., & Pearce, M. T. (2014). The Effect of Musical Training on Auditory Grouping. In *Proceedings of the Seventh International Conference of Students of Systematic Musicology*. London.
- Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrung anhand einer deutschen Stichprobe. *Musicae Scientiae*, *18*(4), 423–447. <https://doi.org/10.1177/1029864914541851>
- Schellenberg, E. G. (1996). Expectancy in melody: tests of the implication-realization model. *Cognition*, *58*(1), 75–125. [https://doi.org/10.1016/0010-0277\(95\)00665-6](https://doi.org/10.1016/0010-0277(95)00665-6)
- Schellenberg, E. G. (1997). Simplifying the Implication-Realization Model of Melodic Expectancy. *Music Perception: An Interdisciplinary Journal*, *14*(3), 295–318. <https://doi.org/10.2307/40285723>
- Schmid, H.-J., & Günther, F. (2016). Toward a Unified Socio-Cognitive Framework for Saliency in Language. *Frontiers in Psychology*, *7*. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4974243/>
- Schöön, D., Regnault, P., Ystad, S., & Besson, M. (2005). Sensory Consonance. *Music Perception: An Interdisciplinary Journal*, *23*(2), 105–118. <https://doi.org/10.1525/mp.2005.23.2.105>
- Schröger, E., Bendixen, A., Denham, S. L., Mill, R. W., Böhm, T. M., & Winkler, I. (2014). Predictive regularity representations in violation detection and auditory stream segregation: From conceptual to computational models. *Brain Topography*, *27*(4), 565–577. <https://doi.org/10.1007/s10548-013-0334-6>
- Selfridge-Field, E. (1998). Conceptual and representational issues in melodic comparison. *Computing in Musicology: A Directory of Research*, (11), 3–64.
- Seth, A. K., & Critchley, H. (2013). Extending predictive processing to the body: Emotion as interoceptive inference. *Behavioral and Brain Sciences*, *36*(3), 47–58. <http://dx.doi.org/10.1017/S0140525X12002270>
- Seth, A. K., Suzuki, K., & Critchley, H. D. (2011). An interoceptive predictive coding model of conscious presence. *Frontiers in Psychology*, *2*. Retrieved from <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3254200/>
- Shahin, A. J., Roberts, L. E., Chau, W., Trainor, L. J., & Miller, L. M. (2008). Music training leads to the development of timbre-specific gamma band activity. *NeuroImage*, *41*(1), 113–122. <https://doi.org/10.1016/j.neuroimage.2008.01.067>

- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423, 623–656.
- Shi, Y. Q., & Sun, H. (1999). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. CRC press. Retrieved from https://books.google.co.uk/books?hl=en&lr=&id=LEjJYki9U0wC&oi=fnd&pg=PA1&dq=shi+sun+predictive+coding&ots=KuPt0vWthJ&sig=VL1uxyoAxC-IBSOvzhuRxXH_TKs
- Silva, S., Barbosa, F., Marques-Teixeira, J., Petersson, K. M., & Castro, S. L. (2014). You know when: Event-related potentials and theta/beta power indicate boundary prediction in music. *Journal of Integrative Neuroscience*, 13(01), 19–34. <https://doi.org/10.1142/S0219635214500022>
- Silverstein, S. (2013). Schizophrenia-related phenomena that challenge prediction error as the basis of cognitive functioning. *The Behavioral and Brain Sciences*, 36(3), 49–50.
