The Structure of Cortical Representations of Music and Speech

A THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

NARAYAN SANKARAN

FACULTY OF MEDICINE

THE UNIVERSITY OF SYDNEY

2019

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature: [Redacted] Name: NARAYAN SANKARAN
Abstract

We continually and effortlessly make meaning of the sonic world around us. In music and speech, complex auditory percepts arise when the sensory encoding of physical stimulus properties interact with the neural substrates of domain-specific structural knowledge. For example, in tonal music, the psychological representation of a pitch is determined not only by its afferent sensory properties, but also by its functional relationship with the musical context. Although behavioral, psychophysical, neuroimaging and neurophysiological research over the past several decades have advanced our understanding of how this phenomenon emerges, we are yet to fully explain the neural mechanisms that underlie our ability to extract meaningful information from an acoustic waveform arriving at the ears.

Over the course of three separate experiments, this thesis examined the representational dynamics of musical pitch (chapters 4 & 5) and speech (chapter 6) in human cortex. In each instance, a set of multiple stimuli of interest were presented to listeners while recording their Magnetoencephalographic (MEG) or Electroencephalographic (EEG) activity. In the first two experiments, stimuli comprised a set of musical tones presented within a Western Tonal context. In the third experiment, stimuli comprised eleven different phonemes – the smallest contrastive unit of speech capable of changing word level meaning. Examining multiple stimuli within the same experimental session enabled the neural representation to be characterized based on its collective dissimilarity structure. To measure the dissimilarity between evoked cortical response patterns corresponding to two given stimuli, multivariate pattern analysis (MVPA) was applied to “decode” the stimulus listeners heard from their underlying cortical activity. For each pairwise stimulus combination, decoding accuracy was used as a proxy for their representational distance in the brain. To evaluate the extent to which these neural distinctions honored sensory, acoustic, or perceptual features, empirical M/EEG-based dissimilarities were compared with predictions stemming from various acoustic, peripheral, or perceptual models of stimuli within the framework of Representational Similarity Analysis (RSA).
In experiment 1, we sought to assess the extent to which pitch “class” (i.e. the harmonic function served by a tone within its musical context) could be decoded from neuronal population activity. Using trained musicians as subjects, the MEG activity elicited by four different “probe tones” following a brief tonal context was recorded. Stimuli comprised a set of four pitch-classes whose harmonic and perceptual properties make them strong candidates for observing a clear representational structure in the brain. Two pitch-classes (the tonic and dominant) were “in-key” and perceptually stable within the prevailing context, while the other two pitch-classes (the minor 2nd and augmented 4th) were “out-of-key” and highly unstable. Using MVPA, we observed that the cortical responses to stable and unstable pitch classes were highly separable from one another. To a lesser extent, the brain also distinguished between the two stable classes. However, we found that neural distinctions between the unstable classes were relatively weak, suggesting that in the absence of a clear harmonic schema, the brain’s representation of pitch converges. These neural distinctions were best accounted for by a model based on the standard tonal hierarchy – indicating that the difference in population coding of different pitches in cortex honored the differences in their perceived stability.

The aims of experiment 2 were to complete the characterization of the neural representation of musical pitch for all twelve pitch-classes. Additionally, we examined the temporal dynamics with which sensory representations of pitch (based on acoustics) interface with higher-level representations based on the tonal schema of Western music. Given the high temporal resolution of MEG, the representational dynamics of musical pitch was probed by applying a sliding classification window; training and testing a new classifier at each time point in the neural epoch. Two different models significantly predicted neural dissimilarities at different peristimulus time windows. Beginning 100 ms after onset, cortical distinctions were explained by differences in the fundamental frequency of tones. However, consistent with the findings of experiment 1, from 200 ms onwards the brain’s representation reflected the hierarchy of perceived stability. In addition to examining the brain’s representation of pitch within one key, we also measured the relationship between different major keys in the cortex. In music theory, distances between different keys are described by the well-known circle of fifths. Research suggests that the cognitive basis of “tonality” rests in the pattern of dissimilarities between individual pitches, and
we therefore reasoned that two keys should be related to the extent that they impose a similar structure between individual tones. Indeed, when transposing the original neural distinctions to different musical keys, and correlating the collective dissimilarity structures with one another, the circle of fifths was recovered. The results of experiments 1 and 2 therefore provide a direct link between the complex perceptual structure of tonal music and its underlying origins in the cortex.

Similar to musical pitch, speech perception arises from the mapping of a continuous acoustic signal onto perceptually discrete learnt categorical representations. The broad goals of experiment 3 were to understand the neural processing transformations that occur in the ascending auditory pathway enabling a noisy and highly variable acoustic signal to be mapped onto an invariant representation of a given phoneme. Using similar stimulus decoding methods to those used in previous experiments, we characterized the dynamic representation of a set of eleven consonants based on their evoked EEG activity. Results indicated that cortical dissimilarities between consonants were commensurate with their articulatory distinctions, particularly their manner of articulation, and to a lesser extent, their voicing. To examine the relationship between consonant representations at the auditory periphery and cortex, MVPA was also applied to modelled auditory-nerve (AN) responses of consonants, and time-evolving AN-based and EEG-based dissimilarities were compared with one another. Cortical distinctions between consonants in two periods of activity, centered at 130 ms and 400 ms after onset, aligned with their peripheral dissimilarities in distinct onset and post-onset periods respectively. In relating speech representations across articulatory, peripheral and cortical domains, we further the understanding of crucial transformations in the auditory pathway underlying our ability to perceive speech.

In sum, by measuring cortical stimulus representations in a dynamic fashion and relating them to representations across acoustic, peripheral, and perceptual domains, this thesis furthers our understanding of crucial transformations in the auditory pathway that underlie our ability to perceive speech and music.
Acknowledgements

I first emailed Simon Carlile in 2012, somewhat lost and disenchanted with the available career choices I saw before me. I told him that I was interested in “the intersection between acoustics and neuroscience” without knowing exactly what that may entail. “The intersection is greater than you suppose”, Simon replied. In retrospect, this initial exchange epitomises Simon’s most salient features as a supervisor - a depth of knowledge, a warmth and generosity that knows no bounds, and an enthusiasm for whatever the pursuit. From Simon I learnt a particular brand of scientific inquiry that stresses not only rigour, but engagement and joy. Most importantly, Simon's unending faith in my abilities has in-turn shaped my own sense of self-worth as a scientist. Perhaps this is the greatest gift a mentor could give.

To William Forde Thompson and Thomas Carlson, my de facto mentors at Macquarie University, a most sincere thank you. Bill, from you I learnt what it looks like to possess a truly encyclopaedic knowledge of one's field. Tom, the series of “decoding workshops” you held in 2015 opened my eyes to a whole universe of problem solving within cognitive neuroscience that has profoundly changed the course of my professional and academic life.

To everyone I worked with at the Starkey Hearing Research Centre in Berkeley, CA - Ganesh, Sridhar, Swapan, Christophe - you immediately welcomed me into the family and exposed me to applied problems in speech & hearing neuroscience that have invigorated me ever since. For this I cannot thank you all enough.

My trajectory in reaching this point is far from self-constructed. The privilege I enjoy today is entirely the result of a vast support network of friends, family and partners. In particular, to both my parents, who understood the value of education and made unimaginable sacrifices to ensure that education was had, thank you for your unwavering love, and the many moments of emotional and material support throughout.
## List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>Auditory Cortex</td>
</tr>
<tr>
<td>AN</td>
<td>Auditory Nerve</td>
</tr>
<tr>
<td>aSTG</td>
<td>Anterior Superior Temporal Gyrus</td>
</tr>
<tr>
<td>BCI</td>
<td>Brain Computer Interface</td>
</tr>
<tr>
<td>BM</td>
<td>Basilar Membrane</td>
</tr>
<tr>
<td>CF</td>
<td>Characteristic Frequency</td>
</tr>
<tr>
<td>CM</td>
<td>Confusion Matrix</td>
</tr>
<tr>
<td>DSM</td>
<td>Dissimilarity Matrix (used synonymously with RDM)</td>
</tr>
<tr>
<td>DSS</td>
<td>Denoising Source Separation</td>
</tr>
<tr>
<td>ECD</td>
<td>Equivalent Current Dipole</td>
</tr>
<tr>
<td>ECoG</td>
<td>Electrocorticography</td>
</tr>
<tr>
<td>EEG</td>
<td>Electroencephalography</td>
</tr>
<tr>
<td>EOG</td>
<td>Electrooculogram</td>
</tr>
<tr>
<td>ERAN</td>
<td>Early Right Anterior Negativity</td>
</tr>
<tr>
<td>ERP</td>
<td>Event Related Potential</td>
</tr>
<tr>
<td>F0</td>
<td>Fundamental Frequency</td>
</tr>
<tr>
<td>FDR</td>
<td>False Discovery Rate</td>
</tr>
<tr>
<td>FFR</td>
<td>Frequency Following Response</td>
</tr>
<tr>
<td>fMRI</td>
<td>Functional Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>fT</td>
<td>Ferro-Tesla</td>
</tr>
<tr>
<td>HCT</td>
<td>Harmonic Complex Tone</td>
</tr>
<tr>
<td>HG</td>
<td>Heschl’s Gyrus</td>
</tr>
<tr>
<td>HI</td>
<td>Hearing Impaired</td>
</tr>
<tr>
<td>ICA</td>
<td>Independent Components Analysis</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>IHC</td>
<td>Inner Hair Cell</td>
</tr>
<tr>
<td>ITS</td>
<td>Inferior Temporal Sulcus</td>
</tr>
<tr>
<td>LDA</td>
<td>Linear Discriminant Analysis</td>
</tr>
<tr>
<td>MDS</td>
<td>Multidimensional Scaling</td>
</tr>
<tr>
<td>MEG</td>
<td>Magnetoencephalography</td>
</tr>
<tr>
<td>MMN</td>
<td>Mismatch Negativity</td>
</tr>
<tr>
<td>MVPA</td>
<td>Multivariate Pattern Analysis</td>
</tr>
<tr>
<td>NH</td>
<td>Normal Hearing</td>
</tr>
<tr>
<td>OHC</td>
<td>Outer Hair Cell</td>
</tr>
<tr>
<td>PAC</td>
<td>Primary Auditory Cortex</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Components Analysis</td>
</tr>
<tr>
<td>PET</td>
<td>Positron Emission Tomography</td>
</tr>
<tr>
<td>pIFG</td>
<td>Posterior Inferior Frontal Gyrus</td>
</tr>
<tr>
<td>pMTG</td>
<td>Posterior Middle Temporal Gyrus</td>
</tr>
<tr>
<td>pSTG</td>
<td>Posterior Superior Temporal Gyrus</td>
</tr>
<tr>
<td>pSTS</td>
<td>Posterior Superior Temporal Sulcus</td>
</tr>
<tr>
<td>pT</td>
<td>Pico-Tesla</td>
</tr>
<tr>
<td>RDM</td>
<td>Representational Dissimilarity Matrix (used synonymously with DSM)</td>
</tr>
<tr>
<td>RSA</td>
<td>Representational Similarity Analysis</td>
</tr>
<tr>
<td>SAM</td>
<td>Spiral Array Model</td>
</tr>
<tr>
<td>SPL</td>
<td>Sound Pressure Level</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SQUID</td>
<td>Superconducting Quantum Interference Device</td>
</tr>
<tr>
<td>STG</td>
<td>Superior Temporal Gyrus</td>
</tr>
<tr>
<td>STH</td>
<td>Standard Tonal Hierarchy</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>TPST</td>
<td>Tonal Pitch Space Theory</td>
</tr>
<tr>
<td>VCV</td>
<td>Vowel-Consonant-Vowel</td>
</tr>
<tr>
<td>VOT</td>
<td>Voice Onset Time</td>
</tr>
</tbody>
</table>
# Table of Contents

Abstract .............................................................................................................................. 2

Acknowledgements ........................................................................................................ 5

List of Acronyms ............................................................................................................. 6

Chapter 1: Introduction ................................................................................................... 12
  1.1 Defining the research problem .............................................................................. 12
  1.2 Thesis structure .................................................................................................. 15
  1.3 Authorship Attribution Statements ..................................................................... 16
  1.4 References ......................................................................................................... 17

Chapter 2: General Review .............................................................................................. 19
  2.1 The auditory system ............................................................................................. 19
    2.1.1 The Auditory Periphery ............................................................................... 19
    2.1.2 The Auditory Nerve .................................................................................... 22
    2.1.3 Auditory Cortex .......................................................................................... 23
  2.2 Pitch ..................................................................................................................... 25
    2.2.1 The neurophysiological basis of pitch ......................................................... 25
    2.2.2 Psychological and music-theoretic relations between pitches ...................... 29
    2.2.3 Psychological and music-theoretic relations between musical keys .............. 39
    2.2.4 Neural correlates of perceived relations between pitches and keys ......... 43
  2.3 Speech ................................................................................................................... 51
    2.3.1 Articulatory (& acoustic-phonetic) descriptors .......................................... 52
    2.3.2 Closed-set speech recognition tasks and the Confusion Matrix ................. 54
    2.3.3 Categorical speech perception ..................................................................... 56
    2.3.4 Peripheral processing of speech .................................................................. 57
    2.3.5 Human cortical encoding of speech ............................................................ 58
  2.4 References ......................................................................................................... 62

Chapter 3: General Methodology .................................................................................... 78
  3.1 Apparatus .......................................................................................................... 78
  3.2 Data analysis ...................................................................................................... 80
    3.2.1 Preprocessing (dimensionality reduction and de-noising) ......................... 81
    3.2.2 Multivariate pattern analysis (MVPA) ......................................................... 85
    3.2.3 Representational Similarity Analysis ......................................................... 89
  3.3 References ......................................................................................................... 90

Chapter 4: Experiment One ............................................................................................. 93
  4.1 Abstract ............................................................................................................. 93

List of Acronyms. 9
6.3.2 Stimuli ........................................................................................................ 141
6.3.3 Apparatus .................................................................................................... 142
6.3.4 Procedure & design .................................................................................... 143
6.3.5 Analysis ....................................................................................................... 143
6.4 Results ............................................................................................................ 151
6.5 Discussion ....................................................................................................... 160
6.6 References ..................................................................................................... 162

Chapter 7: Concluding remarks ........................................................................... 168
  7.1 Future implications and applications of cortical measurements of speech
      processing ....................................................................................................... 170
  7.2 Determining the relevant representational unit of speech ................................ 173
  7.3 Further consideration of the cortical mechanisms supporting musical pitch
      perception ....................................................................................................... 175
  7.4 References ..................................................................................................... 176

8. General Appendix ............................................................................................. 180
Chapter 1: Introduction

1.1 Defining the research problem

How does the brain extract meaningful information from an acoustic waveform arriving at the ears? This long-standing question, central to the study of music and speech perception, motivates the series of experiments comprising this thesis. The question is particularly non-trivial considering that, in music and speech alike, perception is not merely the consequence of a faithful encoding of the sensory signal. Two everyday examples provide powerful evidence in support of this claim: Firstly, two acoustically identical tones heard in different musical contexts may bear little resemblance to one another. This percept arises because the two tones, despite being physically identical, serve different harmonic functions within their respective musical frameworks and are thus psychologically distinct. Secondly, a given word or phoneme may be recognized when uttered across a range of different accents, styles, rates, coarticulations, or amidst different background noises. That is, despite considerable variance in the acoustic signal arriving at the ears, the auditory system consistently detects the same speech sound.

From these and other examples, one can surmise that music and speech perception cannot be explained by a simple process of sensory transduction. Instead, a complex series of processing transformations takes place in the ascending auditory pathway (Rauschecker, 1998). At some point, afferent sensory information makes contact with stored internal representations acquired from prior experience. In the former example of the two acoustically identical pitches, these internal representations take the form of a harmonic schema that detail the pitch-structure of tonal music (Lerdahl & Jackendoff, 1985; Deutsch & Feroe, 1981). In the latter example, they reflect learnt linguistic representations of the invariant sensory features that define a given phoneme (Johnson, 2004), and the categorical boundaries that delineate different phonemes (Liberman et al., 1957). Ultimately, these transformations serve to map a continuous and highly variable acoustic signal onto the perceptually discrete and hierarchically structured sound elements that comprise meaningful music and speech (Patel, 2010).

In order to comprehensively study this transformation, it is useful to compare stimulus representations at various stages of auditory processing – acoustic, peripheral, cortical and
perceptual. For example, how does the evoked cortical response to a given speech sound relate to its response properties at earlier stages of neural processing, such as at the output of the auditory nerve synapse? Such comparisons would provide significant insight into how sensory signals morph into perceptually meaningful representations. However, without a correspondence mapping between responses across the different domains and measurement methods, direct comparisons are not straightforward. This is particularly true when considering neural-perceptual comparisons – how would one objectively relate properties of a neural activation pattern with a psychometric function?

A solution is provided by Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008). Within the RSA framework (figure 1), the representation of a collective set of stimuli under study (e.g. multiple phonemes or pitch-classes) are characterized based on their dissimilarity structure at each stage of processing. This structure can be visualized as a representational dissimilarity matrix (RDM). RSA captures both the information represented in a population code and the format in which it is represented (Kriegeskorte et al., 2013). Crucially, because the level of representation is based on dissimilarity rather than absolute encoding features, we may compare various processing stages by simply correlating their respective RDMs, thereby obviating the need for a correspondence mapping.

Using electroencephalographic (EEG) and magnetoencephalographic (MEG) recordings, this thesis empirically characterized the representational geometry of musical-pitch (i.e. tones occurring within a musical context) and speech-sounds in human cortex. Multivariate Pattern Analysis (MVPA) was used to obtain a measure of dissimilarity between the cortical responses evoked by two different stimulus-classes (Haxby et al., 2014). Specifically, Linear Discriminant Analysis (LDA; Duda et al., 2012) was applied to the E/MEG data corresponding to two stimulus classes. The accuracy with which classifiers were able to discriminate between the responses of two different stimuli approximates the dissimilarity in their underlying neuronal population codes, thus providing a measure of their cortical dissimilarity. In this fashion, binary LDA was applied to each pairwise combination of stimuli under study, and the collective set of classification accuracies defined the cortical RDM. Using RSA, cortical RDMs were then compared with known perceptual representations or modelled sensory representations.
Figure 1. Methodological overview. Acoustic waveforms corresponding to stimuli of interest (musical-pitch or phonemes) arrive at the ears and impinge upon the cochlea. Stimuli are encoded at various stages throughout the ascending auditory pathway from periphery to cortex. The resultant percept is highly structured. Rather than examining changes in the absolute encoding features of one given stimulus from stage to stage, we focus on quantifying the dissimilarity across a set of stimuli (labelled A-D in the figure), visualized as a representational dissimilarity matrix (RDM). In this fashion, the collective representational structure of the stimuli may be compared across acoustic, peripheral, cortical, and perceptual domains.

The neural processing of speech has received considerably more attention than that of music. Either empirically or via phenomenological modelling efforts, vast bodies of research have characterized speech or speech-like responses at various stages of processing, including at the auditory nerve (Delgutte, 1980; Zilany & Bruce, 2006; Shamma, 1985a; Shamma, 1985b), brainstem (Russo et al., 2004; Bidelman et al., 2013), cortex (Mesgarani et al., 2014; Chang et al., 2010; Khalighinejad et al., 2017; DiLiberto et al., 2016) and at the perceptual level via psychophysical measurements (Miller & Nicely, 1955; Phatak et al., 2008). However, research that aims to quantitatively measure and relate representations across these stages, using the same set of stimuli and cohort of listeners, remains scant. As such, a comprehensive examination of
the neural processing transformations underlying robust speech perception remains an outstanding goal of research.

In western tonal music, the perceived structure amongst the twelve musical pitch classes is hierarchically organized (Krumhansl, 2001; Krumhansl & Kessler, 1982; Lerdahl, 1988). As with speech, a deficit exists in our understanding of how such a “tonal hierarchy” emerges from sensory processing. In-fact, despite being the perceptual hallmark of western tonal music (Krumhansl, 2001), research is yet to explicitly characterize the representation of the tonal hierarchy in human cortex. A major goal of this thesis was therefore to evaluate whether the seminal models of music psychology, positing a hierarchical organization of pitch, are reflected in the underlying neuronal population coding in the cortex.

### 1.2 Thesis structure

This dissertation comprises a series of three experiments that are presented in chapters 4, 5 and 6 respectively. Before doing so, Chapter 2 provides a general overview of prior research relevant to the processing of musical-pitch and speech. Chapter 3 describes the general methodology common to all three studies. The experiments contain material that has already been published or is in the review process (see section 1.3 below). Whereas chapters 4 and 5 concern the representation of musical pitch and utilize MEG, chapter 6 describes an experiment probing the cortical representation of consonants using EEG. Finally, chapter 7 provides concluding remarks that situate the research within broader goals in music and speech perception and proposes several future lines of research. An appendix will include additional manuscripts that report research outside the scope of the thesis that were nonetheless conducted during the period of PhD candidature.
1.3 Authorship Attribution Statements

This dissertation contains material previously published (reformatted where appropriate).

Chapter 4 of this thesis is published as: Sankaran, N., Thompson, W. F., Carlile, S., & Carlson, T. A. (2018). Decoding the dynamic representation of musical pitch from human brain activity. *Scientific reports*, 8(1), 839. I co-designed this study, collected and analyzed the data, and wrote the draft of the manuscript.

Chapter 6 of this thesis has been accepted for publication in the Journal of the Acoustical society of America as: Sankaran, N., Swaminathan, J., Micheyl, C., Kalluri, S., Carlile, S. (In press) Tracking the dynamic representation of consonants from auditory periphery to cortex. I co-designed this study, collected the data, analyzed the data, and wrote the draft of the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

______________________________
Narayan Sankaran
October, 2018

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

______________________________
Simon Carlile
October, 2018
1.4 References


Chapter 2: General Review

2.1 The auditory system

Both music and pitch perception arise from a complex analysis of the acoustic signal at various levels of auditory processing. In order to describe the neurophysiological mechanisms behind such operations, a brief description of the human auditory system is first needed. An effort is made to only describe information of relevance to the encoding of music and speech.

2.1.1 The Auditory Periphery
The peripheral auditory system in humans is comprised of the outer, middle and inner ear, and is shown in figure 2. The outer-ear is responsible for filtering incoming sound in a direction-dependent manner. While this plays a crucial role in spatial perception, it serves no relevant function in the context of the current work, and will therefore not be discussed here.

![Figure 2](image source: Wikimedia commons. File: Blausen_0328_EarAnatomy)

Figure 2. The structure of the peripheral auditory system in humans showing the outer, middle and inner ear. (image source: Wikimedia commons. File: Blausen_0328_EarAnatomy)
Once incoming sound enters the external ear canal and impinges upon the tympanic membrane, the middle ear is responsible for transferring these acoustic vibrations to the cochlear. This is done via the mechanical action of three small bones, collectively referred to as the ossicles. The smallest of the ossicles contains a footplate that makes contact with the oval window—a membrane-covered opening in the bony wall of the cochlear. Importantly, whereas the middle-ear is air-filled, the cochlear is filled with nearly incompressible fluids (endolymph and perilymph). Given that the two mediums exhibit different levels of resistance, the middle-ear effectively behaves as an impedance-matching device; concentrating the energy distributed across the relatively large area of the tympanic membrane towards the relatively small area of the oval window. Because of this effective area ratio, the lever advantage of the ossicles, and the modes of vibration of the tympanic membrane itself, the middle ear forms an extremely efficient mechanical amplifier, with research suggesting 18:1 acoustic gains and an optimal transmission around 1KHz (Rosowski, 1991) or from 500Hz - 4KHz (Aibara et al., 2001).

Within the inner-ear, the cochlear is a spiral structure containing 2.5 turns in humans and is responsible for the electromechanical transduction of acoustic pressure waves into neural codes. As this spiral arrangement is not crucial to understanding cochlear function, we may describe its “unwound” configuration (figure 3) to more easily illustrate its properties. The basilar membrane (BM) runs along the length of the cochlear. The dynamic response of the BM to incoming sound of primary interest for the neural encoding of sound in the periphery. Incoming pressure waves produce motion of the oval window, resulting in a pressure wave that travels down the length of the cochlear, in-turn causing the BM to move. This motion takes the form of a travelling wave along the length of the BM from its base to its apex. Crucially, the BM has mechanical properties that vary considerably along its length; whereas the basal end (near the oval window) is relatively narrow and rigid, the apical end is wider and less rigid. As a consequence, the seminal studies of Von Békésy (1960) demonstrated that the position of maximal displacement along the BM depends systematically on the frequency of the stimulating sound – whereas the basal end of the BM responds primarily to higher frequency content, the apical end responds primarily to lower frequency content. In this fashion, the BM behaves as a Fourier analyzer; spectrally decomposing the acoustic waveform by mapping its frequency content onto ‘place’ along the BM. This “tonotopic” arrangement of sound is maintained throughout the ascending auditory
pathway up to and including primary auditory cortex (PAC; Kaas et al., 1999; Pickles, 1988). Following from the studies conducted by Von Békésy, subsequent work has more precisely defined the tuning properties of the BM in both live animals and human cadaver ears (Cooper, 1999; Leonard & Khanna, 1984; Robles et al., 1986; Ruggero, 1992).

Electromechanical transduction occurs within the organ of corti, a structure resting on the BM. Its anatomical and functional properties are best described by considering a cross-section of the cochlear, shown in figure 4. The organ of corti consists of approximately 3500 inner hair cells (IHC) and 12000 outer hair cells (OHC) arranged along the length of the BM. Following from the frequency response described above, different hair cells are activated in response to different frequency ranges depending on their position along the BM. Whereas IHCs function to transduce the mechanical movement of the BM into neural activity, OHCs behave as a cochlear amplifier, accepting efferent input from the central auditory system. Small fibers (stereocilia) protrude from the hair cells, and when incoming sound is received, deflections along the basilar membrane cause stereocilia to brush against the immobile tectorial membrane (see figure 4), resulting in the opening of stretch-sensitive transduction channels. The influx of K+ ions from the surrounding
endolymph causes depolarization of the hair cell and results in a receptor potential between the inside and outside of the cell. In-turn, this prompts the release of neurotransmitters that trigger the firing of action potentials in the neurons of auditory nerve (AN) fibers. This polarization-depolarization process repeats with each period of the stimulating sound, resulting in temporal sequences of firing in AN fibers that transmit information to the central auditory system.

**Figure 4.** Cross-section of the cochlea showing the organ of corti, where sound is transduced from travelling pressure-waves into neural activity.

(image source: Wikimedia commons. File: cochlea_crossection.png)

### 2.1.2 The Auditory Nerve

Each AN fiber contains approximately 30 000 neurons that transmit information from the inner-ear to the central nervous system. Single-unit electrode recordings that measure spikes in individual fibers have demonstrated several important properties of neuronal activity in the AN. Firstly, fibers exhibit background firing in the absence of sound, and fall into three basic categories (low, medium, high) based on their spontaneous firing rate (Liberman, 1978). Secondly, each fiber has a characteristic frequency (CF) to which it responds best (Sachs &
Young, 1980; Kim & Molnar, 1979). Research tracing the tuning curves of single neurons has determined that this frequency selectivity follows from the restricted portion of the BM from which each fiber derives its output (Liberman, 1982), thus preserving the tonotopy originating from the BM. Lastly, and most importantly in the context of the current work, neural spikes in the AN exhibit “phase-locking”, firing at a given phase in the period of a stimulating waveform. As a result, the temporal pattern of firing is regular in response to a periodic sound, occurring at integer-multiples of the period of the waveform. Phase-locking has been demonstrated up to 6kHz in the cat (Johnson, 1980) and 3.5kHz in the guinea pig (Palmer & Russell, 1986). Note, however, that these studies also demonstrate that the “strength” of phase locking rapidly declines at frequencies above 1kHz, and that human psychophysical sensitivity seems to mirror this threshold. As a result, the effective range of phase-locking is thought to lie primarily at frequencies below 1kHz. Above these frequencies, variability in the exact instant an action potential is initiated becomes comparable to the period of the waveform, producing a temporal smearing of spikes and resulting in a loss of phase-locking (Palmer & Russell, 1986).

2.1.3 Auditory Cortex

The ascending auditory system is responsible for transcoding the low-level neural firing patterns of AN fibers into high-level representations of auditory objects. To achieve this transformation, neuroimaging and neurophysiological studies support the notion that ascending information is hierarchically processed via a series of nuclei that progressively extract more abstract auditory features from the neural signal. Given the complexity of the auditory pathway, it is well beyond the scope of the current thesis to give an exhaustive description of early and mid-brain computations (for a review see Palmer, 1995). Instead, we will provide a basic description of the structure and function of auditory cortex.

The Auditory Cortex (AC) in humans is bilaterally located in the temporal lobes and comprises Heschl’s gyrus (HG) and superior temporal gyrus (STG). Descriptions often divide the AC into three regions comprised of: a core or primary region (PAC), which is surrounded by “belt” and lateral “parabelt” regions (secondary or non-primary ACs). Consistent with the frequency-based arrangement at the cochlear, at least one tonotopic axis exists in PAC – whereby neurons are spatially arranged according to the frequencies to which they are maximally responsive. The
existence of cortical tonotopy was demonstrated by several early studies showing that the
cortical depth of evoked neuromagnetic field sources depended systematically on the frequency
of the stimulating tone (Romani, Williamson & Kaufman, 1982; Pantev et al., 1988; Pantev et
al., 1995). Cortical tonotopy has also been verified using positron emission tomography (Lauter
et al., 1985) and functional Magnetic Resonance Imaging (Talavage et al., 2004), and has been
found to exist in nonhuman primates (Merzenich & Brugge, 1973). Generally, PAC has been
found to contain neuronal populations that encode faithful stimulus representations,
demonstrating a (somewhat) linear response to primary physical sound attributes such as
frequency and amplitude. In contrast, secondary areas contain neurons responsive to more
complex features such as band-passed noise and species-specific vocalizations (Rauschecker,
Tian & Hauser, 1995). In humans for example, the STG performs spectrotemporal processing
and extracts early phonological representations of speech sounds (Leonard & Chang, 2014;
Mesgarani et al., 2014). In music, the STG shows sensitivity to the spectral motion (Thivard et
al., 2000) relationships between pitches (i.e. pitch-intervals), constituting an early melodic or
harmonic analysis, beyond simple pitch extraction (Zatorre, Evans & Meyer, 1994). After basic
frequency and amplitude information is extracted in PAC, recent models of auditory cortical
processing propose that the flow of information branches into separate ventral and dorsal streams
in a similar fashion to the visual system (Hickok & Poeppel, 2007; Hickok & Poeppel, 2004;
Romanski et al., 1999; Rauschecker, 1998). While the ventral stream has an established
involvement in mapping sounds onto object identity (the “what” stream), the functional role of
the dorsal stream is less clear. Some argue that, similar to visual processing, the dorsal stream
constitutes the auditory processing of spatial information (the “where” stream) (Rauschecker,
1998; Romanski et al., 1999; Kaas & Hackett, 1999), while within the language domain, Hickok
& Poeppel (2007; 2004) posit that the dorsal stream is critical for auditory-motor integration, and
that together, the two systems provide a mechanism for the maintenance of simultaneous
auditory and motor representations of speech necessary for both perception and production.
2.2 Pitch

Pitch is defined as “that property of sound that enables it to be ordered along a scale going from low to high” (Acoustical Society of America Standard Acoustical Terminology, cf. Randel, 1978). Along with timbre and loudness, pitch comprises one of the primary perceptual dimensions of sound; playing a vital role in the perception of both music and speech. Within music, sequential and simultaneous patterns of pitch define melody and harmony respectively. Within speech, pitch-contours form the basis of speech prosody, and in tonal languages such as Mandarin, pitch-contours convey explicit meaning. Though the perceived pitch of a sound is strongly related to the acoustic repetition rate – or “periodicity” - of its waveform, pitch is an entirely subjective entity, with many caveats complicating a simple mapping between acoustics and perception.

In this review, we will progressively build towards the context of natural music-listening. First, we consider isolated tones; describing their acoustic properties and the physiological mechanisms thought to underlie their percept. Next, the perceived relations between individual tones will be considered, before discussing the rich psychological and theoretical structure of pitch when embedded within the harmonic framework of western tonal music. Finally, we will consider pitch in the cortex, reviewing the research seeking to uncover the brain-basis for perceived musical pitch-structure.

2.2.1 The neurophysiological basis of pitch

Pure-tones

For pure-tones, the perceived pitch of a sound is highly correlated with its physical frequency. Two mechanisms readily explain the process by which the early auditory system encodes pure-tone frequency. Firstly, due to the spectral decomposition of sounds in the cochlear, frequency may be represented by its place of maximal excitation along the BM. In-turn, this theory assumes that the neural code to different frequencies lies in the responses of populations tuned to different characteristic frequencies (CFs). Alternatively, the frequency of a pure-tone could be determined solely by a temporal code arising from the phase-locked timing of spikes in the auditory nerve. Other theories have been proposed; for example, Loeb et al. (1983) describe an
array of coincidence detectors that determine the relative phase at two given points along the basilar membrane, which would in-turn represent stimulus frequency. A similar mechanism was also proposed by Shamma (1985a, b). Most of the additional theories may be viewed as second-order computations based on either place or temporal information, and as such, place and temporal codes constitute the most fundamental theories that explain the peripheral encoding of pure-tone frequency.

Harmonic Complex Tones

Though pure-tones have theoretic utility, one cannot fully account for naturalistic pitch-perception by considering pure-tones alone. In everyday environments, Harmonic Complex Tones (HCTs) are the most ubiquitous and important class of pitch-eliciting sounds. HCTs have periodic waveforms that repeat at the rate of a fundamental frequency (f0) and can be decomposed into a series of sinusoidal components that occur at integer multiples of the f0 (figure 5, top panel). Despite differences in their spectrum, an HCT and a pure-tone will be perceived as having the same pitch if the f0 of the HCT is the same as the pure-tone. Remarkably, the HCT need not contain any energy at the F0 for its pitch to be readily perceived, so long as the higher-order components are present. This phenomenon of the “pitch of the missing fundamental”, present from early stages of human development (Clarkson & Clifton, 1985) and occurring also in nonhuman primates (Tomlinson & Shwarz, 1988), proves important in disambiguating the potential mechanisms behind complex pitch perception (see below). In primate secondary AC, pitch-selective neurons that respond equally to pure-tones and missing fundamental HCTs have been identified, providing a neural correlate for pitch constancy across acoustically variant sounds (Bendor & Wang, 2005).

The brain’s ability to resolve the individual frequency components within a harmonic complex (the “frequency selectivity” of the auditory system) plays a crucial role in the mechanisms thought to underlie complex pitch perception. Frequency selectivity has primarily been studied via psychophysical masking experiments (Oxenham & Wojtczak, 2010). Results from such studies postulate a bank of behaviorally-defined auditory filters (figure 5, panel 2nd from top) thought to arise as a consequence of cochlear filtering (Shera et al., 2010). Importantly, the width of auditory filters increases with increasing CF. As a consequence, low-order harmonic
components are exclusively represented within a single auditory filter and are therefore “resolved”, while higher-order harmonics that interact within the same auditory filter are “unresolved”. The effects of auditory filtering can be viewed in the frequency domain as a long-term average “excitation pattern” (figure 5, panel 3rd from top). This may be thought of as a schematic representation of the frequency response of the peripheral auditory system. Resolved harmonics produce peaks in the excitation pattern, whereas unresolved harmonics produce no distinct peaks. Alternatively, we may consider the time-varying output of each auditory filter (figure 5, bottom panel). The output of filters that represent resolved harmonics have time-varying waveforms that are similar to pure-tones of the same frequency, whereas unresolved filter outputs have complex amplitude modulated waveforms that reflect the interaction of the multiple frequency components.

Figure 5. Peripheral representations of a 100Hz complex tone. Moving down from the top illustrates the following – Panel 1: The spectrum of a HCT consists of frequency components at
integer multiples of the $f_0$. Panel 2: Behaviorally-defined bank of auditory filters that are thought to arise from cochlear filtering. Panel 3: The excitation pattern can be thought of as the response (neural or mechanical along the BM) of the peripheral auditory system. Frequency components that are uniquely processed within an auditory filter channel (“resolved” components) produce distinct peaks in the excitation pattern, while higher-frequency components that are not uniquely processed within one channel produce no such peaks (“unresolved” components). Panel 4: Time-domain representation of filter outputs. The output of unresolved channels produces an amplitude modulation at the rate of the $f_0$.


How might the auditory system compute the pitch of a harmonic complex tone? We may restate this problem as: how is the $f_0$ of an HCT extracted, regardless of whether or not it is physically present? As demonstrated by the missing fundamental, the perceived pitch does not necessarily correspond to the position of maximal displacement along the BM. As such, complex pitch perception cannot be readily explained by classical place theory. Models generally fall into two categories depending on whether they propose a spectral or temporal code for inferring pitch. The spectral category of models propose that the auditory system first encodes the frequencies of individual resolved harmonics, before a process of neural “pattern recognition” combines information about individual components to determine an estimate of the overall $f_0$ (Schouten et al., 1962; Terhardt, Stoll & Seewan., 1982; Terhardt, 1974; Goldstein, 1973). In contrast, temporal models propose that the $f_0$ is extracted directly from the temporal patterns generated by unresolved harmonics (Schouten et al., 1962). This assumes that the amplitude modulation at the output of unresolved filter channels produces regular peaks in neural activity, and that the brain is capable of measuring the period of these peaks. The debate over which code underlies complex pitch perception has continued into recent times. While some recent models (e.g. Meddis & O’Mard, 2006) and neural measurements (Cariani & Delgutte, 1996; Griffiths et al., 1998) suggest a purely temporal code – based on the relative timing of spikes in the auditory nerve, a growing body of behavioral evidence suggests that the preservation of frequency-to-place mapping may be crucial. For example, Oxenham, Bernstein & Penagos (2004) dissociated temporal from place information using “transposed stimuli”, in which low-frequency temporal information is delivered to locations on the cochlear tuned to high frequencies. When tonotopy
was disrupted in this fashion, pitch-perception was poor. Furthermore, Oxenham et al. (2011) showed that pitch perception exists beyond the putative upper limit of phase-locked spike timing in humans, suggesting the breakdown of temporal information does not constrain the ability to perceive complex pitch.

In sum, the human auditory system displays a remarkable ability to extract pitch, an entirely subjective entity, from a periodic signal. Because the same pitch percept may be produced by two spectrally differing signals, it is clear that pitch emerges from complex processing mechanisms that go beyond a faithful representation of the stimulus. While physiological, neural, and behavioral research has advanced our understanding of what these mechanisms may be (i.e. place and temporal codes), there is still significant uncertainty regarding the precise neural code underlying perceptual pitch constancy. One possibility is that two pitch discrimination systems concurrently exist, and that this redundancy facilitates pitch perception amidst complex and noisy situations where masking may render one system unreliable.

2.2.2 Psychological and music-theoretic relations between pitches
The psychological quality of a pitch depends largely on its relationship with other pitches. This has motivated numerous attempts, spanning decades of research, to represent the structural relations between pitches. This section first reviews the key geometric models that describe such pitch-relations. Key to understanding these relations is the notion of a “pitch-space”, whereby a topological model is constructed in order to visually convey musical and perceptual relations. In the abstract representational space of these models, the distance between two tones is proportional to their psychological dissimilarity. After describing prominent geometric representations of pitch, we will turn our focus towards models that describe the hierarchical structure of pitch-relations within a musical context.