- Singh, P. G., & Bregman, A. S. (1997). The influence of different timbre attributes on the perceptual segregation of complex-tone sequences. *Journal of the Acoustical Society of America*, 102(4), 1943–1952. <https://doi.org/10.1121/1.419688>
- Skoe, E., & Kraus, N. (2013). Musical training heightens auditory brainstem function during sensitive periods in development. *Frontiers in Psychology*, 4. <https://doi.org/10.3389/fpsyg.2013.00622>
- Slaney, M., Weinberger, K., & White, W. (2008). Learning a metric for music similarity. In *International Symposium on Music Information Retrieval (ISMIR)*. Retrieved from <https://www.slaney.org/malcolm/yahoo/Slaney2008-MusicSimilarityMetricsISMIR.pdf>
- Snyder, J. S., Gregg, M. K., Weintraub, D. M., & Alain, C. (2012). Attention, awareness, and the perception of auditory scenes. *Frontiers in Psychology*, 3. <https://doi.org/10.3389/fpsyg.2012.00015>
- Southwell, R., Baumann, A., Gal, C., Barascud, N., Friston, K., & Chait, M. (2017). Is predictability salient? A study of attentional capture by auditory patterns. *Phil. Trans. R. Soc. B*, 372(1714), 20160105. <https://doi.org/10.1098/rstb.2016.0105>
- Spada, D., Verga, L., Iadanza, A., Tettamanti, M., & Perani, D. (2014). The auditory scene: An fMRI study on melody and accompaniment in professional pianists. *NeuroImage*, 102(Part 2), 764–775. <https://doi.org/10.1016/j.neuroimage.2014.08.036>
- Spielmann, M. I., Schröger, E., Kotz, S. A., & Bendixen, A. (2014). Attention effects on auditory scene analysis: Insights from event-related brain potentials. *Psychological Research*, 78(3), 361–378. <https://doi.org/10.1007/s00426-014-0547-7>
- Steele, K. M., Ball, T. N., & Runk, R. (1997). Listening to Mozart does not enhance backwards digit span performance. *Perceptual and Motor Skills*, 84(3 Pt 2), 1179–1184. <https://doi.org/10.2466/pms.1997.84.3c.1179>
- Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: evidence from subjective, physiological, and neural responses.

- Journal of Cognitive Neuroscience*, 18(8), 1380–1393.
<https://doi.org/10.1162/jocn.2006.18.8.1380>
- Stewart, L., Verdonschot, R. G., Nasralla, P., & Lanipekun, J. (2013). Action–perception coupling in pianists: Learned mappings or spatial musical association of response codes (SMARC) effect? *The Quarterly Journal of Experimental Psychology*, 66(1), 37–50.
<https://doi.org/10.1080/17470218.2012.687385>
- Strait, D., & Kraus, N. (2011). Playing Music for a Smarter Ear: Cognitive, Perceptual and Neurobiological Evidence. *Music Perception: An Interdisciplinary Journal*, 29(2), 133–146. <https://doi.org/10.1525/mp.2011.29.2.133>
- Strait, D. L., Kraus, N., Skoe, E., & Ashley, R. (2009). Musical Experience Promotes Subcortical Efficiency in Processing Emotional Vocal Sounds. *Annals of the New York Academy of Sciences*, 1169(1), 209–213. <https://doi.org/10.1111/j.1749-6632.2009.04864.x>
- Strait, D. L., Parbery-Clark, A., Hittner, E., & Kraus, N. (2012). Musical training during early childhood enhances the neural encoding of speech in noise. *Brain and Language*, 123(3), 191–201. <https://doi.org/10.1016/j.bandl.2012.09.001>
- Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18(6), 643.
- Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9), 403–409. <https://doi.org/10.1016/j.tics.2009.06.003>
- Summerfield, C., Monti, J. M., Trittschuh, E. H., Mesulam, M.-M., & Egner, T. (2008). Neural repetition suppression reflects fulfilled perceptual expectations. *Nature Neuroscience*, 11(9), 1004.
- Sussman, E., Ritter, W., & Vaughan, H. G. J. (1999). An investigation of the auditory streaming effect using event-related brain potentials. *Psychophysiology*, 36(1), 22–34.