**Geometric Representations**
The most basic geometrical arrangement of tones is the unidimensional rectilinear scale that describes the psychological dimension of *pitch-height*. This reflects a simple correspondence, whereby physical increases in f0 produce subjective increases in pitch-height. Because musical pitches are a logarithmic function of physical frequency, tones are typically arranged from low to
high along a log-frequency continuum. The perceptual existence of pitch-height has been experimentally verified in several ways. Studies have examined listeners ratings of perceived relatedness, demonstrating that tones proximate to one another on the logarithmic frequency scale tend to be rated as more closely related (Levelt, Van de Geer, & Plomp, 1966; Krumhansl, 1979). Other studies have examined perceptual grouping of melodic sequences, showing that adjacent tones far apart in log-frequency tend to lie at phrase boundaries, while grouped phrases tend to contain small log-frequency intervals (Brown & Cooke, 1994; Hamaoui & Deutch, 2006; Dowling, 1973; Bregman, 1994). These perceptual findings are closely corroborated by a statistical analysis of musical compositions across various cultures. Melodies generally contain adjacent tones that are close in log-frequency, thereby suggesting they are more musically and functionally related than tones distant in the log-frequency scale (Knopoff & Hutchinson, 1978; Dowling & Harwood, 1986).

Though rectilinear scales of pitch-height can account for the similarity of tones that are close in f0, both music-theorists and psychologists would argue that the mental representation of pitch is vastly more complex than a one-dimensional continuum along which tones are ordered from low to high (Piston, 1978; Schoenberg, 1969; Deutsche, 2013). Pitches derive additional psychological meaning based on their structural relationship to other pitches. Certain pairs of tones form musically prominent intervals and appear similar to one another despite being distant in pitch-height. The most obvious example of this is the perceptual equivalence between tones separated by an octave. Two tones separated by an octave have frequencies whose ratios are 2:1. The perceptual equivalence of octaves have been confirmed by various behavioral studies (Kallman, 1982; Lockhead & Byrd, 1981; Allen, 1967), for example, the tendency among listeners with absolute pitch to place a note in the incorrect octave (Miyazaki, 1989), and appears across cultures (Patel, 2010). Furthermore, octave equivalence has been found to exist in infants (Demany & Armand, 1984) as well as in animals (Blackwell & Schlosberg, 1943). Octave equivalence gives rise to the circular dimension of pitch (figure 6); that of pitch-class or pitch-chroma, which forms a distinct perceptual entity from the linear dimension of pitch-height. Tones separated by semitones (the smallest possible interval amongst the chromatic pitches of Western tonal music) are adjacent on the chroma circle. Progressing chromatically through the
twelve pitch-classes corresponds to step-wise movements along the circle such that an octave traverses the full circle and folds back on itself.

Figure 6. The pitch-chroma circle.
(Image source: Wikimedia commons; File: Pitchclassspace.png)

To provide a geometric representation that simultaneously accounts for both the linear and circular dimensions of pitch-perception, various studies have proposed the helical arrangement shown in figure 7 (Shephard, 1982; Ruckmick, 1929; Pikler, 1966; Bachem, 1950; Hahn & Jones, 1981). In this arrangement, pitch-height is represented along the axis of the spiral while the chroma-circle is projected onto the perpendicular plane. Moving along the helical trajectory traverses the various chromatic classes while simultaneously moving up/down in pitch-height. Notably, octaves within any given pitch-class (e.g. C2 – C3) are proximate to one another in the representational space.
Octaves however are not the only interval with special musical and psychological significance. Perceptually related musical tones that ‘fit’ with one another are referred to as *consonant*, while musical tones that are at odds with one another form *dissonant* intervals. In its ability to account for properties of consonance and dissonance, the chroma-circle may be misleading. For example, tones separated by the musical interval of a fifth (e.g. C and G) have a prominent functional relationship in musical composition (Schoenberg, 1969; Lerdahl, 1988; Lerdahl, 2004) and are closely perceptually related (Krumhansl, 1990) and yet on the chroma-circle its constituent tones are distant from one another. Similarly, the minor second is the most dissonant interval (e.g. C and C#) yet its constituent tones are represented adjacent to one another on the chroma-circle. Evidence suggests that these perceptual properties of consonance or dissonance arise from intrinsic properties of the tones themselves. Consonant intervals, for example, contain pitch-values that can be described by simple frequency ratios, such as the octave (2:1) and the perfect fifth (3:2), whereas dissonant intervals tend to be expressed by more complex ratios, such as the minor second (16:15). McDermott, Lehr, & Oxenham (2010) showed that the preference for consonance was related to *harmonicity* – that is, the extent to which the individual frequency components of both complex tones comprising the interval are integer multiples of a plausible common fundamental frequency, and thus essentially comprise one single harmonic series. To account for properties of consonance and dissonance, subsequent research has sought to more accurately describe pitch-relations while still making use of symmetric topological models.
Pitch-height has been modelled in combination with a dimension reflecting the proximity of tones separated by fifths. This produced a double-helix (Shephard, 1982). Assuming octave-equivalence, and thereby eliminating pitch-height, Shepard (1982) simultaneously modelled octave-space and fifth-space to produce a four-dimensional torus, referred to as a “melodic map”. Similarly, modelling the proximity of major fifths and thirds simultaneously produced another torus known as the “harmonic map” (Shephard, 1982; Balzano, 1980; Longuet-Higgins & Steedman, 1971). Chew (2000) expressed the harmonic map in a “spiral array”, which conveys the proximity of major thirds along the vertical axis and perfect fifths along the trajectory of a helix. Modelling three spaces simultaneously (pitch-height, octave-space, fifth-space) would produce a five-dimensional solution consisting of a double-helix wrapped around a helical cylinder. The addition of other consonant intervals would produce even higher-dimensional solutions. Various non-symmetric approaches, for example the 2-dimensional grids described in Cohn (1997), Cohn (1998), Helmholtz (1912) and Balzano (1980), have been attempted to topologically model perceptual and theoretic pitch-relations.

Thus far, we have not been explicitly concerned with the broader musical context in which tones are appearing. The perceived relationships between pitch-classes described above, in intervals such as the octave and fifth, may arise from properties of consonance intrinsic to the tones themselves, without any consideration of the broader tonal-harmonic framework in which they appear. However, in music, the prevailing tonal context has a tonality or key. The perception of tonality is an emergent phenomenon, arising from the recognition of specific pitch-relationships and the pitch-distribution of a musical passage. A formal definition of the cognitive basis of tonality lies in the perceptual hierarchy it imposes upon tones (Krumhansl, 1990), which will be discussed in the next section. However, for now it is important to note that tonality provides a harmonic schema that governs the processing of all subsequent pitches, and depending on the key of the tonal context, each pitch-class is now assigned a unique harmonic function. Crucially, the perceived relatedness between two tones is no longer simply determined by their semitone-interval. Instead, the psychological relationship between two tones forming an acoustically fixed interval varies depending on their respective harmonic functions within the tonal context (Krumhansl, 1979; Lerdahl, 2004).
Arguably the most seminal work quantifying and geometrically modeling the perceived relations between pitches in a tonal context is that of Krumhansl (1979). On each trial, listeners were presented with a tonal context (either a chord or scale) followed by a pair of tones. Listeners were then asked to rate the similarity of the two tones within the tonal system suggested by the context. Every ordered pairwise combination of tones in the set of twelve chromatic tones was tested in this fashion. To illustrate the collective structure of perceived similarity between all tones, non-metric Multidimensional Scaling (MDS) was applied to the similarity data (Kruskal & Wish, 1978). In essence, MDS spatially depicts the similarity ratings while optimally preserving the original structure inferred from the behavioral ratings (see methodology section below for more detail on MDS). In the MDS solution, the distance between classes is proportional to their “dissimilarity”, providing an intuitive visualization of the data. The idealized 3-dimensional MDS solution uncovered by Krumhansl took the form of a cone, with the twelve pitch-classes distributed on the surface. Pitch-classes that were “in-key” (they belonged to the scale in which the context was in) were situated closest to the apex of the cone. In particular, three classes serving harmonic functions called the tonic, dominant, and mediant were located closest to the apex. In the key of C major, these three classes correspond to the tones C, G and E respectively and together they comprise the C major triad chord. Their proximity to one another in the representational space of the model suggests that they are perceptually similar. Conversely, “out-of-key” pitch-classes were distributed around the basal end and are therefore perceptually dissimilar to one another and also to the in-key classes. Lastly, the twelve pitch-classes were organized according to their pitch-height along the circumference of the circular dimension perpendicular to the axis of the cone. Figure 8 demonstrates these perceived relations for four pitches that are labelled by their harmonic function (the tonic and dominant are in-key and belong to the major triad, while the minor 2nd and augmented 4th are out-of-key). By making explicit the perceived relatedness between pitch-classes, the conical model forms a cornerstone in our understanding of the cognitive processing of musical pitch.
**Hierarchical Representations**

As mentioned above, the structural relations between pitch-classes within Western tonal music can be explained by the phenomenon of tonality. A hallmark and defining feature of tonality is the *tonal hierarchy*, in which each pitch-class occupies an ordered position according to its status within the tonal-harmonic framework. This status is conferred in several ways. Compositionally, certain pitches appear more frequently, for longer durations, and in positions of rhythmic stress (Lerdahl, 1988). Additionally, these pitches tend to occur at the beginning and end-points of important phrase boundaries, and serve as the “resolution” to periods of melodic tension. The compositional treatment of these prominent pitches is also reflected in human perception, whereby the same group of privileged pitches are more anticipated within a tonal context (Huron, 2006) and are perceived to have greater tonal *stability* than other pitches (Krumhansl & Kessler, 1982). Given such a tiered structure amongst pitches, it is unsurprising that prominent models choose not to represent pitch-relations as geometrical structures, but rather they conceptualize the configurational properties between pitches as a hierarchical structure (Lerdahl, 1988; Lerdahl, 2004; Lerdahl & Jackendoff, 1985; Krumhansl & Shephard, 1979; Krumhansl & Kessler, 1982; Krumhansl & Schmuckler, 1990; Deutsch & Feroe, 1981; Temperley, 1999). In describing the structure of the hierarchy below, it is important to note the correspondence between music-theoretic accounts (based, for example, on a statistical analysis of Western
music) and perceptual accounts (based, for example, on behavioral ratings of stability). As such, for the purposes of this review, the relevant work across theoretic and behavioral domains will be described as instantiations of the same tonal hierarchy.

A schematic illustration of the putative hierarchical structure amongst the twelve pitch-classes is shown in figure 9 for the key of C major. This arrangement is made explicit in the “basic space” of the *Tonal Pitch Space Theory (TPST)* proposed by Lerdahl (1988; 2004). Pitch-classes are distributed across 5 levels of representation. Moving down in levels corresponds to descending the hierarchy. One single pitch, referred to as the *tonic*, serves as the central pitch or cognitive anchor to which the whole tonal system is oriented. In the key of C major, the pitch-class $C$ serves as the tonic and is represented at the highest level in figure 9. The tonic is emphasized in composition by its prominent placement both harmonically and rhythmically across a range of musical styles (Krumhansl, 1990). The next most prevalent pitch-class is the *dominant*, represented at level $b$ in figure 9. The dominant is located 7 steps above the tonic, together forming the musical interval of a perfect fifth, and corresponds to the pitch-class $G$ in the key of C major. Descending another level, all three tones of the major triad are represented ($C$, $G$ and $E$). At the fourth level, all pitch-classes belonging to the scale of the musical key (all “in-key” tones) are represented ($C$, $D$, $E$, $F$, $G$, $A$, $B$). At the lowest level of the hierarchy, all twelve chromatic pitch-classes are represented, including the “out-of-key” tones.

| level a: tonic | $C$ |
| level b: fifth | $C$ $G$ |
| level c: major triad | $C$ $E$ $G$ |
| level d: diatonic | $C$ $D$ $E$ $F$ $G$ $A$ $B$ |
| level e: chromatic | $C$ $C#$ $D$ $D#$ $E$ $F$ $F#$ $G$ $G#$ $A$ $A#$ $B$ |

**Figure 9.** Schematic representation of the hierarchical structure amongst musical pitches, adapted from the “basic space” outlined in Lerdahl (1988).
The TPST described above is purely theoretical, however, as already mentioned, its predictions are closely corroborated by behavioral data. A first attempt to measure the tonal hierarchy in listeners was made by Krumhansl & Shephard (1979). The authors introduced the “probe tone” method, in which ascending or descending scales were presented as a tonal context. The scale sequences were incomplete, stopping at the penultimate scale degree. Sequences were then followed by a single probe-tone. For example, the format of a trial using the ascending C major scale as a tonal context would be [C, D, E, F, G, A, B, ?]. All twelve chromatic pitch-classes were used as probe-tones and listeners rated each pitch-class based on how well it “completed” the context sequence. Results demonstrated some of the important hierarchical properties present in figure 9. For example, the tonic was consistently rated most highly, irrespective of the musical training of the listener. However, many of the finer details of the hierarchy purported by music-theorists were not evident from these initial experiments, particularly amongst listeners with moderate-to-no musical training. Seeking an improved measurement of listeners’ internal representation of a tonal hierarchy, Krumhansl & Kessler (1982) also utilized the probe-tone method, but now provided a range of different tonal contexts in order to measure a hierarchy that was generalizable across musical mediums. All contexts implied the same musical key and were comprised of either complete scales (including the tonic in the first and last scale degree), major triad chords that contained the tonic in root position, or three different 4-chord sequences comprising common musical cadences that firmly establish a given key. Because the probe-tone was no longer a “completion” to any of these preceding contexts, listeners were now instructed to rate how well each pitch-class “fit” with the musical context. The goodness-of-fit of each tone was construed as a direct reflection of its perceived tonal stability. The stability ratings, averaged across all contexts and listeners, showed a remarkably similar hierarchical structure to the theoretical predictions of the TPST. The profile of average ratings given to all twelve different pitch-classes are collectively referred to as the standard tonal hierarchy (STH). Assuming the key of C major, the highest stability rating was given to the pitch-class C, followed by G and then E. The other in-key pitches (D, F, G, A, B) received the next highest ratings and were all rated similarly. Lastly, the five lowest ratings corresponded to the five out-of-key pitch-classes (C#, D#, F#, G#, A#). The experiment was also repeated for minor tonal contexts and produced a standard tonal hierarchy for minor keys. Further research has shown that these distinctions
between pitch-classes emerge gradually in listeners throughout the course of their musical and cognitive development (Krumhansl & Keil, 1982). The STH has subsequently been used as the basis for several musical key-finding algorithms (Krumhansl & Schmuckler, 1990; Huron & Parn cott, 1993; Temperley, 1999) and forms another cornerstone in our understanding of the cognitive basis of tonality. In particular, it should be noted that the hierarchical organization amongst pitch-classes is not merely a perceptual by-product of tonal music, but rather the defining feature of tonality itself (Hyer, 2002; Bamberger & Hernandez, 2000). In this sense, the STH – with its specific pattern of relative stability across tones – constitutes the perceptual “signature” of a musical key. Finally, an important methodological detail of the probe-tone study of Krumhansl & Kessler (1982) is the authors’ use of circular or “Shepard” tones (Shepard, 1964). These tones are synthesized with octave-spaced frequency components that gradually decay at both the low and high ends of the distribution. Because of this synthesis, circular tones produce an ambiguous percept of pitch height while remaining distinct in their chroma. For this reason, the authors were able to state with reasonable certainty that the hierarchy measured was largely driven by the schematic effects of tonality rather than differences in pitch-height. Additionally, precluding pitch-height ensured that the measured profile of stability ratings was invariant under transposition to a new musical key. The importance of this capability will become clear in the next section.

Statistical learning & computational modelling of pitch structure

While the work of Krumhansl and colleagues establishes the psychological reality of a tonal space, further work has sought to understand the principles by which such a reality emerges from an audio signal through computational modelling efforts. Most notably, behavioural reaction time data from melodic priming experiments have been used to construct physiologically plausible models that detail the transformation of raw audio signals into increasingly complex object-based representations using principles spanning a sensory-cognitive continuum (see Collins et al., 2014 for a review of empirical priming studies and a computational model). These models reveal that listeners’ reaction times are predicted through a combination of sensory features – such as periodicity pitch distributions – and learnt knowledge that projects melodic information onto geometric structures of tonal space. This suggests that processing of tonal
music involves multiple representational stages that span a continuum – from afferent sensory representations to those reflecting learnt cognitive schema. Interestingly, high-level tonal variables were consistently found to explain more of the variance in reaction times than periodicity variables, suggesting a greater contribution of learnt tonal representations in shaping melodic processing.

A similar strand of research has sought to examine the principles underlying listeners’ acquisition of tonal structure. Specifically, research has focused on constructing network models that implicitly learn the statistical and correlational structure of music through mere exposure to sequences of tones (Tillmann, Bharucha & Bigand, 2000). When tested against empirical data dealing with tone, chord, and key relationships, these networks are found to provide a parsimonious account of a variety of dependent variables (e.g. relatedness or memory judgements), suggesting that the representation of tonal structure emerges from networks in the brain that heuristically learn the statistical structure of tonal music.

These computational modelling efforts provide compelling explanations of the empirical findings relating to tonal-music processing, and hint at some fundamental principles – such as statistical learning - that may underlie the acquisition of the psychological representation of tonal pitch structure.

2.2.3 Psychological and music-theoretic relations between musical keys

Just as the relations between individual pitches have been measured or modelled, significant attempts have been made to quantify the extent to which two tonal musical keys are more or less related to one another. Just as pitch-relations were represented via geometric models in a “pitch-space”, key-relations are similarly described via geometric models in a “tonality-space”.

The question of inter-key proximity is motivated by the compositional practice of modulation, in which a piece of music moves from one tonal region to another. Due to the ubiquity of modulation within much tonal-harmonic music, maps of key relationships have a long-standing tradition within music-theory more so than psychology (see Werts, 1983 for a review). Generally, these relationships have been expressed using circles onto which all the major and
minor keys are placed (Schoenberg, 1969). Adjacent positions on the circles correspond to proximate keys while nonadjacent keys correspond to more distant modulations. Considering only the twelve major keys, distances are most commonly represented by the “circle of fifths” shown in figure 10. In this arrangement, the tonics of neighboring keys are separated by the interval of a perfect fifth. For example, the tonic of C major (C) forms a perfect fifth with the tonic of G major (G). The theoretical basis for the close relationship between keys separated by a fifth lies in the fact that their corresponding scales share all but one pitch. For example, the set of pitches comprising the scale of C major (C, D, E, F, G, A, B) is the same as the set of pitches comprising the G major scale with one exception (G, A, B, C, D, E, F♯). As such, the circle of fifths can be viewed as a theoretic set of relations between keys based on their shared pitches; the number of movements along the circle indicates the number of differing scale-tones for the two keys in question.

Figure 10. The circle of fifths geometrically depicts the relationship between the twelve major keys. Neighboring keys on the circle have tonics separated by the interval of a perfect fifth. Labels can therefore either be interpreted as musical keys, or the local tonics of those respective keys.
Minor keys are also related to one another along a circle of fifths, however, in considering the combined relationship between all twenty-four major and minor keys, obtaining a low-dimensional representation becomes considerably more challenging. Schoenberg (1969) developed charts of local key regions. For a given major or minor key, the charts detail the relationship between the closest neighboring keys that fall within its perceptual “control”. The charts embody several additional circular patterns (beyond the circle of fifths) that emerge when describing major-minor key relationships. For example, each major key was strongly associated with a “parallel” and a “relative” minor key, and these major-minor relationships also fold back to form their own closed circles. Werts (1983) proposed a model of key relationships that was based upon an analysis of progressions modulating from one key to another across a large corpus of tonal-harmonic music. The model took the geometric form of a torus that simultaneously modelled all the circular patterns of inter-key distances previously described in Schoenberg’s key charts – specifically, the major and minor circle of fifths, and the parallel and relative major-minor circles.

Another strand of modelling has focused on deriving inter-key relations through transforms of pitch-space (Chew, 2000; Lerdahl, 1988; Lerdahl & Jackendoff, 1985). These models are generative in the sense that they use the known structure that exists between variables to compute an estimate of an unknown structure. Applied to the current problem of modelling tonal structure, generativity enables musical-key representations to emerge from the transformation of smaller constituent components such as chord or pitch representations. The spiral array model (SAM; Chew, 2000) and tonal pitch space theory (TPST; Lerdahl, 1988) models detail pitch-structure, but both models are generative, providing a parsimonious account of the proximity amongst pitches, chords, and tonal regions within one modelling framework. The SAM is a topological computational model in which the “harmonic map” described earlier (comprised of adjacent tones separated by major thirds and perfect fifths along vertical and horizontal axes respectively) is realized along a helix. The model enables a calculation of the “center of effect” for any pitch-collection on the spiral (such as a major-scale). Using these mathematical centers, the distance between higher level elements such as tonalities can be calculated. In this fashion, the SAM maps the relations between pitches, chords and keys within a single representational space. Rather than a topological model, the TPST considers the proximity of chords and tonal
regions by assessing the number of elements in the *basic space* (figure 9) that are re-arranged under the application of chordal and regional transformation rules. Notably, each application of the regional rule transforms the diatonic major pitch-collection to a new diatonic arrangement a fifth above. Consistent with the theoretical models described earlier, repeated application of the rule recovers the circle of fifths. Extending the rule to both major and minor diatonic pitch-collections produces a “regional torus” that contains similar structural relationships to those described in the key charts of Schoenberg (1969) and toroidal model of Werts (1983).

In addition to the above theoretical models, behavioral research has provided an empirical approach to uncovering the perceived relations between musical keys. To review the most significant finding in this domain, we return to the ‘probe-tone’ study conducted by Krumhansl & Kessler (1982). Firstly, as described earlier, ratings of perceived stability were collected for each pitch-class within the framework of a single major and minor key, resulting in the major and minor STH. Evidence suggests that the cognitive basis of tonality arises from this pattern of relative stability (Krumhansl, 1990; Zatorre & Krumhansl, 2002). With this in mind, the authors reasoned that two keys are related to the extent that they impose similar patterns of stability on the individual pitch-classes – that is, keys that are closely related should have similar hierarchies. Although stability ratings were reported in the keys of C major/minor, the results are generalizable, and would therefore remain invariant to transposition. For example, the rating given to the pitch-class C in C major should be the same as that given to G in G major (because both C and G function as the tonic in C major and G major respectively). Thus, rather than performing the exhaustive task of collecting ratings in every major and minor key, which would have resulted in a [12 x 24] study design, the authors transposed the STH to different keys; keeping the class-labels fixed but shifting the entire profile by the appropriate number of steps. It is important to note that this approach was possible, in part, because tones were synthesized to diminish the effects of pitch-height, thereby ensuring that representations were driven by the harmonic function of each tone within the key of the context, and not by its absolute frequency value. The relatedness of two keys was now simply a matter of assessing the similarity of their two respective profiles. To achieve this comparison, the two profiles in question were correlated (where relatedness equals 1 – correlation). Remarkably, for major keys, the pattern of correlations corresponded perfectly to their distances along the circle of fifths. For both major
and minor keys, the MDS solution was 4-dimensional and comprised of two circles: the first two dimensions emphasized the major and minor circle of fifths, while the last two dimensions emphasized the relative and parallel major-minor circle. Combining the two circular arrangements produced a torus that had many structural features in common with theoretical models of key-relations discussed above. Specifically, the major and minor circle of fifths each wrapped around the torus three times. As a result, every major key was flanked by its relative minor key on one side (e.g., C major and a minor) and its parallel minor on the other side (e.g. C major and c minor).

In sum, research spanning the past several centuries sought to quantify the psychological and musical relatedness of different keys. Theoretical accounts of inter-key distances are derived from fundamental notions of harmony, an analysis of tonal music, and through generative models that first consider the structure of individual pitch-relations within one key. An empirical account has also been obtained by correlating the behaviorally-determined STH across different keys. The correspondence across theoretical and empirical models is striking. For example, the key charts mapped by Schoenberg (1969) correspond to local regions in the toroidal MDS solution to correlational distances in Krumhansl & Kessler (1982), which in-turn is remarkably similar to the toroidal model of Werts (1983). Because of this correspondence, we may summarize the results emerging from these various strands of research as the following: (1) The relatedness of the twelve major or twelve minor keys of Western tonal music are best described by their distances along the major and minor circle of fifths, respectively; (2) associations among the twenty-four major and minor keys can be represented as a “tonality surface” projected onto the doughnut shape of a torus.

2.2.4 Neural correlates of perceived relations between pitches and keys

Having reviewed seminal theoretical and behavioral models that are thought to explain the perception of musical pitch and tonality, we now turn our attention to the underlying neural activity responsible for generating such a percept. In reviewing relevant literature below, there appears to be clear evidence for a stored neural representation of hierarchical tonal structure. However, research that explicitly characterizes the representational geometry of pitch-relations,
thereby providing the neural analogue of aforementioned psychological models, remains scant – thus motivating the first two studies of this thesis.

**Neural Representations of Pitch-Height**

As described earlier, the rectilinear psychological dimension of *pitch-height*, along which tones are arranged by their log-frequency value, accounts for the fact that physical increases in the fundamental-frequency of a tone produces subjective increases in their pitch-height. Recall that the auditory cortex is organized tonotopically, whereby neurons are arranged spatially according to the frequencies to which they are maximally responsive (Romani et al., 1982; Pantev et al., 1988; Lauter et al., 1985). One may conclude that tonotopy provides an inherent neural representation of pitch-height, whereby the extent to which two tones are distant in pitch-height would be reflected in the dissimilarity of their neural spatial encoding. However, a tonotopic frequency arrangement does not provide an explicit representation for the pitch of complex tones, which contain higher order frequency components. It could be the case, for example, that complex pitch is represented in the distributed activity of neuronal populations. Several MEG studies suggest that this is not the case - providing strong evidence for the existence of a tonotopic *pitch* map (Pantev et al., 1989; Langner, 1997). Pantev et al. (1989) first noted that the depth of the equivalent current dipole (ECD) for the M100 wave of an Auditory-evoked magnetic field increases with the logarithm of the frequency of the stimulating pure-tone (Romani et al., 1982; Pantev et al., 1988; Lauter et al., 1985). The authors then showed that there was no difference in the depth of the ECD between a pure tone and a missing-fundamental complex with the same implied f0. It was concluded that the tonotopic representation at cortex is different from that at the cochlear, whereby cortical tonotopy reflected virtual (i.e. subjective) pitch rather than spectral content. Langner et al. (1997) used a similar approach, examining the arrangement of the generators of the M100 in auditory cortex. Rather than using only two missing-fundamental complex tones and two pure-tones (as was the case in the study of Pantev et al., 1989), the authors included six exemplars of each. In testing this extended stimulus set, the authors also concluded that a tonotopic arrangement of complex pitch exists in the cortex. However, unlike Pantev et al. (1989), their results suggested that the tonotopic pitch-map coexists in an orthogonal arrangement to the tonotopic frequency map. Nevertheless, there is clear evidence for the tonotopic mapping of pitch – a purely subjective entity – in the cortex, and
Neural Representations of Consonance and Dissonance

Continuing with the process of reviewing the neural analogues of perceptual pitch-properties, we next consider the neural basis of consonance and dissonance. Recall that certain tone-pairs whose frequencies can be expressed as simple mathematical ratios, such as the octave (2:1) and the perfect fifth (3:2), form consonant intervals and are perceptually closely related. Conversely, intervals expressed through complex mathematical ratios, such as the minor second (16:15), are dissonant and their constituent tones are distantly related. Evidence suggests that the neural basis for these perceptual preferences originates at early stages of auditory processing and is rooted in the fundamental processing constraints of the auditory system (Zatorre, 2005; Trainor, 2008). Animal studies have demonstrated that the magnitude of phase-locked activity associated with processing musical intervals in the auditory nerve (Tramo et al., 2001), inferior colliculus (McKinney et al., 2001), and primary auditory cortex (Fishman et al., 2001) is correlated with behavioral measures of their consonance or dissonance. In the human brainstem, Bidelman & Krishnan (2009) measured the frequency-following response (FFR) for various musical intervals. For consonant intervals, the encoding in brainstem was more robust, signified by higher measures of “pitch salience” (i.e. enhanced representation of the temporal and harmonic properties of pitch), than the encoding of dissonant intervals. Differential processing of consonant and dissonant intervals has also been found in the cortex using functional Magnetic Resonance Imaging (fMRI; Minati et al., 2009) and intracranial EEG (Foo et al., 2016). Using fMRI, Moerel et al. (2015) uncovered evidence for the neural basis of octave equivalence, showing that cortical neuronal populations were selectively tuned to multiple frequencies in octave spacing, thereby suggesting that the cortical representation of tones was organized more-so according to pitch-chroma, above and beyond pitch-height. A similar mapping of pitch-chroma in auditory cortex has been found using EEG (Briley et al., 2012).

Neural Representations of Musical Pitch

To investigate cortical representations of pitch within a musical context, research conducted during the past several decades has made use of auditory event-related potentials (ERPs). By
recording a subject’s electroencephalogram (EEG) or magnetoencephalogram (MEG) during listening, ERPs measure the summed electrical or magnetic activity across entire neuronal populations in response to a physically sounded stimulus. An advantage to this approach is afforded by the superior temporal resolution of EEG/MEG. Because of this, one can analyze the structure and time-course of ERPs elicited in response to particular auditory events in fine detail. Different components of the ERP, elicited during different time windows after onset of the auditory event, are associated with different cognitive processes. An example of an ERP is shown in figure 11. The paradigms involved in ERP studies often examine pitch-processing as it occurs within melodic sequences. In such a melodic framework, hierarchical pitch-structure manifests as a system of rules, or a music-syntax, that governs the sequential pattern of unfolding pitches (Lerdahl & Jackendoff, 1985; Patel, 2003). It is important to note the correspondence between the tonal hierarchy and rules of music-syntax: tones that are hierarchically prominent (and perceptually stable) tend to be syntactically regular and more ‘expected’, while hierarchically low (unstable) tones are syntactically irregular and ‘unexpected’. The same is true of chords: syntactically-irregular chords are built using a local tonic that is unstable. Thus, by identifying aspects of neural processing sensitive to violations of music-syntax, the ERP research described below provides implicit evidence of tonal-structure in the brain, though it falls short of explicitly characterizing the geometric structure of the collective representation of musical pitches.

![Figure 11. Example of an Event-related Potential (ERP).](image)

ERPs represent the average time-locked neural activity elicited in response to an auditory stimulus. Different components such as
the N1 and P3 are thought to index specific cognitive processes that occur at specific latencies following the onset of an auditory event. (image source: Wikipedia commons. File: ComponentsofERP.svg).

Several pioneering ERP studies sought to identify whether syntactically unexpected (out-of-key) tones were processed differently from syntactically expected (in-key) tones in the cortex (Besson & Macar, 1987; Besson et al., 1994; Besson & Faïta, 1995, see Besson & Schön, 2001 for a review). In these studies, familiar Western classical melodies and unfamiliar musical phrases that followed the rules of tonal harmony were presented to both musicians and non-musicians. Three different versions of melodies were constructed by altering the final note: phrases either ended with the most congruous (expected) pitch, a pitch that was in-key but not the most expected, or an out-of-key pitch that was therefore perceived as unstable and wrong. When analyzing the ERPs associated with these phrase endings, unexpected notes clearly elicited a central to posterior late positive component peaking between 500 and 600 ms (identified as the P600). Importantly, the amplitude of the P600 was commensurate with the degree of incongruence: the most unexpected (perceptually unstable) tones elicited the largest P600, suggesting a direct correspondence between the magnitude of the P600 and the hierarchical position of the stimulating tone. Furthermore, the P600 was larger in musicians than non-musicians, and for familiar rather than unfamiliar musical phrases, suggesting that the P600 amplitude reflects the degree to which the syntactic violation is explicitly known. Lastly, the P600 has also been elicited by out-of-key tones in the middle of unfamiliar phrases (Brattico et al., 2006), and in a range of similar paradigms that present syntactically irregular tones (Verleger, 1990; Paller et al., 1992; Besson et al., 1998; Miranda & Ullman, 2007) or chords (Patel et al., 1998). Taken together, these findings suggest that the P600 reflects a general cortical sensitivity to music-syntactic processing. In fact, the P600 has been similarly elicited in response to deviations of linguistic syntax (Patel et al., 1998), suggesting that it may reflect the violation of more general rule-based structure in the perception of sound sequences (Koelsch, 2011).

Independent of the P600, earlier components of the ERP sensitive to music-syntactic violations have been identified. Janata (1995) showed that unexpected chords presented at the end of sequences elicited a positive component at 300ms (P300) in the resultant ERP. The magnitude of
the elicited P300 was thought to index the degree of “unexpectedness” or dissonance of the chord. A similar P300 effect was found in response to syntactically irregular chords by Regnault et al. (2001). Koelsch et al. (2000) presented listeners with sequences comprised of five chords that were either syntactically regular or contained an irregularity in the third or fifth position within the sequence. Importantly, these irregular chords were consonant when played in isolation, but were rendered ‘abnormal’ only by the harmonic framework in which they occurred. Furthermore, the degree to which the deviant chord sounded irregular was dependent on its placement within the sequence. That is, even though the irregular chord was physically identical when placed in third and fifth positions, due to the structure of the chord progression, the chord sounded more unexpected when placed in the fifth position of the sequence than the third position. Interestingly, syntactically irregular chords, when presented in the fifth position only, elicited an ERP-component with negative polarity over frontal electrodes with a right-hemisphere weighting and a peak latency 150 – 200 ms. In both its scalp topology and latency, the response was strongly reminiscent of the language-syntax sensitive early left anterior negativity (ELAN; Friederici, 2002), albeit more right-hemisphere weighted. Thus, Koelsch et al. (2000) termed the response the early right anterior negativity (ERAN). Subsequent studies have confirmed that the ERAN is robust across a range of similar musical contexts (Loui et al., 2005; Leino et al., 2007), with peak latencies ranging from 150 – 350 ms after the onset of an irregular tone or chord depending on the specific paradigm. When the position of the irregular chord in a sequence is unpredictable, the response often occurs at longer latencies and with a more anterior-temporal distribution (Koelsch & Mulder, 2002; Patel et al., 1998). When elicited due to syntactically-irregular tones amidst melodies, the ERAN may have a shorter latency than when elicited by irregular chords (Koelsch & Jentschke, 2010). Though the ERAN appears similar to other ERP components, it is thought to reflect the operation of distinct cognitive mechanisms. For example, while the ERAN indicates the initial detection of music-syntactic irregularities, the P600 is thought to reflect a later process of re-analysis, repair, and integration into the broader harmonic context (Koelsch, 2009; Koelsch, 2011). In addition, at first the ERAN might seem to resemble the abstract feature Mismatch Negativity (MMN), elicited due to deviations of pitch-contour in a sequence of repetitive short melodic phrases (Saarinen et al., 1992). However, the cognitive mechanisms underlying the two responses differ. The MMN is based on a representation of inter-sound regularities extracted online from the current acoustic environment.
In contrast, the ERAN reflects the violation of rule-based regularities that are stored in long-term memory, reflecting the listeners’ prior exposure to the pitch-structure of western tonal music (Koelsch, 2009).

Even earlier music-syntax sensitive ERP components have been identified. Examining responses evoked by the endings of unfamiliar audio-visually presented melodies, Tervaniemi et al. (2003) found that the amplitude of an early negativity (the N100) was modulated according to the degree of expectedness. Generally, the N100 is elicited by acoustic onsets (Näätänen & Picton, 1987) and is thought to reflect the formation of sound feature representations in auditory cortex. Notably, evidence suggests that the N100 is modulated by prior exposure to sounds, over both immediate and long-term time-scales (Brattico et al., 2003; Menning et al., 2000; Pantev et al., 1998; Pantev et al., 2001), suggesting that the N100 may well reflect the early neural processing of learnt tonal schema.

The studies above report several independent ERP components – the P600, P300, ERAN, and N100 that demonstrate a sensitivity to music-syntactic violations, providing direct evidence that the cortical processing of pitch is influenced by the broader structure of the music in which it appears. However, none of the aforementioned studies have explicitly probed the brain’s representation of hierarchical pitch-structure. To this end, Krohn et al. (2007) presented listeners with sequences of tones comprising a major scale. The tonic occurred more frequently than the other scale degrees. Comparing the ERPs evoked by each tone that functioned as a non- tonic scale-degree, the authors found that the N100 was largest in response to the fifth scale-degree pitch, which is the highest amongst non-tonic tones in the tonal hierarchy. Based on these results, the authors concluded that the N100 may be sensitive to tonal pitch-structure within melodies.

To summarize the conclusions of ERP-research, there is strong evidence to suggest that the neural processing of pitch is influenced by the tonality of the musical context, beginning as early as 100ms after pitch-onset. The N100 may be modulated by the hierarchical position of the evoking-pitch, while the ERAN, P300 and P600 all show sensitivity to the extent to which the evoking tone or chord is syntactically unexpected, thus implicitly suggesting a stored internal representation of tonal structure.
Outside the domain of ERP-based research, studies have identified the neuroanatomical underpinnings of music-syntax processing using both fMRI (Minati et al., 2008) and intracranial EEG recordings (Sammler et al., 2013), implicating a bilateral cortical network involving areas such as superior temporal gyrus, inferior frontal gyrus, and premotor cortex. However, in all prior work investigating the neural processing of musical pitch, no research known to the author explicitly seeks to uncover a representational structure in the cortex. That is, prior research falls short of quantitatively measuring the relatedness between the individual pitch-classes. If such a research-goal were to be accomplished, it would provide the neural analogue to the aforementioned psychological descriptions embodied in geometric and hierarchical models. Perhaps the only neural study whose results directly relate to a topological model of tonal-space is that of Janata et al. (2002). In that study, musically trained listeners heard a melody that systematically modulated through all twelve major and twelve minor keys while having their fMRI activity recorded. While hearing the melody, listeners were required to perform two simultaneous detection tasks – one for timbre and another for tonality deviations. Using such a dual-task paradigm that required different perceptual analyses of the same attended melody allowed the authors to isolate variance in the fMRI activity associated with changes in tonal-structure. Using a regression analysis, the authors identified a tonality sensitive cortical network in the rostromedial prefrontal cortex and reconstructed a “tonality surface” – a pattern of activation associated with each musical key. When the tonality surfaces corresponding to all 12 major and 12 minor keys were correlated with one another, the pattern of relationships uncovered contained striking similarities to those present in perceptual and theoretic descriptions of inter-key relationships. Specifically, the tonality surface of a given major or minor key was highly correlated with that of keys a fifth above, thereby recovering the major and minor circle of fifths in the cortex. Additionally, the tonality surface corresponding to each major key was highly correlated with that of its relative minor (and to a lesser extent its parallel minor) key, constituting another circular pattern of inter-key relations that, together with the major & minor circle of fifths, builds the toroidal geometry used to describe perceptual associations between all musical keys.
While the results of Janata et al. (2002) contributes to our understanding of cortical representations of tonal structure, several questions remain. Firstly, presumably due to the relatively slow response of fMRI, the analysis was performed at the level of musical keys, leaving the question of individual pitch-class relations within one key unanswered. Importantly, theoretic and cognitive descriptions state that tonality arises from the differences in perceived stability across individual tones, rather than the general accumulation of information across a musical passage (Krumhansl, 1990). As such, an accurate measurement of tonal structure in the cortex must involve a fine-grained measurement of neural responses at the level of individual tones. As such, the lack of temporal-resolution in fMRI activity prevents the characterization of a neural structure that offers a fair comparison with perceptual and theoretic models that are founded on relations between individual pitch-classes. Motivated by these unanswered questions, the first two studies described in this thesis introduce a novel methodology focused on uncovering the cortical distinctions between individual pitch-classes. In doing so, a representation of tonality in the human cortex is revealed, and a direct correspondence between music-theoretic, perceptual, and neural descriptions of the harmonic organization of tonal music is established.

2.3 Speech

There exist several reasons for jointly considering representations of music and speech within this thesis. Prior research has revealed remarkable similarities between the cognitive workings of the two modalities. Firstly, evidence suggests that the processing of both tonal and linguistic syntactic information requires contributions from the same neuronal populations, precipitating both the “shared syntactic integration resource hypothesis” (Patel, 2003; Patel et al., 1998) and “syntactic equivalence hypothesis” (Pohlenz, 2014). Secondly, like music, language involves transforming a continuous acoustic signal into a series of perceptually discrete elements. Whereas continuous music signals are mapped to discrete internal representations of tones and chords, continuous speech signals are mapped to discrete internal representations of phonemes or syllables. In both music and speech, these perceptually discrete elements are hierarchically organized to form meaningful structured sequences (Patel, 2003). Whereas individual tone and
chord representations are organized to form coherent melodies and progressions, individual phonemes are organized to form higher-level representations of intelligible words, phrases, and sentences. Finally, because both modalities (music & speech) operate on such clearly defined perceptually discrete units (tones and phonemes respectively), the machine-learning based analytical approach we adopted for studying the collective representational structure of pitch-classes can be directly transferred to the study of a closed set of phonemes (see general methodology in chapter 3).

Of the different phonemes comprising the English phonetic inventory, this thesis is specifically concerned with consonants due to their key role in speech intelligibility and susceptibility to degraded perception in listeners with hearing loss (Phatak et al., 2009; Walden and Montgomery, 1975; Bilger and Wang, 1976). Here, we review relevant work examining the representation of consonants in both perceptual and neural domains. In doing so, we find that while research characterizes consonant representations at each level of processing (acoustic, peripheral, cortical, and perceptual), studies that seek to relate these various stages of processing with one another, thereby characterizing the nature of transformations of speech representations, remains scant.

2.3.1 Articulatory (& acoustic-phonetic) descriptors
Phoneticians and linguists classify different phonemes by features of the articulatory process used to generate those sounds, such as their voicing, manner of articulation and place of articulation. These articulatory descriptors are correlated with the acoustic-phonetic properties of speech, and thus the two terms may be used interchangeably. As we shall describe in this section, these articulatory properties are the underlying property of speech discriminated by listeners, and as such they provide a basis for the formation of perceptual speech sound categories. That is, phonemes that share common articulatory features tend not only to be acoustically similar, but these articulatory groupings are also reflected in the observed perceptual similarity patterns between consonants. Three key articulatory descriptors together fully distinguish a consonant: (1) voicing, (2) manner of articulation, and (3) place of articulation. The voicing of a given phoneme describes whether the vocal folds undergo periodic vibration when producing the sound. For example, the consonant /p/ is unvoiced while /b/ is voiced. Rather than falling into discrete categories, consonants lie along a continuum of voice onset time (VOT). The manner of
articulation describes the interaction and configuration of the speech-organs (tongue, lips, vocal folds and tract). For example, “plosives” are consonants in which the vocal tract is completely occluded by the tongue blade, lips, or glottis, thus preventing any airflow. “Nasals” are those phonemes in which the vocal tract is occluded but airflow is directed through the nose.

Part (a) of the diagram shows the manner of articulation categories. The categories are not necessarily mutually exclusive (one consonant may belong to multiple categories). Lastly, the place of articulation describes the location at which the vocal tract is obstructed by an articulator in order to produce a given consonant. For example, alveolar consonants (e.g. /t/ and /z/) have a place of articulation situated near the alveolar ridge, while bilabials are consonant

**Figure 12.** Phonemes and their associated manner of articulation categories. (Image Source: Wikipedia Commons. File: 654px-IPA-euler-manners-features.png)

Different manner of articulation categories are shown in Figure 12. Note that the categories are not necessarily mutually exclusive (one consonant may belong to multiple categories). Lastly, the place of articulation describes the location at which the vocal tract is obstructed by an articulator in order to produce a given consonant. For example, alveolar consonants (e.g. /t/ and /z/) have a place of articulation situated near the alveolar ridge, while bilabials are consonant
sounds articulated by pressing the top and bottom lips together (e.g. /b/, /p/, /m/). Figure 13 displays several places of articulation within the human vocal tract.

**Vocal Tract**

**Articulators and Places of Articulation**

![Vocal Tract Diagram](Image source: Wikimedia Commons. File: Tract.svg)

**Figure 13.** Different places of articulation within the human vocal tract. (Image source: Wikimedia Commons. File: Tract.svg)

### 2.3.2 Closed-set speech recognition tasks and the Confusion Matrix

Speech recognition has primarily been studied using closed-set recognition tasks, in which listeners are tasked with identifying the phoneme that was sounded on a trial, given a set of multiple possible phonemes. The results are typically summarized in confusion matrix (CM). CMs are particularly useful as they readily illustrate the emergence of confusion patterns when listening environments have low SNRs or are degraded with low-pass filtering. These confusion patterns represent systematic misidentifications between phonemes, in-turn revealing subsets of
highly confusable phonemes. Using such a paradigm, Miller & Nicely (1955) conducted their seminal work quantifying perceptual confusions among consonants. Speech tokens were presented to listeners at various signal to noise ratios (SNRs) and under several low-pass and high-pass filter conditions. Particularly at low SNRs and in low-pass conditions, confusions between consonants fell into predictable patterns as determined by several articulatory features. Furthermore, across SNRs, it was found that each articulatory feature functioned as a separate perceptual channel, whereby the perception of one feature was little affected by the perception of others. Later work sought to build on these results. For example, using the original confusion data from Miller & Nicely (1955), Allen (2005) analyzed consonant confusions as a function of SNR. When viewing confusion patterns in this way, recognition performance was found to be linear as a function of the *articulation index* - a measure of speech intelligibility related to the SNR (Allen, 1994). Phatak et al. (2008) repeated the original experiment of Miller & Nicely (1955), seeking to more robustly characterize confusion patterns using modern experimental techniques. These updated results and analysis not only corroborated the original findings, but also revealed several interesting insights at the level of individual consonant perception. Firstly, in certain cases it was found that in addition to degrading intelligibility, lower SNRs can serve to perceptually morph one consonant into another. Secondly, orthogonal to the predicted clustering of consonants according to their articulatory features, there exists significant variability in the perception of the same consonant across different utterances. This manifests as variance in (1) confusion heterogeneity, whereby competitors within the confusion group of a consonant vary, or (2) confusion threshold, whereby the SNR threshold at which a confusion group emerges varies. These results suggest that specific perceptual “events” unique to an individual utterance can significantly alter the recognition of a given consonant. Finally, closed-set consonant recognition paradigms have demonstrated that hearing-impaired (HI) listeners, typically with high-frequency sensorineural hearing-loss, make significantly more errors than their normal-hearing (NH) counterparts (Walden and Montgomery, 1975; Bilger and Wang, 1976; Doyle et al., 1981; Dubno et al., 1982; Gordon-Salant, 1985), motivating metrics that quantify the differences in confusion data between NH and HI listeners, such as the *consonant loss profile* (Phatak et al., 2009). In the context of the current thesis, the confusion data generated from this body of psychophysical work provides a quantitative characterization of the perceived structure between consonants, with which neural data can then be compared.
2.3.3 Categorical speech perception

As mentioned, various transformations in the auditory pathways map a continuous and variable acoustic speech signal onto perceptually discrete phonemic units. Epitomizing this transformation is the phenomenon of categorical speech perception (Liberman et al., 1957; Liberman et al., 1967; Studdert-Kennedy et al., 1970; Pisoni, 1971). For example, the consonants /b/ and /p/ are differentiated by their position along the acoustic continuum of VOT. As such, using speech synthesis techniques it is possible to gradually morph one phoneme into the other. Liberman et al. (1957) showed that listeners hearing sounds varying along this acoustic continuum always perceived either /b/ or /p/. That is, incremental changes in the synthesized phoneme produced an abrupt change in perceived category rather than a gradual transition. Similar categorical performance profiles along other acoustic continuums (e.g. between /b/ and /d/) were observed. Based on these results, the authors concluded that categorical speech perception emerges from the articulatory movements required in order to produce the phonemes. For example, the natural vocal apparatus in humans is not capable of producing the sounds located in-between the voiced /b/ and voiceless /p/. Thus, the proposed “motor theory of speech perception” maintained that the sensory feedback from articulating a speech sound was also responsible for mediating its perception (Liberman & Mattingley, 1980). However, criticisms of the theory abound. In particular, research has shown similar categorical recognition patterns for non-speech stimuli (Lane, 1965), including musical pitch (Krumhansl & Shephard, 1979; Siegel & Siegel, 1977a; Siegel & Siegel, 1977b).

Subsequent research has sought to determine the neural basis of categorical phoneme perception (Molfese, 1978; Bidelman, Moreno & Alain, 2013; Chang et al., 2010). In particular, Chang et al. (2010) used direct cortical surface recordings to examine the neural representation of synthesized phonemes that gradually varied in small and even acoustic increments between /ba/, /da/ and /ga/ in human pSTG. Indeed, the relatively continuous acoustic continuum along which syllables varied produced distinct and invariant neuronal population responses. Furthermore, neurometric and psychometric functions indicated similar category boundaries. To examine the functional level in the auditory pathway at which neural activity first exhibits categorical response behavior, Bidelman, Moreno & Alain (2013) recorded neural activity generated at both
cortical and subcortical levels in response to vowel sounds synthesized along an /a/ - /u/ continuum. The authors found that properties of the auditory brainstem response reflected features of the stimulus waveform, faithfully encoding the acoustics of speech, but not higher-level abstract phonetic categories. In contrast, and consistent with Chang et al. (2010), later cortical responses contained information reflecting the distinct perceptual phoneme categories that listeners reported hearing. These findings thus highlight the critical transformations of the neural representation of speech taking place between the brainstem and early auditory cortical areas, in this case representing the acoustic-phonetic mapping underlying categorical speech perception.

2.3.4 Peripheral processing of speech

Beginning in the 1970’s, research examined the earliest stages of neural speech processing by recording the activity of single neurons in the auditory nerve fibers of animals (cats, rodents, monkeys) in response to speech and speech-like stimuli (Kiang and Moxon, 1974; Kiang, 1975; Hashimoto et al., 1975). These single-unit recordings were shown to contain information regarding speech in both the average discharge rates and the temporal discharge patterns. For example, single-unit activity encodes rapid changes in amplitude and spectrum that are phonetically important (Delgutte, 1980), as well as information specific to phonetic feature categories, such as vowels (Delgutte & Kiang, 1984a), nasals (Deng & Geisler, 1987), and fricatives (Delgutte & Kiang, 1984b). Of particular relevance for the early encoding of consonants, studies demonstrated that at points of abrupt acoustic onset, such as the release burst for stop consonants, the instantaneous discharge rate of auditory-nerve neurons displayed a prominent peak followed by a gradual decay – the latter of which represents adaptation (Delgutte, 1980; Delgutte and Kiang, 1984c). Neural adaptation in the auditory nerve plays a key role in pointing to important spectrotemporal events in the speech signal and increasing the temporal precision for representing the onset transients that are prominent in many consonants (e.g. plosives and fricatives). These physiological studies have motivated several models of peripheral speech processing (Zilany & Bruce, 2006; Cooke et al., 1993) and led to research examining speech processing at subsequent pre-cortical stages such as the cochlear nucleus (Palmer et al., 1986; Kim et al., 1986; Blackburn and Sachs, 1990; Winter and Palmer, 1990) and brainstem (Russo et al., 2004). It should be noted that, although the studies mentioned here are
conducted on animals, evidence suggests that at the phonetic level, speech perception is supported by general auditory mechanisms that display adaptation towards certain classes of stimuli (Bregman, 1994). Furthermore, the organization of peripheral and brainstem auditory pathways appears strikingly similar across many mammalian species (Moore, 1987). As such, it is likely that examining single-unit responses in non-human mammals offers a great deal of insight into the early stages of human speech perception.

2.3.5 Human cortical encoding of speech

The functional neuroanatomy of speech in human cortex has been the subject of inquiry for well over a century, starting with the discovery that patients with lesions of the left STG experienced language comprehension disorders (Wernicke, 1874). Inspired by these initial findings, the Wernicke-Geschwind language model described the propagation of information along the arcuate fascicle fiber, which connected the two “language centers” of the brain (Wernicke’s and Broca’s areas), providing a basic account of the functional anatomy thought to support both speech comprehension and production (Geschwind 1970). In subsequent decades, the Wernicke-Geschwind model has been recognized as overly simplistic. Informed by evidence from human neuroimaging studies, current views propose that speech perception is supported by an information processing hierarchy in distributed regions of the temporal lobe, each supporting somewhat discrete (yet largely interdependent) aspects of linguistic encoding (Hickok & Poeppel, 2007; Rauschecker, 1998).

In detailing the cortical circuitry underlying speech perception, most current conceptions first propose that frequency and amplitude information from A1 is fed to the posterior STG, which performs a basic spectrotemporal analysis that encodes the most fundamental features of speech (Hickok & Poeppel, 2007). This view is supported by a vast body of neuroimaging studies (see Price, 2010) that show converging evidence that superior temporal areas support spectrotemporal and early phonological processing. Neurons in STG have been found to respond to speech over other sounds (Binder et al., 2000; Boatman et al., 1997), even in the absence of any linguistic task demands (Vouloumanos et al., 2001; Liebenthal et al., 2005), suggesting that they play a fundamental role in early cortical speech processing. Moreover, neuronal populations in STG demonstrate a broadly tuned frequency selectivity that is rapidly malleable (Nourski et al., 2012),
allowing spectrotemporally complex stimuli (e.g. phonemes) to be encoded by relatively local cortical populations.

Perhaps the most significant advances shaping our current understanding of speech representations in the temporal lobe come from research examining the high-gamma band (70-150Hz) of direct cortical surface, or electrocorticographic (ECoG), recordings (Mesgarani et al., 2014; Pasley et al., 2012; Mesgarani & Chang, 2012; Koshkoo et al., 2018; Chang et al., 2010; Leonard et al., 2016; Hullett et al., 2016). In particular, Mesgarani et al. (2014) recorded ECoG activity from listeners as they heard natural speech. When examining the mean responses evoked by individual phonemes within the running speech, the authors found distinct selectivity for phonetic features at different electrode sites across STG. That is, rather than being selective for individual phonemes, STG electrodes exhibited response selectivity to specific groups of phonemes. For example, while one electrode responded primarily to plosives (e.g. /p/, /t/, /k/), another electrode showed selective responses to sibilant fricatives (e.g. /s/, /sh/, /f/). Recent work has also characterized the spatial organization of human STG in response to speech, finding a robust anterior-posterior distribution of spectrotemporal modulation tuning (Hulett et al., 2016). That is, while pSTG is tuned for speech sounds with large temporal variations but relatively constant spectral energy, aSTG is tuned for sounds with slow temporal modulations but a high degree of spectral variation. Together, these findings demonstrate the crucial role the STG plays in encoding acoustic-phonetic speech properties. Interestingly, evidence from ECoG recordings also suggests that the STG might participate in higher-level processes such as those underlying categorical phoneme perception (Chang et al., 2010) and phonemic restoration (Leonard et al., 2016).

Other research has made use of scalp-based recordings; measuring auditory evoked potentials (Agung et al., 2006) or whole-brain EEG activity (Khalighinejad et al., 2017; Di Liberto et al., 2015; Lalor & Foxe, 2010) during speech listening. While EEG lacks the spatial resolution and superior SNR of ECoG, measurements are no longer anatomically restricted to a portion of the cortical surface and therefore may capture more distributed components of linguistic processing. Khalighinejad et al. (2017) examined differences in mean evoked EEG responses corresponding to each phoneme occurring within a natural speech stream. Several interesting results were
uncovered. Firstly, using simple F-statistics, the authors showed that the similarity of phoneme-evoked responses reflected whether or not the evoking phonemes belonged to the same manner of articulation category. For example, ERPs elicited in response to nasals (e.g. /m/ and /n/) were similar to one another, but distinct from that of vowels. Secondly, acoustic differences between phonemes were repeatedly represented in the cortical activity at three distinct time intervals centered at 120 ms, 230 ms and 400 ms after phoneme-onset. Lastly, it was found that different time intervals were responsible for jointly encoding speaker identity and phonetic information. Taken together, these results were thought to provide evidence for the dynamic transformation of low-level acoustic features to abstracted speech representations along the auditory pathway. Di Liberto et al. (2016) used a stimulus reconstruction approach to examine the relationship between continuous speech and the corresponding evoked neural activity. Interestingly, reconstruction accuracy was highest when continuous speech was represented by both a categorical labeling of phonetic features and low-level spectrotemporal information. Additionally, the neural activity became increasingly discriminative for phonetic feature categories at longer latencies – in line with categorical perception emerging as a result of hierarchical processing. Thus, complementary to ECoG research, the results of EEG-based experiments provide evidence of a hierarchical processing system whose successive stages extract increasingly abstract and categorical features of the acoustic input.

In general, speech-evoked activity shows a direction of information flow spreading from posterior to anterior temporal areas (Lerner et al., 2011; DeWitt et al., 2012; Leff et al., 2008). This so-called ventral stream for speech perception is supported by anatomical connections between the superior temporal plane and anteroventral temporal areas. After early phonetic feature representations emerge in STG, current theories view the ventral stream as the primary pathway for transforming acoustic sensory signals into abstract representations of phonemes, words and sentences (Hickok & Poeppel; 2007). STG is connected to the posterior superior temporal sulcus (pSTS), which encodes phonological-level features such as phonemes. In turn, pSTS is directly connected to the posterior middle temporal gyrus (pMTG) and inferior temporal sulcus (ITS), which are thought to be the “lexical interface” where abstract linguistic representations are held.
Particularly in the left-hemisphere, posterior temporal areas are connected to a distinct but concurrently activated network that is thought to be responsible for translating acoustic speech signals into articulatory representations (Hickok & Poeppel, 2007). This dorsal stream consists of a sensorimotor interface in the Sylvian fissure at the parieto-temporal boundary, which is then connected to an articulatory network in frontal areas such as posterior inferior frontal gyrus (pIFG) and premotor cortex. The division of speech processing into ventral and dorsal streams is not only of theoretic relevance. Many speech perception tasks (including that reported in the current thesis) involve only sub-lexical tasks, such as syllable discrimination. In contrast to paradigms that require recognition of running speech, these tasks presumably do not activate neural processing pathways that make contact with the mental lexicon, only requiring that the listener maintain an active sub-lexical representation of the heard syllable. As such, Hickok & Poeppel (2007) propose that sub-lexical tasks are supported to a greater extent by dorsal stream circuitry, while recognition tasks rely more on ventral networks.

In sum, human neuroimaging and electrophysiological recordings over recent decades have been instrumental in shaping our current understanding of the cortical organization of speech processing, suggesting a hierarchical system where increasingly abstract representations are extracted from the signal. It should be stressed, however, that no agreed upon model of speech perception currently exists and many of the views presented above continue to be challenged by new empirical evidence. Two particular holes in our current understanding come to mind. Firstly, research is yet to provide an adequate explanation of verbal short-term working memory, which would in turn provide a solution to the sequencing and temporal binding problems. Secondly, research is yet to provide compelling evidence of a “lexical interface” or “word area” in cortex. It thus remains to be seen how cortex retains information across the time-spans necessary for speech perception, and how the sensory representation of speech gains access to the mental lexicon needed to extract meaning. These, and many other unanswered questions loom large in the study of the neurobiology of language.
2.4 References


cortex revealed by simultaneous magnetic and electric recordings. Electroencephalography and clinical neurophysiology, 94(1), 26-40.


Chapter 3: General Methodology

In this chapter, details relating to the experimental apparatus and analytical methods will be described. Only those methodological points that are not fully covered in the respective methods sections of subsequent chapters will be discussed here. The reader is referred to those subsequent chapters for methodological descriptions that specifically pertain to the approaches used in each corresponding study.

3.1 Apparatus

Electroencephalography (EEG) and Magnetoencephalography (MEG) are noninvasive techniques for recording neuronal activity from the human cerebral cortex. The outer layer of the cerebral cortex contains around $10^{10}$ neurons that transmit information via a signaling network consisting of $10^{14}$ interconnected synapses. When external-sensory or endogenous signals are processed by the brain, a portion of this vast cortical network is activated. In the dendrites of individual neurons, a primary intracellular current is generated during synaptic transmission due to the movement of ions along their chemical gradients. Simultaneously, a passive extracellular current is established in the surrounding cortical medium completing a loop of ionic flow. Both EEG and MEG signals originate from the net current generated by thousands (at least 50,000) of such neurons operating in concert. For example, pyramidal neurons have a palisade arrangement with dendritic axes parallel to one another and perpendicular to the cortical surface, such that their electric and magnetic fields are recordable from distances beyond the scalp (da Silva, 2013). Computation of the resultant electric field is achieved using Maxwell’s equations, and application of the right-hand rule determines the mutually orthogonal magnetic field (Hämäläinen et al., 1993).

The neuromagnetic signals generated by cortical neuronal ensembles may be as small as 10 ferroteslas (fT) and as large as 1 picotesla (pT), with typical values ranging between 50—500 fT. Such signals are considerably smaller than the surrounding ambient magnetic field, and their measurement thus poses a significant engineering challenge. Detection of the weak neuromagnetic flux is noninvasively achieved using an array of magnetometers situated in a
helium dewar, which, in-turn, is housed within a magnetically shielded room. The most commonly used magnetometers are superconducting quantum interference device gradiometers (SQUIDs; Zimmerman et al., 1970). The reader is referred elsewhere for an extensive description of SQUIDs (Ryhänen et al., 1989). However, one important design feature worth mentioning is the flux-transformer within SQUID units. The system on which experiments in the current thesis were conducted uses coaxial type first-order gradiometers, in which the transformer consists of a lower pickup coil and a compensation coil. The two coils are equal in area but wound in opposition. Using such a coil configuration means that the system is sensitive to inhomogeneous field changes but remains relatively unresponsive to spatially uniform changes in the background. As a result, the unwanted effects of external magnetic noise are significantly reduced.

In addition to being non-invasive, a major advantage offered by both EEG and MEG is the superior temporal resolution (less than 1 ms) relative to alternative neuroimaging approaches such as functional magnetic resonance imaging (fMRI) or positron emission tomography (PET). For determining the neural source of activity within the cortex, MEG outperforms EEG. This is because the scalp-measured electric field is distorted by inhomogeneity in the skull, making accurate determination of source area from EEG measurements difficult. In contrast, the currents in the skull and scalp are weak contributors to the magnetic field. With MEG under optimal circumstances, sources activated in the cerebral cortex can be spatially discriminated with a resolution of 2-3mm, provided assumptions can be made such that a unique solution to the inverse problem exists. Briefly, the inverse problem refers to the mathematical finding that no unique solution exists to the location of source-currents responsible for generating an externally measured electromagnetic field (Von Helmholtz, 1853). Instead, in order to overcome the inverse problem and estimate a field source, researchers must implement source models such as equivalent current dipoles (ECD), in which a least squares solution between theoretic and measured field patterns is found (Hämäläinen et al., 1993). As this thesis is more concerned with when stimulus-related information is held in cortex during the time-course of sensory processing, and less with where, no further description of source modelling will be provided here.
The first two studies of this thesis, presented in chapters 4 and 5 respectively, recorded neural activity on human subjects using the MEG system housed at the Australian Research Council Centre for Excellence in Cognition and its Disorders on the campus of Macquarie University (Sydney, Australia). The whole-head MEG system (Model PQ1160R-N2; KIT, Kanazawa, Japan) consists of 160 coaxial first-order gradiometers with a 50 mm baseline, and a magnetic field resolution of 4 fT/Hz or 0.8 (fT/cm)/Hz. The reader is referred to Kado et al. (1999) and Uehara et al. (2003) for a detailed description of the current MEG system.

3.2 Data analysis

A broad overview of the analysis pipeline is schematically illustrated in figure 14. While each analytical step displayed in the figure will be described in detail within this chapter, the reader may benefit from a brief description at this stage. First, a set of stimuli (a collection of different musical pitches in chapters 4 & 5, or a set of consonant speech tokens in chapter 6) are presented to listeners whilst their EEG/MEG activity is recorded. Neural data are first pre-processed (dimensionality reduction & de-noising). Next, we use discriminant classifiers that attempt to "decode" (i.e. predict) the stimulus category listeners heard from the corresponding neural activity. We assume that the linear decoding schemes currently implemented are effective in estimating the degree to which the neuronal population codes of different stimulus classes differ. As such, the accuracy with which two stimulus classes can be decoded provides a quantitative measure of their cortical dissimilarity. Parallel to this neural analysis, several perceptual, sensory and acoustic models are developed that attempt to explain neural dissimilarities. Finally, stimulus models and empirical neural observations are compared with one another within the framework of representational similarity analysis (Kriegeskorte, 2008). The reader is referred to the methods sections in chapters 4-6 for details regarding the stimulus, experimental design and recording sessions. Here, we turn our attention to more specifics of the analysis, describing the individual steps comprising the EEG and MEG analysis pipeline below.
Figure 14. A schematic overview of the analysis pipeline. A set of experimental auditory stimuli are presented to listeners while recording their corresponding E/MEG activity. The goal of the neural analysis is to characterize cortical distinctions between stimuli by applying MVPA to the corresponding neural response patterns elicited. In parallel, several models are developed that attempt to account for the observed neural distinctions. Models predictions are compared with neural observations using the framework of representational similarity analysis.

3.2.1 Preprocessing (dimensionality reduction and de-noising)

The time-series data resulting from M/EEG recordings are high-dimensional, with as many dimensions as there are sensors/electrodes. As such, a characterization of the neural representation of a sensory stimulus is performed in a multivariate feature space. Here, we use multivariate pattern analysis (MVPA; Haxby, Connolly & Guntupalli, 2014), which distinguishes multivariate patterns of neural activity associated with different stimuli (see below). However, before this stage, the neural data must be pre-processed.

Two broad aims of pre-processing exist. Firstly, we seek to reduce the dimensionality of the dataset to the minimal number of dimensions necessary to sufficiently represent stimulus-related variance in the signal. Low-dimensionality representations are advantageous in several ways. They reduce the likelihood that learning models will over-fit (i.e. fit to noise) in the data during classification training (see below). Additionally, low-dimensional spaces boost classification performance by reducing the sparsity of the feature space. Secondly, we seek to “de-noise” the signal so as to retain variance of interest while diminishing or discarding non-signal related

Chapter 3: General Methodology. 81
variance caused by biological artifacts and system noise. Often, the techniques applied to achieve these two aims are mutually inclusive. That is, the variance partitioning techniques used to reduce dimensionality also effectively de-noise the data (and vice versa). One possible approach is to preserve the most informative channels and discard the rest (Hanke et al., 2009; De Martino et al., 2008). For example, Isik et al. (2014) select MEG channels that contain significant stimulus-related information using an ANOVA test. In contrast, the current studies of this thesis implemented unsupervised, data-driven approaches to reduce dimensionality and de-noise the data. These approaches are briefly described below.

**Principle Components Analysis (PCA)**

PCA is an orthogonal transformation that takes as its input the original sensor-channel data and outputs an equal number of linearly uncorrelated components (Jackson, 2005). The resultant principle components (PCs) are ordered according to their power (i.e. the amount of variance they each explain). The results of PCA can be understood both in terms of the component *scores* (the PC values) and their *loadings* (the linear weights that define the mapping between original sensor-space and PC-space). To achieve the orthogonal transformation, the current implementations of PCA used the singular-value decomposition algorithm (Abdi & Williams, 2010), which was applied to the mean-removed multichannel data matrix. The application of PCA to MEG and EEG data has several distinct advantages. Firstly, by retaining components that account for the majority of variance in the data (we imposed a variance threshold of 99% in the studies reported in this thesis), we significantly reduce the dimensionality of the data (see chapters 4 & 5). Secondly, by treating time points as observations and sensor channels as features, PCA operates on data across an entire neural epoch while still preserving the temporal dimension. Thirdly, PCA is effective at partitioning artifacts due to eye-blinks, movement, and other noise sources into separate components (Grootswagers et al., 2017) which are then suppressed by classifiers because they presumably contain no discriminatory information between different stimulus-classes. Finally, because of the orthonormal constraint under which PCA operates, the different PCs are uncorrelated with one another, and they therefore enable the use of a class of classifiers that assume no feature covariance (e.g. naïve Bayes). As a result, classifiers are relatively easy to interpret, and are both simple and fast in their implementation.
Independent Components Analysis (ICA)

Similar to PCA, Independent Components Analysis (ICA) applies a transform to multichannel data that results in an equal number of components that are linear combinations of original sensor channels. The goal of ICA is to optimize a set of spatial filters to be maximally statistically independent from one another based on temporal covariance. Thus, if applied successfully, ICA will separate multiple independent sources of variance (such as oculomotor from stimulus-evoked responses) into different independent components (ICs). The current implementation of ICA used the logistic infomax algorithm to decompose the data into ICs (Bell & Sejnowski, 1995). Particularly in EEG research, ICA has been shown to be effective at isolating variance due to oculomotor responses (Hoffmann & Falkenstein, 2008). In chapter 6 of the current thesis, the EEG recording montage included vertical and horizontal electrooculogram (EOG) channels, which captured a large degree of the eye blink related variance that presumably also contaminates EEG channels. ICA was therefore applied to the full montage (including EOG channels), and ICs reflecting the oculomotor signal were identified (both visually and based on their correlation with the EOG channels). The results of such a process are displayed in figure 15 below. Oculomotor ICs were removed before data was projected back into EEG channel space to obtain cleaned EEG data.

Figure 15. (A) Vertical and horizontal EOG channel timeseries. (B) Two Independent components (ICs) during the same time-region as in (A). The correlation between these ICs and EOG channels is evident. (C) The scalp topology defined by the linear weights of the two respective ICs.
Denoising Source Separation (DSS)

To further improve the signal-to-noise ratio (SNR) of measurements, a blind source separation method known as denoising source separation (DSS) may be applied (de Cheveigné & Simon, 2008; Särelä & Valpola, 2005). Rather than sorting data into multiple independent sources, as ICA does, the goal of DSS is to create a spatial filter that partitions the neural data into components that reflect stimulus-related and stimulus-unrelated components. The spatial filter is created by first defining a bias function as the average neural response across trials. This bias function is applied to the data such that reproducible variance across trials is reinforced while stimulus unrelated variance is not. The filter weights are then found by submitting the output of the bias function to PCA. Finally, applying the resulting rotation matrix to the spatially normalized data results in a series of DSS components. These DSS time-series are ordered in terms of power, and the first N components that reflect the signal of interest can be retained, while the other noise-related components discarded. Figure 16 below displays the filter weights across the scalp associated with the first DSS component (i.e. the linear combination of sensor channels that explain the most reproducible variance), for one subject from the phoneme-decoding study presented in chapter 6. Note the strongest weights over bilateral temporal lobe areas associated with auditory processing.

Figure 16. Scalp topology of filter weights associated with the first DSS component for one subject from the phoneme-decoding study presented in chapter 6.
3.2.2 Multivariate pattern analysis (MVPA)

Having de-noised the dataset and reduced its dimensionality as far as possible, we now perform discriminative analyses that measure the distinctions between neural responses corresponding to different stimulus classes. Multivariate pattern analysis (MVPA) is an umbrella term referring to a set of neural analysis techniques that consider the activation of multiple variables (e.g. multiple E/MEG channels or components), rather than treating each independently and measuring their activation strengths one-by-one. MVPA is used to build what are referred to as “decoding” models in the sense that they output a prediction of stimulus category using neural data as the input. In contrast, “encoding” models perform the reverse process – mapping a stimulus onto the neural response domain (see Holdgraf et al., 2017 for a review of common encoding and decoding approaches in electrophysiology). Within cognitive neuroscience, decoding models are built by training classifiers to identify patterns within neuroimaging data and associating those patterns with corresponding experimental stimulus conditions (Haxby, Connolly, & Guntupalli, 2014). Though a diverse set of classifiers exist within the machine learning literature, within cognitive neuroscience, linear decoding models are commonly implemented because of their relative interpretability (Schwarzkopf & Rees, 2011) and the finding that neural decoding is often well approximated by linear transformations (Eliasmith & Anderson, 2003). Different classifiers make different underlying assumptions about the data, and therefore the choice of classifier can be critical. Commonly occurring within neuroscience applications of MVPA are correlation classifiers (Haxby et al., 2012) or discriminant classifiers (Carlson et al., 2003) such as linear support vector machines (SVM) and Linear Discriminant Analysis (LDA). Although SVM is a popular choice in fMRI due to its ability to deal with a large number of features, LDA has been shown to be optimal in many circumstances that involve human electrophysiological data with smaller feature spaces (< 50) than fMRI (Grootswagers et al., 2017). For this reason, the current thesis implements LDA for neural classification.

Similar to PCA, LDA attempts to find a linear combination of input variables or “features” (i.e. neural time-series) that best explain the data by solving the eigenvalue problem. However, whereas PCA does not differentiate between different classes of stimuli, LDA explicitly attempts to find linear feature combinations that best differentiate the activity corresponding to different
classes. To achieve this, several assumptions are first made (see Duda, Hart & Stork, 1973).

Firstly, given a set of features $\mathbf{x}$ that correspond to a known stimulus class $y$, LDA assumes that the conditional probability distribution functions corresponding to two different classes $P(\mathbf{x} | y = A)$ and $P(\mathbf{x} | y = B)$ are both normally distributed, with mean and covariance parameters $\left( \mu_A, \Sigma_A \right)$ and $\left( \mu_B, \Sigma_B \right)$. Secondly, LDA makes the assumption of homoscedasticity (i.e. that the covariances of classes are equal; $\Sigma_A = \Sigma_B$). Under these assumptions, the decision criterion $c$ for assigning given input features to a class can be expressed as a dot product:

$$\mathbf{w} \cdot \mathbf{x} > c$$

(1)

where the linear weight vector:

$$\mathbf{w} = \Sigma^{-1}(\mu_A - \mu_B)$$

(2)

Figure 17 illustrates this formulation in geometric terms (displayed in 2-dimensions for illustrative purposes). The set of feature values $\mathbf{x}$ will be assigned to class $y$ if they are located on a certain side of the separating hyperplane whose location is defined by the threshold criterion $c$ and lies perpendicular to the vector $\mathbf{w}$.

![Figure 17](image)

**Figure 17.** Geometric illustration of linear discriminant analysis, which finds a separating hyperplane that optimally distinguishes between two classes of data in their given feature space.

A schematic of the MVPA workflow is illustrated in figure 18. Classification was always performed in a pairwise fashion - training and testing on two given stimulus classes at a time. A
separate discriminative analysis was performed for decoding the neural activity corresponding to each pairwise combination of stimuli in the experimental set. For a given class, multiple repeat trials during experimental testing resulted in multiple neural observations, which were then divided into independent training and testing sets. In all studies reported in this thesis, generalization of classifiers was evaluated using k-fold cross validation to maximize use of the data (k = 10). Specifically, data corresponding to the two classes being classified were randomly assigned to 10 bins of equal size. Next, nine of the bins were pooled together and used to train the classifier, and the trials in the remaining bin were used to test the classifier. This procedure was repeated 10 times such that each bin was utilized for testing exactly once. The reported accuracy is the mean across all 10 cross-validated classification runs. Within a given cross-validation fold, decoding performance was assessed by calculating the balanced accuracy (the mean of the accuracy across both classes individually). Importantly, classifiers were constrained to uniform priors so as not to bias predictions in cases where the number of samples were unmatched across the two classes.

Figure 18. Schematic of the linear classification pipeline, demonstrated for the classifying the activity corresponding to two different phonemes: /k/ and /sh/. (A) Multiple repeat trials for each stimulus class are presented to listeners, and corresponding neural activity is divided into independent training and testing datasets. (B) Training a linear model entails finding the optimal decision boundary that separates the distributions corresponding to two stimulus classes in the
feature space of the dataset. (C) Independent un-labelled and shuffled test data is submitted to the classifier to assess the accuracy with which category labels can be recovered (i.e. predicted) by the model.

Because the sampling rate of E/MEG data offered a fine enough temporal resolution to probe the dynamics of auditory processing (data was sampled at either 100 Hz or 200 Hz), the classification procedure described above was applied with a sliding time window, whereby a new classifier was trained and tested at each time point in the neural epoch (figure 19). This enabled the emergence of stimulus related information in the cortex to be dynamically tracked. For example, the latency at which classification accuracy first rises above chance indicates when discriminable information between two stimulus classes first emerges, while the latency of peak accuracy indicates when the two stimulus representations are maximally distinct. Furthermore, using a non-singular window-size (i.e. > 1) provided classifiers with information about the temporal response structure. More specifically, each time point within a window was used as an additional input feature in classification. As a result, classification was not only based on spatial response information at a given time, but also on the local temporal structure of the response within the window.

**Figure 19.** (A) Multi-dimensional neural response data is classified (B) using a sliding time window, resulting in a decoding curve that tracks the emergence and temporal dynamics of discriminable stimulus-related information in cortex.
3.2.3 Representational Similarity Analysis

A primary goal of this thesis is to use the measured cortical response differences from MVPA to test perceptual, sensory, and computational models of auditory information processing. To achieve this, we use Representational Similarity Analysis (RSA; Kriegeskorte, Mur & Bandettini 2008; Kriegeskorte & Kievit, 2013; Nili et al., 2014). As described in the above section, MVPA was used to estimate a set of cortical response dissimilarities resulting from discriminating between the response patterns of every pairwise combination of classes in the experimental stimulus set. The geometry of this collective representational structure can be expressed as a diagonally symmetric representational dissimilarity matrix (RDM), where the different classes index the rows and columns of the RDM. Intuitively, the RDM behaves as a look-up table for assessing the extent to which two stimuli have distinct neuronal population codes. To visualize the geometry of the representational structure in a low-dimensional space, Multi-dimensional scaling (MDS; Borg & Groenen, 2005; Kruskal & Wish, 1978) may be applied to the RDM. MDS attempts to solve the problem of representing the distance structure between $n$ objects in a low-dimensional space. To achieve this, MDS computes a configuration of $n$ points which optimizes the goodness of fit (i.e. minimizes the “stress”) of the solution (Kruskal, 1964). As such, the MDS solution to a given RDM is one in which spatial distances between stimulus classes approximate their neural representational distance.

Next, candidate RDMs may be constructed in which each cell expresses a given model’s prediction regarding the dissimilarity between the two corresponding stimuli indexing that cell’s row and column. For example, considering the representational geometry of a set of tones, one prediction may be that the cortex represents each tone based on its fundamental frequency ($f_0$). A candidate RDM may therefore code the difference in $f_0$ between the twelve tones. Alternatively, each tone’s representation may be based on spectral features, in which case a candidate RDM would compute the distance between the corresponding spectrograms of tones. Yet another hypothesis may be that the brain represents each tone based on its higher-level harmonic function within a musical context, in which case the candidate RDM may use behavioural data to compute listeners’ perception of tonal-harmonic differences between tones. Each candidate RDM may then be correlated with the time-evolving empirical (neural) RDM. In this fashion, specific
hypotheses relating to the structure of the neural representation are tested in addition to probing
the dynamics of when such a structure emerges in the cortex.

An important feature of RSA is the framework it provides for evaluating the statistical
significance of a model’s power to explain observed representational structure. Before
correlating empirical and candidate RDMs, it is helpful to ascertain the expected range of
correlations in a quantitative data-driven manner. Put differently, low RDM correlations may be
due to either (1) a failure of the model to explain neural representational structure or (2) a large
degree of noise inherent in the neural measurement. Any assessment of the significance in a
model’s ability to explain observed structure must therefore be made in light of these two
potential sources of misfit. RSA deals with this issue by computing a “noise ceiling” (Nili et al.,
2014). The noise ceiling evaluates the inter-subject coherence in neural RDMs to estimate lower
and upper bounds on the expected correlation between a “true” model and the average neural
RDM. Accordingly, if the correlation between a candidate and empirical RDM fails to reach the
lower bound of the noise ceiling, then the model insufficiently explains neural structure.
However, if the correlation value closely tracks the noise ceiling, the model may be assumed to
provide an adequate fit to the data.