<https://doi.org/10.1017/S0048577299971056>
- Sussman-Fort, J., & Sussman, E. (2014). The effect of stimulus context on the buildup to stream segregation. *Frontiers in Neuroscience*, 8. <https://doi.org/10.3389/fnins.2014.00093>
- Szeto, W. M., & Wong, M. H. (2003). A stream segregation algorithm for polyphonic music databases. In *Database Engineering and Applications Symposium, 2003. Proceedings. Seventh International* (pp. 130–138). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1214920
- Temperley, D., & Sleator, D. (2001). *The Melisma Music Analyzer*.
- Temperley, David. (1999). What’s Key for Key? The Krumhansl-Schmuckler Key-Finding Algorithm Reconsidered. *Music Perception: An Interdisciplinary Journal*, 17(1), 65–100. <https://doi.org/10.2307/40285812>
- Temperley, David. (2007). *Music and probability*. Cambridge, MA, US: MIT Press.
- Temperley, David. (2008). A probabilistic model of melody perception. *Cognitive Science*, 32(2), 418–444. <https://doi.org/10.1080/03640210701864089>

- Temperley, David. (2009). A Unified Probabilistic Model for Polyphonic Music Analysis. *Journal of New Music Research*, 38(1), 3–18. <https://doi.org/10.1080/09298210902928495>
- Temperley, David. (2010). Modeling common-practice rhythm. *Music Perception*, 27(5), 355–376. <https://doi.org/10.1525/mp.2010.27.5.355>
- Temperley, David. (2013). Computational models of music cognition. In D. Deutsch & D. (Ed) Deutsch (Eds.), *The psychology of music (3rd ed.)*. (pp. 327–368). San Diego, CA, US: Elsevier Academic Press.
- Temperley, David. (2014). Probabilistic Models of Melodic Interval. *Music Perception: An Interdisciplinary Journal*, 32(1), 85–99. <https://doi.org/10.1525/mp.2014.32.1.85>
- Tervaniemi, M. (2009). Musicians—Same or Different? *Annals of the New York Academy of Sciences*, 1169(1), 151–156. <https://doi.org/10.1111/j.1749-6632.2009.04591.x>
- Thomassen, S., & Bendixen, A. (2017). Subjective perceptual organization of a complex auditory scene. *The Journal of the Acoustical Society of America*, 141(1), 265–276. <https://doi.org/10.1121/1.4973806>
- Thompson, S. K., Carlyon, R. P., & Cusack, R. (2011). An objective measurement of the build-up of auditory streaming and of its modulation by attention. *Journal of Experimental Psychology: Human Perception and Performance*, 37(4), 1253–1262. <https://doi.org/10.1037/a0021925>
- Thompson, W. F., Graham, P., & Russo, F. A. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica*, 2005(156), 203–227. <https://doi.org/10.1515/semi.2005.2005.156.203>
- Toch, E. (1923). *Melodielehre: ein Beitrag zur Musiktheorie*. Berlin: Max Hesse.
- Toiviainen, P. (Ed.). (2007). Similarity perception in listening to music. *Musicae Scientiae*, 11(1_suppl).
- Tougas, Y., & Bregman, A. S. (1990). Auditory streaming and the continuity illusion. *Perception & Psychophysics*, 47(2), 121–126. <https://doi.org/10.3758/BF03205976>
- Toussaint, M. (2009). Probabilistic inference as a model of planned behavior. *ResearchGate*, 3. Retrieved from https://www.researchgate.net/publication/251685706_Probabilistic_inference_as_a_model_of_planned_behavior
- 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.