In sum, RSA provides a framework to evaluate each model’s capacity to predict the observed
neural representation at the level of dissimilarity, enabling a determination of which features
(e.g. perceptual, sensory etc.) the underlying neural code honors, and which it disregards. For
further information regarding specific analytical details and evaluation of statistical significance,
the reader is referred to the methods sections of each subsequent chapter.

3.3 References


separation and blind deconvolution. Neural computation, 7(6), 1129-1159.


10. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. Reviews of modern Physics, 65(2), 413.


Chapter 4: Experiment One

Decoding the neural representation of musical pitch from human brain activity

4.1 Abstract

In music, the perception of pitch is governed largely by its tonal function given the preceding harmonic structure of the music. While behavioral research has advanced our understanding of the perceptual representation of musical pitch, relatively little is known about its representational structure in the brain. Using Magnetoencephalography (MEG), we recorded evoked neural responses to different tones presented within a tonal context. Multivariate Pattern Analysis (MVPA) was applied to “decode” the stimulus that listeners heard based on the underlying neural activity. We then characterized the structure of the brain’s representation using decoding accuracy as a proxy for representational distance, and compared this structure to several well established perceptual and acoustic models. The observed neural representation was best accounted for by a model based on the Standard Tonal Hierarchy, whereby differences in the neural encoding of musical pitches correspond to their differences in perceived stability. By confirming that perceptual differences honor those in the underlying neuronal population coding, our results provide a crucial link in understanding the cognitive foundations of musical pitch across psychological and neural domains.
4.2 Introduction

Context is critical to perception. In music, two physically identical tones heard in different contexts may bear little resemblance. This distinction arises because the tonality or key of the musical context assigns a unique function to each pitch (1). The two tones, despite being acoustically identical, differ in their tonal function and are therefore perceptually distinct. More generally, tonality describes the tendency for pitch relationships to be oriented around a central pitch termed the tonic (2). This organization establishes a hierarchy of perceived stability amongst the pitch-classes. The tonic occupies the most stable position, and other classes vary in perceived stability depending on their harmonic relationship to the tonic. Krumhansl (1979) (3) investigated the mental representation of musical pitch by measuring the perceived similarity between different pitch-classes. From this work, a geometric model was derived that places each pitch-class on the surface of a cone. By making explicit the perceived relatedness between pitch-classes, the conical model forms a cornerstone in our understanding of the cognitive processing of musical pitch. The model predicts that in-key and out-of-key classes are distinct from one another, occupying two distant regions of the representational space. The in-key classes are situated near the apex of the cone and are thus perceptually similar to one another. Conversely, the out-of-key classes are dispersed around the basal end of the cone, and are therefore perceptually distant from one another, and from the in-key classes.

In the neural domain, studies have identified features of pitch-evoked cortical responses (4-11), and described the anatomical and functional brain regions implicated in the neural processing of tonal structure (12-17). Recording electroencephalographic (EEG) activity during melodic listening, Brattico et al. (2006) (4) found that out-of-key pitches elicited an early pre-attentive cortical negativity, suggesting that tonal properties of musical pitch are automatically processed in the cortex. Recording EEG from trained musicians, Krohn et al. (2007) (5) found that the amplitude of pitch-evoked response components were modulated based on the perceptual stability of the evoking pitch-class, suggesting a stored representation of hierarchical pitch structure in cortex. While these studies flag the presence of tonal-schematic processing relatively little is known about the explicit representational structure of musical pitch in the cortex.
To examine the relatedness of different musical keys in the brain, changes in fMRI activity have been measured as a musical passage modulates across multiple tonal regions (18). While the relatively slow fMRI responses are sensitive to the gradual accumulation of pitch-distributional information across a musical passage, evidence suggests that the cognitive basis of tonality arises from the perception of individual pitches given a tonal context (19). We therefore measured highly temporally-resolved responses to individual tones presented within a musical context. By examining distinctions in the brain’s response to various pitches of differing tonal function, the current study provided a neural analogue to prior psychological models and evaluated their specific predictions: how distinctly does the brain represent each pitch-class relative to one another, and well do these neural distinctions align with perceptual differences between musical pitches?

To answer these questions, we focused on a set of four pitch-classes whose harmonic and perceptual properties make them strong candidates for observing a clear representational structure in the brain (20-21). Two pitch-classes (the tonic and dominant) were highly stable scale notes within the prevailing key, while the other two pitch-classes (the minor 2nd and augmented 4th) were out-of-key and highly unstable. Each tone was presented within a tonal context to trained musicians (Figure 20). Magnetoencephalography (MEG) was used to measure the neural response patterns evoked by each class. We then used Multivariate Pattern Analysis (MVPA) to “decode” the pitch that listeners heard based on the underlying neural activity. Using decoding accuracy as a proxy for representational distance, we characterized the structure of the neural representation of musical pitch and compared this structure to several perceptual and acoustic candidate models. The observed neural representation was best accounted for by a model derived from the Standard Tonal Hierarchy (20), indicating that differences in the neural encoding of musical pitch correspond to differences in their perceived stability.
**Figure 20. Experimental Paradigm.** Each trial consisted of a four-chord tonal context followed by a single probe-tone. Stimuli were piano tones. Each chord and tone was 650ms in duration and a silent interval of 650ms separated the last chord and probe-tone. Contexts were either in the key of C major (top) or F# major (bottom). Subsequent probe-tones were C4, G4, F#4 or C#4. When the context was in the key of C major, the former two probe-tones were “in-key” (tonic & dominant), while the latter two probe-tones were “out-of-key” (augmented 4th & minor 2nd). When the context was in the key of F# major, this mapping reversed.

4.3 Results

4.3.1 Decoding pitch-class from neural activity

Our goal was to map a set of pitches to points in representational space based on the similarity of their neuronal population activity. To achieve this, we trained a machine-learning classifier to measure the neural discriminability between all pairwise combination of tones at each time-point in the MEG data. The accuracy with which the classifier could discriminate between neural
responses for a given pair of tones provided a measure of their neural representational distance; and the complete set of pairwise distances defined the geometry of the stimulus’ representational structure in the brain (22-23). This approach affords greater power over straightforward activation-based measures (such as ERPs) because it preserves the rich dimensionality of the measured cortical response patterns, allowing a more fine-grained characterization of neural activity. Moreover, the MEG data was sampled at 200Hz (see methods), and by applying the analysis to each time point, we were able to track the emergence of the brain’s representational structure across time.

Perceptually, in-key pitches sound stable with respect to the tonal context, while out-of-key pitches sound unstable (19, 21). Is this psychological distinction evident neurally? To answer this question, we sought to discriminate between responses corresponding to in-key tones from that of out-of-key tones. Figure 21A displays the time-varying decoding performance averaged across subjects, tonal contexts, and pairwise combinations of in-key/out-of-key tones. As expected, classification performance at onset (t=0) was at chance (50%) because stimulus-related information was yet to activate the cortex. However, by 150ms we could successfully discriminate between the neural activity corresponding to in-key and out-of-key pitches, with maxima in classification accuracy occurring at 350ms.

We next considered the brain’s distinction between the two in-key tones by evaluating the classifiers ability to discriminate between evoked responses corresponding to the tonic and dominant. To this end, we significantly decoded the tonic-dominant distinction from 150 ms onwards (figure 21B). In this case however, the peak accuracy was lower than when discriminating between in-key/out-of-key tones, suggesting that the in-key tones are more similar to each other than they are to out-of-key tones. Lastly, we attempted to decode activity corresponding to the two out-of-key tones (augmented 4th/minor 2nd). Here, decoding performance failed to rise above chance classification for any sustained period (figure 21C), suggesting that the neural distinction between the out-of-key tones is relatively weak. Indeed, when examining the time-averaged decoding performance (figure 21D), we found that the neural distinction between the out-of-key tones was significantly weaker than both the in-key/out-of-key (Z = 2.97, p = 0.006) and the tonic/dominant (Z = 2.41, p=0.03) distinctions.
Figure 21. Decoding pitch-class from MEG activity. Neural distinctions were probed at each time-point from -100ms to 1000ms relative to onset of each probe-tone. Performance is averaged across both tonal contexts (C major and F# major) and all subjects. Statistically significant time points are indicated by the black points underneath each decoding curve (p<0.01; Wilcoxon sign-rank test, corrected by controlling the false discovery rate). Shaded regions indicate standard errors. (A) Classification accuracy when discriminating between in-key and out-of-key pitch-classes. (B) Accuracy for decoding the two in-key classes (tonic/dominant). (C) Accuracy for decoding the two out-of-key classes (minor 2nd/augmented 4th). (D) Time-averaged decoding performance for each of the distinctions assessed in (A-C) over 250 – 600ms (the period of maximal context-related effects; see section 4.3.4). Significance is indicated by asterisks where * p <0.05; ** p<0.01 (Bonferroni corrected Wilcoxon sign-rank test).

4.3.2 Characterizing the neural representational structure of pitch-class
We next examined the brain’s collective representation for all four pitch-classes. The decoding performance for all pairwise combinations of tones defines a multidimensional geometrical structure in neural representational space, which can be summarized in a **Representational Dissimilarity Matrix** (RDM) (22, 24). The four different pitches are indexed along the rows and columns of the RDM, with each cell of the matrix indicating the measured neural dissimilarity.
Figure 22A shows the RDM for the data time-averaged across 250ms – 600ms, a period chosen based on the period of maximal context-related information in the brain (see section 4.3.4). We used multidimensional scaling (MDS) to visualize the structure of the RDM in two dimensions (figure 22B; see methods for details). In this arrangement, the distances between different pitch-classes directly relate to the measured neural dissimilarities, providing an intuitive and data-driven illustration of the representational structure. Most evident within this geometry is the brain’s distinction between the in-key and out-of-key tones, indicated by their locations to the right and left of the structure respectively. The geometry of the MDS solution also makes evident the neural distinction between the two in-key tones (tonic and dominant), while the two out-of-key classes are more similarly represented in the brain.

4.3.3 Comparing neural dissimilarities with perceptual and acoustic models

We next tested several candidate models that might explain the currently observed neural representation. Two models were derived from behavioral data; one based on inter-class Euclidean distances in the Conical model, thus reflecting the perceived similarity between pitch-classes (2) and another derived from the Standard Tonal Hierarchy (21), where dissimilarities correspond to the difference in perceived tonal stability between all pairs of pitches. To assess the extent to which sensory differences might account for current neural dissimilarities, we tested a model of Spectral Distance. Specifically, spectrograms for each probe-tone were first extracted using a biologically inspired model of the auditory periphery (25). We then computed the Euclidean distance between each pairwise combination of spectrograms. Next, we tested a candidate model based on the Spectral Overlap between the tonal context and each probe-tone. Although the context and probe-tones were separated by a 650 ms silent period in the current study, models of auditory short-term memory involve time constants of up to 4 seconds (26-27). Thus, the neural distinctions between probe-tones may have been driven by the sensory memory of the context. Spectral overlap was computed by finding the reciprocal of the Euclidean distance between spectrograms corresponding to the context and each probe-tone. Dissimilarities were then determined by calculating the difference in spectral overlap for each combination of context and probe-tone, averaged across the two tonal contexts. Finally, we evaluated the hypothesis that listeners may simply be representing each probe-tone in terms of its pitch-value alone (i.e. fundamental frequency). We therefore tested a model - termed Proximity - that was based solely
on the differences in semi-tone pitch-height between each pair of probe-tones. Each model makes explicit predictions about the structure of the brain’s representation of pitch-class and can be expressed as a candidate RDM (figure 22C).

Figure 22. Representational similarity analysis of pitch-class. (A) Neural dissimilarities summarized in a time-averaged Representational Dissimilarity Matrix (RDM) (B) Multidimensional scaling (MDS) applied to the time-averaged neural RDM provides an intuitive visualization of the representational structure of musical pitch in the brain. (C) The time-varying neural structure is indexed with a new RDM at each time-point and compared with 5 candidate models. (D) Time-varying correlation (Kendall’s Tau rank-order) between the observed neural structure and each of the candidate models. Significance is indicated by the points below the curves (p<0.05; randomization test; FDR corrected). Shaded regions indicate standard errors. (E) Time-averaging the neural-to-candidate correlations over 250 – 600 ms reveals that neural dissimilarities are significantly correlated with the Tonal Hierarchy (p<0.05; randomization test),
with the average correlation closely tracking the noise-ceiling (indicated by the shaded region). Bars indicate standard errors.

To evaluate each model’s capacity to predict the observed neural representation, we constructed a set of time-varying RDMs from the MEG decoding accuracies and compared this time-evolving representation to each model’s prediction using Representational Similarity Analysis (22, 28-29). Figure 22D shows the time varying correlation between each candidate RDM and the time varying neural RDMs, averaged across both tonal contexts. We found that neural dissimilarities were significantly correlated with the Tonal Hierarchy model during a period starting around 250ms, suggesting that the brain’s representation of the stimuli is best explained by differences in perceived stability between the four classes. Conversely, the all other RDMs provided little to no explanatory power. Tonal Hierarchy model correlations closely tracked the noise ceiling (24), indicating that the model provided a good fit despite inherent noise in the MEG data. Finally, to summarize the modelling, we computed average correlations over the putative period of maximal context-processing (250 - 600ms; see section 4.3.4). Again, the Tonal Hierarchy model was the best fit for the data (figure 22E); the average correlation is both significant and close to the noise ceiling.

4.3.4 Dissociating tonal schema from acoustics

In addition to modelling, we wished to verify that the currently observed representation of pitch-class was based upon the neural processing of tonal-schema rather than afferent acoustic information. Our experimental paradigm deliberately included two different tonal contexts to re-map the harmonic function of each given probe-tone (see figure 20). Specifically, tones that were in-key in C major were out-of-key in F# major (and vice-versa). This enabled us to examine the brain’s distinction between two acoustically identical tones that were preceded by different contexts (thus differing solely in their pitch-class). Decoding ‘across-context’ in this fashion allowed us to examine neural distinctions between stimuli with identical sensory features, thereby isolating the effects of tonality.

The results, averaged across the four probe-tones, are shown in Figure 23. The tonal function could clearly be decoded from the neural response patterns of two acoustically identical tones.
This distinction in the brain first emerged at 160ms, with performance peaking at 58% correct for a sustained period from 250 - 600ms. Crucially, performance was similar or better than when decoding pitch in the presence of acoustical differences (section 4.3.1). This result, in addition to the failure of spectral distance and proximity models to account for neural dissimilarities (Figure 22 D-E), suggests that the representation of pitch-class (Figure 22A) reflects primarily the processing of tonal schema in the brain.

![Graph](image)

**Figure 23. Decoding acoustically identical probe-tones.** Applying MVPA to decode the pitch-class from physically identical tones that were preceded by different tonal contexts.

Note that stimuli in Figure 23 were preceded by acoustically different contexts. As a result, there are indications of significant decoding before the onset of the probe-tone (i.e. at time-points less than zero). Crucially however, decoding accuracy is around chance at time zero. This result, in addition to the failure of the spectral overlap model to explain neural distinctions (Figure 22D-E), suggests that the silent period separating context and probe-tone (650ms) was sufficiently long enough to ensure that the sensory processing of the context did not influence the subsequent evoked response to probe-tone. As such, we believe that current neural distinctions between pitch-classes were driven by the processing of probe-tones within the tonal framework imparted by the context.
4.4 Discussion

We characterized the brain’s representation for a set of four pitches that differed in their tonal function by examining differences in their evoked MEG response patterns using multivariate pattern analysis (MVPA). Consistent with prior Event Related Potential (ERP) research examining the cortical processing of tonal structure, these neural distinctions first emerged from 150ms and were maximal from approximately 250 – 600ms (7, 30). We uncovered a neural representation that placed in-key and out-of-key classes in distant regions of representational space. The in-key classes were also distant from one another. Conversely, the out-of-key classes were closely related. Because the collective representation was based on dissimilarities between classes, we could directly compare the observed neural structure with prior behavioral and acoustic models without the need for a direct correspondence mapping (Figure 22C). Our principal finding was that neural dissimilarities were correlated with differences in the Standard Tonal Hierarchy (19, 21), suggesting that the difference in perceived stability between two pitch-classes can be conceptualized in terms of representational distance in the brain.

At a fundamental level, a common finding across behavioral and neural domains is that the representational distance between two tones forming an acoustically fixed interval varies depending on their respective classes within the tonal context. However, the nature of this variation is corroborated somewhat differently. The conical model (3) suggests that in-key tones are proximate and out-of-key tones are distant from one another, whereas the current results suggest the opposite. Though seemingly contradictory, the two findings may not be mutually exclusive. The current method models the ‘distinctiveness’ of pitch-classes based on the dissimilarity of their evoked neural features. Because the dominant has a distinctive harmonic function relative to the tonic, it is likely that trained listeners possess a schema with which to separate the two pitches. In contrast, the augmented 4th and minor 2nd have no such functional harmonic relationship. It seems, therefore, that in the absence of a clear schema, the brain’s representation of the two out-of-key pitches converges. In the conical model however, distance reflects the degree to which two tones are perceived to be musically associated. In such a framework, the tonic and dominant may indeed be related. Both classes have a high probability of joint occurrence within tonal passages or as constituent pitches of chords (31). Conversely,
out-of-key classes may be considered musically unrelated; the occurrence of an out-of-key tone is likely to be followed by a resolution to the nearest in-key (rather than another out-of-key) pitch-class. By distinguishing between these two concepts – the distinctiveness of two pitches as opposed to how well they musically fit – the current study helps to clarify the complex nature of representational distance.

It should also be noted that the current investigation presented single tones following a tonal context, whereas the conical model arose from similarity ratings of tone pairs presented within a tonal context. Nonetheless, the conical model provides a general account of the psychological similarity of pitch-classes. Therefore, it is valuable to assess the extent to which this model can be generalized beyond the specific behavioral methodology from which it arises. As such, we tested the hypothesis that perceptually dissimilar tones (as described by the conical model) also have dissimilar patterns of brain activation when presented in isolation.

In order to further understand the neural basis of tonality, future work should extend the stimulus set to include all twelve pitch-classes so that more precise modelling may begin to establish the optimal combination of perceptual, sensory and acoustic features that explain the brain’s response to tonal pitch. For example, it may be the case that when all pitch-classes are considered, a mixed model of sensory and schematic features may be needed to account for neural distinctions.

We assessed the ability of classifiers to discriminate between MEG responses evoked by tones differing in their tonal functions. Using classification accuracy as a measure of representational distance, we characterized a representation for a set of musical pitches and showed that their collective representational structure correlates with the respective differences in their perceived tonal stability. Our results provide a crucial link between musical pitch perception and the underlying neural activity from which it materializes. Music psychology has long held that the cognitive basis of tonality can be derived from the “Tonal Hierarchy” – the pattern of stability across different pitches within a musical context. The current results strengthen this notion by showing consistency in the relations between tones across neural and perceptual domains.
4.5 Materials and Methods

4.5.1 Experimental Design
Thirteen trained musicians with a minimum of 5 years of formal training (Mean = 11.8 years, SD = 3.0) were recruited from the Sydney Conservatorium of Music, the Australian Institute of Music and Macquarie University. The sample size was not pre-determined, but rather testing was terminated once trends in the decoding analyses displayed sufficient statistical power (see statistical analysis below). All subjects reported having no known hearing loss or brain abnormalities, and did not possess absolute pitch. The study was approved beforehand by the Human Research Ethics Committee at Macquarie University (REF 5201300804) and all methods were carried out in accordance with the stated guidelines. Informed consent was obtained prior to testing, after all experimental details and potential risks were explained.

Each trial consisted of a tonal context followed by a probe-tone. Contexts were either in the key of C major or F# major, and consisted of four major chords written in four-part harmony outlining an I-IV-V-I harmonic progression. Subsequent probe-tones were either C4 (261.6Hz), G4 (392.0Hz), F#4 (370.0Hz) or C#4 (277.2Hz). Within each key, two versions of the tonal context were presented: one in which chords contained tones that were also probe-tones (for example, in a C major context the C4 and G4 were both physically present in the preceding chords), and an alternate version in which these constituent tones were transposed an octave above or below their original position in the chords (i.e. a chord inversion). The inclusion of this alternate tonal context enabled us to assess the effect, if any, of the acoustic spectral overlap between context and probe. We found no significant difference in classification performance when the data were divided into the ‘f0-overlap’ and ‘no-f0-overlap’ conditions (Wilcoxon sign rank test, False Discovery Rate corrected p > 0.05). All reported results are therefore based on an analysis of trials grouped across the two versions of the tonal context.

Stimuli were piano tones recorded at 44.1kHz. Tones were sampled in Max/MSP (Cycling ’74, San Francisco, CA) to form chords and probe-tones that were 500ms in duration with an additional 150ms decay. Following the last chord of the context, a silent period equivalent to one beat in the tempo of the passage (650ms; roughly 92 beats per minute) was inserted. This
temporal separation was intended to prevent the sensory processing of the context from influencing the evoked response to probe-tones (see figure 23), while also maintaining metric regularity. Prior to testing, all probe-tones were passed through a time varying loudness model (32) to normalize for differences in perceived loudness. For each tone, the maximum short-term loudness (STL_{max}) was computed and normalized to the mean value of all four tones. Differences in STL_{max} between the four probe-tones did not exceed 3 phones.

Each participant's MEG data were collected in a single hour-long session. The session was subdivided into 8 testing blocks separated by one-minute breaks, during which subjects watched and listened to a movie. Each block consisted of 80 trials in a single tonal context (C major or F# major), with adjacent blocks alternating between the two keys. The two versions of each tonal context (with and without shared probe-tone pitches) were presented in randomized fashion with equal probability within a block. ERP studies indicate that increases in the probability of syntactically irregular trials results in decreased effect sizes (7). We therefore opted for an in-key to out-of-key presentation ratio of 6:4, resulting in a total of 192 in-key and 128 out-of-key observations within each tonal context. After each trial, participants responded as to whether the probe-tone was ‘in-key’ or ‘out-of-key’, registering their response by pressing one of two buttons. This was done to ensure participants were attending to the stimuli (33). Participants used their left and right thumbs to register the two different responses. The assignment of in-key/out-of-key to left/right button was interchanged every two blocks to control for the effect, if any, of motor activity on the measured neural responses. Once the response was registered, inter-trial-intervals were randomly jittered between 0.5 – 1 seconds. Before testing, subjects completed a training session consisting of 20 trials with an identical behavioral task to that of the MEG recording session. Feedback was provided after each training trial and the experimenter ensured that subjects could perform the task (using a threshold of ≥90% correct) before proceeding to the MEG recording session. No trial-by-trial feedback was provided during the MEG recording; however, subjects were informed of their accuracy after each block. On average, subjects responded correctly on 96% of the trials (SD = 5.9%).
4.5.2 Apparatus
Data were collected with a whole-head MEG system (Model PQ1160R-N2; KIT, Kanazawa, Japan) consisting of 160 coaxial first-order gradiometers with a 50 mm baseline (34-35). Prior to recording, five marker coils were placed on the participant's head, their positions were registered and the participant's head shape were measured with a pen digitizer (Polhemus Fastrack, Colchester, VT, USA). MEG data was bandpass filtered online from 0.1 – 200Hz using first-order RC filters and digitized at 1000Hz. Participants were in a supine position in the scanner and were instructed to direct their gaze at a fixation cross. Both the fixation cross and experimental instructions were projected by an InFocus IN5108 LCD back projection system (InFocus, Portland, Oregon, USA) to a screen located above the participant at a viewing distance of 113 cm. Sound stimuli were delivered via Etymonic ER-30 insert headphones at a sampling frequency of 44.1kHz.

4.5.3 MEG Pre-processing
All pre-processing was performed in MATLAB. Data was epoched from 0.1s before to 1s after onset of probe-tones. To improve the signal to noise ratio while still retaining temporal resolution, the MEG data was downsampling to 200Hz with a low-pass Chebyshev Type 1 filter. Applying Principal Components Analysis (PCA) has been found to be an efficient pre-processing step for optimizing (or near-optimizing) data for MEG decoding analyses. In a single step PCA reduces the dimensionality of the data, and obviates the need for additional artefact rejection or de-noising procedures, as the classifiers can “learn” to suppress nuisance variables isolated by PCA, e.g. eye-blinks and environmental noise (36). In the present study, PCA was applied to each participant’s dataset of [640 trials x 160 channels x 1100ms] and the first n components accounting for 99% of the variance for each subject were retained for the decoding analysis. On average, PCA reduced the dimensionality of the dataset from 160 sensor channels to 26 principle components.

4.5.4 Time-series Decoding Analysis
Multivariate pattern analysis (MVPA) of MEG data was performed in MATLAB. For each set of pre-processed data, we used a naïve Bayes implementation of linear discriminate analysis (LDA) (37) to perform single-trial classification for each pairwise combination of pitch-classes at each
time-point. Generalization of the classifier was evaluated using k-fold cross validation with a 9:1 training to test ratio. In this procedure, the MEG data for all trials corresponding to the two classes being decoded were randomly assigned into 10 bins of equal size, with matched numbers of observations across the two classes in each bin. Nine of the bins were pooled to train the classifier, and the trials in the remaining bin were used to test the classifier. This procedure was repeated 10 times such that each trial was included in the test bin exactly once. Decoding was performed with a sliding time-window to assess the time-varying ability of classifiers to discriminate between neural activity corresponding to two given pitch-classes. To provide more observations in each classification run, we used a window size of 25ms and a step-size of 5ms. This meant that each classification run was based on data from the 5 most recent points in the timeseries. Because of the temporal ‘smearing’ associated with such windowing, the reported onset times for significant decoding are conservative estimates. Classification was also performed using various longer window-sizes (50ms and 100ms) and was found to produce an identical representational structure between stimuli.

4.5.5 Representational Similarity Analysis (RSA)
Decoding every pairwise combination of the four pitch-classes occurring within the two tonal contexts resulted in an 8x8 Representational Dissimilarity Matrix (RDM) for each subject at each time point. From this, we created a single 4 x 4 RDM (Figure 22A) that was averaged across subjects, the two tonal contexts, and time-points from 250-600ms (the period of maximal context decoding - see Figure 23). Multidimensional Scaling (MDS; Kruskal’s normalized stress criterion) was used to illustrate the structure of the RDM in two dimensions. MDS aims to spatially represent the RDM whilst preserving the original distances as much as possible. The loss function or stress of the solution indicates how faithfully MDS preserves the distances. Typically, the stress is minimized with higher-dimensional solutions. In our case however, the 2-D solution produced negligible stress (stress = 8.5819 x10^-7), indicating that the dimensionality was sufficient to visually depict the RDM. We constructed five model RDMs based on perceptual and acoustic properties that may account for the neural dissimilarities observed. The Conical RDM was based on the Euclidean distances between pitch-classes in Krumhansl (1979) (3). The Tonal Hierarchy RDM was based on the difference in stability ratings reported in Krumhansl & Kessler (1982) (21). Stimulus spectrograms were computed by passing the raw
audio through a model of the auditory periphery (25). The model consisted of three main stages: (1) a cochlear filter bank comprised of 128 asymmetric filters uniformly distributed along a logarithmic frequency axis, (2) a hair cell stage consisting of a low-pass filter and a nonlinear compression function, and (3) a lateral inhibitory network approximated by a first-order derivative along the tonotopic axis followed by a half-wave rectifier. We constructed an RDM of Spectral Distance between probe-tones by calculating the Euclidean distance between the respective spectrograms across all 128 frequency bands. For each context, a Spectral Overlap RDM was computed by calculating the differences in the Euclidean distance between the context spectrogram and each of the probe-tone spectrograms. The resultant RDM had similar patterns of dissimilarity for both tonal contexts (C major and F# major). For this reason, we averaged the Spectral Overlap RDM across both contexts for further analysis. Finally, the Proximity RDM was based on the semitone difference in pitch-height between probe-tones. Using the RSA framework (22, 24), we studied the brain’s emerging representation by comparing each model RDM with our empirical time-varying MEG RDM.

4.5.6 Statistical Analysis

All statistical analyses were performed on all subjects (N=13). Decoding performance is reported in terms of balanced accuracy (the mean of percent-correct in class A and percent-correct in class B). Time-series and time-averaged decoding performance was tested for significance using a two-sided Wilcoxon signed rank test. To correct for multiple comparisons, we controlled the false discovery rate (FDR) (38-39) with $\alpha = 0.05$. Correspondence between neural and model RDMs was assessed by computing Kendall’s Tau$_A$ (i.e., a rank-order correlation) (24) for each time point and each subject, producing a time-varying correlation between each model and MEG data. Significance of model correlations at each time-point was assessed using randomization testing. Briefly, the class labels on RDMs being compared were randomly re-assigned before computing correlation, and this process was repeated 10,000 times to define the null distribution at each time-point. Significant time-points were identified as those whose correlation values lay outside the 95% confidence intervals of the null distribution. Multiple comparisons were accounted for by controlling the false discovery rate ($\alpha = 0.05$). We used the ‘noise ceiling’ as a benchmark for testing model performance (24). The noise ceiling estimates the magnitude of the
expected correlation between the “true” model RDM and the empirical RDM given the noise inherent in the measurement.

4.6 References


34. Kado, H. *et al.* Magnetoencephalogram systems developed at KIT. *IEEE transactions on applied superconductivity, 9*(2), 4057-4062 (1999).


Chapter 5: Experiment Two

The dynamic emergence of tonal structure in human cortex revealed through decoding MEG responses to musical pitch.

5.1 Abstract

Tonal music is characterized by a hierarchical structuring of pitch. Although this hierarchy has been perceptually well documented, its neural basis remains poorly understood. The present study sought to determine whether the hierarchical distinctions between the different pitch-classes of Western Tonal Music are reflected in their neuronal population coding differences. Listeners were presented with each pitch-class within a tonal context whilst having their magnetoencephalographic (MEG) activity recorded. We used multivariate pattern analysis (MVPA), assessing the accuracy with which decoders could classify the MEG activity corresponding to different pitch-classes, in order to estimate their cortical dissimilarity. Decoding was performed using a sliding window to capture the moment-to-moment representational dissimilarity between pitch-classes, and time-evolving neural dissimilarities were then compared with the predictions of various sensory and perceptual models. While distinctions 100 ms after tone onset reflected differences in fundamental frequency, by 200 ms the representational structure observed in cortex mirrored the ‘tonal hierarchy’ of perceptual stability. Additionally, when transposing the neural dissimilarities to different keys, the collective structural differences between keys aligned with their distances on the ‘circle of fifths’, further verifying that the measure neural representational structure corroborates the perceptual structure of Western Tonal Music.
5.2 Introduction

In musical systems throughout the world, pitch is organized hierarchically (Lerdahl, 1992). Depending on the prevailing key or tonality of a musical passage, certain pitch-classes occur more frequently and occupy positions of melodic, harmonic and rhythmic prominence (Vos & Troost, 1989). Perception mirrors this compositional hierarchy in that those privileged pitch-classes also have greater stability – a quality reproducible both within and across listeners (Krumhansl & Shephard, 1979; Krumhansl & Kessler, 1982). For example, within the Western tonal context of C major, the first scale degree (C) is maximally stable and heads the hierarchy. This is followed by the fifth and third scale degrees (G and E respectively), the other scale tones (D, F, G, A, B), and finally the non-scale or “out-of-key” tones (C#, D#, F#, G#, A#). This collective structure is referred to as the Tonal Hierarchy (figure 24A).

Figure 24. (A) The tonal hierarchy based on listener’s ratings of perceived stability reported in Krumhansl & Kessler (1982). (B) The “circle of fifths” conveying the relatedness between the different major keys.
Despite its principal role as the organizing schema behind such a ubiquitous musical system, the neural substrates supporting the Tonal Hierarchy remain unknown. After primary auditory cortex extracts basic frequency information from the acoustic signal, a representation of complex pitch is thought to emerge in lateral auditory regions (Patterson et al., 2002; Zatorre et al., 1994; Griffiths et al., 1998; Patterson et al., 2001). How does this isolated sensory representation then acquire the perceived attributes of musical pitch? The surrounding musical context must be integrated, and cortical populations reflecting an internal knowledge of Western tonal structure must be recruited. The current study sheds light on this process by evaluating whether cortical populations respond to pitch in a fashion that precipitates the organization observed in the Tonal Hierarchy. Both lesion and neuroimaging studies have identified regions implicated in processing both melodic (Lee et al., 2011; Fedorenko et al., 2012) and harmonic (Klein & Zatorre, 2011; Foo et al., 2016) structure, while electrophysiological research has identified cortical response components sensitive to the hierarchical status of the evoking tone (Krohn et al., 2007; Brattico et al., 2006). More recently, Sankaran et al. (2018) showed that hierarchical properties of pitch can be decoded from multivariate patterns of Magnetoencephalographic (MEG) activity, even when the tones were acoustically identical but occurring in different contexts, providing a novel framework for understanding how the perceived structure of musical pitch may be mirrored in the cortex. Despite these advances, empirical work is yet to map the neural representational space of the twelve pitch-classes and explicitly test the predictions of perceptual & music-theoretic models. Specifically, do distinctions in the cortical patterns evoked by musical pitch honor those predicted by the Tonal Hierarchy? If so, what are the temporal dynamics with which afferent sensory representations interface with higher-level musical representations?

To answer these questions, we recorded MEG activity while listeners were presented with each pitch-class within a Western harmonic context. Multivariate Pattern Analysis (MVPA; Haxby et al., 2014) was used to decode pitch-class from the corresponding MEG activity. The accuracy with which decoders could discriminate between the response patterns of different pitch-classes provided an intuitive measure of their representational distance in the brain. Furthermore, as
MEG responses were highly time resolved (see methods), MVPA was applied using a sliding peristimulus time window in order to track the temporal evolution of these representations. Finally, measured neural dissimilarities between pitch-classes were compared with predictions from relevant perceptual, theoretical, and sensory models.

5.3 Results

Results are derived from MEG recordings during the presentation of twelve different “probe-tones” that spanned all pitch-classes within an octave range following a C major context (see methods). Discriminant classifiers attempted to decode the MEG activity of two different tones at each time-point in the neural epoch (from -100 ms to 1000 ms relative to onset), and the resulting curve of time-varying accuracy provided a dynamic estimate of the dissimilarity in their neuronal population codes. Applying this classification procedure to every pairwise combination of the twelve different tones, we characterized the dynamic representational structure of musical pitch in cortex.

We first assessed the average decoding performance across all pairwise combinations of pitch-class at each time sample in the neural epoch (figure 25A). Prior to onset of probe-tones (t = 0), accuracy was at chance (50%) as stimulus-related information was yet to activate cortex. However, significant distinctions between pitch-classes first emerged in neural activity approximately 100 ms after tone onsets, were maximal at 250 ms, and remained above chance for the remaining extent of the neural epoch.

Next, we studied the dissimilarity between specific tones whose acoustic or perceptual properties generated explicit predictions regarding their representational distance. Firstly, as tones acoustically differed from one another, we reasoned that their distinctions in cortex may be commensurate with their fundamental frequency ($f_0$) separation, which we term *pitch-height* (PH). Decoding performance was therefore examined for pairwise combinations of tones...
grouped based on whether their PH difference was small (1–4 semitones), medium (5–7 semitones), or large (8–11 semitones). A period from approximately 100 to 250 ms was found in which the above hypothesis held true (figure 2B). For example, cortical distinctions between tones that had large PH differences (blue curve) significantly exceeded those between tones that had small PH separation (red curve). Secondly, in addition to acoustic differences, tones differed in their perceived stability given the preceding musical context. We therefore hypothesized that distinctions in their cortical encoding may honor their perceptual differences, embodied by the Standard Tonal Hierarchy (STH) of stability. If so, MEG decoding performance would be greatest for tones located at opposite ends of the hierarchy, and poorest for tones that are hierarchically equal. In general, results confirmed this hypothesis (figure 2C). MEG responses to the most stable tone [C] were highly distinct from those of the unstable tones [F#, G#, D#, A#, C#] (green curve), but less discriminable from those of the second and third most stable tones [G and E respectively] (orange curve). Additionally, classifiers performed poorly when attempting to distinguish the neural activity of unstable tones from one another (purple curve). These results suggest that the extent to which the cortical activity elicited by two tones differ corresponds to the difference in their position within the STH. Unlike the earlier neural distinctions based on pitch-height, the correspondence between decoding accuracy and hierarchical distance only emerged approximately 200 ms after onset and persisted throughout the duration of the neural epoch.

Figure 25. Temporal decoding of tones from evoked MEG responses. The time axis in all plots are aligned to onset of tones. (A) Average classification accuracy for decoding all pairwise combinations of the twelve tones. (B) Average classification accuracy when decoding tone-pairs grouped based on their pitch-height separation: large (8-11 semitones; blue), medium (5-7 semitones; green), and small (1-4 semitones; red).
The neural distinctions imparted from MVPA suggest that two pitch-classes are representationally dissimilar in the cortex to the extent that their perceived stability differs. We sought to explicitly test this hypothesis within the framework of *representational similarity analysis* (RSA; Kriegeskorte, 2008). First, the set of accuracies resulting from decoding the MEG responses of all pairwise stimulus combinations were indexed in a time-varying *representational dissimilarity matrix* (RDM; figure 26A). For a given subject and time point, each cell of the diagonally symmetric RDM indexes the dissimilarity between the two pitch-classes corresponding to that cell’s row and column. Examining the mean inter-subject correlation (figure 26B) confirmed the structure of RDMs were consistent across listeners for the majority of post-onset time points in the neural epoch. Next, we evaluated the predictions of several models that attempt to explain the dissimilarity structure observed in the MEG-based RDMs. Each model was coded as a candidate RDM that makes explicit predictions about the expected dissimilarities between pitch-classes (figure 26C). One candidate RDM was based on the *Standard Tonal Hierarchy (STH)*, where distances between classes corresponded to their respective differences in perceived tonal stability, using the behavioral ratings reported in Krumhansl & Kessler (1982). To test the hypothesis that neural distinctions between pitch-classes were driven by differences in each tone’s fundamental frequency, we coded a *Pitch Height (PH)* candidate RDM based on semitone differences in absolute pitch values. We also
tested two sensory models: one based on the *Spectral Distance (SD)* between tones, and another based on their *Spectral Overlap (SO)* with the preceding musical context (see methods for details). Each listener’s neural RDM at every time point was compared to the four candidate RDMs using a rank-order correlation measure, resulting in four curves tracking neural-model correlation across time (figure 26D). In line with our hypothesis, observed cortical distinctions were significantly correlated with a candidate model of the Tonal Hierarchy over a broad temporal extent, beginning at 190 ms. Crucially, this correlation closely tracked the noise ceiling (Nili et al., 2014), indicating that the Tonal Hierarchy offered the maximal amount of explanatory power possible for any model given the measurement noise inherent in the MEG data (see methods). Interestingly, neural dissimilarities at earlier times in the course of pitch-processing (100 – 250 ms) were significantly predicted by pitch-height differences. These findings are consistent with dominant conceptions of melodic perception, in which fundamental frequency extraction occurs at earlier stages of auditory processing, before the higher-level analysis of musical structure (Koelsch, 2011). Interestingly, the transition from a neural representation reflecting acoustics to one reflecting the tonal hierarchy was not discrete. Instead, a period existed (from 190 – 250 ms) in which both models (STH and PH) were significantly correlated with neural RDMs, suggesting an intermediary period during which the brain holds a combined representation of a pitch’s fundamental frequency and tonal class. To more intuitively visualize these results, we first averaged neural-model correlations in three different time bins (figure 26E): a period before stimulus-specific information was present in cortical activity (-100 to 100 ms), an initial post-onset period during which the cortical structure was most strongly correlated with the pitch-height RDM (100 to 200 ms), and a later period during which it most strongly reflected the Tonal Hierarchy (200 to 1000 ms). The time-averaged neural RDMs corresponding to each of the three bins are displayed in the top panels of figure 26 F-H. Finally, nonmetric multidimensional scaling (MDS) was applied to each neural RDM in order to obtain a 2-dimensional solution that spatially illustrates the representational space (figure 26 F-H bottom panels). In general, the MDS solutions visually reflect the organization of pitch by height during the 100 – 200 ms window, and by perceived stability during the 200 –1000 ms window. For example, C and C# are proximate in pitch-height (separated by only one step) and are therefore closely located to one another in figure 26G. In contrast, they are situated at opposing ends of the continuum of perceived stability, and are correspondingly distant from one another in figure.
More generally, the organization of the MDS solution in figure 26H spatially illustrates many fundamental properties of the Tonal Hierarchy. For example, traversing the space from right to left reveals the structure of the hierarchy, with the most stable pitch-class (C) situated on the right side, closest to the next most stable classes (G and E) but distant from the unstable classes (F#, G#, D#, A#, C#) that are clustered in the lower left corner. Prior behavioral research has underscored the perceptual primacy of the hierarchical arrangement of musical pitch (Krumhansl & Kessler, 1982). Our findings provide evidence of its origins in the cortex.
Figure 26. Representational similarity analysis of pitch-class. (A) Neural representational similarity matrix (RDM) indexing measured cortical dissimilarities between pairs of pitch-
classes at each time point in the neural epoch. (B) The mean rank-order correlation between the RDMs of individual listeners (N=18). Significant time points are indicated underneath the curve (p<0.05; randomization test, FDR corrected). (C) Four different candidate RDMs based on models that attempt to explain neural dissimilarities. (D) Rank-order correlations between each model and neural RDMs at each time point. Shaded regions indicate standard errors across listeners. Significant time points are indicated by colored markers beneath curves (p<0.05; Wilcoxon sign-rank tests, FDR corrected). (E) For visualization purposes, neural-model RDM correlations were averaged across three different peristimulus time bins. (F-H) Average neural RDMs (top) and multidimensional scaling solutions (bottom) for the three regions in E. Colormaps indicate pitch-height (low to high) or perceived stability (unstable to stable) in G and H respectively.