- Trainor, L. J., & Trehub, S. E. (1992). A comparison of infants' and adults' sensitivity to Western musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, 18(2), 394–402. <https://doi.org/10.1037/0096-1523.18.2.394>
- Tsang, C. D., & Conrad, N. J. (2011). Music Training and Reading Readiness. *Music Perception: An Interdisciplinary Journal*, 29(2), 157–163. <https://doi.org/10.1525/mp.2011.29.2.157>

- Uhlig, M., Fairhurst, M. T., & Keller, P. E. (2013). The importance of integration and top-down salience when listening to complex multi-part musical stimuli. *NeuroImage*, *77*, 52–61. <https://doi.org/10.1016/j.neuroimage.2013.03.051>
- Uitdenbogerd, A. L., & Zobel, J. (1998). Manipulation of music for melody matching. In *Proceedings of the sixth ACM international conference on Multimedia* (pp. 235–240). ACM. Retrieved from <http://dl.acm.org/citation.cfm?id=290776>
- van der Weij, B., 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>
- van Noorden, L. (1975). *Temporal Coherence in the Perception of Tone Sequences*. Technical University Eindhoven, Eindhoven.
- Varèse, E., & Wen-Chung, C. (1966). The liberation of sound. *Perspectives of New Music*, *5*(1), 11–19.
- Vines, B. W., Krumhansl, C. L., Wanderley, M. M., & Levitin, D. J. (2006). Cross-modal interactions in the perception of musical performance. *Cognition*, *101*(1), 80–113. <https://doi.org/10.1016/j.cognition.2005.09.003>
- Vliegen, J., Moore, B. C. J., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *Journal of the Acoustical Society of America*, *106*(2), 938–945. <https://doi.org/10.1121/1.427140>
- Volk, A. (2008). The Study of Syncopation Using Inner Metric Analysis: Linking Theoretical and Experimental Analysis of Metre in Music. *Journal of New Music Research*, *37*(4), 259–273. <https://doi.org/10.1080/09298210802680758>
- Volk, A., Chew, E., Hellmuth Margulis, E., & Anagnostopoulou, C. (2016). Music Similarity: Concepts, Cognition and Computation. *Journal of New Music Research*, *45*(3), 207–209. <https://doi.org/10.1080/09298215.2016.1232412>
- von der Malsburg, C., & Schneider, W. (1986). A neural cocktail-party processor. *Biological Cybernetics*, *54*(1), 29–40.
- Vuust, P., Ostergaard, L., Pallesen, K. J., Bailey, C., & Roepstorff, A. (2009). Predictive coding of music – Brain responses to rhythmic incongruity. *Cortex*, *45*(1), 80–92. <https://doi.org/10.1016/j.cortex.2008.05.014>
- Vuust, P., & Witek, M. A. G. (2014). Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Frontiers in Psychology*, *5*, 1111. <https://doi.org/10.3389/fpsyg.2014.01111>
- Wang, D., & Brown, G. J. (2006). *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press.
- Wang, D. L., & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Transactions on Neural Networks*, *10*(3), 684–697.
- Wang, DeLiang, & Chang, P. (2008). An oscillatory correlation model of auditory streaming. *Cognitive Neurodynamics*, *2*(1), 7–19.

- West, K., & Lamere, P. (2007). A model-based approach to constructing music similarity functions. *EURASIP Journal on Applied Signal Processing*, 2007(1), 149–149.
- Winkler, I., Denham, S., Mill, R., Bohm, T. M., & Bendixen, A. (2012). Multistability in auditory stream segregation: a predictive coding view. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1591), 1001–1012. <https://doi.org/10.1098/rstb.2011.0359>
- Winkler, István, Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, 13(12), 532–540. <https://doi.org/10.1016/j.tics.2009.09.003>
- Womelsdorf, T., & Fries, P. (2007). The role of neuronal synchronization in selective attention. *Current Opinion in Neurobiology*, 17(2), 154–160. <https://doi.org/10.1016/j.conb.2007.02.002>
- Zarcone, A., van Schijndel, M., Vogels, J., & Demberg, V. (2016). Saliency and Attention in Surprisal-Based Accounts of Language Processing. *Frontiers in Psychology*, 7. <https://doi.org/10.3389/fpsyg.2016.00844>
- Zendel, B. R., & Alain, C. (2009). Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience*, 21(8), 1488–1498. <https://doi.org/10.1162/jocn.2009.21140>
- Ziv, J., & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Transactions on Information Theory*, 24(5), 530–536.