Theoretical and empirical accounts suggest that distances between the 12 major keys of Western Tonal Music are conveyed by the “circle of fifths” (figure 24B), and that this structure emerges directly from the relationship between individual pitches (Krumhansl & Kessler, 1982). Thus, to further verify that the observed MEG distinctions between pitches corroborates the perceptual structure of Western Tonal Music, we used the MEG dissimilarities between individual tones to derive a measurement of inter-key distances in cortex. We utilized a procedure that constitutes the neural analogue of the behavioral paradigm in which listeners ratings of the perceived stability of individual tones is used to generate key distances. In order to preserve neural activity reflecting tonal-schematic processing while minimizing the more stimulus-driven representations of PH, we utilized MEG data from 250–1000 ms, during which only the STH model significantly predicted cortical dissimilarities (see above). First, neural RDMs across this window were averaged to obtain a single RDM for each subject. To convert this representation from one that is dissimilarity-based to one that is position-based, we applied MDS in 11 dimensions to each subject’s neural RDM, noting that n objects will always perfectly fit into n-1 dimensions (Kruskal & Wish, 1978). Next, we adapted the procedure formerly used to obtain perceptual key distances from individual tone ratings (Krumhansl & Kessler, 1982) to the neural domain. Namely, we reasoned that two keys were closely related to the extent that the neural representation of each constituent pitch-class was similar. To generate pitch representations...
within a new key, the original configuration of 12 pitch-classes in 11-D space (in the key of C major) was shifted the appropriate number of steps entailed by that modulation. For example, to translate the representational structure from C major to G major, the location of each tone was shifted 7 semitone steps such that “C” occupied the position formerly occupied by “G”, “C#” occupied that of “G#”, and the process repeated for all twelve classes. Finally, the dissimilarity between two keys was computed as the mean Euclidean distance across all twelve pitch-translations. Application of this procedure to all pairwise combinations of twelve major keys resulted in an inter-key RDM in which the rows and columns correspond to different musical keys and each cell coded the corresponding distance between the keys indexed by that cell’s row and column. The average inter-key RDM across subjects is displayed in figure 27A. For ease of visualization, RDM rows and columns are ordered such that musical keys progress in intervals of a fifth (i.e. 7 semitone steps). Consistent with the circle of fifths, the most proximate keys were separated by fifths (e.g. C major and G major), while keys separated by 6 steps were most distant (e.g. C major and F# major). Comparing the neural inter-key RDM with a candidate inter-key RDM based on the circle of fifths (figure 27B), we found this association to be significant (Kendall’s Tau_A = 0.26; p =0.002). As such, current results provide a correspondence between cortical, perceptual and music-theoretic measurements of key distance.

Figure 27. Inter-key relationships. (A) Neural RDM indicating the pairwise distances between the twelve different major keys. (B) A candidate RDM based on the “circle of fifths”. Note that RDM labels now correspond to different keys.
5.4 Discussion

We dynamically mapped the representational structure of Western musical pitch-classes in human cortex. This was achieved by recording the spatiotemporal patterns of magnetic activity evoked by each chromatic pitch presented within a musical context. Traversing the time-course of cognitive processing, we tracked how the representation evolved from one based on afferent sensory information to one based on internal schematic knowledge. Specifically, the structure of MEG dissimilarities between pitches were explained by their differences in fundamental frequency during an earlier period of processing, but their hierarchical musical properties from 200 ms onwards. Interestingly, the transition from a sensory to schematically driven representation was not discrete - taking place gradually over a 100 ms window. Future work should further elucidate the mechanistic computations involved in integrating current sensory input with internally stored tonal knowledge. A better understanding of this music-specific phenomenon may shed light on domain-general cognitive processes whereby complex percepts arise through the interaction between external sensory information and previously acquired structural knowledge held in long-term memory systems.

5.5 Methods

5.3.1 Participants
Eighteen subjects, each with a minimum of 5 years of formal music training (mean = 11.9 years), were recruited through the Sydney Conservatorium of Music and Macquarie University to partake in the study. All subjects reported having no known hearing loss or brain abnormalities and did not possess absolute pitch. The study was approved by the Human Research Ethics Committee at Macquarie University (REF 5201300804) and all methods were carried out in
accordance with the stated guidelines. Informed consent was obtained prior to testing, after all experimental details and potential risks were explained.

5.3.2 Apparatus
Data were collected with a whole-head MEG system (Model PQ1160R-N2; KIT, Kanazawa, Japan) consisting of 160 coaxial first-order gradiometers with a 50 mm baseline (Kado et al., 1999; Uehara et al., 2003). Prior to recording, each participant’s head shape was measured with a pen digitizer (Polhemus Fastrack, Colchester, VT, USA) and the positions of five marker coils on the surface of the scalp were registered. During recording, MEG data was bandpass filtered online from 0.1 – 200Hz using first-order RC filters and digitized at 1000Hz. Participants were in a supine position within a magnetically shielded room containing the MEG sensors. During experimental trials, they were instructed to direct their gaze at a fixation cross. Both the fixation cross and experimental instructions were projected by an InFocus IN5108 LCD back projection system (InFocus, Portland, Oregon, USA) to a screen located above the participant at a viewing distance of 113 cm. Sound stimuli were delivered via Etymonic ER-30 insert headphones at a sampling frequency of 44.1kHz.

5.3.3 Experiment Design
The experimental design was similar to that of a previous study from our lab (Sankaran et al., 2018). Subjects were presented with a series of trials while having their MEG activity recorded. Each trial consisted of a tonal context followed by a single tone (hereafter referred to as the probe-tone). The tonal context consisted of four major chords written in four-part harmony outlining an I-IV-V-I harmonic progression in the key of C major. The context and probe-tone were separated by a silent period equivalent to one beat (650ms, 92 bpm). This temporal separation was introduced in order to prevent the sensory processing of the context from contaminating evoked responses to probe-tones whilst maintaining metric regularity. The subsequent probe-tone was one of twelve possible notes spanning the chromatic range between F#3 (185Hz) and F4 (349Hz). This range was chosen to minimize the average pitch-distance between probe-tone and preceding context. On each trial, presentation of probe-tones was randomized but constrained to avoid repeated presentation across adjacent trials. To ensure...
participants were attending to stimuli (Louie et al., 2005), after each trial participants judged whether the probe-tone heard was ‘in-key’ or ‘out-of-key’, registering their response by pressing one of two buttons. Participants were instructed to use their left and right thumbs to register the two respective responses. The mapping of in-key/out-of-key to left/right button was interchanged every two blocks to control for the effect, if any, of motor activity on the measured neural responses. No trial-by-trial feedback was provided during the MEG recording. On average, subjects responded correctly on 78% of the trials (SD = 16.3%). All trials, including those with incorrect responses, were included in the subsequent neural analysis (VanRullen, 2011). Once the response was registered, inter-trial-intervals were randomly roved between 0.5 - 1 second. Before testing, subjects completed a training session consisting of 20 trials with an identical behavioral task to that of the MEG recording session. Feedback was provided after each training trial and the experimenter ensured that subjects could perform the task (using a threshold of ≥ 75% correct) before proceeding to the MEG recording session. Each participant's MEG data were collected in a single hour-long session. The total experiment comprised 672 trials, yielding 56 neural observations of each of the 12 probe-tones. Testing was divided into 8 blocks of 84 trials each, separated by breaks of approximately one-minute.

5.3.4 Stimuli

All stimuli were piano tones recorded at 44.1kHz. Individual tones were sampled in Max/MSP (Cycling ’74, San Francisco, CA) to construct chords and probe-tones that were 500ms in duration with an additional 150ms decay. Prior to testing, all probe-tones were passed through a time varying loudness model (Glasberg & Moore, 2002) to normalize for differences in perceived loudness. For each tone, the maximum short-term-loudness ($STL_{\text{max}}$) was computed and normalized to the mean value of all four tones. Differences in $STL_{\text{max}}$ between all probe-tones did not exceed 3 phones.

5.3.5 MEG pre-processing

Pre-processing of MEG data was performed in MATLAB. Data corresponding to each participant was first epoched from 100 ms before to 1000 ms after onset of probe-tones and then downsampled to 100 Hz with a low-pass Chebyshev Type 1 filter. Downsampling was found to
improve the signal-to-noise ratio while still retaining a suitable level of temporal resolution to examine the time course of neural pitch-processing. Next, Principal Components Analysis (PCA) was applied to the dataset of each participant. We implemented “spatial” PCA, whereby sensor channels comprised input features. We retained components that cumulatively explained 99% of the variance. On average, PCA reduced the dimensionality of datasets from 160 sensor channels to 28 principle components (SD = 5.4). PCA has been found to be an efficient pre-processing step for optimizing data for MEG decoding analyses (Grootswagers et al., 2017). In a single step, PCA reduces the dimensionality of the data, and obviates the need for additional artefact rejection or de-noising procedures, as the classifiers can “learn” to suppress nuisance variables isolated by PCA, e.g. eye-blinks and environmental noise.

5.3.6 Time-series classification

To measure the neural dissimilarity of two tones, we applied Multivariate pattern analysis (MVPA; Haxby, 2014), whereby a binary classifier learns features of the evoked neural activity that best distinguishes two different tones. MVPA was applied to each subject’s pre-processed dataset using MATLAB. Prior to classification, we averaged the MEG responses of 2 trials of the same pitch-class in order to boost the overall SNR of classification (Grootswagers et al., 2017). We used a naïve Bayes implementation of linear discriminate analysis (LDA; Hart et al., 2001) to perform classification for each pairwise combination of tones. Generalization of the classifier was evaluated using k-fold cross validation with a 9:1 training to test ratio. Specifically, MEG data corresponding to the two classes being classified were randomly assigned to 10 bins of equal size, with a balanced number of observations from each class in every bin. Next, nine of the bins were pooled together and used to train the classifier, and the trials in the remaining bin were used to test the classifier. This procedure was repeated 10 times such that each bin was utilized for testing exactly once. The reported accuracy is averaged across all 10 cross-validation folds. A sliding classification time-window was used on the neural time-series, resulting in a curve of classifier accuracy across time that tracks the dynamic emergence of stimulus-related information in the cortex. Each classification run at a given time point was performed on data across a 50ms window and classification runs traversed the neural epoch in 10ms steps. Importantly, the neural response at each adjacent time point within the 50 ms window mapped
onto a new dimension in the classification feature space. In this fashion, classifiers not only discriminated between responses based on their spatial activation patterns, but also their fine-grained temporal response structure (i.e. spatiotemporal classification).

5.3.7 Representational similarity analysis of pitch-class

The MEG responses corresponding to every pairwise combination of the twelve tones were classified using the MVPA method described above, resulting in a 12x12 diagonally symmetric Representational Dissimilarity Matrix (RDM) for every subject and time point. The mean inter-subject RDM correlation (figure 26B) was evaluated at each time sample by averaging the rank-order correlation (Kendall’s TauA) across all pairwise combinations of individual subjects’ RDMs (N=18). Significance was assessed through randomization testing. Briefly, on each randomization run (n=1000), the columns of subjects’ RDMs were randomly permuted, resulting in a noise floor. Significance was based on the true mean correlation rising above the 95% distribution of the noise floor, corrected based on the FDR. MEG RDMs for each subject and time point were compared with four candidate RDMs computed from various perceptual and sensory models of pitch. Candidate RDMs were as follows: (1) An RDM based on the Standard Tonal Hierarchy (STH) was constructed in which each cell coded the difference in perceived stability between the two corresponding tones using the major-profile ratings reported in Krumhansl & Kessler (1982). (2) In order to test the hypothesis that MEG dissimilarities reflected the difference in each tone’s fundamental frequency (f0), we constructed a Pitch-Height (PH) RDM, in which each cell corresponded to the semitone difference in f0 for the two tones in question. (3) To assess whether neural dissimilarities between tones reflected their sensory differences, a Spectral Distance (SD) RDM was constructed in which each cell corresponded to the Euclidean distance between the 128-channel stimulus spectrograms of two tones. Spectrograms for each tone were extracted by passing the raw audio through a biologically inspired model of the auditory periphery (Chi & Shamma, 2005). The model consisted of three main stages: (i) a cochlear filter bank comprised of 128 asymmetric filters equally distributed in log-frequency, (ii) a hair cell stage consisting of a low-pass filter and nonlinear compression function, and finally (iii) a lateral inhibitory network modelled as a first-order derivative along the tonotopic axis followed by a half-wave rectifier. (4) Lastly, although the tonal context and
probe-tone were separated by 650ms (see experiment design), models of auditory short-term memory involve time-constants of up to 4 seconds (Leman, 2000; Huron & Parncutt, 1993). Thus, it was possible that neural dissimilarities between tones were driven by the sensory memory of the context. To test this possibility, we constructed a *Spectral Overlap (SO) RDM*. First, the context stimulus waveform was passed through the auditory peripheral model described above in order to obtain a context spectrogram. Next, the Euclidean distance between the context and each probe-tone was calculated from their respective spectrograms. Each cell in the SO RDM was then coded as the difference in spectral distance between context and probe-tone for the two tones in question. Additional perceptual models were considered - for example that outlined in the “basic space” of the *Tonal Pitch Space Theory* (Lerdahl, 2004). However, the candidate RDMs arising from such models shared an identical rank-order structure to that of the STH and were therefore precluded from the analysis. Using the framework of Representational Similarity Analysis (RSA; Kriegeskorte, 2008), we studied the brain’s emerging representation by comparing each candidate RDM with the empirical time-varying MEG RDM (see statistical analysis below).

### 5.3.8 Representational similarity analysis of musical keys

Inter-key distances in the cortex were derived by adapting the analytical approach established in Krumhansl & Kessler (1982). First, neural RDMs from 250 – 1000 ms were averaged to obtain a single time-averaged neural RDM for each subject. In order to geometrically express the distances between different pitch-classes as points in high-dimensional space, we applied nonmetric MDS to the time-averaged RDMs of each subject. Because *n* objects will always fit perfectly into *n*-1 dimensions (Kruskal & Wish, 1978), MDS solutions were obtained in eleven dimensions. Accordingly, the stress of the solution was equal to zero in each instance, indicating that the MDS decomposition perfectly preserved distance information in the RDMs. To transpose the representational structure obtained using MDS into different keys, we replicated the technique used by Krumhansl & Kessler (1982), albeit in eleven dimensions instead of one. Specifically, the twelve points corresponding to each pitch-class were each shifted by the appropriate number of steps implicated by the transposition. For example, to transpose from C major to G major (seven semitone steps), the point in MDS space representing the first scale-degree was shifted from its original position of ‘C’ to that originally occupied by ‘G’. The point
originally occupied by ‘C#’ was shifted to that of ‘G#’, and this process was repeated for all twelve pitch-classes. The distance between two keys was then defined as the mean Euclidean distance between the original and new positions of all twelve tones. In this fashion, distances were computed between all twelve major keys, resulting in a neural inter-key RDM for each subject and time point. In the inter-key RDM (figure 27A), rows and columns correspond to different musical keys and each cell codes the corresponding distance between two keys. Finally, neural inter-key RDMs of each subject were correlated with a candidate inter-key RDM based on the “circle of fifths” (figure 27B).

5.3.9 Statistical Analysis

Classifier performance at each time point was evaluated in terms of balanced accuracy (i.e. the mean accuracy across individually evaluated classes). Group level significance (N = 18) was evaluated using two-sided Wilcoxon sign-rank tests (p<0.05). Multiple comparisons were corrected by controlling the false discovery rate (FDR; Benjamini & Yekutieli, 2001; Nichols, 2012) with α = 0.05. Correlations between neural and model RDMs were assessed by computing a rank-order correlation measure (Kendall’s Tau; Nili et al., 2014). We used the ‘noise ceiling’ as a benchmark for testing model performance (Nili et al., 2014). The noise ceiling uses inter-subject variance in RDMs to estimate the magnitude of the expected correlation between a “true” model RDM and the empirical RDM given measurement noise. Mean inter-subject RDM correlations were evaluated for significance at the 95% confidence level using randomization testing (FDR corrected). On each successive randomization run (n = 1000), subject RDM columns were randomly shuffled before being correlated with one another. Multidimensional Scaling (MDS) was performed using Kruskal’s normalized stress 1 criterion.

5.4 References


5.5 Appendix

This section contains supplementary analyses not included in the main body of the experimental manuscript.

Supplementary analysis 1: Effects of musical training

Figure S1 shows that the years of formal musical training subjects received was not significantly correlated with the average accuracy to which their MEG data could be decoded (averaged across all pairwise combinations of pitch-class from 250 – 1000 ms; p > 0.5).

![Figure S1](image)

**Figure S1.** Years of musical training (x-axis) plotted against average decoding accuracy (y-axis).

Supplementary analysis 2: Regression analysis to control for acoustic differences
As the twelve tones were all acoustically different from one another, we attempted to remove the variance in the MEG data driven by pitch-height and examine tonal class in isolation using multiple linear regression. Specifically, we assume that, for each subject $s$ at each time-point $t$, the MEG RDM is comprised of the sum of an RDM reflecting pitch-height (PH) plus a residual RDM.

$$RDM_{MEG}(s, t) = RDM_{MEG}^{PH}(s, t) + RDM_{MEG}^e(s, t)$$  \hspace{1cm} (1)$$

The pitch height-related component of MEG activity can be estimated using a linear regression model with the pitch-height model RDM as a predictor:

$$RDM_{MEG}^{prox}(s, t) = \beta_0(s, t) + \beta_1(s, t) \cdot RDM_{PH}$$  \hspace{1cm} (2)$$

Finally, we subtract the pitch-height MEG RDM from the original MEG RDM to obtain our residual MEG RDM. Figure S2 shows the correlation of this residual MEG RDM with model RDMs. As all the correlation curves remain the same as in the original analysis (see figure 26), we can be fairly confident that the influence of pitch-height was orthogonal to the variance related to processing tonal pitch-class. It should be stressed however that there are problems with this analysis – namely that assumptions of regression are violated within an RSA framework. For example, cells of the RDM are not independent from one another, as it is possible to reconstruct a “missing” cell using the surrounding RDM structure. Nevertheless, this analysis gives an indication that the presence of acoustic differences (i.e. pitch-height differences) did not significantly alter the correspondence between the tonal hierarchy RDM and MEG dissimilarities.
Figure S2. Correlations between model RDMs and the residual MEG RDM after controlling for variance predicted by pitch-height.
Chapter 6: Experiment Three

Tracking the dynamic representation of consonants from auditory periphery to cortex.

6.1 Abstract

In order to perceive meaningful speech, the auditory system must recognize different phonemes amidst a noisy and variable acoustic signal. To better understand the processing mechanisms underlying this ability, evoked cortical responses to different spoken consonants were measured with electroencephalography (EEG). Using Multivariate Pattern Analysis (MVPA), binary classifiers attempted to discriminate between the EEG activity evoked by two given consonants at each peri-stimulus time sample, providing a dynamic measure of their cortical dissimilarity. To examine the relationship between representations at the auditory periphery and cortex, MVPA was also applied to modelled auditory-nerve (AN) responses of consonants, and time-evolving AN-based and EEG-based dissimilarities were compared with one another. Cortical dissimilarities between consonants were commensurate with their articulatory distinctions, particularly their manner of articulation, and to a lesser extent, their voicing. Furthermore, cortical distinctions between consonants in two periods of activity, centered at 130ms and 400ms after onset, aligned with their peripheral dissimilarities in distinct onset and post-onset periods respectively. In relating speech representations across articulatory, peripheral and cortical domains, we further the understanding of crucial transformations in the auditory pathway underlying our ability to perceive speech.
6.2 Introduction

To discriminate between utterances of the words bark and park, the auditory system must differentially encode the phonemes /b/ and /p/. Furthermore, since either word may be spoken across a range of different styles, accents, articulatory contexts, rates and background noises, the auditory system must extract the acoustically invariant features of a given phoneme from a highly variable signal. Despite numerous advances in our understanding of the neural mechanisms underlying noise-robust speech perception, the signal processing transformations in the auditory pathway that convert a noisy, variable acoustic input into intelligible speech are not fully understood.

Employing closed-set recognition tasks, decades of psychophysical research has examined the perceived distinctions between different consonants for normal hearing (NH) and hearing-impaired (HI) listeners in both quiet and under degraded listening conditions (Miller & Nicely, 1955; Phatak et al., 2008; Phatak et al., 2009; Walden & Montgomery, 1975; Phillips et al., 2010; Wang et al., 1978; Bilger & Wang, 1976; Doyle et al., 1981). An analysis of the resulting confusion matrices reveals the existence of perceptually similar consonant sets, or ‘confusion groups’. These perceptual distinctions are often aligned with those based on articulatory descriptions, such as the voicing, manner, and place of articulation of the consonant (Phatak et al., 2008; Miller & Nicely, 1955; Allen, 1994). For example, under certain listening conditions unvoiced plosives such as (/p/, /t/, /k/) are often confusable with one another in NH and HI English-speaking listeners, as are certain fricative consonants (/ʃ/, /s/, /f/), whereas consonant pairs that span different manner of articulation categories (e.g. /p/ and /ʃ/) are perceptually distinct.

How these perceptual and articulatory distinctions are reflected in cortex is a question of significant theoretical and translational interest. Prior studies have used intracranial high-density electrode arrays, or electrocorticography (ECoG), to measure cortical activity during speech listening (Dichter et al., 2016; Mesgarani et al., 2014; Chang et al., 2010; see Leonard & Chang, 2014 for a review). Mesgarani et al. (2014) showed that different electrode sites across superior
temporal gyrus responded to different subsets of phonemes in a manner consistent with their articulatory properties (particularly their manner of articulation). Chang et al. (2010) uncovered neural evidence of categorical phoneme perception, whereby stimuli varying along an acoustic continuum produced neurally distinct categories. In leveraging the fine spatial and temporal resolution offered by ECoG, these studies provide compelling evidence for the neural basis of phoneme perception. However, as the recording procedure is highly invasive - involving specialized neurosurgical procedures - it remains unfeasible for investigating speech processing in many real-world conditions, sampling large listening populations, or easily embedding in Brain Computer Interface (BCI)-based listening technologies. Furthermore, ECoG recordings are typically anatomically restricted to a portion of the cortex, and may therefore fail to capture more distributed components of the neuronal population coding involved in speech (Huth et al., 2016).

Other studies have examined the encoding of speech using Electroencephalography (EEG) – a non-invasive recording technique (DiLiberto et al., 2015; Khalighinejad et al., 2017). In particular, Khalighinejad et al. (2017) examined differences in the event-related potentials (ERPs) corresponding to each phoneme within a continuous speech stream. In line with prior perceptual and neural research, phonemes with similar phonetic features evoked similar cortical responses (e.g. ERPs elicited by vowels were similar to one another, while dissimilar from that of plosive consonants). While these distinctions were resolved enough to delineate broad categories of phonemes, it remains to be seen whether EEG offers the finer resolution necessary to elucidate distinctions among consonants, which are most impacted by noisy listening conditions and hearing impairment.

The current study sought to robustly evaluate the cortical representation of consonants using EEG, and differed from prior work in several key aspects. Rather than continuous speech, listeners were presented with a closed set of consonants in vowel-consonant-vowel (VCV) format. This stimulus choice has several advantages. Firstly, current findings can be related most directly to the extensive body of psychophysical results employing closed set consonant recognition tasks. Secondly, by minimizing the influence of higher-order semantic and lexical effects that exist in running speech, VCVs provide more control over potential explanatory...
variables and facilitate an isolated measurement of the neural activity associated specifically with phoneme perception. Secondly, though VCVs provided more control than running speech, we also sought to measure an ecologically valid neural representation that honors the acoustic variability of real world listening. To manage this trade-off, individual VCV tokens within each consonant category contained a high degree of utterance variability (see methods for details). In addition to the stimulus design, the current study took a novel analytical approach; applying Multivariate Pattern Analysis (MVPA) to neural responses. Specifically, binary classifiers were trained and tested on their ability to discriminate between the EEG activity evoked by two different consonants. Within this framework, two consonants are cortically dissimilar to the extent that classifiers can accurately discriminate between their evoked activity. In contrast with the ERP-based methods used previously (Khalighinejad et al., 2017), this approach enabled a multivariate characterization of spatiotemporal neural features, without averaging across all sensor channels and repeat observations. Finally, to quantify distinctions between consonants at the auditory periphery, consonants were first modelled in terms of their auditory nerve (AN) synapse output, and MVPA was then applied to the AN responses. Examining the relationship between time evolving EEG-based and AN-based consonant representations provided a global view of how sensory information is transformed into higher-level cortical representations.

6.3 Method

6.3.1 Participants

Eighteen paid participants were recruited through the Starkey Hearing Research Center (Berkeley, CA). Data from three subjects were excluded from the analysis due to high impedance values or excessive line noise. The remaining fifteen subjects (10 female; mean = 26.1 years, SD = 10.4) had normal hearing as confirmed by clinical audiometry (ANSI/ASA S3.21-2004 R2009).
6.3.2 Stimuli

A set of consonants were chosen whose perceptual and articulatory properties were such that explicit hypotheses could be formed regarding their dissimilarity structure. Specifically, the set spanned several perceptual confusion groups (with multiple exemplars in each), and also clustered similarly when grouped by their articulatory properties (see table 1). Consonants were \{/b/, /d/, /g/, /p/, /t/, /k/, /m/, /n/, /s/, /ʃ/, /f/\} presented in the medial position of vowel-consonant-vowel (VCV) syllables. Each consonant was presented in five different vowel contexts, each with six different speaking styles that varied in either *rate* (fast, slow), *effort* (loud, soft) or *intonation* (statement, question). Additionally, each token contained three natural repetitions, resulting in 90 unique utterances of each consonant [5 vowels x 6 styles x 3 repetitions]. Stimuli were part of the Oldenburg Logatome Speech Corpus (Wesker et al., 2005) and were spoken by a single German female speaker. In pilot-testing, stimuli were selected such that the non-English background of the speaker did not introduce perceptible phonetic deviations from that of American-English. Speech tokens were normalized to 99% amplitude, low-pass filtered with 8 kHz cut-off frequency, and presented at 48 kHz with 16-bit resolution. Temporal boundaries of phonemes within each VCV utterance were labelled by the *Munich Automatic Segmentation System* (MAUS), which used a procedure similar to Hidden Markov Model forced alignment approaches (see Wesker et al., 2005 for details).
<table>
<thead>
<tr>
<th>Consonant</th>
<th>Voicing</th>
<th>Manner of Articulation</th>
<th>Place of Articulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>/b/</td>
<td>voiced</td>
<td>plosive</td>
<td>bilabial</td>
</tr>
<tr>
<td>/d/</td>
<td>voiced</td>
<td>plosive</td>
<td>alveolar</td>
</tr>
<tr>
<td>/g/</td>
<td>voiced</td>
<td>plosive</td>
<td>velar</td>
</tr>
<tr>
<td>/p/</td>
<td>unvoiced</td>
<td>plosive</td>
<td>bilabial</td>
</tr>
<tr>
<td>/t/</td>
<td>unvoiced</td>
<td>plosive</td>
<td>alveolar</td>
</tr>
<tr>
<td>/k/</td>
<td>unvoiced</td>
<td>plosive</td>
<td>velar</td>
</tr>
<tr>
<td>/f/</td>
<td>unvoiced</td>
<td>fricative</td>
<td>labiodental</td>
</tr>
<tr>
<td>/s/</td>
<td>unvoiced</td>
<td>fricative</td>
<td>alveolar</td>
</tr>
<tr>
<td>/ʃ/</td>
<td>unvoiced</td>
<td>fricative</td>
<td>postalveolar</td>
</tr>
<tr>
<td>/m/</td>
<td>voiced</td>
<td>nasal</td>
<td>bilabial</td>
</tr>
<tr>
<td>/n/</td>
<td>voiced</td>
<td>nasal</td>
<td>alveolar</td>
</tr>
</tbody>
</table>

Table 1. Consonants used in the current stimulus set and their corresponding articulatory features.

6.3.3 Apparatus

Testing took place in a sound attenuating booth. Subjects were seated in front of a monitor and keyboard. Stimuli were generated in MATLAB (Mathworks, Natick, MA) running on a PC. Audio was output at 48 kHz to ER2 insert headphones (Etymotic Research, Elk Grove Village, IL) at a constant SPL of 60 dB. EEG recordings were conducted on a BrainVision actiCHamp system (Brain Products, Munich, Germany) and digitized at 1000 Hz. The montage consisted of 96 electrode sites plus 4 EOG channels (2 horizontal, 2 vertical) and was referenced to a nose electrode. A trigger with a pulse width of 100 ms and amplitude of 0.1 V was used to ensure accurate presentation timing.
6.3.4 Procedure & design

Each unique utterance was presented three times, yielding 270 observations of each consonant and a total of 2970 VCV syllables over the entire experiment. Testing was divided into four blocks separated by five minute breaks and totaled approximately one hour. Consonant presentation was randomized but constrained to avoid repeat presentations of a given VCV token. Inter-stimulus-intervals were roved between 500 - 750 ms. During each block, subjects were instructed to keep their gaze focused on a fixation cross in the center of the screen. To ensure that subjects were attending to the speech tokens, 10% of the stimuli were followed by a visually presented word. Upon viewing the word, subjects responded as to whether the consonant contained in the middle of the word was congruent or incongruent with the VCV heard immediately prior by pressing one of two buttons. Subjects used one finger from each hand to register the responses. To ensure that neural activity associated with the motor response did not contaminate neural activity associated with acoustic-phonetic processing, words were only displayed on the screen 800 ms after VCV offset. As a further precaution, the mapping of true/false to response hand (left/right) was counterbalanced across the four blocks.

6.3.5 Analysis

Pre-processing

Pre-processing of EEG data was performed at the individual subject level and implemented in MATLAB using EEGLAB v.13. Data was first high-pass filtered using a zero-phase 3000-point FIR filter with a low-frequency edge of 1 Hz. Large movement artefacts were removed by visually inspecting the continuous time series and rejecting affected regions. Sensor channels were removed if: (1) they contained visually apparent non-biological signals such as 60 Hz line noise, (2) had a mean RMS value above 100 μV, or (3) an impedance above 40 kΩ. Continuous data was then subject to independent components analysis (ICA) using the logistic infomax algorithm to decompose the data into statistically independent components based on temporal covariance (Bell & Sejnowski, 1995). Independent Components that corresponded to ocular artefacts were identified based on correlation with EOG channels and inspection of scalp topology. These oculomotor components were removed before data was projected back into the
EEG sensor space. Next, each data channel was low-pass filtered with a high-frequency cut-off of 40Hz and normalized to have zero mean and unit variance. Continuous data was then epoched from -200 to 600 ms relative to the onset of the consonant. Epoched data was then downsampled to 200 Hz after filtering with an 8th order Chebyshev Type I low-pass filter to avoid aliasing. Lastly, epochs in which any channel had a maximum absolute value greater than 10 z-score normalized units were removed. On average, the combined processes of channel and epoch rejection led to 12% of the data being excluded from further analysis. To further boost the SNR, a blind source separation technique called Denoising Source Separation (DSS; deCheveigné & Simon, 2008; Särelä & Valpola, 2005) was applied to the epoched data. Using a criterion of response reproducibility across trials, DSS computes a set of linear spatial filters that attempts to partition evoked variance from measurement noise. Critically, these linear transformation matrices are computed on neural data pooled across all stimulus conditions so as not to artificially introduce differences that appear as experimental effects. This technique has been demonstrated to be effective in de-noising cortical responses to speech (Ding et al., 2016; Peelle et al., 2012; Ding & Simon, 2013; Ding & Simon, 2011). After decomposing the EEG channels, we retained the first DSS component (shown for one subject in Figure 28). This component constituted the linear combination of original sensor channels with the greatest reproducible power across all trials.
Figure 28. **Time-series corresponding to the first DSS component for one subject.** Responses corresponding to each consonant are averaged across all repeat observations for illustrative purposes only. Units are arbitrary as the data were normalized beforehand. In all time-series plots, \( t = 0 \) corresponds to the onset of the consonant. Line thickness reflects the \( \pm 1 \) standard error around the mean.

**Classification**

Classification was performed at the individual subject level using the first DSS component obtained from pre-processing. Conceptually, we sought to map the eleven consonant classes to points in representational space, where the distance between two classes corresponds to the dissimilarity in their cortical activity. To quantify this dissimilarity, we applied Multivariate Pattern Analysis (MVPA; Haxby et al., 2014), whereby a binary classifier learns features of the
neural activity that best distinguishes two different classes. Within this framework, the accuracy
with which classifiers can discriminate between the cortical activity of two given consonants
provides an intuitive measure of their representational distance in the brain (Kriegeskorte et al.,
2008). Importantly, classification was supervised such that neural responses to different
utterances of the same consonant were given the same class label. As a consequence, classifiers
naturally extracted the invariant neural features of a given consonant despite their acoustic
variability across utterances. Additionally, as classifiers only utilize neural information that aids
discrimination between different classes, evoked potentials arising due to the preceding vowels
(e.g. the P1-N1-P2 complex) should logically be disregarded because all exemplars were the
same in this respect. Classification was performed using a naïve-Bayes implementation of Linear
Discriminant Analysis (LDA; Duda et al., 2012), with 10-fold cross-validation to maximize use
of the data while maintaining independence of training and testing sets. Given the relatively high
temporal resolution of the data (see section II.E.1), neural responses were classified using a
sliding time window that proceeded in adjacent 5 ms steps (Grootswagers et al., 2017). This
resulted in a curve of classifier accuracy across time that tracked the dynamics of consonant-
related information in the cortex. We hypothesized that the information responsible for
distinguishing different consonants would not only be contained in the spatial patterns of neural
responses, but also in the temporally evolving structure of each evoked response. Thus, for two
given consonants, we attempted to discriminate between their neural responses in three different
classification analyses using sliding time-windows of 20 ms, 50 ms, and 100 ms respectively.
Importantly, the response at each adjacent time sample within a window mapped onto an
additional dimension in the classification feature space (e.g. a 100 ms window and a sampling
rate of 200 Hz resulted in a 20-dimensional feature space). In this fashion, with every increase in
window size classifiers were provided with additional temporal response structure, and by
comparing classification performance across the different window sizes, we were able to assess
the relative gains afforded by this additional information. To boost the classification SNR, the
neural responses of six trials within each consonant category were averaged prior to
classification. In order to minimize the bias of classifiers toward any one type of utterance (e.g. a
question rather than statement speaking style), averaging was constrained such that each of the
six trials pooled for averaging contained stimuli with a different utterance style and/or vowel
context.
The binary classification procedure described above was repeated for all pairwise combinations of consonants and for all subjects. To assess the significance of classification at the group level (figure 30 A-D), Wilcoxon sign-rank tests were carried out at each time point and corrected for multiple comparisons by controlling the false-discovery-rate (FDR; Benjamini & Yekutieli, 2001). At the level of a single subject (figure 30 E-F), significance was determined using randomization testing. Specifically, classification was performed with randomly permuted class labels and the procedure was repeated 1000 times. The performance of these randomized models defined a classification noise floor, and time points were deemed significant when the true classification performance exceeded the 95 percentiles of the randomized runs (FDR-corrected for multiple comparisons).

**Cortical representation**

Results were evaluated within the framework of *Representational Similarity Analysis* (Nili et al., 2014; Kriegeskorte et al., 2008). The classification performance for all pairwise combinations of consonants defined a representational structure that can be visualized as a dissimilarity matrix (DSM), averaged across all subjects and time points (figure 31A). The DSM is diagonally symmetric, with the different consonant classes indexing the rows and columns. Each cell indicates the dissimilarity of EEG responses corresponding to the two referenced classes. To spatially illustrate the structure of the cortical representation, multidimensional scaling (MDS) was applied to the values in the mean DSM using Kruskal's normalized stress 1 criterion to obtain a solution in 3 dimensions (figure 31B). To evaluate the extent to which these cortical distinctions between consonants reflected their acoustic-phonetic features, the EEG DSM was compared with three model DSMs. Specifically, each model DSM coded the binary (1/0) dissimilarity between consonants based on their membership within the three respective articulatory feature groups: (1) manner of articulation, (2) place of articulation and (3) voicing. Model DSMs reflecting a combination of these articulatory feature groups were also constructed by summing the relevant articulatory DSMs. To initially evaluate their predictive capacity, each model DSM was correlated with the time-averaged EEG DSM. Next, to examine the temporal dynamics of the relationship between articulatory features and cortical responses, those
articulatory features whose DSMs were significantly correlated with the average EEG DSM were also compared with time-varying EEG DSMs. Correlations were computed using a rank-order measure (Kendall's Tau; Nili et al., 2014). Additionally, to account for any covariance between model DSMs themselves, partial correlations were utilized, whereby the relationship between a given model and EEG DSM was assessed while controlling for other models. Significance of model correlations was assessed using Wilcoxon sign-rank tests. Time-averaged comparisons were Bonferroni corrected to control for multiple comparisons, while the time-evolving correlation analysis controlled the FDR.

Covariance across subjects

To compare the DSMs of individual subjects with one another, we used rank-order correlation (Kendall's TauA). This resulted in an nSubjects x nSubjects matrix in which each cell indicated the correlation between the two respective subjects’ DSMs. We then averaged one half of the diagonally symmetric matrix to obtain the mean inter-subject correlation. This process was repeated at each time-point to obtain a curve of average correlation across time (figure 32). Significance at each time-point was determined using randomization testing, whereby the row and column labels of individual DSMs were randomly shuffled before each repeat correlation (n = 1000). The 95% confidence intervals of these randomized runs determined the correlation noise floor and significance was corrected for multiple comparisons by controlling the FDR.

Auditory-nerve modelling

In order to probe transformations in consonant representations occurring between auditory periphery and cortex, stimuli were modelled in terms of their discharge rate at the output of the auditory-nerve (AN) synapse. This was achieved using a phenomenological model of the auditory periphery (Zilany & Bruce, 2006; Zilany et al., 2009; Zilany et al., 2014). The AN model has been rigorously validated against real physiological AN responses to both simple and complex stimuli, including tones, broadband noise, and speech-like sounds (see Heinz 2010 for a detailed review of the AN model). Model threshold tuning curves have been well fit to the characteristic frequency (CF; the frequency at which nerve fibers respond to the lowest sound...
dependent variation in bandwidth for normal-hearing cats. Many properties associated with nonlinear cochlear tuning are captured by the AN model, including compression, suppression, broadened tuning, and best-frequency shifts with increases in sound level. The stochastic nature of AN responses is accounted for by a nonhomogeneous Poisson process that was modified to include the effects of both absolute and relative refractory periods. Although the Zilany and Bruce (2006; 2007; 2014) model was chosen for this study, the results presented here do not depend on this choice and several other AN models exist that would be expected to produce similar results (see Lopez-Poveda, 2005 for a review).

We chose to model consonants at the AN synapse stage of auditory processing for several reasons. Firstly, onset activity followed by neural adaptation in the AN synapse plays a crucial role in the coding of transients, which are prominent in many consonants. For example, figure 29 demonstrates the pronounced coding of consonants at the output of the AN synapse relative to their response earlier in the auditory pathway along the basilar membrane. Secondly, as transformations of the speech signal occurring between acoustic-to-AN stages have already been well documented (Delgutte & Kiang, 1984a; Delgutte & Kiang, 1984b), we chose to focus instead on later transformations occurring between the output of the AN synapse and the cortex. VCVs were first resampled to 100kHz and windowed using a linear rise and fall time. The signal was then decomposed into 128 ERB-scale frequency bands between 100Hz and 8kHz. Neural adaptation at the AN-synapse stage was modelled using both exponential and power law dynamics (see Zilany et al., 2009 for more details). For each VCV utterance, spectrograms of AN-synapse activity (synaptograms) were computed by finding the energy in each frequency channel across a 10ms frame with a 50% overlap between adjacent frames.
Figure 29. Modelled peripheral responses for utterances of /apa/ (row 1) and /ata/ (row 2).

Cochleograms (column A) model the response along the basilar membrane while synaptograms (column B) model the discharge rate at the output of the auditory-nerve synapse. In the AN synapse, neural adaptation following vowel onset facilitates the coding of consonants relative to vowels, reflected in the pronounced response to consonants in column B relative to column A.

Auditory-nerve representation

We sought to derive distances between the different consonants based on the similarity of their respective AN responses at each time point, constituting the peripheral analogue of the EEG DSMs. To achieve this, MVPA was now applied to the peripheral data, whereby binary classifiers were both trained and tested on their ability to discriminate between the AN-responses
of a given consonant-pair. First, the 90 unique AN synaptograms corresponding to different utterances of the same consonant were treated as “observations”, and each frequency-channel x time synaptogram was epoched in an identical fashion to the EEG data (from -200ms to 600ms relative to consonant onset). We then applied principle components analysis (PCA) to reduce the dimensionality of AN datasets. PCA reduced the 128 frequency-channels to 18 components while retaining 99% of the variance. Pairwise classification of consonants was then performed in an identical fashion to the prior EEG analysis – using a linear discriminant classifier with 10-fold cross validation. Importantly, the temporal characteristics of the sliding classification window were identical to the earlier EEG analysis (100ms window moving in 5ms steps), ensuring that comparisons made between cortical and peripheral domains were temporally unbiased. To test significance of AN classification, a noise-floor was generated by repeatedly attempting to classify stochastically generated AN spike-trains (n = 500). The classification of all pairwise combinations of consonants based on their respective AN-responses resulted in peripheral DSMs at each time point (Figure 33B).

**Cortical-peripheral covariance**

To examine the relationship between stimulus representations in the cortex and periphery, EEG-based DSMs at every time-point were compared with AN-based DSMs at every time point using rank-order correlation (Kendall’s Tau_A). This resulted in a timepoints x timepoints matrix of correlation values indexing the degree of covariance between the two domains at various time lags (figure 34).

**6.4 Results**

Our primary goal was to determine whether distinctions between different consonants were observable in their corresponding evoked EEG activity. Visually examining the average neural responses for each consonant in one subject (figure 29), we found that mean responses were ill-fit for elucidating any distinctions between different consonants. Because of this, we adopted a multivariate machine-learning approach, whereby classifiers were trained and tested on their
ability to discriminate between responses based on fine-grained spatiotemporal patterns of activation (see methods).

On average, we found that classifiers could successfully discriminate between the neural activity evoked in response to two different consonants. Figure 30A displays the average classification accuracy for discriminating between all pairwise combinations of consonant-evoked responses across all participants. Before onset, accuracy is at chance (50%) because consonant-related information is yet to activate the cortex. However, classifier performance rises after onset and peaks within approximately 200 to 300 ms, suggesting that, on average, cortical distinctions between consonants were maximal during this period. We found that average classification performance improved monotonically with increases in the classification-window size (from 20 ms to 100 ms), underscoring the discriminant utility provided by additional temporal information (see section II.E.2). Critically, the differences in classification performance across individual consonant-pairs were highly correlated across the three different window sizes (all correlations had Pearson’s r ≥ 0.96; p < 0.001), indicating that changes in the window size did not alter the relative pattern of discriminability between consonants, only the absolute performance of the classifiers. For this reason, all subsequent analyses are based on a 100 ms classification window.

Next, we tested whether patterns of cortical similarity between consonants reflected their perceptual and articulatory groupings. We hypothesized that discriminability would be poorer (lower accuracy) when classifying EEG responses of two consonants within the same perceptual confusion group, but greater when classifying responses of consonant-pairs that span multiple groups. For our stimulus set, prior literature indicate the following confusion groups: unvoiced plosives (/p/, /t/, /k/), voiced plosives (/b/, /d/, /g/), fricatives (/s/, /ʃ/, /f/) and nasals (/m/, /n/) (Miller & Nicely, 1955; Allen, 2005; Phatak et al., 2008). The average classification performance, assessed separately for within and across group consonant pairs, is shown in figure 30B. At the group level, results were consistent with the hypothesis stated above. That is, performance was higher, on average, when discriminating between responses of consonant pairs belonging to different confusion groups. This was also evidenced at the level of individual consonant pairs. For example, EEG responses to /k/ and /s/ were highly dissimilar from one
another (figure 30C), but results suggest that the two nasals (/m/ and /n/) evoked more similar patterns of cortical activation (figure 30D). The same trends persisted at the single subject level (figure 30 E-F), suggesting that the above group-level results reflect veridical neural distinctions rather than noise. Nevertheless, the low signal-to-noise ratio of single-subject classification, as indicated by the 95 percentile bands, precludes a strong interpretation of current results at the resolution of an individual listener.

**Figure 30. Results of MVPA.** (A) Group mean classification accuracy across time, averaged over all pairwise combinations of consonants, using three different classification-window sizes. (B) Group mean classifier performance when discriminating between the neural responses of phoneme pairs that fall within the same acoustic-phonetic category (red) and across different categories (blue). The inset bars display performance averaged across all time-points where ***
signifies p<0.001. (C) Group-mean accuracy for discriminating between the neural activity of one across-category consonant-pair and (D) one within-category pair. Shaded regions in panels A-D indicate standard errors across participants. Significance is indicated by lines underneath curves (Wilcoxon sign-rank test (p<0.05); FDR corrected). (E-F) Single subject classifier performance assessed for the same respective phoneme-pairs as C-D. Shaded regions indicate the 95-percentiles of randomized classification runs with lines underneath indicating significance.

Using the accuracy of each pairwise classification as a measure of cortical dissimilarity between two given consonants, we visualized the collective representational structure in tabular form as a *dissimilarity matrix* (DSM) averaged across all time-points (figure 31A; hotter color corresponds to greater dissimilarity). Again, clustering of consonants in a manner consistent with perceptual groupings is evident in the DSM. For example, /k/ is similar to /t/ and /p/, but distant from /f/, /s/ and /ʃ/. We used multidimensional scaling (MDS) to more intuitively illustrate the structure of the DSM (figure 31B). MDS attempts to optimally preserve the structure of the DSM, therefore the distance between two consonants in the MDS solution can be construed as their representational distance in the cortex. Visual inspection of the MDS solution revealed that the nasals (/m/ and /n/) and sibilant fricatives (/s/ and /ʃ/) clustered together in distinct regions of representational space. To a lesser extent, voiced (/b/, /d/, /g/) and unvoiced (/p/, /t/, /k/) plosives were distinguished in MDS space. To explicitly test whether EEG dissimilarities between consonants reflected their articulatory features, we compared the geometric structure of the EEG DSM with several articulatory model DSMs (Figure 31 C-E) that coded dissimilarity as a binary [1/0] measure of a given consonant-pair’s membership within the [different/same] articulatory feature group, respectively. Correlations between the time-averaged EEG DSM and each model DSM (as well as all combinations) are shown in figure 31F. While the DSM coded for place of articulation (P) did not significantly predict cortical distinctions, DSMs coded for both manner of articulation (M) and voicing (V) - and any combination including at least one of the two - were significantly predictive of EEG DSMs. Given this, we next sought to examine the dynamics of the relationship between the significant articulatory feature DSMs (M/V) and time-evolving EEG DSMs. As the M and V model DSMs were also likely correlated with each other, we accounted for their covariance by running partial correlations with the EEG DSM at each time point (figure
Results indicated that M was the dominant predictor of cortical distinctions at early latencies in the time course (50 to 150 ms post consonant onset), whereas V was significantly predictive over a later broad period beginning at 200 ms and extending to 450 ms. Notably, after returning to noise floor, M DSMs were again correlated with EEG DSMs during a subsequent period such that both models (M & V) significantly predicted the cortical representation at relatively long latencies, centered 400 ms after consonant onset.

Figure 31. Cortical representation of consonants. (A) EEG dissimilarity matrix (DSM) detailing the representational structure of consonants in the cortex. Values are averaged across all time-points in the neural epoch. (B) 3-D Multidimensional scaling (MDS) solution to the values in the DSM. (C-E) Model DSMs each coding the binary dissimilarity of consonants based on

Chapter 6: Experiment Three. 155
three articulatory features: manner of articulation (M), place of articulation (P), and voicing (V).

(F) Correlations between the time-average EEG DSM and all articulatory model DSMs (including combinations). Bars indicate standard errors across participants. * p < 0.05, ** p < 0.01, *** p < 0.001 (Wilcoxon sign-rank test, Bonferroni corrected). (G) Correlation between time-evolving EEG DSMs and two model DSMs (Manner & Voicing). Shaded regions indicate standard errors across participants. Time points during which the mean correlation is significantly different from zero are indicated by colored markers underneath (Wilcoxon sign-rank test (p < 0.05); FDR corrected).

We next examined the extent to which the observed representational structure was consistent across listeners. Figure 32 displays the mean correlation between individual subject DSMs at each time-point. A strong inter-subject correlation suggests a relatively ‘universal’ representation across listeners. Before onset, we hypothesized that DSMs will be relatively uncorrelated, as consonant-related information is yet to be represented in cortex. However, across normal-hearing listeners, we expected a consistent representational structure to emerge following onset. Indeed, this appeared to be the case; whereas correlation failed to rise above the noise floor before onset, inter-subject correlation was significantly high from onset until approximately 300ms. Interestingly, we also found subsequent periods of relative consistency re-emerging across listeners (from 400 to 500ms and again, to a lesser extent, from 550 to 600ms).

![Figure 32. Inter-subject correlation of DSMs.](image)

Mean correlation between the DSMs of individual subjects at each time-point. The shaded region indicates the upper and lower bounds of the 95% confidence intervals resulting from randomization testing. Time points during which the average inter-subject correlation was greater than the shaded region (FDR corrected) are indicated by black lines.
Lastly, we examined how the time evolving representational structure of consonants in the cortex relates to their sensory representation at earlier stages of auditory processing. First, raw audio stimuli for each VCV were passed through a phenomenological model of the auditory periphery to estimate the discharge rate at the output of the auditory nerve (AN) synapse. For each VCV, the model produced a time-series of AN-synapse activity across 128 different frequency bands (see section II.E.5). We sought the most equitable comparison between modelled AN and observed EEG responses. Thus, in order to estimate AN-based dissimilarities between consonants, the same implementation of MVPA previously used to classify EEG responses was applied to AN-responses. To do this, the 90 modelled AN-responses corresponding to each unique utterance of a consonant were treated as “observations” on which to train and test a binary classifier.

The average accuracy for discriminating between the AN-responses of all pairwise consonant-combinations is shown in figure 33A. Performance rose to a maximal accuracy of approximately 90% for a sustained period from 50 to 200 ms. On average, performance began to rise 150 ms before consonant onset, indicating the presence of consonant-related information in the acoustic signal preceding the labelled onset time, likely due to anticipatory coarticulations (see section IV). The DSM in figure 33B illustrates the peripheral representation resulting from classifying all pairwise combinations of AN-responses, averaged across all time points. Though consonants are generally more resolved from one another than in their cortical counterpart, the peripheral DSM still shows clear similarity structure, with clustering along the diagonal intuitively suggesting that phonemes with similar acoustic-phonetic properties share similar sensory features. The MDS solution to the peripheral DSM (figure 33C) provides a visually intuitive description of this representational structure.
Figure 33. Modelled representation of consonants at the auditory periphery. (A) Average auditory nerve (AN) synapse classification accuracy across all pairwise combinations of consonants. The shaded region indicates the noise-floor determined by classifying stochastic AN spike trains multiple times (n=500). The black line beneath the curve indicates time-points for which classification accuracy is significantly above the noise-floor (FDR-corrected). (B) Time averaged peripheral DSM resulting from classification of each pairwise-combination of AN responses. (C) Three-dimensional MDS solution to the time-averaged peripheral DSM.

Cortical DSMs (based on classification of measured EEG responses) at each time point were then compared with peripheral DSMs (based on classification of modelled AN responses) at all time points. This produced a temporal covariance matrix (figure 34), in which each cell indicates the correlation between the peripheral and cortical DSMs at time-points indexed by that cell’s row and column respectively. Visual inspection clearly revealed that regions of high correlation were located in the lower right half of the matrix (below the diagonal). Intuitively, this suggests a latency, whereby stimulus features represented at the periphery take time to reach the cortex. Additionally, we found two distinct periods of high correlation. Firstly, peripheral DSMs during the 100 ms window approaching onset (reflecting a mixture of anticipatory coarticulations and
onset bursts) were correlated with cortical DSMs at a later period, from 50 to 200 ms after onset. Note that the above-chance classification of AN-responses prior to \( t = 0 \) (figure 33A) is consistent with the existence of consonant-specific information in peripheral DSMs during this pre-onset period. Secondly, we observed a later correlation, whereby the pattern of peripheral dissimilarities during a broad period from 0 to 300 ms were correlated with cortical dissimilarities around 400 ms. Importantly, the large vertical spread of this correlation is consistent with the relatively large variance in the physical duration of consonants across the entire stimulus set (median = 140 ms, inter-quartile range = 80 to 190 ms).

**Figure 34. Covariance matrix of cortical and peripheral representations.** Cortical and peripheral DSMs at each time point are indexed along the x and y-axis respectively. Each cell in the matrix indicates rank-order correlation between the two corresponding DSMs. EEG and AN DSMs during regions of interest are displayed to the sides of the x and y-axis respectively. Green
bars along axes indicate the median consonant duration across the entire stimulus set, with errors indicating the interquartile range (i.e. 25%-75%).

6.5 Discussion

A major aim of the current study was to evaluate whether distinctions between consonants could be observed based on their corresponding evoked EEG activity. To this end, MVPA was used to discriminate between the spatiotemporal patterns of neural activity elicited by consonants. In general, the extent to which classifiers could discriminate between the responses of two consonants was commensurate with their perceptual and articulatory distinctions (figure 30B; figure 31). In particular, consistent with prior cortical studies, we found that the differences in consonants’ manner of articulation was most strongly predictive of cortical distinctions (Mesgarani et al., 2014). Several pieces of evidence suggest that the observed EEG dissimilarities reflects veridical distinctions in the cortex rather than measurement noise. Firstly, the overall pattern of dissimilarities that emerged after physical onset of consonants was consistent across subjects (figure 32). Secondly, by generating a classification noise floor, we determined that the relative discriminability between consonants at the individual subject level generally corroborated group trends (figures 30 E-F). However, given the low SNR, no other assertions are currently made regarding the significance of individual subject classification.

Concerning the feasibility of EEG for probing real-time speech representations, results suggest further work aimed at increasing the SNR of measurements is necessary in order to elucidate phoneme-level distinctions in the individual listener.

To obtain a peripheral representation of the speech stimuli, MVPA was also applied to modelled auditory-nerve (AN) responses to VCVs. In assessing AN classification accuracy, we found that, on average, performance began to rise 150ms before consonant onset, indicating the presence of consonant-related information in the acoustic signal preceding the labelled onset (figure 33A). This is consistent with anticipatory coarticulations, whereby properties of the target consonant heavily modify the formant trajectories of the preceding context (Martin & Bunnell; 1981,
Martin & Bunnell, 1982). However, it is important to note that temporal boundaries between phonemes are often acoustically ill-defined and were determined here using a probabilistic model (Wesker, 2005). Given the potential ambiguity in transition boundaries, it seems likely that AN classification runs approaching $t = 0$ were influenced by the onset transients of consonants in addition to coarticulations. The combined influence of co-articulatory and onset features on AN classification during this “pre-onset” period cannot be dissociated, and the influence of each are likely to vary with properties of the given consonant-pair in question (for example, strong onset bursts are characteristic of unvoiced plosives, while voiced plosives may contain substantial coarticulations in addition to onset bursts).

Modelling and classifying AN responses to VCVs served our second aim, which was to examine the relationship between the same speech stimuli represented at the cortex and the peripheral auditory system. Importantly, dissimilarity analyses of EEG and AN responses were locked to the same temporal grid, enabling an unbiased comparison of peripheral and cortical representational structures as they temporally unfolded. In doing so, we uncovered two distinct cross-temporal regions in which cortical distinctions aligned with stimulus distinctions encoded at the periphery (figure 34). Firstly, the peripheral representation of co-articulatory and onset features was reflected in the cortex 100 to 150ms after onset. This time course is consistent with the latency of event-related potential components sensitive to acoustic onsets in speech. For example, the speech-evoked mismatch negativity (MMN), thought to cortically index pre-attentive speech discrimination, is typically elicited 100 to 250ms after the onset of a deviant speech sound (Martin et al., 2008). Additionally, the strong representation of onset features in the cortex is consistent with research showing that cortical coding is dominated by onset responses (Rabinowitz et al., 2013). The temporal covariance of EEG and AN structures also revealed a later period of correlation, whereby peripheral encoding in the 250 ms following onset was reflected in the cortex at 400 ms. It seems probable that the perception of a consonant is emergent, arising from the encoding of acoustic features throughout its time course, from its coarticulation with the preceding vowel to its offset. The cortical activity at 400 ms thus likely reflects the encoding of acoustic properties post-onset. Additionally, the relatively large vertical
spread of this correlational region suggests that the cortex only represents these differences after the gradual accumulation of sensory evidence.

Finally, a further scrutiny of the correspondence between EEG-based and AN-based representations highlights important cortical transformations occurring in the ascending auditory pathway. Firstly, if the cortex merely represented the acoustic properties of speech in a linear fashion, albeit with a latency in processing time, we would expect the temporal covariance analysis (figure 34) to feature a single unbroken ‘strip’ of correlation that was right-shifted with respect to the diagonal. Clearly this is not the case. Instead, the structure of EEG-AN correlations suggests that cortical processing is divided into two independent and temporally distinct onset and post-onset stages. Furthermore, the two correlational regions were not adjacent to one another in the cortical domain. Instead, a period from 200 to 350 ms exists in which EEG DSMs were not strongly correlated with any AN DSMs. Interestingly, this period coincides with the time at which the average EEG-classification curve was maximal (200 to 300 ms; figure 30A). Taken together, these results indicate that the time at which cortical distinctions between consonants were greatest did not coincide with a period in which those distinctions strongly reflected sensory differences. This may be evidence of further computations performed in cortex, beyond a linear analysis of acoustic input, thus enabling such distinctions to emerge. Future work should focus on more explicitly testing the hypotheses arising from these observations. Additionally, although the current study was limited by the number of observations within a given consonant category, future work should characterize the neural representational structure of both within-category (across various utterance types) and across-category distinctions within a unified representational space. By comparing time-evolving representational structures across peripheral and cortical domains, we lay the groundwork for better understanding the neural transformations involved in perceiving meaningful speech.

6.6 References


Chapter 7: Concluding remarks

This thesis has explored the neural representational space occupied by musical pitch and consonants as they manifest in the multivariate patterns of auditory evoked M/EEG activity. Broadly speaking, we sought to assess whether relatively straightforward linear decoding schemes were capable of recovering stimulus categories in the cortex. We clearly found this to be the case – linear discriminant classifiers could significantly decode the pitch-class or consonant-category from response patterns in evoked MEG and EEG activity respectively. Given the distinctions evident in cortical activity, a more targeted aim was to assess the extent to which the collective representational structure recovered by such decoding schemes was consistent with sensory and perceptual models. We were able to achieve this goal by way of Representational Similarity Analysis (Kriegeskorte et al., 2008; Nili et al., 2010), operating at the level of dissimilarity in order to directly compare representational structures across multiple domains.

A universal analytical feature of this thesis is its use of the temporal resolution of M/EEG to gain insight into the representational dynamics of auditory processing. For example, in chapter 4, we found that a model based on perceptual characteristics of musical pitch offered significant explanatory power of cortical data only 200 ms after tone-onset. In chapter 5, we refined this finding by showing that an earlier period of pitch processing existed prior to 200 ms in which the representational structure of musical pitch first reflected an acoustically-driven attribute – it’s fundamental frequency – before it’s perceptual tonal-hierarchical properties emerged later in the time course of neural processing. In chapter 6, temporal dynamics were considered in both peripheral and cortical representations, and the cross-temporal correlation of the two structures provided insight into the dynamics with which peripheral features of speech were represented in cortex. Thus, by harnessing the high temporal resolution of neural data, this thesis characterized the dynamics of representational spaces in cortex, and in turn deepened our understanding of the time-course with which external auditory information produces the complex and highly structured percepts of musical pitch and speech.
It is important to consider whether any of the findings presented in this thesis, particularly those relating to the neural representation of pitch measured in Chapters 4 and 5, may be readily explained by alternate mechanisms to those hypothesized. As mentioned in Chapter 2, the perceptual hierarchy of stability shares a direct correspondence to the “expectedness” of each tone. Perhaps the information being decoded that enables two tones to be separated in neural space is simply the difference in their “surprisal” or a similar dimension of expectation violation. However, the neural substrates of expectation violation cannot be regarded as mutually exclusive from the residual activity comprising the brain’s representation of tonal pitch. As sensory information ascends the processing hierarchy of the nervous system, activations reflect the increasing degrees of complexity with which stimulus representations are mapped onto object-level representations – including probabilistic representations that reflect how statistically likely the incoming sensory input is. Therefore, even though we may be able to partition out “surprisal” responses into statistically orthogonal components of activity, it is inherently part of the neural representation of tonal pitch structure. Secondly, during the experiment, listeners were instructed to make an explicit judgment of whether the heard tone was in or out of key. One potential explanation for current effects is that the neural activity reflecting such an explicit decision process was being decoded from the signal, rather than the representation of pitch per se. However, because this decision involved only two broad categories (in/out) and was uncorrelated with the perceptual dimensions being discriminated (every pairwise tone-combination), the decision related activity should run orthogonally to the dimensions of interest in shaping the geometric structure of neural RDMs, and therefore pose a minimal potential for confounding effects.

As the main findings of each experiment have already been discussed at length in previous chapters, the remainder of this section is dedicated to providing potential applications of the research and contextualizing results within the broader long-term goals of research that seeks to understand the neural mechanisms underlying music and speech perception.
7.1 Future implications and applications of cortical measurements of speech processing

A practical motivation for examining the representation of consonants at multiple stages of the auditory pathway lies in its utility for understanding how signal transformations in normal hearing (NH) listeners produce noise-robust speech perception, and for then determining where this signal gets corrupted in hearing-impaired (HI) listening. In order for both of these goals to be realized, two major tasks remain. Firstly, while the current results measure peripheral and cortical dissimilarity structures, the perceptual dissimilarity structure remains missing. This was primarily due to time and logistical constraints on testing. Thus, future work must complete the characterization of signal transformations in NH listeners by running psychophysical experiments to generate confusion matrices which can be directly compared with peripheral and cortical dissimilarities. Ideally all three representational stages (peripheral, cortical and perceptual) should be characterized using the same stimulus set and conducted on the same cohort of individuals, thus enabling direct and unbiased comparisons between each processing stage. Secondly, the experiment must be repeated in a cohort of HI listeners. Importantly, the peripheral model utilized in the current experiment (Zilany et al., 2009) contains parameters that can be customized to simulate outer and inner hair cell loss, thus enabling an individual’s hearing loss profile to be modelled. As displayed in figure 36 below, dissimilarity matrices (DSMs) can then be compared across various stages of processing (vertical comparisons), and to that of corresponding stages in NH listeners (horizontal comparisons). To this end, we have embarked on a follow up study that has sought to characterize consonant representations in hearing impaired listeners, combining psychophysics with EEG to generate perceptual confusion matrices that can be directly compared with EEG DSMs (Su, Sankaran & Swaminathan, 2018). The study utilizes the same stimulus set and analytical approaches as those implemented in this thesis. Such a design may offer significant advances in our ability to diagnose and understand the underlying causes of hearing impairment. This is particularly true when clinical audiometry fails to identify symptoms, as is the case in hidden hearing loss (Liberman, 2015).
Figure 36: Schematic of future representational similarity analysis of speech across stages in the auditory pathway and listening populations. Future work should examine the representational structure of the same set of phonemes at various stages of processing in both normal hearing (NH) and hearing impaired (HI) populations. Comparisons across different stages of the auditory pathway (vertical arrows) and across the two populations (horizontal arrows) will help determine both the signal changes that underlie noise-robust speech perception, and how hearing loss disrupts this process.

A major aim of the current examination of speech was to evaluate whether phoneme-level distinctions are evident in the evoked activity of individual listeners using EEG – a relatively low-cost and accessible neural measurement device. While current results suggest that EEG is currently plagued by signal-to-noise ratios too low for such distinctions to be reliably found, solving this problem remains an important future goal with obvious allure, particularly if such distinctions can be measured in real-time (Moses et al., 2016). For example, decoding speech from cortical activity in real-time could provide a path to efficient linguistic communication for those with locked-in syndrome (Birbaumer, 2006; Oken et al., 2014).
A real-time instantiation of an EEG-based speech decoding system also has potential applications in the realm of audiological assessment and hearing-aid fitting (figure 37). Specifically, evidence suggests that the audiometric curve alone – the current standard for evaluating appropriate hearing-aid gains – is not a perfect predictor of speech intelligibility (Rankovic, 2002; Rhebergen et al., 2005). As such, the fitting procedure could benefit from speech specific tests that optimize fitting parameters for speech intelligibility. Though this can be achieved behaviorally by conducting psychophysical tests to generate confusion matrices for a given set of fitting parameters, the process would be time-intensive and exhaustive for the patient, with repeat testing needed each time parameters change. Thus, psychophysical speech tests are unlikely to be feasible within the time-span practical to audiological assessment. On the other hand, extracting the representation of speech directly from cortex using EEG would provide a rapid alternative. Patients listening to phonetically transcribed speech could have neural phoneme “maps” generated in real-time within a given level of confidence. In turn, a clustering analysis of this map could be used to provide an output that determines a more optimal set of parameters that separate the highly confusable clusters in neural representational space, and the process could be iteratively applied so as to maximize the cortical distance between phonemic categories. This iterative feedback loop could use computational approaches to optimize those parameters that maximize neural separation, coupled with feedback from the patient. Although such an application seems relatively straightforward to implement computationally, the ultimate feasibility of such practical uses of real-time speech decoding will hinge largely on the availability of low-impedance scalp-level neural measurement devices in the future.
7.2 Determining the relevant representational unit of speech

In chapter 6, we chose to probe the neural representations of consonants, thereby treating the phoneme as the fundamental perceptual unit of speech underlying recognition and intelligibility. This seems an intuitive choice, as the phoneme is the smallest contrastive unit of speech capable of changing word-level meaning (e.g. the words “altitude” and “aptitude”). Indeed, current conceptions of speech perception propose that the superior temporal gyrus (STG) – the core region implicated in acoustic-phonetic processing of speech – consists of a series of neural populations tuned to different phonetic or spectrotemporal features that deterministically respond each time the appropriate feature is encountered in the speech stream (figure 35A; Hickok & Poeppel, 2007; Mesgarani et al., 2014). However, upon more thorough consideration, it could well be the case that these neural populations do not operate at the level of a phoneme, but instead have longer temporal receptive windows that show tuning to particular biphones, triphones or other larger perceptual units such as syllables or words. This claim is based on several factors. Firstly, even in the case of a phoneme, the underlying acoustic features cannot be recognized immediately as they arrive (that would require that perception be instantaneous). Rather, sensory evidence must be gradually accumulated over time. Given that such a non-zero
temporal encoding window already exists, there is no logical reason to assume that it is then limited to the smallest unit of speech-sound. Secondly, behavioral studies have attempted to determine the perceptual unit of speech by way of auditory detection and recognition masking studies (Massaro, 1972; Liberman et al., 1967). Such studies implicate the syllable, not the phoneme, as the perceptual unit. Furthermore, in contrast to its conventionally viewed role as a sensory processing region, multiple lines of research implicate the STG as a perceptual auditory region, participating in phenomena such as phonemic restoration and categorical speech perception (Leonard et al., 2016; Chang et al., 2010). Thus, it seems plausible that the STG contains neural populations tuned to perceptually relevant speech units, thus making syllables a likely candidate over phonemes (figure 35B). If this were the case, it would additionally provide a solution to the phonological sequencing (Dehaene et al., 2015) and temporal binding (Engel & Singer, 2001) problems in speech.

Figure 35: Two contrasting views on the representational unit of speech in neural populations in the superior temporal gyrus (STG). (A) Traditional views assert that neurons in STG are tuned to particular phonetic features and respond deterministically to those features in the speech stream. Thus, each of the three displayed populations respond selectively to each phoneme in the word “pin”. These individual representations are then thought to be integrated in a higher level “word area” (e.g. “Broca’s area” - the pars opercularis of the inferior frontal gyrus). (B) In an alternate view, neurons in STG have longer temporal receptive windows that are tuned to larger units of speech – syllables.
7.3 Further consideration of the cortical mechanisms supporting musical pitch perception

While the series of experiments presented in chapters 4 and 5 elucidate the time-course with which representations of musical pitch transform from those based on sensory features to those aligned with perceptual properties, they fall short of a specific description of the neural mechanisms by which this transformation occurs. As described in earlier chapters, feature extraction in auditory cortical areas underlies the early transformation of acoustic features (such as frequency information) into complex pitch representations (Zatorre, 1988; Bendor & Wang, 2005), and these sensory regions are also implicated in the early melodic and harmonic analysis of pitch-relations (Tramo et al., 2002; Warren et al., 2003). How are these extracted stimulus features integrated with a stored knowledge of tonal schema to produce the ultimate percept of musical pitch - with its associated continuum of stability? One possibility is that while melodic feature extraction occurs, this information enters an auditory short-term memory store. The nature of this memory system is up for debate – one option is that it occurs through local recurrent connections in the temporal lobe that feed information back to sensory populations and provide the network with an online memory of recent stimulus history (e.g. a memory of the context in which a tone occurs). The appeal of such an architecture is that memory becomes intrinsic in the sensory representation of “current” input, providing a parsimonious representation of the tone plus its prior context, rather than requiring access to a separate memory store (Ellman, 1990; 2004; Hasson et al., 2015). While this network is actively updated by incoming input, projections from other cortical regions combine with the sensory representation to produce a percept. One possible source of these projections is pars opercularis in the inferior frontal gyrus (i.e. Broca’s area) due to its involvement in both music (Maess et al., 2001; Koelsch 2006; Koelsch et al., 2002; Tillmann, Janata, & Bharucha, 2003; Janata et al., 2002) and linguistic (Caplan, 2006) syntactic processing. The nature of this interaction between frontal and temporal regions is another open question – do the top-down projections “act on” the sensory representation or is activity from both regions convolved and read out by downstream populations? Finding the answers to such mechanistic questions will prove vital not only to understanding the basis of musical pitch perception, it will also provide domain-general insight into how the brain combines current information about the external environment with prior
internal structural knowledge to produce an eventual percept. Fortunately, with current high-density intracranial electrophysiological recording methods, such fine-grained mechanistic explorations of neural information processing are now possible.

### 7.4 References


8. General Appendix

The following section contains additional bodies of research unrelated to the main research questions of this thesis that were completed the period of PhD candidature.
Perceptual grouping of brief patterns in melodic contour revealed through response times

Narayan Sankaran*, Simon Carlile*

a. Auditory Neuroscience Laboratory, Department of Physiology and Bosch Institute, School of Medicine, University of Sydney, New South Wales, Australia
Abstract:

This study examined the perceptual grouping of basic patterns in melodic contour using a 2 alternative forced choice reaction time paradigm. Sequences comprised of three tones were presented to listeners. The first two tones defined an Implicative Interval (II) which independently varied in its magnitude (small or large) and relative melodic direction (congruent or incongruent) to the immediately proceeding Continuation Interval (CI). Upon hearing the final tone, listeners made a speeded decision regarding the direction of the CI (experiment 1) or detected whether or not any pitch change occurred within the CI (experiment 2). In both experiments, responses were faster for sequences in which the directions of the CI and II were congruent rather than incongruent, and in which IIs were small rather than large, though these effects were stronger in experiment 1. Additionally, response times were inversely correlated with the magnitude of the CI. Results are discussed in relation to endogenous predictive and exogenous attentional processes.

Keywords: melodic contour, perceptual grouping, expectancy, reaction time, implication-realization
1. Introduction

Before the influence of tonality or any other formally learned schema, the cognitive processing of melody is likely influenced by bottom-up, perceptual mechanisms relating to the ways in which the auditory system groups patterns of melodic contour - the rising and falling relationship in pitch between successive notes (Narmour, 1992; Bregman, 1994; Trainor & Zatorre, 2009). The aim of the current study was to investigate such early low level processing by presenting listeners with brief melodic structures comprised of two adjacent melodic intervals that independently varied in their magnitude and relative melodic direction, using reaction times (RT) as a dependent measure.

Several emergent, context-dependent factors affect the way in which the auditory system groups streams of sound into coherent sequences. Within a musical context, certain melodic structures are thought to instantiate specific predictions in the listener regarding the anticipated outcome (Huron, 2006). In considering this phenomenon of Musical Expectancy, Narmour (1992) proposed the Implication-Realization (IR) model to describe the induction of expectations from specific patterns in melodic contour. The model details several rules that are largely derived from the analysis of musical form and the notion of auditory Gestalten (Meyer, 1956; Köhler, 1970). Two rules state that a listener can expect a continuation in the same direction following a small melodic interval, but a reversal following a large one, and expect intervals in general to be small rather than large. Evidence suggests that these principles may arise from innate dispositions towards proximity, parsimony and regression to the mean (see Jones, 1990) as well as through listeners heuristically learning the statistical properties of musical sequences (Von Hippel & Huron, 2000; Von Hippel, 2002).

Alternatively or concurrently, a salience-driven process may influence grouping, whereby certain melodic features capture exogenous attention more readily than others. Outside a musical context, prior studies have shown that the accuracy with which listeners identify the direction of frequency changes was reduced when the direction of the immediately preceding, task-irrelevant,
interval was incongruent with that of the judgement interval (Mathias, Micheyl & Bailey, 2010; Semal & Demany, 2006; Mathias, Bailey, Semal & Demany 2011). Additionally, the thresholds for these direction identification judgements increased proportionately with the magnitude of the preceding interval. Mathias et al. (2011) suggest that such “sequential interference” results from some listeners being unable to ignore the initial interval. Thus sensitivity to frequency changes may be reduced to a degree commensurate with the magnitude and relative direction of the immediately preceding frequency leap.

Here, two reaction time (RT) experiments were performed with both implementing a two alternative forced choice (2AFC) paradigm. Sequences of three tones comprising two musical intervals were presented to listeners. The initial *implicative interval* (II; between T0-T1) and subsequent *continuation interval* (CI; between T1-T2) independently varied in their magnitude *(small/large)* and direction *(congruent/incongruent)* relative to one another. A speeded decision regarding the CI was made immediately following the onset of the T2, enabling a quantitative measurement of the preceding interval’s influence on ‘current’ information processing. In experiment 1, listeners identified the CI direction; in experiment 2, listeners detected whether or not any pitch change occurred within the CI.

Within a predictive framework, neural processing time may be inversely correlated with expectedness. That is, the characteristics of the first interval (II) may induce a prediction, guiding endogenous attention to the expected spectrotemporal region. Doing so would produce greater sensitivity and thus facilitate RTs when the second interval (CI) fulfils the expectation. Alternatively, characteristics of the II may sequentially interfere with perception of the CI in a manner similar to that of the earlier described frequency direction identification studies (Mathias et al., 2010; Semal & Demany, 2006; Mathias et al., 2011). This may produce variations in response latency, behaving as a proxy for frequency sensitivity. Rather than dissociating between effects corresponding to endogenous predictive and exogenous salience-driven signals, the current study simply seeks to examine the effects of contour shape on early melodic processing, reflecting the operation of low-level grouping principles that may be generalizable across all auditory stimuli (Juslin & Laukka, 2003; Patel, 2008).
The perception of different structures in melodic contour has been examined by prior behavioural studies (Von Hippel, Huron, & Harnish, 1998; Eerola, Toiviainen, & Krumhansl, 2002; Von Hippel, 2002; Schellenberg, 1997; Cuddy & Lunney, 1995; Graves, Micheyl, & Oxenham, 2014). However, most paradigms afford listeners the time to engage in retrospective analysis before responding on a subjective scale (for an overview of common procedures see Huron, 2006: chapter 3). Evidence suggests that with such sustained retrospection, listeners integrate ever-more complex factors into their judgement, reflecting the broader context in which the musical selection was presented (Bailes, Dean, & Pearce, 2013). As such, the exact structure of melodic patterns predisposed for grouping by the early auditory system remains unclear. RTs have been successfully used to probe music-syntactic processing mainly via implicit priming paradigms (Janata & Reisberg, 1988; Tillmann & Bharucha, 2002; Marmel, Tillmann, & Dowling, 2008), however there use as a tool to probe low-level melodic grouping remains scarce. Here, the RT approach assumes that the time pressure of a speeded response minimized retrospective appraisal, thereby diminishing the influence of high-level schematic knowledge and ensuring the measurement of fundamentally bottom-up processes. To further prevent such top-down confounds, the current stimulus design impeded the across-trial emergence of any singular tonal region (see methods).

2. Materials & Methods

2.1 Design & Procedure.

The current experimental design was adapted from previous studies (Cuddy & Lunney, 1995; Graves et al. 2014). All melodic stimuli corresponded to discreet keynotes of the western musical system. Trials consisted of sequences with three tones (Fig. 1A). The first two tones (T0-T1) established an implicative interval (II); the second and third tones (T1-T2) defined a continuation interval (CI). Two separate speeded judgements were made upon onset of the final tone. In experiment 1, subjects identified the pitch direction of the CI (up/down/same). In experiment 2, subjects detected whether T1 and T2 were the same or different. Eight different IIs were
presented in a blocked Latin-square design. These were either small (±2, ±3 steps) or large (±9, ±10 steps), corresponding to the ascending and descending forms of a major second, minor third, major sixth, and minor seventh respectively. Within each block, the same II was followed by CIs that were randomly selected without replacement from a set containing all 25 chromatic intervals spanning 12 steps below to 12 steps above the second tone. The pitch of the second tone in each sequence was randomly fixed at either C₄ or F♯₄, adjusting the first and third tones based on this selection. These two keynotes (the musical interval of a tritone) are highly unrelated in the western tonal profile (Krumhansl, 1990) and vacillating between them aimed to destabilize the top-down emergence of a tonal centre over the course of the experiment. Each block began with three practice trials and prior to testing, subjects completed a brief training routine to ensure a consistent speed-accuracy level. After each block, subjects were prompted to take short breaks before continuing. In addition, a separate testing session consisted of contours with unison IIs (0 semitones) followed by the same 25 CIs presented randomly without replacement.

Figure 1: (A) Schematic of the RT paradigm with concurrent neurobiological processes. (B) Four contour structures resulting from possible combinations of small or large IIs followed by congruent or incongruent CIs and the additional ‘Unison II’ condition. Arrows in shaded regions indicate the possible melodic range of final tone. (C) Example of ex-Gaussian PDF (red line) fit to Vincentized RT distribution quantiles (see section 2.3.1).
2.2 Stimuli and apparatus.

Harmonic complexes were synthesized using customizable software (Max 6, Cycling ’74, San Francisco CA) and shaped with spectral envelopes determined by applying a Gaussian weighting function to the amplitudes of individual harmonics (spectral centroid = 800Hz, SD = 200Hz). Onsets and offsets for all three tones were gated with 20ms raised-cosine ramps and had durations of 1200, 400 and 800ms respectively. These durations were chosen, in keeping with prior studies utilizing the same stimulus design, to create a sense of simple quadruple or duple time (Graves et al., 2014; Cuddy & Lunny, 1995). Stimuli were presented at a constant SPL of 60dB through headphones (Sennheiser, Old Lyme, CT). Subjects registered the categorical responses by tapping piezoelectric sensors whose signals were digitally analysed in Max 6. Both audio-output and response-input signals were sampled at 96kHz via an RME Fireface 400 interface. Prior system testing indicated that measuring responses in this way yielded low system latency with sub-millisecond precision. Subjects were instructed to use their preferred hand-finger combination to register responses, maintaining this configuration for the entire duration of the experiment.

2.3 Analysis

Responses corresponding to trials with unison CIs were excluded from the analysis as the response bias created by their relatively infrequent occurrence confounded the measurement (in experiment 1 the probability of a unison trial within a block was 1/25 compared with 12/25 for both ‘up’ and ‘down’ trials). The remaining possible combinations of IIs and CIs resulted in 5 distinct contour structures (Fig. 1B): small IIs followed by CIs in the congruent direction (SC); large IIs with congruent CIs (LC); small IIs with incongruent CIs (SI), large IIs with incongruent CIs (LI), and unison IIs followed by any non-unison CI.

Because reaction time (RT) response distributions typically have a positively skewed, unimodal shape (Luce, 1986; Balota & Yap, 2011), spurious long responses could overlap with genuinely
long RTs. As such, normalizing the distribution by trimming the skewed tail to calculate Gaussian parameters may also remove effects that reflect the underlying process of interest. Indeed, other psychophysical research utilising RT as the dependent variable suggests that applying standard Gaussian summary statistics to RT distributions may result in findings that reflect artefacts of the analysis itself (Ratcliff, 1979; Heathcote, Popiel, & Mewhort, 1991; Meyer, Osman, Irwin, & Yantis, 1988; Whelan, 2010). To test the distribution, a chi-squared goodness-of-fit test indicated that RT distributions of ten out of twelve subjects differed significantly from Gaussian in experiment 1 of the current study, \( \chi^2 (5, N=192) > 11.93; p < 0.0178 \) for 10 subjects.

To manage such statistical complications, the method adopted here involved parameterization of the shape of the entire response distribution (see section 2.3.1). Firstly, the issue of outliers due to transient inattention needed to be managed. For the reasons outline above, we chose to adopt a liberal method of outlier identification in which responses were only deemed spurious if they fell below 150ms (an implausibly fast reaction time) or above two standard deviations from each participant’s mean RT. The use parametric statistics for outlier removal served only to increase statistical power. An analysis conducted on all responses yielded exactly the same trends as that identified without the outliers (see below), but with a considerably larger variance due to the sensitivity of random resampling to data in the distribution tails (see section 2.3.1). Removing outliers in this fashion excluded 4.4% of the data.

2.3.1 Melodic Contour Shape

Responses were represented relative to each subject’s median RT (denoted by a value of 1), permitting comparisons between all 5 contours despite unmatched experimental cohorts (see section 3.1). To aggregate responses across subjects, accounting for individual variation whilst preserving the overall distribution shape, data was pooled across subjects using a procedure known as Vincentizing (see Ratcliff, 1979 for a more detail description): for every subject, RT distributions corresponding to each of the five contour conditions (pooled across all CIs) were sorted and grouped such that 12 bins contained an equal number of responses (8.33% quantiles).
Each respective quantile was then averaged across all subjects to produce aggregated group distributions (Fig. 1c). *Ex-Gaussian* probability density functions (PDF) were model-fit to the Vincentized distributions (see figure 1) using a *maximum likelihood estimation* procedure implemented in the DISTRIB Toolbox for MATLAB (Lacouture & Cousineau, 2008).

The *ex-Gaussian* is produced by the convolution of Gaussian and exponential functions and has been shown to accurately model empirical response distributions for a variety of neuropsychological tasks (Epstein et al., 2011), including choice reaction time (Leth-Steensen, Elbaz, & Douglas, 2000; Geurts et al., 2008) and Go/No Go (Vaurio, Simmonds, & Mostofsky, 2009) paradigms. The model has three descriptive parameters; μ and σ correspond to the Gaussian mean and variance respectively while τ represents the mean and variance of the exponential decay. The *ex-Gaussian* mean and variance are $\mu + \tau$ and $\sigma^2 + \tau^2$ respectively (Ratcliff, 1993; Heathcote et al., 1991). Prior to fitting the model, Kolmogorov-Smirnov goodness-of-fit tests indicated that RT distributions for every *subject x contour shape* combination did not significantly differ from *ex-Gaussian* (All $Z$s <0.13, $p$s >0.41). Finally, parameter estimates were non-parametrically bootstrapped (Efron & Tibshirani, 1994) by randomly resampling with replacement from the original distributions for each subject and condition, repeating the Vincentizing and curve-fitting procedure with every successive bootstrapping iteration ($N_{\text{bootstrap}} = 2000$).

2.3.2 Continuation Interval

The above *ex-Gaussian* analysis for each individual *CI size x contour shape* was not possible due to the prohibitive number of trials that would have been required. Instead, for every subject the median RT at each CI size was calculated and normalized to their respective median across all responses.

2.4 Ethics Statement
Written informed consent was provided and experiments were approved by the Human Research Ethics Committee of the University of Sydney (HREC number 791).

3. Experiment 1: Pitch Direction Identification

3.1 Subjects

Thirteen subjects (6 male, mean music training = 8.9 years, SD = 6.1) were recruited to participate in the non-baseline portion of the experiment. Five of these subjects (3 male) participated in the additional ‘Unison II’ condition.

3.2 Results

The ex-Gaussian mean ($\mu+\tau$) and variance ($\sigma^2+\tau^2$) of RT distributions for each stimulus condition are shown in Figure 2. Significance values are based on the 95% confidence intervals derived from non-parametric bootstrapping. Significance limits were Holm-Bonferroni corrected for multiple comparisons (Abdi, 2010). Only significant comparisons after correction are reported.
Figure 2: The mean ($\mu+\tau$; top) and variance ($\sigma^2+\tau^2$; bottom) of ex-Gaussian model fits to distributions corresponding to the 5 different contour structures in experiment 1 (n = 13). SC = small congruent, LC = large congruent, SI = small incongruent, LI = large incongruent. All 24 non-zero CIs were presented for ‘Unison II’ contours. Parameter values are plotted relative to the median RT across all trials. Error bars are the 95% confidence intervals resulting from bootstrapping.

Mean RTs for contours with unison IIs were significantly longer than trials with congruent IIs and CIs (Unison II - SC: p<0.0001; Unison II - LC: p=0.0002). Trials with LI contours had significantly longer mean RTs than all 4 other contour types (p<0.001 in all 4 comparisons). Additionally, the mean of SI contours was significantly longer than that of SC (p=0.0068). Both the congruent contours also had significantly smaller variances in their RT distributions than that of Unison II contours (Unison II - SC: p=0.0011; Unison II - LC: p=0.0012). The variance
(Figure 2, bottom panel) of LI contours was significantly larger than both congruent contours (LI – SC: p=0.0025; LI – LC: p=0.0031). Incorrect responses occurred at rates of 7.0%, 2.5%, 3.2%, 6.1% and 7.3% for Unison II, SC, LC, SI and LI contours respectively. A one-way Kruskal-Wallis analysis found these differences to be approaching significance ($\chi^2=9.05$, p=0.06). However, since errors were positively correlated with the trends in mean and variance (c.f. Figure 2), the results are assumed to be void of speed-accuracy confounds. That is to say, the fast RTs were also associated with the smallest variance and the lowest error rates.

To assess the effects of the CI, RTs were first tested for significant asymmetry about zero semitones. To achieve this, RTs corresponding to each ascending CI were subtracted from those corresponding to its respective descending interval (for example, RTs when the CI = +5 were subtracted from RTs when the CI = -5). A one-sample t-test at each of the 12 magnitudes confirmed that the mean of these RT differences did not significantly differ from zero, indicating that no significant asymmetry existed between rising and falling CIs (all $p$s > 0.16; Bonferroni adjusted). The group mean of median RTs collapsed across negative and positive CIs are shown in figure 3. For the purposes of a matched-sample analysis, the Unison II condition was removed. A two-factor repeated-measures ANOVA found significant main effects for contour shape ($F(11,121)=11.39; p<0.0001$) and CI ($F(3,33)=12.34; p<0.0001$). The interaction between contour shape and CI was insignificant ($F(33,363)=1.304; p=0.13$).
Figure 3: Group means of the median RTs collapsed across ascending and descending CIs for the four contour structures with non-zero CIs.

Pairwise comparisons across CI magnitude indicate that, in general, the median RT for contours with smaller CIs were longer than that of larger CIs. For example, RTs of trials with CIs of 1 and 2 steps were significantly longer than those of [4, 5]*, 6**, and [7, 8, 9, 10, 11, 12] **** steps (where * indicates p<0.05, **p<0.01, ***p<0.001, ****p<0.0001; Tukey’s multiple comparison test).

Overall results from experiment 1 suggest that both the size and relative direction of the preceding interval (II) influenced the latency in identifying the direction of the subsequent interval (CI), reflected in a longer ex-Gaussian mean and variance for incongruent contours with larger IIs. Additionally, the group mean of median RTs was longer for smaller CIs.

4. Experiment 2: Pitch Change Detection

Recent work suggests that identifying frequency direction and detecting frequency change may be perceptually separable processes (Mathias et al., 2011). Additionally, findings from cortical lesion studies support the anatomical independence of the two tasks (Johnsrude, Penhune, & Zatorre, 2000; Tramo, Shah, & Braida, 2002; Warrier & Zatorre, 2004). These studies suggest that perceiving frequency-direction recruits right auditory cortical regions, whereas simply detecting a frequency change may involve sub-cortical regions alone (Foxton, Weisz, Bauchet-Lecaignard, Delpuech, & Bertrand, 2009; Johnsrude et al., 2000). Based on this, if the effects observed in experiment 1 persist in a pitch-change detection task, they may in part reflect pre-cortical processing mechanisms.

4.1 Subjects
Eleven subjects (5 male) with an average 7.4 years of musical training (SD = 5.3) were recruited to participate. Six of these subjects (2 male) participated in the ‘Unison II’ block.

### 4.2 Design and Procedure

Upon hearing the onset of the final tone in each trial, subjects registered a speeded response as to whether the second and third tones (comprising the CI) were the same or different. Eight non-zero IIs used in experiment 1 were presented in a blocked Latin square design (see section 2.1) in addition to a separate ‘Unison II’ testing block. Each block consisted of 50 trials (2 practice). In half, the trials contained Unison CIs (zero semitones) while the other half contained the 24 non-zero chromatic intervals spanning -12 to +12 steps from the second tone, presented randomly without replacement.

### 4.3 Results

Only trials with non-zero CIs were analysed as described in section 2.3. The ex-Gaussian mean and variance of RT distributions corresponding to each contour shape are shown in figure 4.
The mean RT for trials with LI contours was significantly longer than that of all other contour shapes (all $p$s<0.0034; Holm-Bonferroni corrected). Additionally, the variance of the RT distribution for LI contours was significantly greater than that of Unison II contours ($p = 0.0034$). Mean error rates did not significantly differ across presentation of the 5 different contours (Kruskal-Wallis; $X^2 = 2.54$, $p = 0.64$) again suggesting that the data is void of speed-accuracy confounds. Overall, differences in the mean and variance of RT distributions corresponding to different contour shapes is less pronounced than those observed in experiment 1, however trends suggest a similar effect of II size and relative direction across direction identification and change detection tasks.

**Figure 4:** The mean (top) and variance (bottom) of ex-Gaussian model fits to distributions corresponding to the 5 different contour structures (see Figure 2 for details).
In analysing the effects across CIs, one-sample t-tests performed on the difference in RTs between ascending and descending forms of each CI magnitude indicated that the mean RTs corresponding to rising intervals of +5 (p=0.0173) and +7 (p=0.0433) semitones were significantly less than their respective descending forms (intervals of -5 and -7 semitones, respectively). Thus rather than collapsing data across ascending and descending CIs, group means of median RTs across all 24 signed CIs were preserved in the analysis and are displayed in figure 5.

![Figure 5: Group means of individual median RTs across all 24 ascending and descending CIs.](image)

A two-way repeated measure ANOVA found a significant main effect of CI (F(23,230) = 3.79, p<0.0001) and contour shape (F(3,30) = 4.58, p=0.0094). The interaction between factors was insignificant (F(69,690) = 1.23, p=0.105). A Tukey’s multiple comparison test found that median responses to contours with a CIs of 1 and -1 semitone were significantly larger than those with 5*, 6*, [7, 10]** and 11** semitones.

5. Discussion & Conclusion

The aim of the current study was to measure the bottom-up processing of melodic contour in order to better understand the operation of low-level perceptual grouping mechanisms, using speeded response times (RTs) in an attempt to minimize the cognitive, top-down processing of melody. Here, several of the observed effects are considered as the result of the processing of both endogenous predictive and exogenously salience-driven signals.
In experiment 1, mean responses were faster and more accurate for trials in which the directions of the CI and II were congruent rather than incongruent. This may reflect a low-level prediction that predisposes the auditory system to use the repetition of frequency direction as a fundamental grouping cue. That is, an initial rising interval may have biased the listener to anticipate another rising interval. This finding is consistent with that of studies demonstrating perceptual continuity when listeners are presented with frequency glides interrupted by brief occlusions (Crum and Hafter, 2008). Additionally, auditory event-related potential (ERP) studies have demonstrated that the violation of contour shape (Paavilainen, Degerman, Takegata, & Winkler, 2003; Saarinen, Paavilainen, Schröger, Tervaniemi, & Näätänen, 1992) in a sequence of established standard contours reliably elicit the mismatch negativity (MMN). The MMN has been explained by the notion of a neural memory trace and is generally regarded as evidence of early predictive mechanisms (Rohrmeier & Koelsch, 2012). Since the first two tones of the current study already define a melodic direction, it seems plausible that the facilitation of directionally congruent contours resulted from the operation of similar low-level predictive process. Accordingly, in trials with unison IIs the listener presumably had the strongest expectation for another unison interval to follow, constituting a basic repeating pattern, as evidenced by the ‘classic’ frequency-oddball MMN (Näätänen, Schröger, Karakas, Tervaniemi, & Paavilainen, 1993). However, as only non-zero CIs were analysed (due to the response bias created by the available task choices), such trials consistently denied expectations. Accordingly, as a result of such repetition violation, trials with Unison IIs had longer mean RTs than trials with congruent contours in both experiments (though not significantly so in experiment 2).

Models of musical expectancy (see section 1) propose that listeners expect reversals following large melodic leaps (Narmour, 1992; Schellenberg, 1997). However, mean RTs were generally longer for LI contours compared with all other contours, and responses were consistently faster for contours with smaller rather than larger IIs. This apparent discrepancy has two likely explanations. Firstly, an organizational system that predicts facilitated perception of LI contours may reflect a high-level mode of music listening. Indeed, evidence suggests that these rules may arise epiphenomenally from constraints imposed by melodic range or tessitura (Von Hippel & Huron, 2000) or by listeners heuristically learning the statistical properties of music (Von
Hippel, 2002), rather than innate or ‘hard-wired’ dispositions. Secondly, predictive effects may have been saturated by exogenous attentional capture, whereby larger IIs distracted listeners from the task. Consistent with the effects on frequency direction thresholds from Mathias et al. (2011), the extent to which such “sequential interference” produced observable effects in RTs was proportional to the magnitude of the interval preceding the judgement (the II in the current study). Interestingly, the correlation between the two experiments (Mathias et al., 2011 & the current study) suggests that neural processing time may behave as a proxy for frequency sensitivity.

Another fundamental principle of musical expectancy is that listeners expect successive continuations in a melodic sequence to be proximate in pitch (Narmour, 1992). Interestingly, mean responses were longer for smaller CIs than larger ones, regardless of the magnitude or relative melodic direction of the preceding interval. As such, predictive mechanisms are unlikely to account for the observed effects across CI size. Instead, responses seem to be salience-driven, whereby larger frequency leaps more readily capture exogenous attention and are therefore processed faster. Additionally, the smaller CIs may have increased task-difficulty in experiment 1 due to conceptual confusion regarding the appropriate labelling of pitch changes as ‘up’ or ‘down’. To examine strictly predictive processes in the future, experiments may make use of implicit priming paradigms in which the task runs orthogonal to the variable of interest (e.g. examining whether RTs for identifying the timbre of the last note are facilitated for certain contour patterns).

Both the magnitude and relative direction of the preceding interval influenced the processing latency of ‘real-time’ melodic information. In experiment 1, identification of contour direction was faster (ex-Gaussian means and variances were smaller) for sequences which maintained the same melodic direction and in which the magnitude of the first interval was small rather than large. The same general trends were observed in a change-detection task (experiment 2), albeit less pronounced. Given the evidence for functional and anatomical specificity in perceiving the signed direction of a frequency-change versus merely detecting
the unsigned difference (Johnsrude et al., 2000; Tramo et al., 2002; Mathias et al., 2010), the similar effects across both experiments may reflect low-level grouping principles generalizable to all auditory processing (McDermott, Lehr, & Oxenham, 2008; Graves et al., 2014).

Throughout the present study, effect sizes were not large in perceptual terms due to the high levels of noise inherent in simple behavioural measures such as choice reaction times. However, they still reflect significant differences in processing mechanisms. As such, great care needs to be taken in analysing RT data to ensure that results are sensitive to the processes of interest. Lastly, RTs were used under the assumption that they negate top-down, retrospective processing that situates the melodic fragment within a broader musical framework governed by schematic knowledge, such as tonality. This assumption, however, remains unverified, and future studies should focus on further isolating bottom-up from top-down melodic processing. Current findings may help to further decode musical emotion, and provide insight into the perception of other low-level emergent auditory phenomena such as speech prosody (Juslin & Laukka, 2003; Patel, 2008).

6. Acknowledgements

The authors would like to thank the many subjects, the reviewers for their time and attention and Ms. Francesca Meliton for her assistance with data collection. This work was supported by the Australian Research Council Grants DP110104579 http://www.arc.gov.au/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

7. References


Luce, R. D. (1986). Response times (No. 8). Oxford University Press.


Effects of Virtual Speaker Density and Room Reverberation on Spatiotemporal Thresholds of Audio-Visual Motion Coherence

Narayan Sankaran*, Johahn Leung, Simon Carlile
Auditory Neuroscience Laboratory, School of Medical Sciences, The University of Sydney, Sydney, NSW, Australia

Abstract

The present study examined the effects of spatial sound-source density and reverberation on the spatiotemporal window for audio-visual motion coherence. Three different acoustic stimuli were generated in Virtual Auditory Space: two acoustically "dry" stimuli via the measurement of anechoic head-related impulse responses recorded at either 1° or 5° spatial intervals (Experiment 1), and a reverberant stimulus rendered from binaural room impulse responses recorded at 5° intervals in situ in order to capture reverberant acoustics in addition to head-related cues (Experiment 2). A moving visual stimulus with invariant localization cues was generated by sequentially activating LED's along the same radial path as the virtual auditory motion. Stimuli were presented at 25°/s, 50°/s and 100°/s with a random spatial offset between audition and vision. In a 2AFC task, subjects made a judgment of the leading modality (auditory or visual). No significant differences were observed in the spatial threshold based on the point of subjective equivalence (PSE) or the slope of psychometric functions (β) across all three acoustic conditions. Additionally, both the PSE and β did not significantly differ across velocity, suggesting a fixed spatial window of audio-visual separation. Findings suggest that there was no loss in spatial information accompanying the reduction in spatial cues and reverberation levels tested, and establish a perceptual measure for assessing the veracity of motion generated from discrete locations and in echoic environments.

Introduction

Various experiments have sought to determine the nature of the spatiotemporal integration window for audio-visual motion [1–3]. To probe this question, studies typically deliver moving auditory stimuli using an array of sequentially activated speakers in free-field [4–7], or over headphones by measuring Head Related Impulse Responses (HRIRs) and rendering a Virtual Auditory Space (VAS) [8]. Irrespective of the delivery method, there are a number of unresolved issues in the process.

In the generation of acoustical motion, moving the sound source itself mechanically has the advantage of real-world coherence [9,10]. However, physical constraints such as background motor noise, restricted speeds and limited spatial extents present numerous disadvantages experimentally. Instead, the percept of motion is usually created by sequentially activating discrete stationary sound-sources. Whether these are physical speakers placed in free field arrays or stimuli rendered in VAS via the measurement of HRIRs [see methods], the changes in acoustical cues are quantized, resulting in a loss of spatial information. While the resulting moving stimulus may be perceived as spatially continuous, other psychophysical consequences of this reduction in cue density remain unclear. This is an important consideration given that a clear understanding of the mechanisms underlying auditory motion perception remain outstanding. Typical step-sizes utilized in auditory motion studies range from approximately 2° to 6° [4,6,11]. Intuitively, a perceptual limit of this quantization can be estimated from the minimum audible movement angle (MAMA), defined as the minimum spatial extent required for a sound to elicit a motion percept [12]. However, reported values differ depending on velocity and spectral content, confounding a systematic description of MAMA across any one parameter. Using moving stimuli generated by stereo balancing a 500 Hz tone across two speakers, Grantham [13] reported MAMAs ranging from 5° to 21° at source velocities of 15°/s and 90°/s respectively. Perrott and Marlborough [10] found MAMAs ranging from 0.9° to 1.6° using a speaker that rotated at 20°/s with a 500 Hz–8 kHz pink noise stimulus. The small but statistically significant difference depended on whether onset and offset cues were provided to the listeners. In contrast, Chandler and Grantham [9] reported a value of 3.6° using 500–10 kHz “wideband” noise delivered by a speaker moving at 20°/s, increasing to 14.4° at a velocity of 90°/s. Taken together, these studies describe a metric that is highly variable, with the only commonality being the increase with velocity. Further complicating the issue, as suggested in Grantham [13] and confirmed in Carlile and Best [14] and Freeman et al. [15] (2014), velocity per se is not a salient cue in auditory motion.
perception. Given the increasing number of auditory motion studies that use a discrete-sequential presentation technique, a goal of the present study is to compare auditory motion perception of the finest spatial discretization (1°) against one that is commonly used (5°) using wide-band stimuli at various velocities.

Experiments often present a moving auditory stimulus that is anechoic. However, everyday environments contain reverberant energy due to sound-reflecting surfaces. Despite its ubiquity, little is known about the perceptual effects of reverberation outside its influence on stationary sound sources [16–19]. Such studies have demonstrated that, though the ratio of direct to reverberant energy (D/R) provides a direct cue to source depth that would be unavailable to the listener under anechoic conditions [20,21], the interference of direct and reflected sound at the listener’s ears can decorrelate the binaural cues, thereby diminishing localization ability [16]. One goal of the current study is to explore the nature of this trade-off when a source is in motion. Rather than utilizing a very echoic environment, where reflected sound obviously diminishes localizability, the current study examines reverberation levels found in typical listening rooms. In doing so, the perceptual impact of reverberation in the most common listening environments can be better understood.

A body of neurophysiological and psychophysical evidence suggests that specific motion-detectors are present at early stages of visual processing [22]. In contrast, there is no similar evidence of similar low level encoding in the auditory periphery [23]. However, various models of auditory motion have been proposed. One such model that is widely quoted is the “snapshot” hypothesis, whereby motion is perceived via the sequential localization and comparison of a number of static snapshots [24]. In this context, reverberation then may also degrade the acuity of motion perception since movement is inferred from the same static cues that reverberation degrades. Consistent with this notion, in a motion detection task, Saberi and Petrosyan [25] reported a rapid deterioration in performance from supra-threshold to chance level as the amount of correlation in the binaural acoustical cues decreased.

Traditionally, investigations into these issues are limited to unimodal approaches. Here, we present audio-visual motion in order to explore the effects of acoustic spatial quantization and reverberation on the spatiotemporal integration window. Models of optimal integration suggest that overall localization uncertainty is minimized via the optimal weighing of each sensory input based on the reliability of their constituent cues [26–28]. Such models account not only for ventriloquism, where visual cues dominate perception, but describe a two-way interaction in which auditory and visual streams concurrently influence each other [29,30]. Other studies have shown this holds for moving audio-visual sources [5,31–33]. In the current study, subjects compared the relative times at which moving virtual auditory and visual targets were perceived to pass the midline.

In Experiment 1, auditory motion was spatially constrained to two step-sizes: a densely sampled 1° and the sparser 5° quantization. This reduction in cue density necessitates a spread of acoustical information from a 1° to a 5° window, which may elicit greater spatial uncertainty. In Experiment 2, relevant room acoustical information was included in the construction of the VAS, using binaural room impulse responses (BRIRs) recorded in situ in 5° step-sizes. If the reverberation perceptually decorrelates the binaural cues, we expected the spatial uncertainty of the reverberant stimuli to be even greater than that of the anechoic stimuli. Throughout this study, the visual stimuli remained unchanged (see methods), ensuring invariant visual localization cues across all auditory conditions. Given this, and the significantly greater spatial resolution of the visual system, the visual stimulus served as a reference, allowing for an unambiguous comparison between acoustic conditions. We thus hypothesized that the reduction in cue density and reverberation would increase task difficulty, making the judgment about which modality was leading harder. This would be reflected by an increase in the spread of the distribution, resulting in greater variance of a fitted Gaussian function (β). The point of subjective audio-visual equality (PSE) was also measured for the three acoustic conditions, though the effects of auditory uncertainty on this parameter are harder to predict. It is important to note that even though vision has a significantly greater spatial resolution, the current study provides insight into the effects of quantization and reverberation through the relative comparisons across acoustic conditions.

### Experiment 1: Quantization of Auditory Space

#### Methods

**Participants.** Six subjects (five male, one female) participated in the experiment. All subjects had normal hearing as confirmed by audiometric screening.

**Ethics Statement.** Written informed consent was provided and experiments were approved by the Human Research Ethics Committee of the University of Sydney (HREC number 15278).

**Stimuli.** The recording procedure and rendering of motion in VAS is briefly outlined below. For a more detailed description see Carlile [8]. Individualized blocked ear HRIRs [14,34] were measured under anechoic conditions by securing microphones in the ear canals using medical grade silicon gel (Polyvinylsiloxane). The subjects’ head was stabilized by a chin-rest and monitored using a head-tracker (InterSense IC3). One-second exponential sine sweep stimuli [35] were presented by a speaker (Audience A3) mounted at the apex of a robotic arm that moved along a radial arc 1 meter from the listener. Measurements were taken from −90° to +90° along the audio-visual horizon in 1° increments.

The responses of the recording microphone and stimulus speaker were then deconvolved from the HRIRs. Figure 1 summarises the process by which moving auditory stimuli were generated. First, a broadband white noise (300 Hz to 16 kHz) of the total trial duration was generated. This was then filtered with a series of bandpass filters (from 400 to 16 kHz, equally spaced at 200 Hz with a bandwidth of 100 Hz) and amplitude modulated at 20 Hz. Such a stimulus provided a high level of modulation coherence so as to encourage perceptual object formation [36]. Finally, each segment of the noise stimulus was convolved with left and right HRIRs corresponding to each recording position (1° or 5° steps), the duration of each segment being determined by the chosen velocity of motion (see below). Subjects indicated (via qualitative feedback) that the auditory stimuli were externalized and easily localizable, which is consistent with our previous findings using similar stimuli [37]. Apparent motion was created by sequentially playing the convolved output corresponding to adjacent HRIR positions along the radial trajectory. Different velocities were generated by changing the duration per segment of noise at each quantized step; e.g. a 100°/s stimulus will have a 10 ms duration time per 1°. The 5° quantized stimulus followed the same procedure, however HRIR positions were constrained to 5° increments and the duration per position was correspondingly increased (i.e. a 100°/s stimulus would have a duration of 50 ms per 5° step). In addition, the final and initial conditions of the convolved signal from adjacent filters were combined in software (MATLAB 8.0, The MathWorks Inc) to ensure a smooth continuous signal. The rendered auditory signal was delivered to a pair of Beyer-Dynamic DT990 open-back headphones via an...
RME Fireface 400 audio interface, using the Psychophysics Toolbox extensions [38–40] to ensure sample-accurate playback timing. All recording and digital processing was performed at a 48 kHz-sampling rate.

To generate the visual stimulus, an array of high-density LEDs spaced by $1.8\ mu$ was arranged in a strip along the same radial path as the rendered auditory stimuli. For each individual LED, the brightness, colour, and activation timing were controlled using a WS2801 integrated controller with microsecond accuracy. An Arduino Mega2560 USB platform connected to a Matlab interface was used to power and control the LED strip. Apparent visual motion was produced by the sequential ‘on-off’ activation of adjacent LEDs along the strip, again, specifying velocity as a function of time per pulse. All subjects reported that apparent visual motion was smooth for all velocities examined, with the stimulus eliciting the percept of a moving line along the radial path.

Playback timing between the auditory and visual stimuli was calibrated by measuring the excitation of two photodiodes placed at various locations along the LED strip while simultaneously recording audio output. In doing so, systematic latencies in LED activation were adjusted to ensure temporal onset accuracy of auditory and visual stimuli (see below).

**Procedure.** Trials consisted of moving auditory and visual stimuli presented along a common radial trajectory with a 1 m radius along the frontal audio-visual horizon. The trajectory subtended $140^\circ$ around the subject whose head was aligned using reference lasers and stabilized by a chinrest (figure 1). The two modalities were temporally aligned, but spatial congruency was varied such that audition with respect to vision was either leading or lagging in the direction of motion. This was done by presenting the visual stimulus along a constant trajectory from $-70^\circ$ to $+70^\circ$ and varying the auditory start and end points to achieve the desired spatial offset. To avoid motion after-effects, the stimuli direction (leftward or rightward) alternated on a trial-to-trial basis.

In a 2AFC task, observers were asked to track the visual stimulus with their eyes and indicate the perceived leading modality as the stimuli crossed the approximate midline, registering their response on a keyboard. For each quantization level ($1^\circ$ vs. $5^\circ$), auditory and visual stimuli were presented at three velocities; 25 $\mu$/s, 50 $\mu$/s and 100 $\mu$/s, resulting in total stimulus durations of 5.8, 2.9 and 1.45 seconds respectively. Auditory and visual stimuli were spatially offset by randomly varying the starting location of the auditory stimuli to one of nine possible values (Table 1). Here, positive offsets indicate a visual lead; negative offsets indicate an auditory lead and zero represents spatiotemporal equality. A testing block consisted of 90 trials (10 repeats per displacement) at a given velocity and for a given acoustical condition. Psychometric functions (PF) were fitted to the results and analyzed as described below. Subject responses were fitted to a cumulative Gaussian distribution using a maximum likelihood estimation function. The lapse rate of the PF fit was maximally limited to 0.06 to account for errors due to stimulus-independent effects [41,42].

From each PF, two values were extracted. Firstly, the Point of Subjective Equality (PSE), here defined as the domain value at the inflection of the cumulative PF. Secondly, the Slope ($\beta$), defined as
the variance of the Gaussian fit. PFs were parametrically bootstrapped based on a maximum likelihood model \[43\] (n = 1000) in order to obtain 95% confidence limits solely for comparing within-subject data. All relevant experimental data is available at http://dx.doi.org/10.6084/m9.figshare.978755 including individual subject data.

Results

PSEs for experiment 1 are shown in Figure 2A for all subjects (see also Table S1). Positive PSEs equate to a physically leading visual stimulus, indicating a perceptual tendency to judge the auditory stimulus as leading when both stimuli had spatiotemporal equality. We refer to this as an auditory lead bias. Similarly, negative PSEs denote a visual lead bias. The results show substantial across-subject variability for a given acoustic condition and velocity. While PSEs were slightly greater in the HRIR 1° condition, this general trend did not reach statistical significance. A repeated-measures ANOVA was performed to examine the effects of both quantization level (HRIR 1° vs. HRIR 5°) and stimulus velocities (25 vs. 50 vs. 100°/s). No significant main effects were observed for quantization level (F = 5.74, p = 0.12) or velocity (F = 2.33, p = 0.15). The interaction between quantization levels and velocity was also insignificant (F = 0.005, p = 0.995).

The group means of the psychometric function slopes (\(\beta\)) are plotted in Figure 2B. Though there was a general trend of higher \(\beta\)s in the HRIR 5° condition, a repeated measures ANOVA showed that the effects of quantization level on \(\beta\) was not statistically significant (F = 5.4, p = 0.07). Interestingly, \(\beta\) was statistically equivalent across the three velocity conditions (F = 0.870, p = 0.45). Furthermore, the interaction between quantization level and velocity was also statistically insignificant (F = 1.05, p = 0.39).

Experiment 2: Reverberant Auditory Motion

Methods

In order to examine the effects of reverberation, Binaural Room Impulse Responses (BRIRs) were measured in situ i.e. in the experimental testing room (17 m², RT 60 ~200 ms), ensuring that a veridical amount of room acoustical information was included in the recordings. BRIR recordings were made as in Experiment 1, with the exception that 5-second exponential sine sweeps were used as the impulse response recording stimuli. These were presented over a Fostex PMO.4n dual-cone speaker that was positioned manually in 5° increments. This longer recording stimulus was necessary to ensure that the relevant reverberant acoustics were properly characterized (see below). The duration of the test stimulus was determined as per the method and velocities of Experiment 1.

Major reflective peaks were found in the first 21 ms of all BRIRs measured (Figure 3), which was preserved and convolved with the input stimulus. Pilot testing confirmed that there was no perceptual difference between stimuli rendered from the entire BRIR versus one which only used the first 21 ms of the filter (i.e. the reverberant tail contained no perceptually significant detail). Further testing and estimation of D/R also verified that the reverberant stimuli contained a salient amount of room information (see Discussion). Motion was then generated as described earlier (see Figure 1) and the experimental procedure followed that of Experiment 1.

Results

PSEs for experiment 2 are shown in Figure 4A (BRIR 5°), plotted alongside PSEs corresponding to the anechoic condition of

![Figure 2. Results for Experiment 1.](image)

(A) Individual PSEs from the two acoustic conditions for all six subjects. Blue squares and red circles denote values from HRIR (1°) and HRIR (5°) respectively. (B) Group mean \(\beta\)s shown in blue and red for HRIR (1°) and HRIR (5°) respectively at the three experimental velocities. Error bars indicate between-participants standard errors. doi:10.1371/journal.pone.0108437.g002

### Table 1. Audio-Visual Spatial displacements.

<table>
<thead>
<tr>
<th>Velocity (°/sec)</th>
<th>(\mu) (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0, ±1.25, ±2.5, ±5, ±10</td>
</tr>
<tr>
<td>50</td>
<td>0, ±2.5, ±5, ±7.5, ±10</td>
</tr>
<tr>
<td>100</td>
<td>0, ±5, ±10, ±15, ±20</td>
</tr>
</tbody>
</table>

At each velocity, congruence between auditory and visual stimuli was offset by one of nine randomized values. Positive offsets indicate an auditory lag; negative values indicate an auditory lead. Zero represents spatiotemporal equality.

doi:10.1371/journal.pone.0108437.t001
peaks are visible in the preserved BRIR. 

Represent direct and reverberant energy respectively. Six early-reflected

condition was also not significant (\(F = 0.75, p = 0.50\)). The interaction between velocity and acoustical

acoustical condition (\(F = 0.10, p = 0.77\)) or velocity (\(F = 1.57, p = 0.26\)). The interaction between velocity and acoustical condition was also not significant (\(F = 0.75, p = 0.50\)).

Group mean \(\beta\) from reverberant conditions are shown in Figure 4B (BRIR 5°), presented alongside the results from experiment 1 (HRIR 5°). Contrary to our expectations, no significant differences were observed between reverberant and anechoic conditions (2x3 repeated measures ANOVA; \(F = 0.11, p = 0.75\)) or stimulus velocities (2x3 repeated measures ANOVA; \(F = 0.49, p = 0.63\)) and the interaction was insignificant (\(F = 1.25, p = 0.33\)).

**Discussion**

Visual cues remained constant throughout both experiments. Given this, our results suggest that the reduction in spatial cue density and the interaural decorrelation imparted by reverberation cause no perceptually significant loss of auditory spatial information, at least across the range of parameter space tested. Additionally, \(\beta\) was invariant across velocity in both experiments, suggesting that the threshold for perceptible audio-visual separation had a constant spatial arc. We discuss these findings in the context of several stimulus and task-related factors.

**Multimodal Interaction.** While this study compared between different acoustical conditions, we took advantage of the perceptual separability between the visual and auditory stimuli to use a visual stimulus as the spatiotemporal reference, thus mitigating any potential acoustical confounds. Stimuli in both modalities were distinct and highly localizable and subjects reported no multisensory integration or fused percept. Prior knowledge of the independent nature of the modalities may also have contributed in preventing any sensory integration. While there is evidence suggesting that cross-modal interactions may occur despite the lack of such integration [4], this has only been shown in a split attention task with short, narrow band stimuli. In this study we provided subjects with access to the full range of auditory localisation cues and emphasised the need to attend to both modalities.

**Velocity Invariant Spatial Window.** The slope of the PF (\(\beta\)) reflects the level of uncertainty in the psychophysical judgement rather than accuracy [41,42]. In this context it reflects the objective difficulty in resolving the location of auditory from visual stimuli. Given this, \(\beta\) enables inferences about the magnitude of the spatiotemporal window of audio-visual separation. Interestingly, results from Experiments 1 and 2 found no significant difference in \(\beta\) across experimental velocities. Given that audio-visual offsets (Table 1) were defined spatially and \(\beta\) values were thus calculated from a PF in the spatial domain, these data suggest that the spatial resolution of the audio-visual system is constant across the parameter space tested in the current study. Because stimulus velocity was constant, this corresponds to a finer temporal window of audio-visual separation for higher velocities. Such a decrease in temporal variability with increasing stimulus velocity is consistent with a previous motion extrapolation study [11] in which subjects registered the arrival times of a moving auditory stimulus crossing a stationary visual fixation. Though not the focus of their study, they found a decrease in the variability of estimated arrival times as stimulus speed increased (from a 160 ms standard deviation at 17°/s to 100 ms at 47°/s, read from their Figures 2 and 3). Furthermore, the lack of significant differences in \(\beta\) across 1° and 5° VAS suggests that the invariance in \(\beta\) between anechoic and reverberant conditions, where both stimuli were quantized at 5°, represents a genuine perceptual threshold, whereby a judgment of the leading modality in the latter condition (echoic vs. anechoic conditions) may have been constrained by the resolving capability of the auditory system rather than a physical limit imposed by the spatial quantization of the stimulus.

**MAMA.** The perceptual relevance of a reduction in the density of acoustic cues can be linked to measures of the MAMA. Findings from prior studies have reported MAMA’s ranging from
1° to 21°, depending on stimulus velocity and bandwidth [9,13]. PSEs in the current study had a magnitude of less than 5°, ranging from −4.41° to 3.6° (Figures 2, 4 and Table S1), and βs did not significantly differ across 1° and 5° VAS (Figure 2). Together, these results suggest that the 5° quantization of auditory space is still sub-threshold, thus resulting in no perceptually significant discretization of auditory motion for the broadband stimuli and velocities tested. Consistent with this, subjective feedback from pilot tests confirmed that motion was perceptually smooth at 5° step-sizes (see also Feinkohl et al. [37]) suggesting that the sparser auditory sampling resulted in no loss of spatial resolution.

Reverberation Level and Room Characteristics. As discussed previously, a clear body of evidence demonstrates that reverberation degrades the quality of acoustic cues utilized for static localization [16]. Consequently, we predicted that reverberation would degrade auditory motion perception and thus alter the spatiotemporal dynamics between vision and audition. In light of this, the lack of significant difference in both the PSE and β between anechoic and reverberant conditions was surprising.

The present study sought to examine reverberation in everyday listening rooms, with less reflected energy then that used in Hartmann [16]. The perceptual quality of reverberation in the BRIR recording (and testing) room (RT₆₀ ~ 200 ms) is therefore of interest. Though the reverberation level of the environment was relatively lower than previous studies, qualitative listening confirmed that the reverberation was perceptible (particularly so in contrast to the anechoic chamber environment); the stimulus had a vastly different sound quality, contained more “presence” and yielded a more externalized percept than the anechoic stimulus. To obtain quantitative evidence of this perceptual difference, we consider the difference in D/R between the two environments. Using techniques outlined by Jeub et al. [44], the D/R of the anechoic and reverberant impulses were estimated to be 20.3 dB and 3.9 dB respectively. Zahorik [45] determined the JND for D/R sensitivity in VAS to be 6 dB, which is substantially lower than the 16.4 dB difference between acoustical conditions found in the present study. This strongly suggests that the BRIRs obtained in our testing room contained a perceptually salient level of reverberation.

Even though reverberation levels in our experiment were above perceptual threshold, geometric properties of the room may be such that the precedence effect remediated the deleterious effects of reverberation on localization [16,46]. The precedence effect refers to the perceptual ability to suppress late-arriving signals in order to extract localization cues in the onset waveform. The mechanism by which precedence operates varies depending on the temporal separation of subsequent signals [47]. When the temporal spacing of direct and reflected signals are proximate (0–1 ms interval), a fused image is observed rather than two separate sounds, and the perceived direction is a complex average of the two waveforms, referred to as localization summation [48]. Note however that the direct and first-reflected peaks of BRIRs in the current study are separated by approximately 2.5 ms (Figure 3). For intervals of this magnitude, direct and reflected waveforms maintain a fused percept but the perceived direction is dominated by the initial signal. In such cases of localization dominance [47], reverberation still holds perceptual weighting, conveying qualitative information about the environment, but directional information is extracted solely from the direct waveform. Even beyond the echo threshold, when fusion ceases and two separate images are heard, discrimination suppression caused by the presence of the direct signal can inhibit processing of the reflected signal’s spatial cues. The echo threshold varies according to several acoustic properties of the surrounds, but widely reported values lie between 3 and 10 ms [47]. Thus, with a temporal delay of 2.5 ms between direct and reflected signals, subjects presented with reverberant stimuli may have recovered direct onset cues due to the combined processes of localization dominance and discrimination suppression. Such a process of echo suppression would result in reverberant stimuli with directional cues akin to anechoic stimuli, accounting for the result of the present study. Supporting this, the environment in which Hartmann [16] showed the disruptive effect of reverberation on static localization was highly echoic, with an RT₆₀ of 4 seconds. Though the precedence effect operates at time periods proximal to onset, research suggests that precedence has a longer time-course for ongoing sounds due to multiple onsets brought about by local energy fluctuations [49,50]. Given that our stimuli consist of a concatenation of multiple discrete signals, it is possible that auditory localization may still have been influenced by mechanisms relating to precedence. An interesting question for future consideration is whether a reverberant source in motion alters the thresholds of fusion, dominance and suppression or gives rise to entirely new perceptual phenomena.

Concluding Remarks

The current study explored the effects of spatial quantization and reverberation on auditory motion perception. In order to do this, three different acoustic stimuli were rendered in VAS: two anechoic stimuli which differed in their spatial cue density, and a reverberant stimulus recorded in situ in order to capture veridical room acoustics. These stimuli were presented with a temporally synchronous but spatially varied co-moving visual stimulus with constant cues, thereby serving as a localization reference. No significant differences were found in the PSE or β between conditions in which the auditory spatial sampling was discretised to 1° and 5° or between conditions in which the auditory stimulus was anechoic and reverberant, suggesting that listeners lacked sensitivity to the quantization and reverberation levels tested in the current study. The MAMA and precedence effect offer potential explanations for these findings. We also found no significant difference between the β at all three velocities, suggesting that the physical audio-visual threshold in order to achieve a perceptual separation at the respective sensory peripheries is spatially invariant. Findings suggest a key role for auditory de-reverberation in processing moving auditory stimuli, informing the development of algorithms implemented in digital hearing aids, automatic speech recognition systems and telecommunications aimed at preserving speech intelligibility in reverberant spaces. The present result also establishes a perceptual measure for assessing the veracity of auditory motion generated from discrete spatial locations and in echoic environments.

Supporting Information

Table S1  PSE values for all 6 participants across the three acoustic conditions and three experimental velocities tested in experiments 1 and 2. (PDF)

Author Contributions

Conceived and designed the experiments: NS JL SC. Performed the experiments: NS. Analyzed the data: NS JL SC. Contributed reagents/materials/analysis tools: NS JL SC. Wrote the paper: NS JL SC.
References

Perceptual thresholds of spatial audio latency in dynamic virtual auditory environments

Narayan Sankaran *1, James Hillis1, Marina Zannoli1, and Ravish Mehra1

1Oculus & Facebook, 1 Hacker Way, Menlo Park, CA 94025

narayan.sankaran@sydney.edu.au

(Dated: April 7, 2018)

Abstract

Generating the acoustic signals that reproduce that of natural environments through headphones remains a significant technical challenge. One hurdle relates to the time taken to update the signal when the observer moves. The end-to-end spatial audio latency (SAL) is the time elapsed between the listener assuming a new position and the updated sound being delivered to their ears. The current study probed perceptual thresholds of SAL. We used a 2- interval-forced-choice paradigm to measure detectability at (±10° and ±60°) azimuths, both with and without the presence of co-located visual stimuli. Overall, mean SAL thresholds were between 128ms and 158ms. Consistent with results from minimum audible motion angle data, thresholds were greater at larger azimuthal positions. The presence of a visual target decreased the slope of the psychometric function relating probability of detection to SAL without significantly shifting its position. A retrospective analysis revealed that listeners who strategically varied the velocity, acceleration and rate of their head rotation were better able to detect SAL in the audio-only condition, suggesting that SAL thresholds will be lower when users move their heads more rapidly and abruptly. Results are discussed in the context of prior research and potential implications for rendering Virtual Reality audio.

PACS numbers: 43.66.-x, 43.64.-q, 43.55.Ka
I. INTRODUCTION

In real world environments, people reflexively orient toward a sound source. This remarkable ability relies on a neural system that uses monaural spectral cues and inter-aural differences in the acoustic signals reaching the two ears to compute sound direction. A primary goal of virtual audio systems is to present signals that recreate these real-world listening conditions. Providing these signals through headphones however remains a significant technical challenge. High fidelity signal reproduction for observers moving through virtual environments requires (1) simulating the interaction between the incoming acoustic signal and the individual’s body and (2) rapidly updating the signal whenever the source or listener moves. With present day technology, updating the signals instantaneously is impossible; it takes time to capture head movements and to recompute the sound signal. As a result, there is inevitable lag between listener movement and signal updating. The research presented here addresses the following question: what system latency is required to make this lag undetectable to listeners?

All of the information available to the auditory system for sound localization is contained in a pair of Head-Related Transfer Functions (HRTFs), one for each ear. These functions define the acoustic transformations for signals emanating from all directions in space that result from interactions with the individual’s anatomy [1]. To simulate direction specific acoustic signals through headphones, the source sound is convolved with the HRTF corresponding to the target location. Thus, as the listener moves, the source signal must be continuously convolved with the new direction-specific HRTF that aligns with the changing position of the listener relative to the object. While systems attempt to perform this task in real-time [2-4], the HRTF-update process is never instantaneous [5]. The end-to-end spatial audio latency (SAL) is defined as the time elapsed between the listener assuming a new position and the signals for that position being delivered to the ears. The total SAL is comprised of latencies in head-tracking, HRTF interpolation and filtering, operating system callback, audio driver and hardware buffering (D/A conversion) and other parts of the signal processing chain.

Much of the prior research on SAL has concentrated on it’s effects on sound localization accuracy [8-12]. For brief stimulus durations (between 125ms and 2 seconds), SAL values as low as 73 ms have been shown to result in an increase in localization error [12]. On the other
hand, for longer stimulus durations, studies report no significant localization impairment with SAL values as large as 150ms [8] and 500ms [9, 10]. However, it has also been shown that it takes longer for the listener to orient to the location of such longer duration signals [12] indicating that it may take longer for the listener to make a localization judgement for SAL > 0.

Outside of the effects of SAL on localization, research has shown that individuals experience a reduction in perceptual object stability and coherence as the SAL increases [6, 7, 36]. Other research has directly measured the detectability of SAL. For broadband stimuli located directly in front of the listener (0° azimuth, 0° elevation), SAL detection thresholds have been reported in the range of 60-95ms [7, 13, 14]. One study [13] compared SAL detection thresholds (DTs) to difference limens (DLs) across the same participants, finding values of 94 ms and 70 ms respectively. Moreover, DTs and DLs were strongly correlated across listeners, suggesting that the DL can be interpreted as the DT when the minimum system latency is sufficiently small. In another study [6], listeners were asked to rate perceived latency on a 25-point scale. Based on these measurements, the authors argued that listeners could only readily perceive SAL values in excess of 250ms. However, as the mapping between the arbitrary rating scale and SAL detectability is unknown, these results are more ambiguous than research employing signal detection approaches to measurement. Lastly, a previous study [7] found SAL detection thresholds in typical listeners to be 80 ms in conditions with a single virtual sound source. However, the addition of a co-located low-latency reference sound reduced SAL thresholds by approximately 25ms, suggesting a variation in the allowable SAL depending on the complexity of the scene. The goal of the current study was to build on prior research by examining two unanswered questions: what are the effects of (1) source location and (2) the presence of a spatio-temporally co-located visual target on sensitivity to SAL?

With respect to the effects of source location on spatial auditory perception, many studies have measured the Minimum Audible Angles (MAA) [15, 16], the minimum spatial separation of sounds that is detectable to the human observer, throughout the sound field. These studies show that MAA increases with azimuth (i.e. further from the midline) from approximately 1°, 4° and 10° at azimuthal locations of 0°, ±60° and ±90° [15, 20]. Other studies used moving sources to measure the Minimum Audible Movement Angle (MAMA) [21]. These are particularly relevant because the signals reaching the ears due to SAL are equivalent to
a sound source that moves with the head during the lag period. MAMA values along the horizontal plane follow a similar trend as that of the MAA, with MAMAs of approximately 5° at the midline, increasing to 30° at ±90° [16, 22].

The azimuthal dependence of MAA and MAMA suggest that SAL detection thresholds may be larger at locations further from the midline. However, the current experimental task differs from the aforementioned literature in that self-motion cues (vestibular, proprioceptive, efferent copy) are all required to perform the task (see methods). Brimijoin and Akeroyd [23] have recently shown that MAMA is smaller during self-motion than source motion. This suggests that the detectability of source motion resulting from SAL may be lower than that predicted by measurements of MAMA. Their task however involved discrimination of relative position of 2 simultaneously presented sounds. Differences between performance for absolute source motion detection with and without head movements has not been measured.

As with audio, dynamic VR systems must rapidly update the spatial location of visual objects, and prior research has investigated perceptual thresholds of visual update latency. Findings from multiple studies indicate that maintaining the perceived stability of virtual visual objects across a range of contexts requires latencies less than approximately 15 ms [41–44]. Developing a complete model of object perception in VR requires understanding how cues combine across multiple modalities in the presence of their respective spatial update latencies. The current study, however, presented a “ground truth” (i.e. zero-latency) visual target to first understand the fundamental changes in SAL perception, if any, that occur due to cross-modal interactions between audition and vision. For this reason, we used a static real-world visual target that was co-located with the sound (see methods for details). Little is known about how the presence of visual information impacts SAL detection. On the one hand, the visual target provides a stable “zero-latency” referent from which any auditory spatial instability could be detected. If used in this way, observers will be more sensitive to SAL in the presence of a visual stimulus. On the other hand, it is known that visual stimuli can “capture” the spatial location of an auditory stimulus [24, 28]. In this case, sensitivity to SAL will decrease in the presence of the visual stimulus. In our experiments, we measured SAL discrimination thresholds at two different azimuthal source locations (±10° and ±60°) with and without the presence of a visual target.
II. METHODS

A. Participants

Twenty participants were recruited to participate. Based on the results of pre-experimental training (see section II.D), two participants were identified as unable to perform the task and excluded from subsequent participation.

B. Apparatus

Experimental testing was conducted in an anechoic chamber (500Hz cutoff frequency, 80dB absorption down to 200Hz, 20dBA SPL noise floor). A hollow cloth spherical dome (radius = 1m) was suspended from the ceiling of the chamber, and participants were seated with their head positioned at the center. Head movements were tracked using an Optitrack prime 17W system. A rigid-body structure with retro-reflective markers was fixed onto the participant’s head using a headband. The reflector positions were tracked by sensors that allowed for 6D tracking of the participant’s head movement with a spatial precision within 1mm. Tracking cameras recorded at 360 frames per second using Motive 1.10 motion capture software.

A schematic of the spatial audio system is shown in Figure 1. The audio stimulus was controlled by an RME Fireface UFX soundcard and delivered through Etymonic ER2 in-ear headphones driven by a Stewart Audio PA-50B amplifier. The system was calibrated to ensure equal output levels in each stereo channel at the output of the headphones. Raw audio stimuli (see section II.C) were spatialized using a generic set of HRTFs from the CIPIC database [29]. These were measured with a 5° spatial sampling density at 25 different azimuths and 50 different elevations (1250 directions). We used barycentric interpolation with minimum phase criterion to construct location-precise HRTFs [30]. To render the spatial audio, the raw audio signal was convolved with the left and right HRTFs. In the minimum SAL case, the convolution (buffer size of 64 with a sampling rate of 48kHz) was based on the current estimate of head position provided by the tracking system. In other instances, the SAL was independently varied from trial-to-trial. To achieve this, each convolved buffer of spatial audio reflecting the participant’s current position was written into a circular buffer. Depending on the temporal extent of SAL, the output frame of audio
was read from a correspondingly delayed position in the buffer.

The experiment was controlled by a PC with a 3.5GHz Intel Processor and 32GB of installed RAM, operating on a 64 bit Windows 10 Enterprise system. Experimental scripts were written in Lua and Matlab, which interfaced with C++ for rendering spatial audio in real-time. The minimum SAL of our system was empirically determined by measuring the time elapsed between the motion of a tracked rigid-body and audio output reflecting this motion (Figure 2). First, a sound source was rendered at an infinite distance (i.e. zero amplitude HRTF filter). The location of the source was set to update only once motion was detected (using a movement threshold of 0.5mm). To trigger motion, the metallic rigid-body was struck with a conductive hammer, completing an electromechanical circuit whose voltage was registered on a digital oscilloscope. Once motion was registered by the tracking system, a signal was sent via a PCIe parallel port card to the oscilloscope. The time elapsed between these two events comprised the tracking-system latency and averaged 7.5ms (SD = 1.8ms) over 30 observations. When the HRTF was updated to reflect the induced motion and convolved with the raw audio signal, the resultant output was registered on the oscilloscope. The time elapsed between the initial circuit completion and audio output comprised the total end-to-end SAL (mean SAL = 14ms, SD = 2.5ms). Importantly, in order to accurately detect onset, we ensured that the amplitude envelope of the audio signal had minimal initial rise time and truncated the file so that no silence preceded the waveform.

Visual stimuli were rendered with an Nvidia GTX 980 video card and presented using
FIG. 2. Schematic of latency measurement test. First, a virtual sound source is initialized at an infinite distance (zero-amplitude sound output). The timings of three events are then recorded by the oscilloscope. (1) A conductive hammer strikes a metal element (attached to the rigid-body receiver), completing an electromechanical circuit. This triggers a voltage that is registered on the oscilloscope. The hammer strike also triggers motion of the receiver, which is subsequently detected by the tracking system. (2) Once movement is registered by the PC, a signal is sent to the oscilloscope via PCIe parallel port. The HRTF then updates to a near-field location with non-zero amplitude and is convolved with the raw audio signal. (3) The onset time of the resulting output waveform is recorded on the oscilloscope. (2) - (1) indicates the latency of the tracking system, while (3) - (2) gives the latency in HRTF update, convolution, and output latency. The end-to-end minimum SAL of our system is (3) - (1) and was found to be 14ms (SD = 2.5ms).

three projectors (Aaxa technologies P300 Lumen HD) with a 60Hz frame rate. Projectors were mounted to the same frame as the Optitrack sensors at the top of the dome and aimed at the spatial positions of the auditory stimuli. Images were projected directly onto the fabric of the dome. The projectors were spatially calibrated by aligning them with measured physical locations on the surface of the dome that were also used as sound source positions. The onset timing of audio and visual stimuli were synchronized by delaying audio output until the next projector frame refresh. When sampled over 100 trials the mean difference in onset times was 538μs (SD = 0.62 ms).
C. Stimuli

In order to address the aims of the current study, the choice of raw audio signal was guided by a desire to create a scenario in which SAL sensitivity was likely to be the greatest. The raw audio was a 5 second steady-sound complex tone consisting of 100 harmonics extending from 100Hz - 10kHz, with a 3 dB per octave roll-off. Onsets and offsets were gated with 5ms raised cosine ramps. Subjective evaluation in pilot tests indicated that this signal encouraged perceptual object formation, enabling SAL to be readily detected better than other stimuli we considered. These included fixed-amplitude white noise, amplitude modulated white noise and spectro-temporally modulated tones. Furthermore, the stimulus was consistent with previous research [13]. Left and right audio levels were presented at 80 dB SPL, calibrated by measuring the SPL of left and right headphone output with an Audio Precision x555 spectrum analyser. Visual stimuli were circular patches filled with dynamic white noise updated at 60Hz and subtending approximately 5° of the visual arc (Figure 3).

D. Procedure and Design

Participants were first oriented towards a fixed central position on the sphere using a head-mounted laser and instructed to make large discrete head rotations (i.e. include a brief pause
at the end of each head turn) for the duration of each trial. Once the participant initiated
a trial by starting a head rotation, the stimuli (either audio-only or audio-visual depending
on the condition) appeared at a predetermined location on the sphere. We used a 2-interval
forced choice task (2IFC), whereby stimuli were presented in two 5-second intervals with a 1-
sec inter-stimulus interval (ISI). The minimum system SAL stimulus (14ms; see section II.B)
was presented in one and the other had the stimulus with additional SAL. The assignment
of this additional SAL to a given interval (first or second) was random. Participants pressed
a button to select the interval they believed contained SAL.

Though descriptors of SAL exist [35, 36], individual differences in the perceptual expe-
rience of SAL are likely so it is difficult to justify a common a priori descriptor. For this
reason, participants were pre-exposed to both stimuli during a familiarization session so they
could learn to recognize the difference between stimuli with minimum SAL and those rendered
with larger SALs. The 10-trial familiarization block had the same within-trial structure as
described above with SALs ranging from 500ms to 50ms. Visual feedback indicated which
of the two intervals contained additional SAL. The familiarization procedure was repeated
until participants reported that they were able to differentiate between the two stimuli. This
familiarization phase did not exceed two 10-trial blocks and lasted approximately 5 minutes.

This familiarization was followed by two 10-trial practice blocks. For the practice and
experimental trials, the magnitude of SAL on each subsequent trial was determined using
the QUEST adaptive Bayesian staircase procedure [31], implemented using the Palamedes
toolbox [32] in Matlab™. The QUEST procedure was initialized in mean mode (also known
as ZEST [37]), with a Gaussian prior distribution (mean = 200ms; SD = 150ms) and a min-
imum step-size of 1ms. These parameters were found to be suitable through initial pilot
testing.

Audio-visual trials were presented in different testing blocks, and were identical to audio-
only trials except for the addition of a zero-latency (spatially fixed) visual target (see section
II.C). The spatially aligned visual and audio stimuli were presented simultaneously after
participants initiated a head turn.

We measured latency detection performance in four azimuthal directions (±10° or ±60°)
at 15° elevation in audio-only and audio-visual conditions. The slight elevation in source
location was due to subjective reports that the HRTFs produced greater externalization
at locations above the horizon. The experiment consisted of 160 trials divided into four
testing blocks of 40 trials. Within each block, two 20-trial staircases [38] corresponding to
the two different locations (±10° or ±60°) were randomly interleaved. Audio and audio-
visual conditions were presented in separate blocks, and each 20-trial staircase was repeated
twice for each of the four conditions (2 locations x 2 modalities). The testing order was
counterbalanced to ensure half of the participants experienced the audio-only condition first
and vice versa. Testing was completed over two one-hour sessions, separated by a minimum
of 2 hours.

One concern about the audio-visual condition was that participants could perform the
task with their eyes closed. To ensure that participants remained visually attentive, 10%
of the trials contained an irrelevant visual task. This was the case in both audio-only and
audio-visual conditions to ensure consistency in cognitive demands across these conditions.
On these trials, four circles subtending 5° of the visual arc were displayed for 300ms at a
random time during stimulus playback at all four stimulus locations (±10° and ±60°).
Instead of indicating latency, participants were instructed to report the color of the circles
(yellow or green). Naturally, these trials were not taken into account for the updating of the
staircase values.

III. RESULTS

Weibull functions (Eq. 1) were fit to the psychometric data,

\[
\Psi(\chi; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda) (1 - e^{-(\chi/\alpha)^\beta})
\]

where \(\Psi\) refers to the proportion of correct responses as a function of the stimulus intensity
\(\chi\). Two of the four fitting parameters were fixed; the ‘guess-rate’ (\(\gamma\)) was set to 0.5 and the
‘lapse-rate’ (\(\lambda\)) to 0.02 to account for stimulus-independent errors. The two free parameters,
\(\alpha\) and \(\beta\), give threshold and slope estimates, respectively, and were determined using a
maximum likelihood criterion [33]. From these fits, we report thresholds as the 80.3%
correct point on the psychometric function.

For each individual participant, Weibull functions were fit to psychometric data pooled
from the two staircases for each condition. The group mean SAL thresholds for the four
conditions are displayed in Figure 4. A 2 x 2 [location x modality] within-subject repeated-
FIG. 4. Group mean thresholds estimated from psychometric fits to individual subject data. Standard errors for each condition are shown and points correspond to individual thresholds for the 18 subjects.

measures ANOVA revealed a significant main effect of stimulus location \( F(1,17) = 5.15; p = 0.036 \). Post-hoc tests indicated that SAL thresholds were significantly higher at ±60° than ±10° in the audio only condition \( (p = 0.038; \text{Bonferroni corrected}) \). Neither the effect of modality \( F(1,17) = 1.99; p = 0.18 \) nor the interaction between modality x location were statistically significant \( (F(1,17) = 0.11; p = 0.74) \). Pooling data across the two stimulus locations, we found that participants’ SAL thresholds were significantly correlated in the audio and audiovisual conditions (Figure 5; \( r = 0.48; p = 0.003 \)), indicating that the same subjects who were highly sensitive to SAL in the audio condition also displayed increased sensitivity in the presence of a visual target.

Within each condition, we observed a high degree of inter-subject variability. For example, in the audio condition at 10° azimuth, participants’ SAL thresholds ranged from 35 to 187ms. Across the 20 participants, standard deviations were 37ms (audio at ±10°), 32ms (audio at ±60°), 58ms (audio-visual at ±10°) and 65ms (audio-visual at ±60°). This variability prevents the simple application of group mean values as a way of setting industry tolerances for SAL. Additionally, while thresholds are informative, the slope of the psychometric function is also required in order to get continuous estimates of the probability that participants detect SAL. However, there were not enough data from each individual to obtain reliable estimates of the slope of each psychometric function. We therefore repeated the analysis by pooling data across all participants and fitting condition-specific psychometric
functions as described above. To pool data, each participant’s data were first normalized to the mean of the latencies for which they responded incorrectly across all conditions. This normalization does not affect analyses of relative differences between conditions. Parameter estimates from psychometric fits are shown in Figure 6 and Figure 7. Significance was tested by way of parametric bootstrapping [34] (Number of bootstraps = 1000), whereby two conditions were significantly different if both the upper and lower 95% confidence intervals of the difference in their bootstrapped parameters were greater or less than zero (i.e. zero did not fall within the 95% bounds of the difference in their bootstrapped values). The reported p-values have been Bonferroni-corrected to account for multiple comparisons. Consistent with the main effect of source location found from analyses of individual data, thresholds at ±60° were significantly higher than at ±10° in the audio-only condition (p=0.004). This, however, was not the case in the audio-visual condition (p = 0.27). Interestingly, while threshold values did not differ between audio-only and audio-visual conditions, the slopes were shallower in the audio-visual condition. This difference, however, was only statistically significant at ±60° (p=0.002 at ±60° and p=0.06 at ±10°).

In addition to psychometric data, we examined participants’ head movements as a func-
FIG. 6. Threshold estimates resulting from psychometric functions fit to normalized data pooled across all participants for each condition. Asterisks indicate significance where **p < 0.01

FIG. 7. Slope estimates resulting from psychometric functions fit to normalized data pooled across all participants for each condition. Asterisks indicate significance where **p < 0.01

tion of SAL. Participants were instructed to make head turns with brief pauses in between, but were otherwise free to adopt a head movement strategy that best enabled them to perform the task. For stimuli with a given SAL, faster and more abrupt head rotations would result in larger spatial errors, and may therefore increase the detectability of SAL. Similarly, we hypothesized that a faster rate of head rotation over the 5 second trial duration would effectively increase participants’ sampling of the stimulus, thereby increasing their sensitivity to SAL. To test these hypotheses, we examined the relationship between SAL thresholds and three different motion parameters: (1) maximum speed, (2) average acceleration and
the number of head turns made per second. To perform this analysis, head-tracking data consisting of timestamps and corresponding positions were used to reconstruct motion trajectories. Motion data was first smoothed using an exponential weighted moving average filter. The three motion parameters were computed by finding, for each trial, the maximum speed (the peak in the absolute value of the smoothed velocity profile), the rate of turns (the number of peaks in the profile divided by the stimulus duration), and the average acceleration. Data from the audio-only and audiovisual conditions were analysed separately.

First we wished to determine whether participants with faster, more abrupt, and more frequent average head rotations were more sensitive to SAL. Figure 8 displays the relationship between each participant’s SAL threshold and each of their motion parameters averaged across all trials within a given condition, pooled across the two locations (±10° and ±60°). For both audio and audiovisual conditions, no significant correlation existed between speed and SAL thresholds, or acceleration and SAL thresholds. Interestingly, we found a slight indication of an inverse relationship between the average rate of head turns made by a participant and their respective SAL thresholds (R = -0.30; p = 0.08 for audio; R = -0.38; p = 0.02 for audiovisual).

Though no clear correlation was found between SAL thresholds and average head-motion parameters, we hypothesized that head-motion may vary as a function of SAL on a trial-to-trial basis. Specifically, participants may have implemented an “online” adaptive strategy; varying the speed, acceleration and rate of their rotations commensurate with the SAL present in the current signal in order to maximize the likelihood of detection. Figure 9 displays data for one participant (pooled across both positions) in the audio condition. Each point represents a trial in which a motion parameter (on the y-axis) is plotted against the SAL in that trial (x-axis). There is a clear inverse relationship between the SAL and the speed, acceleration, and rate at which this particular participant rotated. The strength of these inverse correlations differed across individuals (see below). To quantify the extent to which each participant’s head-motion parameters varied as a function of the SAL, a multiple linear regression model was fit for each participant’s data (Eq. 2),

\[
SAL = b_0 + v_{max} b_1 + a b_2 + r b_3
\]

where \(v_{max}\), \(a\) and \(r\) correspond to the maximum velocity, acceleration, and rate respect-
FIG. 8. The relationship between head-motion parameters and SAL thresholds. Each point represents an individual subject, whose head-rotational speed (top), acceleration (middle) and rate (bottom), averaged across all trials, is plotted against their respective SAL threshold for audio (left) and audiovisual (right) conditions. Data is pooled across both source locations ($\pm 10^\circ$ and $\pm 60^\circ$). Each panel also indicates the associated Pearson correlation coefficient.

具体的。每个模型的R^2值表示运动参数解释的SAL变异量。

回归模型包括所有三个预测变量，被发现优于基于每种组合的仅一个或两个预测变量的模型。如预期的那样，数据解释的多样性在参与者的模型中都存在。总体模型适合度被参与到个体参与者的方式中（通过方差分析），被发现对参与者的12个模型显著（在这些12个线性模型：F(3,76) > 3.94; p < 0.05; 调整R^2 > 0.1; 其中R^2是每个主题的模型的决定系数）在音频-only 条件，和在13个参与者（F(3,76) > 2.88; p < 0.05; 调整R^2 > 0.07 对所有13个模型）在音频视觉条件。对于这些参与者，线性回归分析表明，他们的头部旋转速度，加速度，和速率是SAL的可靠预测器。

接下来，我们想要确定上述逆关系是否反映了参与者用来检测SAL的有效策略。如果他们是，个体-
FIG. 9. Data from one participant indicating the maximum speed (left), mean acceleration (middle) and rate of head turns (right) on each trial as a function of the SAL present in the signal. Significant inverse correlations were observed between the SAL and the rotational speed (r = -0.32; p = 0.004), acceleration (r = -0.58; p < 0.00001) and rate (r = -0.49; p < 0.00001).

Individuals whose regression models had larger R² values would have lower SAL thresholds. Each participant’s R² value is plotted against their respective SAL threshold in Figure 10. In the audio condition, results corroborated our hypothesis above (R = -0.49; p = 0.003). The circles with asterisks represent participants for whom model coefficients were significantly different from 0 at the p < 0.05 level. Different motion parameters were significant predictors of SAL for different participants (represented by the colored triangles in Figure 10). These results indicate that participants who systematically varied their head motion commensurate with the magnitude of SAL across trials were better able to perform the task (i.e. they had lower overall SAL thresholds). The equivalent analysis for the audiovisual data (Figure 10B) showed no such correlation (R = 0.04; p = 0.78), suggesting that, although certain participants rotated their heads as a function of SAL, this functional relationship did not necessarily yield more sensitive SAL detection. Notably, the R² values of participants’ regression models were not significantly correlated across audio and audiovisual conditions (r = -0.32; p = 0.17), suggesting that those who systematically varied their head rotation in the audio condition were not necessarily applying the same strategy in the presence of a visual target.

IV. DISCUSSION

The primary aim of the current study was to establish measures of SAL detectability that reflect a “worst case scenario” for different azimuths and with and without coincident visual...
FIG. 10. (A) In the audio-only condition, the adjusted $R^2$ values from each participant’s linear regression coefficients were inversely correlated with their respective SAL thresholds ($R = -0.49; p = 0.003$). (B) In the audiovisual condition, no such relationship existed between participants’ $R^2$ values and respective SAL thresholds ($R = 0.04; p = 0.78$). Circles with asterisks indicate participants with significant model fits. Colored triangles indicate which of the three motion characteristics were significant predictors in the model.

targets. We did this by asking participants to detect the difference between two targets, one at the minimum 14ms latency of our system and one with added SAL. We also gave them the freedom to strategically make rotational head movements that could help them perform the task. Group mean thresholds (Figure 4) of SAL detectability varied between 128ms (audio-only at $10^\circ$) to 158ms (audio-visual at $60^\circ$). This is considerably larger than the 60-90ms range previously reported for targets in the midline[7, 13]. There was however a large degree of variability across individuals. The relatively larger thresholds and high between-participant variability may be explained by our use of generic HRTFs (whereas previous research used individualized HRTFs). Specifically, each listener may have experienced a distortion of auditory space commensurate with the extent to which their head and pinna morphology differed from that suggested by the generic HRTF. Prior research has shown that this spatial distortion can increase localization errors, particularly front-back confusions and judgements of source-elevation [6, 39, 40]. The distortions in auditory space created by the use of generic HRTFs may have also altered SAL thresholds, however, our results have direct relevance to commercial products where the use of generic HRTFs is common.

With respect to the effects of target location, we found that SAL thresholds were signifi-
cantly smaller for targets at azimuthal angles of 10° than those at 60° in the audio condition. This is consistent with previous work showing that both MAAs and MAMAs increase with source azimuth. The corroboration between our findings and literature on the MAA and MAMA suggests that SAL detection may be rooted in the same sensory processing mechanisms underlying the auditory representation of object position and motion. Given that there is no real-world correlate for SAL, it is plausible that participants would use perceptual estimates of object motion and position to detect SAL. If that were the case, existing MAA and MAMA measures may serve as a proxy for understanding SAL tolerances in VR. Such knowledge generalization is limited however by the fact that SAL detection involves use of self-motion cues, while MAA and MAMA measurements do not. Previous work has shown that motion sensitivity is greater during self-motion than source-motion\textsuperscript{[23]}. Nonetheless, consistency in SAL thresholds and the MAMA as a function of source azimuth provides some evidence that differences in thresholds under self-motion and object motion paradigms may come down to a common scale factor across locations.

Our findings with respect to the presence of a spatiotemporally coincident visual target do not have as clear an interpretation as the effects of target location. As discussed in the introduction, a visual target could act as a stable referent, thereby increasing the likelihood of detecting audio SAL, or it could "capture" the sound location and make SAL detection less likely. Neither hypothesis predicts the shallower slopes of the psychometric functions we observed (see Figure 7). This finding may be a result of visual referent only impacting a portion of the listeners. Several participants did, in fact, report that the audio and visual targets seemed like separate events (i.e. they were not bound into a single perceptual event). Thus, thresholds for these individuals would not have been subject to potential effects of visual capture. While our results are inconclusive, understanding the impact of visual targets on audio SAL detection remains of high practical relevance to setting tolerances for more typical virtual reality applications where rich stimulus environments are the norm. Performing experiments that ensure audio-visual binding will help clarify the impact of a visual target on SAL detection. This is likely to be achieved with the use of naturalistic audio and visual stimuli that have dynamic, spectrotemporally coherent features (e.g. matching audio and visuals of a person speaking). Such studies would also provide insight into the relative impact of audio-visual signal fidelity for human performance (e.g. in target localization) as well as the quality of experiences in virtual environments.
A correlational analysis determined that participants who on-average made more frequent head-turns tended to have lower SAL thresholds. The same inverse relationship did not hold when comparing average rotational speeds or accelerations with SAL thresholds. The only scenario in which it is physically possible for rotations to be more frequent without having greater speed and acceleration is if certain participants made smaller head turns (i.e. their rotations subtended a smaller angle). These results therefore suggest that small and frequent head movements increase sensitivity to SAL, irrespective of speed or acceleration. However, when examining the relationship between motion parameters and SAL on a trial-by-trial basis, we found an inverse relationship between the SAL in a given trial and all three motion parameters. Furthermore, listeners who systematically adapted their motion as a function of the current SAL (Figure 9) had lower audio-only SAL thresholds. This finding further corroborates the idea that the mechanism responsible for SAL detection is the same as that underlying MAMA: for a given SAL, faster and more abrupt head rotations produce a greater spatial discrepancy between the expected and actual direction of a target. The same trial-by-trial inverse relationship between a given motion-parameter and SAL did not result in lower thresholds in the presence of an accompanying visual target. Clearly, the addition of a visual stimulus fundamentally altered the way in which participants moved, thereby complicating any simple functional relationship between motion parameters and SAL. As mentioned above, understanding the impact of visual stimuli on dynamic auditory spatial processing in general remains a goal of future research. Additionally, we note that current results relating head-motion to SAL sensitivity are correlational; falling short of establishing causality in the relationship. Furthermore, it is highly unlikely that the 3 motion predictors are independent. However, intercorrelations between predictors do not negate the primary goal of this analysis, which is not to build a predictive model, but to determine whether characteristics of head motion underlie SAL sensitivity. To this end, the current data strongly suggests that a user’s motion will have implications for the amount of SAL that is perceptually tolerable in virtual reality. Understanding the task dependence of SAL detection would allow developers to balance the risk of creating richer audio environments against the potential of increasing computational time and SAL.
ACKNOWLEDGMENTS

The authors would like to thank Lindsey Neby, Ryan Ebert and members of the audio team at Oculus Research for their help with experimental setup and data collection. We would also like to thank the anonymous subjects for taking part in our user study.

REFERENCES


A. B. Watson and D. G. Pelli, “Quest: A bayesian adaptive psychometric method”, Attention,


Fig. 1. Schematic of spatial audio system.

Fig. 2. Schematic of latency measurement test. First, a virtual sound source is initialized at an infinite distance (zero-amplitude sound output). The timings of three events are then recorded by the oscilloscope. (1) A conductive hammer strikes a metal element (attached to the rigid-body receiver), completing an electromechanical circuit. This triggers a voltage that is registered on the oscilloscope. The hammer strike also triggers motion of the receiver, which is subsequently detected by the tracking system. (2) Once movement is registered by the PC, a signal is sent to the oscilloscope via PCIe parallel port. The HRTF then updates to a near-field location with non-zero amplitude and is convolved with the raw audio signal. (3) The onset time of the resulting output waveform is recorded on the oscilloscope. (2) - (1) indicates the latency of the tracking system, while (3) - (2) gives the latency in HRTF update, convolution, and output latency. The end-to-end minimum SAL of our system is (3) - (1) and was found to be 14ms (SD = 2.5ms).

Fig. 3. Visual stimulus.

Fig. 4. Group mean thresholds estimated from psychometric fits to individual subject data. Standard errors for each condition are shown and points correspond to individual thresholds for the 18 subjects.

Fig. 5. Threshold estimates resulting from psychometric functions fit to normalized data pooled across all participants for each condition. Asterisks indicate significance where **p < 0.01

Fig. 6. Slope estimates resulting from psychometric functions fit to normalized data pooled across all participants for each condition. Asterisks indicate significance where **p < 0.01

Fig. 7. Maximum velocity (left), mean acceleration (middle) and rate of head turns (right) plotted against the SAL for one typical participant.

Fig. 8. (A) In the audio-only condition, the adjusted $R^2$ values from each participant’s linear regression coefficients were inversely correlated with their respective SAL
thresholds ($R = -0.49; p = 0.003$). (B) In the audiovisual condition, no such relation-
ship existed between participants’ $R^2$ values and respective SAL thresholds ($R = \ 0.04; p = 0.78$). Circles with asterisks indicate participants with significant model fits. Colored triangles indicate which of the three motion characteristics were significant predictors in the model.